Title

Deciphering the Function of Single Nucleotide Variants in Post-transcriptional Gene Regulation

Author

Fu, Ting

Publication Date

2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Deciphering the Function of Single Nucleotide Variants in Post-transcriptional Gene Regulation

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Molecular, Cellular, and Integrative Physiology

by

Ting Fu

2022

ABSTRACT OF THE DISSERTATION

Deciphering the Function of Single Nucleotide Variants

in Post-transcriptional Gene Regulation

by

Ting Fu

Doctor of Philosophy in Molecular, Cellular, and Integrative Physiology

University of California, Los Angeles, 2022

Professor Xinshu Xiao, Chair

Single-nucleotide variants (SNVs), such as genetic variants and RNA editing sites, constitute the most prevalent type of sequence variations in the RNA. Genome-wide association studies (GWAS) and global analysis of RNA editing have revealed many genetic variants and RNA editing sites associated with human diseases and complex traits. Yet, the underlying mechanisms of such associations are still missing for most SNVs. In this dissertation, we studied both non-coding and coding SNVs, revealing their functional roles in regulating mRNA abundance and splicing.

Understanding the function of non-coding rare genetic variants remains a major challenge. To fill in this gap, we developed a massively parallel reporter assay, allowing functional testing of 3' UTR variants regulating mRNA abundance in a high-throughput manner. We screened 14,575 rare 3' UTR genetic variants and identified 5,437 functional ones leading to significant changes in mRNA abundance in at least one human cell line. Supported by TCGA

expression outlier analysis and experimental studies, we observed that many rare 3' UTR variants regulate mRNA abundance of cancer-relevant genes.

We next examined a specific set of RNA editing sites that are differentially edited between epithelial and mesenchymal tumors across 7 cancer types in The Cancer Genome Atlas (TCGA). Inspired by the correlations between editing levels and gene expression, we uncovered many 3' UTR RNA editing sites regulating mRNA abundance. Further, we identified the RNA-binding protein ILF3 as a potential regulator of the editing-dependent gene expression change, especially for immune-relevant genes. We showed that multiple RNA editing sites mediate the ILF3 stabilizing effect on the transcripts encoding Protein Kinase R (PKR), a key player in immune response.

In addition to non-coding variants, we further characterized two RNA editing sites in the alternative exon of the gene podocalyxin-like (*PODXL*), a significant clinical indicator for tumor detection and prognosis. Supported by survival analysis in Kidney Renal Clear Cell Carcinoma (KIRC) patients, we discovered dual roles of RNA editing in promoting the loss of function of PODXL in cancer. We hypothesized that exonic RNA editing sites contribute to proteomic diversity through alternative splicing, a previously overlooked function of RNA editing.

The dissertation of Ting Fu is approved.

Amy Catherine Rowat

Dinesh Subba Rao

Yibin Wang

Xinshu Xiao, Committee Chair

University of California, Los Angeles

2022

To my family

TABLE OF CONTENTS

LIST OF MAIN FIGURES

LIST OF SUPPLEMENTARY FIGURES

LIST OF SUPPLEMENTARY TABLES

ACKNOWLEDGEMENTS

This dissertation will not be possible without the contributions and suggestions of many people. I would like to thank my mentor Professor Xinshu Xiao for taking me as her graduate student and supporting my doctoral training financially and mentally. I sincerely appreciate the freedom she has provided me to explore different hypotheses of my research projects, meanwhile learning new experimental techniques and computational skills. I also appreciate her guidance and all the research opportunities she offered that have shaped my scientific career.

I would like to thank Professor Amy Rowat, Professor Dinesh Rao, and Professor Yibin Wang, for serving as my doctoral committee and providing helpful suggestions and comments during committee meetings. Specifically, I would like to thank Professor Amy Rowat for sharing her lab resources with me to access the IncuCyte live-cell analysis system. I would also like to thank Jae Hoon Bahn, Tae-Hyung Kim, Rocky Cheung, Professor Sriram Kosuri, and Professor Chonghui Cheung for their guidance and discussions.

I would like to thank all current and former members of the Xiao lab: Jae Hoon Bahn, Yun Yang, Adel Azghadi, Esther Hsiao, Ashley Cass, Boon Xin Tan, Yi-Wei Sun, Hyun-Ik Jun, Ei-wen Yang, Stephen Tran, Jae-Hyung Lee, Jingyan He, Ling Zhang, Kiku Koyano, Tracey Chan, Giovanni Quinones Valdez, Mudra Choudhury, Christina Burghard, Kofi Amoah, Carlos Gonzales-Figueroa, Jonathan Hervoso, Elaine Huang, Zhiheng Liu, Tadeo Spencer, Ryan Barney, Duncan Croll, Alison King, Melinda Luo, Sari Terrazas, Lindsey Dudley, Jack Dodson, and Ryo Yamamoto. They have made our lab a great place to learn, work, play, and relax. Particularly, I would like to thank Tracey Chan, who has been a great colleague to work with, a patient teacher to learn from, and a close friend to talk to.

Lastly, I would like to thank my parents, Lihong Yang and Meiqing Fu, for their tremendous support and personal sacrifices that allowed me to pursue a PhD overseas. I would also like to thank my friends Min Chai, Yan Cui, Xinwei Huang, and Xiaojing Mao for spending time with me and sharing the laughter and tears.

This work was partially supported by the UCLA Hyde Fellowship and Dissertation Year Fellowship.

Chapter 2 is currently in preparation for publication as “Fu T*, Amoah K*, Chan TW, Bahn JH, Lee JH, Terrazas S, Cheung R, Kosuri S, Xiao X. Massively parallel screen uncovers many rare 3’ UTR variants regulating mRNA abundance of cancer-relevant genes. (*Contributed equally).” T.F., K.A. and X.X. designed the study with inputs from all other authors. T.F., J.H.B., S.T. and R.C. conducted the molecular, cellular, and biochemical experiments. K.A., T.F., T.W.C., and J.H.L. conducted the bioinformatics analyses. X.X. and S.K. provided supervisory inputs. All authors contributed to the writing of the paper. All authors approved the final manuscript. X.X. was the principal investigator.

Chapter 3 is a version of “Chan TW*, Fu T*, Bahn JH, Jun HI, Lee JH, Quinones-Valdez G, Cheng C, Xiao X. RNA editing in cancer impacts mRNA abundance in immune response pathways. Genome biology. 2020 Dec;21(1):1-25. https://doi.org/10.1186/s13059-020-02171-4. (*Contributed equally).” T.W.C., T.F., C.C., and X.X. designed the study with inputs from all other authors. T.W.C., J.H.L., and G.Q.V. conducted the bioinformatics analyses. T.F., J.H.B., and H.I.J. conducted the molecular and cellular experiments. All authors contributed to the writing of the paper. X.X. was the principal investigator.

Chapter 4 is a version of “Fu T, Chan TW, Bahn JH, Kim TH, Rowat AC, Xiao X. “Multifaceted role of RNA editing in promoting loss-of-function of PODXL in cancer. iScience. 2022. https:// doi.org/10.1016/j.isci.2022.104836.” T.F. and X.X. designed the study with inputs from all other authors. T.F., J.H.B., and T.-H.K. conducted the molecular, cellular, and biochemical experiments. T.W.C. and T.F. conducted the bioinformatics analyses. X.X. and R.C.W. provided supervisory inputs. All authors contributed to the writing of the paper. All authors approved the final manuscript. X.X. was the principal investigator.

VITA

Education and employment

2010-2014	B.S., Biosciences University of Science and Technology of China
2014-2015	M.S., Stem Cell Biology and Regenerative Medicine University of Southern California
2015-2016	Resource Employee, Zilkha Neurogenetic Institute University of Southern California
2016-2022	Graduate Student Researcher, Molecular, Cellular, and Integrative Physiology Interdepartmental Program University of California, Los Angeles
2017, 2019	Teaching Assistant, Life Sciences Core Education University of California, Los Angeles

Awards

2021-2022	Dissertation Year Fellowship, Graduate Division University of California, Los Angeles
2021	Poster Award, Institute for Quantitative and Computational Biosciences University of California, Los Angeles
2020-2021	Hyde Fellowship, Department of Integrative Biology and Physiology University of California, Los Angeles

Publications

1. **Ting Fu***, Kofi Amoah*, Tracey W. Chan, Jae-Hoon Bahn, Jae-Hyung Lee, Sari Terrazas, Rocky Cheung, Sriram Kosuri, and Xinshu Xiao. “Massively parallel screen uncovers many rare 3’ UTR variants regulating mRNA abundance of cancer-relevant genes.” In preparation. (*co-first author)

2. Mudra Choudhury, **Ting Fu**, Kofi Amoah, Hyun-Ik Jun, Tracey W. Chan, Sungwoo Park, David W. Walker, Jae Hoon Bahn*, and Xinshu Xiao*. “Widespread RNA hypoediting in schizophrenia and its relevance to mitochondrial function.” Submitted. (*co-correspondence author)

3. **Ting Fu**, Tracey W. Chan, Jae Hoon Bahn, Tae-Hyung Kim, Amy C. Rowat, and Xinshu Xiao. “Multifaceted role of RNA editing in promoting loss-of-function of PODXL in cancer.” *iScience*, 2022

4. Jingyan He, **Ting Fu**, Ling Zhang, Lucy Wanrong Gao, Michelle Rensel, Luke Remage-Healey, Stephanie A. White, Gregory Gedman, Julian Whitelegge, Xinshu Xiao, and Barney A Schlinger. “Improved Zebra Finch Brain Transcriptome identifies novel proteins with sex differences.” *Gene*, in press, 2022

5. Zhiheng Liu, Giovanni Quinones-Valdez, **Ting Fu**, Mudra Choudhury, Fairlie Reese, Ali Mortazavi, and Xinshu Xiao. "L-GIREMI uncovers RNA editing sites in long-read RNA-seq." Submitted, 2022

6. Giovanni Quinones-Valdez, **Ting Fu**, Tracey W. Chan, and Xinshu Xiao. "scAllele: a versatile tool for the detection and analysis of variants in scRNA-seq." *Science Advances*, in press, 2022

7. Tracey W. Chan*, **Ting Fu***, Jae Hoon Bahn, Hyun-Ik Jun, Jae-Hyung Lee, Giovanni Quinones-Valdez, Chonghui Cheng, and Xinshu Xiao. "RNA editing in cancer impacts mRNA abundance in immune response pathways." *Genome biology* 21, no. 1 (2020): 1-25. (*co-first author)

8. Ei-Wen Yang, Jae Hoon Bahn, Esther Yun-Hua Hsiao, Boon Xin Tan, Yiwei Sun, **Ting Fu**, Bo Zhou, Eric L Van Nostrand, Gabriel A Pratt, Peter Freese, Xintao Wei, Giovanni Quinones-Valdez, Alexander E Urban, Brenton R Graveley, Christopher B Burge, Gene W Yeo, and Xinshu Xiao. "Allele-specific binding of RNA-binding proteins reveals functional genetic variants in the RNA." *Nature communications* 10, no. 1 (2019): 1-15.

9. Claudio P. Albuquerque, Eyan Yeung, Shawn Ma, **Ting Fu**, Kevin D. Corbett, and Huilin Zhou. "A chemical and enzymatic approach to study site-specific sumoylation." *PLoS One* 10, no. 12 (2015): e0143810.

CHAPTER 1

Background

1.1 What are single nucleotide variants (SNVs)?

The human genome, composed of four types of DNA nucleotides, encodes fundamental biological information. Although the genome integrity is being strictly controlled by the DNA repair pathways, mutations may still occur and potentially become inheritable if present in germ cells[1]. Thanks to the wide application of whole-genome and whole-exome sequencing in parent-offspring trios, a large number of inherited and de novo mutations have been identified, providing new insights into the occurrence of de novo mutations and their links to disease etiology[2]. Around 4.1 to 5 million nucleotide positions in an individual's genome are different from the human reference genome[3]. According to the 1000 Genomes Project, most human genetic variants are single-nucleotide polymorphisms (SNPs), many of which have a frequency less than 0.5% in the human population[3]. As of June 2022, over 1 billion SNPs have been reported in the dbSNP Human Build 155 release[4].

However, not all SNPs cataloged by the dbSNP are true genetic variants, as some RNA editing sites have been annotated as SNPs due to a failure in recognizing the RNA-DNA differences (RDDs) in these sites[5]. Unlike genetic mutations, RNA editing is a co-/post-transcriptional process that introduces nucleotide substitutions, insertions, and deletions in the RNA transcripts, which are not present in the DNA sequences. The identification of RNA editing sites requires not only variant calling from RNA, but also evidence of RDDs in each sample. RDDs can be identified given matched RNA and DNA sequences[6]. Nonetheless, in the absence

of DNA sequencing data, RNA editing sites can be inferred from RNA sequences alone, taking advantage of the prior knowledge of distinct features of RNA editing sites, such as high allele frequencies in the population[7], tendency to cluster together[8], and random allelic linkage[9]. The most common type of RNA editing is Adenosine-to-Inosine (A-to-I) editing, catalyzed by the adenosine deaminase acting on RNA (ADAR) protein family[10]. Most A-to-I editing sites are found in *Alu* elements, which are primate-specific retrotransposons accounting for ~10% of the human genome[11]. The highly repetitive Alu elements can form long dsRNAs that serve as the targets for ADARs[12,13]. Around 16 million human A-to-I editing events have been recorded in the REDIportal database[14].

Both genetic variants and RNA editing sites can be expressed in the RNA as single-nucleotide variants (SNVs). Although their biogenesis pathways are substantially different, the functional impact of these two types of SNVs, once transcribed, share a high level of similarity. Our work focuses on the function of SNVs in post-transcriptional processes, which is discussed below.

1.2 SNVs in post-transcriptional processes

1.2.1 Deciphering the function of genetic variants, a post-genomic challenge

The discovery of a large number of genetic variants has greatly facilitated gene-disease mapping via linkage analysis[15]. More than 6,000 single-gene disorders with known molecular basis have been reported in the Online Mendelian Inheritance in Man (OMIM) database[16]. For complex diseases resulting from multiple genetic variants and environmental influences, genome-wide association studies (GWAS) were adopted to identify common and less penetrant

variants with disease associations[17]. The NHGRI-EBI database provides a comprehensive catalog of GWAS results, currently encompassing 384,894 associations between genetic loci and human traits[18]. It is important to note that many GWAS SNPs may not necessarily cause the associated traits but, rather, in linkage disequilibrium (LD) with a potentially causal SNP[19]. Following GWAS, candidate causal SNPs may be further predicted using fine-mapping approaches[20]. Although a large number of SNP-trait associations have been reported, the underlying causal variants and their associated functional mechanisms remain largely undetermined. As reported in a previous study[21], 88% GWAS SNPs were in the non-coding regions of the genome, whose functional roles are challenging to decipher. Thus, decoding the genotype to phenotype relationships represents a significant challenge in the post-genomic era.

1.2.2 Approaches to predict functional genetic variants in post-transcriptional regulation

To decode the genotype to phenotype relationships, the first fundamental step is to understand the molecular function of a genetic variant. Post-transcriptionally, the function of a genetic variant can be inferred based on the genomic region. For example, variants within exons can cause amino acid changes and stop-codon gain/loss[22]. In addition, variants located in splice sites could lead to the formation of alternative isoforms with abnormal functions[23]. Compared to coding genetic variants, variants in the non-coding regions, such as untranslated regions (UTRs) and introns, are much harder to decode directly. Mechanism-wise, variants in the non-coding regions can alter the sequences or the accessibility of *cis*-regulatory elements, thus changing their binding preferences with *trans*-factors, which ultimately contribute to phenotypic changes[24]. Fortunately, even with limited knowledge of *cis*-regulatory elements in non-coding regions, taking advantage of the increased availability of genotype-phenotype datasets such as

Genotype-Tissue Expression (GTEx)[25], the linkage between genetic variants and molecular phenotypes (i.e., gene expression and splicing) can be statistically inferred using quantitative trait locus (QTL) analysis[26], allelic-specific linkage analysis[27,28], and machine learning[29,30]. Such association studies have also been used to investigate the regulatory effect of genetic variants on RNA modifications[31,32].

While computational methods can annotate functional variants, experimental validations are needed to establish the causal link between the genotype and phenotype in an appropriate context. To this end, reporter gene assays, such as luciferase reporter assays[33] and green fluorescence protein (GFP) reporter assays[34,35], have been used to characterize the functional effects of genetic variants in gene expression[36] and splicing[37]. This process can be very slow as one locus of interest is being characterized at a time. Facilitated by the development of massively parallel DNA synthesis and sequencing, massively parallel reporter assays (MPRAs) have been developed to measure the function of genetic variants in a high-throughput manner[38]. Depending on the method of readout, MPRAs can be divided into two types. The first type utilizes fluorescence-activated cell sorting (FACS) to detect a normalized florescent expression level (i.e. the intensity ratio of the test fluorescent protein against the control fluorescent protein), which is designed to reflect gene expression[39,40] or splicing activity[41] of a variant. Cells are sorted into different bins representing different activity levels, followed by DNA sequencing to identify the variants that are perturbed in each bin. The second type measures variant activities directly from RNA sequencing reads, sometimes normalized against DNA sequencing reads. For example, the RNA/DNA ratio of a variant can be used to reflect its expression level[42,43]. Also, the splicing activity of a variant can be calculated using percent spliced in (PSI)[44]. With additional cell fractionation steps, MPRAs can also measure the impact of variants on mRNA localization[45].

Exploiting the expanding genome editing toolkit, CRISPR screens can perturb genes in their native genomic context and measure the functional consequences in a more physiologically relevant environment[46]. When combined with high-content readout such as single-cell sequencing, CRISPR screens allow the detection of transcriptome changes caused by genetic perturbations in vivo[46]. In addition to validating known variants, MPRAs and CRISPR screens can also nominate novel functional elements related to the phenotype of interest through random mutagenesis and perturbations[38]. The experimental measurements, together with other genome annotations obtained computationally or experimentally, can be used to train computational models to predict functional genetic variants[47–49].

1.2.3 Functional roles of RNA editing sites

Almost at the same time when genetic variants are transcribed into RNA, RNA editing modifies the sequences of nascent RNA molecules[50]. Thus, RNA editing has profound impact on post-transcriptional regulation, just like genetic variants. In A-to-I editing, the inosine is recognized as guanosine by the cellular translational machinery[51]. The consequential recoding events have been characterized for a few RNA editing sites with important physiological functions. For example, RNA editing in glutamate ionotropic receptor AMPA type subunit 2 (*GRIA2*) causes a codon change from glutamine (Q) to arginine (R), which substantially increases the calcium permeability of AMPA receptors that are responsible for synaptic transmission in the central nervous system[52]. This mechanism is conserved in mice, where a lack of Q/R recoding events led to early-onset epilepsy and death[53]. Another Q/R recoding site in the gene filamin A (*FLNA*) is highly edited in normal hearts, but lowly edited in patients with dilated cardiomyopathy[54]. Transgenic mice with a deficiency in *Flna* editing showed pathological cardiac remodeling[54]. In addition to altering protein coding, most RNA editing sites are located in non-coding regions

such as introns and UTRs[55], actively involved in other aspects of post-transcriptional regulation. RNA editing in introns can modulate pre-mRNA splicing[50]. Notably, RNA editing can create novel splice sites leading to the exonization of *Alu*-exons[56]. When present in non-coding RNAs such as microRNAs (miRNAs) or 3' UTR regions of transcripts, RNA editing can affect miRNA-mediated gene silencing, thus altering mRNA abundace[57,58].

RNA editing also plays a unique role in innate immunity. Upon viral infection, the external viral RNA forms dsRNA in the host cell, which is recognized by dsRNA sensors to initiate immune responses against viruses[59]. Importantly, RNA editing prevents the recognition of self dsRNAs by dsRNA sensors, such as the melanoma differentiation-associated protein 5 (MDA5)[60] and protein kinase R (PKR)[61]. As part of the innate immune system, MDA5 senses cytoplasm dsRNAs and elicit type I interferon responses[62]. PKR, on the other hand, triggers translational shutdown upon dsRNA sensing[63]. While activation of MDA5 and PKR are helpful in suppressing the replication of viruses, these processes are also detrimental to the host cell if elicited in the absence of viruses. Thus, RNA editing is essential in controlling the activation of immune response by self dsRNA, whose dysregulation may result in autoimmune disorders[64]. Indeed, loss of ADAR1, a main regulator of RNA editing, induced abnormal interferon response in multiple systems, including hematopoietic stem cells[65] and liver[66,67]. The loss of *Adar1* in mice led to early embryonic lethality due to elevated interferon response[68,69]. Recent studies in mice also showed that *Adar1* deletion promotes tumor inflammation and sensitivity to immunotherapy, which are likely mediated by MDA5 and PKR[70].

1.2.4 Relevance of RNA editing to human diseases

Normally, RNA editing is dynamically regulated in tissues, where some editing sites are associated with age and gender[71,72]. Distinct from a genetic variant, which would have a fixed allelic ratio (e.g., 0.5 or 1) in a cell, RNA editing usually comes with a more flexible allelic ratio represented by the proportion of edited transcripts over the sum of edited and unedited transcripts covering a specific site. Several global analyses have revealed widespread dysregulation of RNA editing levels (i.e. ratios) in human diseases[73–78]. It remains unknown if the widespread misregulation of RNA editing in human diseases is a cause or a consequence[64], demanding substantial efforts in the functional characterization of RNA editing sites.

RNA editing is upregulated in most cancer types[77,78]. Perturbation assays showed that both ADAR1 and ADAR2 regulate tumorigenesis, but with varied effects[79–81]. ADAR1 is found to be an oncogene in hepatocellular, esophageal, colorectal, and lung cancers, which facilities tumor progression by mediating RNA recoding in AZIN1 and FAK[80,82–84]. ADAR2 is mostly reported as a tumor suppressor, whose editing activities in COPA, IGFBP7, and PODXL have anti-tumor effects in hepatocellular, esophageal, and gastric cancers, respectively[79,81,85]. Moreover, ADARs can edit miRNAs, such as miR-200, miR-21, and miR-381, and alter the expression levels of tumor suppressor genes or oncogenes[86]. Apart from its direct roles in regulating tumor driver genes, RNA editing in cancer also contributes to proteomic diversity[87], which potentially give rise to neo-antigens to elicit immune responses[88].

RNA editing dysregulation is also associated with neurological disorders, cardiovascular diseases, and metabolic diseases[89]. A few RNA editing sites have been characterized in these disorders. For instance, reduced RNA editing in *GRIA2* is observed in the motor neurons of amyotrophic lateral sclerosis patients[90]. Similar associations with lowly edited *GRIA2* are reported in epilepsy, schizophrenia, and bipolar disorder[91,92]. Another example is serotonin 2C receptor (5-HT2cR), whose editing leads to the generation of multiple edited isoforms that are

linked to schizophrenia, depression, bipolar disorder, Prader-Willi syndrome, and diabetes[64,89]. Enhanced RNA editing in the 3' UTR of the cathepsin S (*CTSS*) transcript mediates its stabilization via human antigen R (HuR)[93]. Increased editing and expression of *CTSS* are associated with atherosclerosis[93]. Further, transcriptome-wide analysis of RNA editing in human patients or mouse models reveals dysregulated RNA editing in many neurological disorders: autism spectrum disorder[74], Alzheimer's disease[73], spinal cord injuries[94], Fragile X syndrome[74,95], and schizophrenia[75], to name a few. Some of the RNA editing changes can be attributed to the regulation of ADARs' editing activities. In particular, the decreased RNA editing in Fragile X syndrome may be explained by a lack of Fragile X proteins FMRP, which interacts with ADARs and modulate RNA editing[74,95].

Future efforts on the identification, characterization, and manipulation of RNA editing in human diseases will offer tremendous opportunities in improving diagnostic, therapeutic, and prognostic approaches.

1.2.5 Current challenges in functional characterization of SNVs

MPRAs haven been widely adopted given its feasibility in high-throughput testing of molecular phenotypes in different cell types[96]. Many MPRAs focused on uncovering transcriptional mechanisms, where the design employs a minimal promoter with low basal activity, tailored to capturing genetic variants that enhance, rather than reduce, transcription[97]. Given their transcriptional focus, these MPRAs cannot uncover functional variants that regulate post-transcriptional processes[39,97–99]. In Chapter 2, we developed a massively parallel screen for 3' UTR variants that affect mRNA abundance post-transcriptionally.

Multiple functional characterization assays[99–102] prioritized for the testing of SNPs discovered by GWAS, which possessed adequate power to tackle relatively common genetic variants[103]. Common variants, however, usually have small phenotypic effects compared to rare variants due to negative selection that has shaped the relationship between effect size and minor allele frequency[104]. Rare variants are enriched in the neighboring regions of gene expression outliers across tissues, indicating its role in contributing to large gene expression changes[49]. However, the functional impact of rare variants on post-transcriptional gene regulation has not been studied. Chapter 2 fills in this gap by identifying functional rare variants in 3' UTRs that affect post-transcriptional gene regulation.

Although the functional impact of RNA editing sites in various disease systems is increasing recognized, systematic studies of the molecular function of RNA editing remain scare. In Chapter 3, we explored the function of differential RNA editing during epithelial-mesenchymal transition, a key process underlying tumor metastasis. We identified an editing-mediated mRNA stabilization mechanism for immune-related genes.

Most previous studies of RNA editing focused on sites occurring in coding exons, given their likely role in altering amino acid sequences. However, exonic RNA editing sites rely on intronic complementary sequences to form double-stranded RNA structures, which are recognized by the ADAR proteins[50,52,105]. Thus, exonic RNA editing sites are likely installed preceding RNA splicing, which renders them possible RNA variants that affect splicing regulation. This aspect of RNA editing function was rarely considered. Chapter 4 characterizes two exonic editing sites in PODXL, revealing multi-facet roles of these editing sites in promoting loss-of-function of PODXL in cancer.

CHAPTER 2

Massively parallel screen uncovers many rare 3' UTR variants regulating mRNA abundance of cancer-relevant genes

2.1 Abstract

Elucidating the functional impact of rare genetic variants, especially those in non-coding regions, represents a significant challenge. Here, we developed a massively parallel screen for rare 3' UTR variants (MapUTR) that affect mRNA abundance post-transcriptionally. Using two human cell lines, we assayed the function of 14,575 rare variants and found that 5,437 (37%) led to significant alterations of mRNA abundance in at least one cell line. These variants are enriched in miRNA target sites and binding sites of RNA-binding proteins, attesting to their functional relevance. Importantly, 71% of these variants are located in cancer-related genes. Further, 37 variants are associated with expression outliers in TCGA. Through prime editing, we characterized three variants in cancer-associated genes (*MFN2*, *FOSL2*, and *IRAK1*), confirming their impacts on mRNA stability. Importantly, all three variants significantly altered cellular proliferation, illustrating the effectiveness of MapUTR in pinpointing functional genetic variants.

2.2 Introduction

The mRNA transcript contains untranslated regions (UTRs) in its 5' and 3' end that regulate its stability, localization, and translation following DNA transcription[106,107]. The 3' UTR is usually

longer than the 5' UTR and expanded during evolution, indicating its essential roles in posttranscriptional regulation[108–110]. Indeed, more than 75% UTR variants identified by the genome-wide association studies (GWAS) are located in the 3' UTR region[24]. Most of the 3' UTR regulatory mechanisms are mediated by *trans*-acting factors, such as microRNAs (miRNAs) and RNA-binding proteins (RBPs)[109,111].

The development of massively parallel approaches enables the identification of *cis*-regulatory elements in 3' UTRs which are otherwise hard to predict without knowing the *trans*-factors[39,42,43,112–114]. Genetic variants altering the *cis*-regulatory elements could affect binding of *trans*-acting factors and lead to dramatic changes in gene expression and associated functional pathways. However, despite the increasing knowledge of *cis*-regulatory elements in 3' UTRs, it remains a major challenge to predict the functional impact of genetic variants in 3' UTRs. This is because such variants are surrounded by different genome contexts and may be involved in more than one regulatory axis. To this end, STARR-seq and its modified versions have been utilized to identify causal 3' UTR genetic variants in regulating DNA transcription[98,99,115,116]. In addition, several studies examined the functional impact of 3' UTR variants on mRNA abundance through post-transcriptional regulation[36,42,102]. However, most previous studies focused on tackling the function of common genetic variants[36,42,102]. Little attention has been given to rare 3' UTR variants, which constitute the majority of 3' UTR variants with unknown function. Rare variants are enriched in the neighboring regions of gene expression outliers across tissues (i.e., individuals with extreme expression levels different from the rest of the population), indicating their role in contributing to substantial gene expression changes[49].

Multiple massively parallel reporter assays (MPRAs) have been developed to measure the effects of 3'UTR sequences on mRNA abundance[36,40,42,45,102,114]. Specifically, sequences of interest were synthesized in parallel and cloned into reporter plasmids, then introduced to test

cells. mRNA abundance was estimated from the RNA-seq reads, which were normalized against the DNA-seq reads generated from the plasmids. Barcodes, first introduced by the MPRAs studying DNA transcriptional regulation[38,117–119], were utilized in some 3' UTR MPRAs to 1) label each sequence and 2) serve as additional replicates. However, different MPRA studies vary in barcode length and the number of barcodes per sequence[40,42,102,114]. While including barcodes may enhance the statistical power for the identification of functional variants, barcodes themselves may also be functional and alter mRNA abundance assigned to the test sequences. Thus, having too few barcodes per sequence would increase the false positive rate caused by barcode effects, while having too many barcodes would require deeper sequencing depth, which does not necessarily improve reproducibility when the number is beyond 200[120]. Another limitation in previous 3' UTR MPRA designs is that they cannot differentiate PCR duplicates from natural duplicates[36,40,42,102,114]. One way to solve this issue is to add a unique molecule identifier (UMI) to each molecule before amplifying sequencing libraries[45].

In this study, we developed a massively parallel screen for rare 3' UTR variants (MapUTR) regulating mRNA abundance. We adopted a barcode-free design and showed that MapUTR sensitively captures the regulatory effect of well-known 3' UTR regulatory elements. With MapUTR, we tested 14,575 rare variants, 5,437 of which were functional in at least one of the two cell lines included in this study. We uncovered an enrichment of functional rare variants in cancer-relevant genes, and experimentally characterized the function of 3 variants in altering cellular proliferation. Together, our study demonstrates the effectiveness of MapUTR in capturing functional rare variants in 3' UTRs and the potential contribution of such variants in cancer pathways.

2.3 Results

2.3.1 A method to identify functional 3' UTR variants regulating mRNA abundance

To test functional 3' UTR variants regulating mRNA abundance, we cloned the synthetic DNA oligos containing 3' UTR variants and their flanking sequences (158nt~164nt in total) into the 3' UTR of the *eGFP* gene (Fig. 1A, Methods). Since we are mostly interested in post-transcriptional regulation of mRNA transcripts by the 3'UTR variants, we used the CMV early enhancer/chicken beta actin (*CAG*) promoter[121], which is a strong promoter that drives high gene expression at a similar transcriptional rate for each variant. The plasmid library was electroporated into HEK293 or HeLa cells to test for mRNA abundance (Fig. 1A).

Unlike genome-integrated reporter assays, transient episomal reporter assays are limited in recapitulating the native physiological environment. If a large amount of plasmids were introduced into the cell, transient transfection may cause exhaustion of the cellular machinery. To find the minimum DNA/Cell ratio for this experiment, we performed cell electroporation at different DNA/Cell ratios in HEK293 cells and manually checked the RNA/DNA ratios of three control variants (previously studied in the literature[122–124]). We observed that transfecting 200ng plasmid DNA to 1M HEK293 cells yielded similar RNA/DNA ratios as reported in the literature[122–124] (Table S1). In contrast, lowering the DNA input affected the allelic ratio in DNA or RNA samples, potentially due to a lack of coverage of SNPs in these samples (Table S1). Based on these data, we chose to use 200ng DNA per 1M cells in further experiments. The same DNA/Cell ratio was used for HeLa cells. Compared to other transient MPRA methods which transfected 1ug ~ 5ug DNA per 1M cells[36,102,114,125], the amount of DNA we use in MapUTR avoids cellular machinery exhaustion.

Following electroporation of the plasmid library, total RNA was extracted for RNA-seq library generation targeting the tested 3' UTRs (Fig. 1A, Fig. S1). A DNA-seq library was also generated from the plasmid library (pre-transfection) to allow for normalizations of RNA expression (Fig. 1A). To remove PCR duplicates, we incorporated 15-mer UMIs during the initial stage of DNA/RNA-seq library generation (Fig. S1). In the data analysis steps, UMIs were extracted to enable removal of PCR duplicates, followed by read alignment, data normalization and detection of functional variants (Fig. 1B, Methods).

2.3.2 MapUTR captures functional effects of random mutations within known *cis*-regulatory elements in the 3' UTR

To test the performance of MapUTR, we picked 5 well-known 3' UTR motifs (Table S2) reported in previous literature[42] and mutated each base individually (to each of the 3 possible nucleotides) within the motif and its surrounding regions (22-23 nt on each side, Fig. 1C). The resulting pool of oligo sequences was tested with MapUTR in both HEK293 and HeLa cells. For all 5 motifs, we obtained high-quality sequencing data signified by the low mismatch rate relative to the reference sequence (0.057% on average) outside of the mutated regions (Fig. 1D, Fig. S2A). Since we designed single-nucleotide mutations to include all 3 alternative alleles, the high mismatch rates in the mutated regions are expected. This observation also indicates that the oligo synthesis and subsequent experimental processing steps induced generally low error rates. Next, we calculated the impact of each mutation in the known motifs and their surrounding sequences on mRNA abundance. As shown in Fig. 1E and Fig. S2B, mutations in the functional motifs induced considerable changes in relative mRNA expression, whereas the flanking regions are associated with relatively small mutation-induced variations. This result

supports the effectiveness of MapUTR in capturing biologically relevant post-transcriptional regulatory events.

2.3.3 Identification of functional rare 3' UTR variants with MapUTR

We next applied MapUTR to test the functionality of rare genetic variants in the 3' UTR. From the Exome Aggregation Consortium (ExAC)[126], we extracted 54,959 rare 3' UTR variants defined as those with an adjusted minor allele frequency (adjAF) < 0.01. After removing sequences incompatible with the cloning strategy (i.e., share similarity with restriction enzyme sites or primer sequences, see Methods and Table S3), we selected 14,575 variants to be tested with MapUTR. Among them, 1,032 variants were also reported in clinically relevant databases (ClinVar[127], CIViC[128], COSMIC[129], iGAP[130]) and were therefore prioritized for testing with MapUTR.

We designed synthetic oligonucleotides (200nt in length) harboring the rare 3' UTR variants, their flanking regions (164nt in length), subpool primers and restriction sites. Each variant was located at the center of its flanking sequences, unless adjustment was necessary to avoid including sequences beyond the 3' UTR for variants close to the boundaries (Fig. 2A). We obtained high-quality sequencing data with an average mismatch rate of approximately 0.016% per position relative to the reference sequences, except at the ends of the reads where sequencing errors are known to be more prevalent[131] (Fig. 2B). Upon data processing (see Methods), we calculated an activity score for each reference or alternative allele as the normalized relative read number (RNA/DNA) for that allele. We then compared the impact of the reference and alternative alleles on mRNA abundance, and calculated the relative activity (lnFC) between the activity scores of the alternative allele relative to the reference. The relative

activity of biological replicates were highly correlated, signifying MapUTR's ability to consistently capture important biological events with regard to the regulation of mRNA abundance (Fig. 2C).

Next, in both HEK293 and HeLa cells, we evaluated the difference in RNA abundance between reference and alternative alleles using MPRAnalyze[132]. We tested 14,490 and 14,494 variants, out of which 3,066 (21.2%) and 3,944 (27.2%) altered mRNA abundance significantly in HEK293 and HeLa, respectively (Fig. 2D), with 5,437 (37%) being significant in at least one cell line. These functional rare variants were harbored in 3,487 genes, representing 50.4% of all tested genes. Among the functional variants identified in HEK293 and HeLa, 51.2% and 51.7%, respectively, had higher expression associated with the variant allele, whereas the variant alleles of the remaining 48.8% and 48.3% downregulated mRNA expression.

We next examined the function of 3' UTR variants between the two cell lines. To this end, we correlated the relative activity scores of the 1,573 (28.9% of the 5,437) variants that were functional in both cell lines. A significant correlation was observed, and the majority of variants shared the same direction of change between the two cell lines (Fig. 2E). Thus, the genetic background, rather than *trans*-acting factors, plays a dominant role in determining the function of many 3' UTR variants. Nonetheless, 92 variants showed opposite directions in their relative activity scores, which may indicate potential tissue-specific regulation of RNA abundance in HEK293 and HeLa cells.

2.3.4 Function variants in 3' UTRs alter miRNA target sites

The 3' UTR is known to harbor *cis*-regulatory elements that recruit *trans*-acting factors, usually miRNAs and RBPs, for post-transcriptional regulation of gene expression[107]. To investigate the

potential mechanisms via which rare functional 3' UTR variants affect mRNA abundance, we overlapped the functional variants with miRNA target sites predicted by TargetScan (release 7.2)[133]. In HEK293 and HeLa cells, 62.4% and 62.7% of all functional 3' UTR variants overlapped predicted miRNA target sites, respectively. For these targets, the alternative alleles of the variants are expected to disrupt miRNA targeting and lead to enhanced mRNA abundance. Consistent with this expectation, in both HEK293 and HeLa, we observed a significant bias toward upregulation of mRNA abundance by the alternative alleles (Fig. 3A). As examples, Figure 3B shows two miRNA-target pairs. One miRNA, miR-34b-3p, is predicted to target the *PLIN4* transcript that harbors a rare variant (*rs767768172*) in the miRNA seed match region. MapUTR revealed significantly higher mRNA abundance associated with the alternative allele, consistent with the expected derepression by the miRNA in the presence of the rare variant. Similarly, another rare variant (*rs145078776*) is predicted to disrupt the binding of miR-3180-5p to the *LDHD* transcripts (Fig. 3B. It should be noted that both genes have important disease relevance, with *PLIN4* implicated in skeletal muscle disease[134] and *LDHD* involved in clear cell renal carcinoma[135].

2.3.5 Function variants in 3' UTRs alter RBP binding sites

Next, we investigated the role of RBPs in affecting mRNA abundance of genes harboring rare functional 3' UTR variants. To relate the functional variants to specific RBPs, we first conducted motif analyses using HOMER[136] to identify overrepresented hexamers among sequences that upregulate or downregulate mRNA expression (see Methods). For sequences that downregulate mRNA expression, we identified well-defined destabilizing motifs such as the AU-rich and GU-rich elements (Fig. S3A, S3C). In contrast, CU-rich, CA-rich, and GA-rich elements,

which are known stabilizing motifs[137], were enriched among sequences that upregulate mRNA expression. These results support the validity of the MapUTR experiment.

Next, we associated the above motifs with RBPs using previously published RNA Bind-n-Seq (RBNS)[138] data where binding motifs of RBPs were characterized experimentally (Fig. S4A, B). We then evaluated whether the alternative alleles of each functional variant alter RBP binding using the DeepRiPe model[139] (Methods). On the global level, we observed that the rare functional variants significantly altered RBP binding compared to random controls (Fig. 3C, Methods). We further examined whether the allele-specific effects of the functional variants detected by MapUTR are concordant with predicted RBP binding alteration by DeepRipe. A number of RBPs showed significant concordance. For instance, higher ZFP36 binding to the minimal ARE (UAUUUA) motif was associated with lower mRNA abundance (Fig. 3D), consistent with the destabilizing function of AREs[140]. We also observed that elevated binding of RBP TIA1 to the CUCUUU motif was strongly associated with the augmentation of mRNA abundance, consistent with the stabilizing function of TIA1[141]. Altogether, these findings support the utility of MapUTR to accurately identify functional effects of variants that are explainable by RBP binding.

2.3.6 Disease relevance of functional rare 3' UTR variants

Genome-wide association studies (GWAS) have primarily identified common variations contributing to the etiology of complex traits[103]. Common variants, however, usually have small phenotypic effects compared to rare variants due to negative selection that has shaped the relationship between effect size and minor allele frequency[104]. To understand the disease relevance of functional rare 3' UTR variants identified by MapUTR, we investigated their

association with GWAS SNPs. Briefly, we analyzed functional 3' UTR variants that are in linkage disequilibrium (LD) and within 200kb of known GWAS SNPs. We also sampled both common and rare random SNPs as controls and repeated the GWAS LD analysis. We noticed that ~68% and ~69% of functional 3' UTR variants in HEK293 and HeLa, respectively, are associated with GWAS SNPs and these percentages are significantly higher than the proportion observed in either common or rare control SNPs (Fig. 4A). These findings suggest the potential relevance of 3' UTR variants to diseases.

We next performed a Gene Ontology enrichment analysis with the top 500 variants ranked by their absolute fold changes in MapUTR of either HEK293 or HeLa cells. We observed that genes containing large-effect rare 3' UTR variants are enriched in cell survival-related terms such as cell growth and cell death (Fig. 4B, Fig S5A-B), indicating a close relationship between these functional rare variants and diseases. Next, we asked if the genes containing functional variants are enriched in certain diseases. We extracted gene-disease associations from the DisGeNET database[142]. Interestingly, we found that cancer is the most represented disease (Fig. 4C). Overall, 3,841 (70.65%) functional rare variants are in genes associated with cancer. Although this observation may reflect the fact that cancer has dominated the gene-disease associations reported in DisGeNET, it is still likely that the functional rare 3' UTR variants play a role in tumorigenesis.

To further examine the relevance of functional rare variants to cancer, we examined whether cancer drive genes contain functional MapUTR variants. Among the 568 cancer driver genes in the Integrative OncoGenomics (IntOGen) database[143], 124 had functional MapUTR variants (182 variants). Importantly, 18 oncogenes harbored functional variants leading to an increase of gene expression levels, whereas 33 tumor suppressor genes had variants leading to a decrease of gene expression levels, potentially contributing to tumorigenesis (Fig. 4D). These

results suggested that the functional rare 3' UTR variants identified by MapUTR are closely relevant to human diseases, especially cancer.

2.3.7 Individuals with functional rare 3' UTR variants are gene expression outliers in TCGA

Given the close relevance of functional rare 3' UTR variants to cancer, we wondered if cancer patients carrying the MapUTR variants showed any gene expression changes. We obtained genotype data including both germline and somatic mutations in The Cancer Genome Atlas (TCGA) from Pan-Cancer Analysis of Whole Genomes (PCAWG)[144]. For each MapUTR variant that was found in patients of a certain cancer type, we extracted gene FPKM values in patients with either the reference or the alternative allele and calculated gene expression z-scores for outlier detection. In total, we found 37 functional variants, mostly germline, showing up in patients whose gene expression value is both an outlier and consistent with the directional change reported in MapUTR (Fig. 4E, Fig. S5C). Among them, 12 variants are identified as functional variants in both HEK293 and HeLa cells. Interestingly, the functional variant in the gene *SDF4* was found in gene expression outliers in multiple cancer types (Fig. S5C). Importantly, 10 genes with functional rare 3' UTR variants in TCGA gene expression outliers have been reported to be associated with cancer (Fig. 4E). Together, these results demonstrated the existence of functional rare 3' UTR variants in cancer patients. Supported by the gene expression outlier analysis, we showed that MapUTR identifies causal variants regulating mRNA abundance.

2.3.8 Functional rare 3' UTR variants in *MFN2*, *FOSL2*, and *IRAK1* regulate mRNA stability and cell proliferation

To further relate MapUTR variants with in vivo function, we experimentally tested three variants in cancer-associated genes. The first gene Mitofusin 2 (*MFN2*) encodes a mitochondrial membrane protein regulating mitochondria fusion[145]. MFN2 has anti-tumor effects and is downregulated in multiple cancers[146]. Previous studies found that MFN2 inhibits cell proliferation by suppressing mTORC2/Akt or Ras-NF-κB signaling pathways[147,148]. In MapUTR, we identified a functional rare 3' UTR variant (*rs777822288*) in *MFN2*, leading to a significant increase in mRNA expression (Fig. 5A). We hypothesized that this variant may also play a role in inhibiting cell proliferation. The second gene, FOS Like 2 (*FOSL2*), encodes a protein serving as a subunit of the transcription factor complex AP-1[149]. FOSL2 promotes cell proliferation, migration, and invasion in breast cancer and ovarian cancer[150,151]. We discovered one rare variant (*rs11884725*) in the 3' UTR of FOSL2, which showed higher activity scores (RNA/DNA) compared to the reference allele (Fig. 5A). This variant may facilitate cell proliferation by upregulating *FOSL2* gene expression. The third gene encodes interleukin-1 receptor-associated kinase 1 (IRAK1), involved in toll-like receptor and interleukin-1 signaling pathway[152]. Overexpressed in several cancers, IRAK1 is a therapeutic target, whose inhibition impairs tumor growth and metastasis[152]. We identified a rare 3'UTR variant (*rs782486025*) in *IRAK1* that significantly decreased mRNA expression (Fig. 5A). This variant may serve as an allele-specific 'inhibitor' for IRAK1, thus reducing cell proliferation.

To measure the effect of functional variants in their native genomic context, we utilized prime editing[153–155] to introduce the MapUTR variants into the genome of HEK293T cells, which provide optimal editing efficiency with prime editing (Fig. 5B). In addition, HEK293T is a daughter cell line derived from HEK293 cell line, in which all three MapUTR variants were identified as functional candidates (Fig. 5A). Thus, HEK293T cells likely possess the *trans*-factors required for the MapUTR variants to be functional and are therefore chosen as the cell

line for prime editing validation. The genome-edited bulk HEK293T cells were diluted and plated to isolate single-cell clones for both the reference and variant alleles in each gene (Fig. 5B). To avoid potential bias due to off-target effects in a specific single-cell clone, we picked 4 ~ 6 single-cell homozygous clones for each allele (Fig. 5C-E).

Through quantitative reverse transcription PCR (qRT-PCR), we measured the mRNA stability of each gene in the single-cell clones by treating the cells with actinomycin D (ActD) to block cell transcription[156] for different time periods (2 h, 8 h, and 24 h) (Fig. 5F-H). Starting at 2 h post-ActD treatment, we observed significantly lower mRNA expression levels of *IRAK1* in the single-cell clones with *IRAK1* variant alleles compared to those with *IRAK1* reference alleles (Fig. 5H). This observation is consistent with MapUTR, in which the *IRAK1* variant allele had a lower activity score (RNA/DNA) compared to the reference allele (Fig. 5A). For *MFN2* and *FOSL2*, we observed a significant increase in mRNA expression levels in clones with the variant alleles at 8 h or 24 h post-ActD treatment (Fig. 5F, G), which are consistent with the higher MapUTR activity scores in the variant alleles for these two genes (Fig. 5A). These results confirm that the MapUTR variants in *MFN2*, *FOSL2*, and *IRAK1* regulate mRNA stability in HEK293T cells.

We next examined the functional impacts of the 3 MapUTR variants on cellular phenotype. To this end, we performed cell proliferation assays using the single-cell clones with either reference or variant alleles of *MFN2*, *FOSL2*, and *IRAK1* (Fig. 5B). We found that single-cell clones with the variant alleles of all three genes showed significantly altered cell proliferation profiles (Fig. 5I-K). Specifically, we observed reduced cell proliferation in clones with variant alleles of *MFN2* (Fig. 5I) and *IRAK1* (Fig. 5K), as well as increased cell proliferation in clones with variant alleles of *FOSL2* (Fig. 5J). Importantly, the directions of the cell proliferation change are consistent with the expected consequence of each variant, based on their effects on mRNA

stability and previous studies on the roles of *MFN2*, *FOSL2*, and *IRAK1* in cell proliferation[148,151,152]. Together, these findings support the in vivo function of MapUTR variants in HEK293T cells.

2.4 Discussion

Rare variants constitute the majority of human genetic variants[3]. Yet, little is known about the function of non-coding rare variants due to their scarcity in individuals, which has made rare variant association tests challenging[103]. In this study, we introduced MapUTR, a massively parallel reporter assay with optimized designs to identify 3' UTR variants regulating mRNA abundance post-transcriptionally. Complementary to the existing studies on common 3' UTR variants[36,42,102], we tested 14,575 rare 3' UTR variants and identified 5,437 (37.3%) rare variants altering mRNA abundance in HEK293 or HeLa (or both). These functional variants are of high disease relevance supported by their associations with GWAS SNPs and TCGA expression outliers. Interestingly, many functional rare MapUTR variants are located in cancer-related genes. Further, we characterized three variants in cancer-associated genes, demonstrating their functional impact on mRNA stability and cell proliferation in the native genome context.

In our optimized design, a strong promoter was incorporated to increase the basal activity of DNA transcription, allowing the detection of post-transcriptional effects of 3' UTR variants in both directions. We also utilized UMI during library generation to avoid potential PCR amplification bias. Compared to a previous study in which over 90% of functional variants showed increased expression[97], our assay reported an unbiased detection of both upregulating (51.5%, mean of two cell lines) and downregulating (48.5%, mean of two cell lines) functional variants. Similar to a previous 3' UTR MPRAs that measured the functional impact of *cis*-

regulatory elements on mRNA stability[42], MapUTR was able to capture alterations of mRNA abundance caused by random mutations within known functional motif (Fig. 1E). To best mimic the physiological relevant environment, we employed a minimal DNA/Cell ratio during cell electroporation. Consistent with the MapUTR measurements, we showed that three functional variants altered mRNA stability in their native genomic context. Importantly, we also observed altered cell proliferation when the functional variants are introduced, which may be explained by the mRNA stability alteration, given its important role in controlling cell proliferation[157,158].

Rare variants are enriched in gene expression outliers across tissues compared to non-outliers[49,159], indicating their pivotal role in regulating gene expression. Among all tested rare variants (14,575) in our experiments, 5,437 (37.3%) led to significant mRNA abundance alterations in at least one cell line, with an average of 24.2% variants identified as functional in each cell line. This prevalence of functional variants is much larger than previously reported for common 3' UTR variants[160], where 19.45% (2,368 out of 12,173 total) functional variants were reported to be significant in at least one of six human cell lines, with an average of 5.7% functional rate in each cell line. While the functional proportion may be arbitrary resulting from differences in the methods applied to call functional variants in these two assays, a similar cutoff of an adjusted p-value less than 0.1 was applied in both assays, with the MapUTR having an additional cutoff at |lnFC|>=0.1, determined using random mutagenesis control variants in 5 well-known motifs. The higher proportion of functional variants in the rare variant screen may reflect the large effects of rare variants on gene expression, which may be explained by the purifying selection during evolution[104]. This observation is in line with a previous study that assessed the contribution of alleles from different allele frequencies to gene expression in lymphoblastoid cell lines[161], revealing a higher contribution of rare variants to gene expression heritability.

It is hypothesized that rare variants contribute to the missing heritability of complex diseases[162]. In our study, we found that around 70% of functional rare 3' UTR variants are associated with GWAS SNPs, suggesting potential contributions of these rare variants to human traits. Interestingly, many functional rare 3' UTR variants (~70%) are in cancer-relevant genes. Genes harboring large-effect rare variants are enriched in Gene ontology terms of cell growth and cell death (Fig. 4B, Fig. S4A-B), which are often dysregulated in cancer cells. Altered gene expression, regulated by either DNA transcription or mRNA stability[163], is one of the major changes in cancer[164]. In a previous study on rare predisposition variants across 33 cancer types, around half of the variants located in tumor repressor genes or oncogenes are associated with low or high gene expression, respectively[165]. Notably, our study identified many functional rare 3' UTR variants that may contribute to cancer by regulating mRNA abundance. For example, 182 functional rare 3' UTR variants are in cancer driver genes (Fig. 4D). In addition, we found 37 functional variants in TCGA gene expression outliers. Particularly, we experimentally validated three rare variants in cancer-associated genes (*MFN2*, *FOSL2*, and *IRAK1*), confirming their functional roles in regulating mRNA stability and cell proliferation. Future studies characterizing these functional rare 3' UTR variants will help to elucidate their causality in cancer.

In conclusion, we present MapUTR, an optimized massively parallel reporter assay that can be used to identify functional 3' UTR variants regulating mRNA abundance post-transcriptionally. With simple adaptions such as cell fractionation, MapUTR can also be utilized to identify functional 3' UTR variants regulating mRNA localization. In addition, the usage of MapUTR is not limited to genetic variants. Other types of RNA variants, such as RNA editing sites, can also be tested in MapUTR. We demonstrated that many functional rare MapUTR variants are cancer relevant. In general, the discoveries from MapUTR will facilitate prioritizing

candidates for causality characterizations, explaining heritability for complex diseases, and training computational models for predicting functional 3' UTR variants.

2.5 Methods

2.5.1 Design of DNA oligos with random mutations within well-known motifs

Known 3' UTR *cis*-regulatory elements (Table S2) were chosen from previous literature[42]. To test if MapUTR could capture the regulatory effects of these motifs, we designed oligos containing random mutations at every base within the regulatory motif region as well as its flanking regions (22-23nt upstream and 22-23nt downstream). Each oligo is 200nt in length, with 158nt being the actual 3' UTR sequences containing the variant of interest. The rest of the oligo contain forward primer binding site (21nt), reverse primer binding site (15nt), and restriction enzyme site (6nt) for cloning. All oligos were included in chip3 and synthesized by Twist Bioscience.

2.5.2 Design of DNA oligos containing rare 3' UTR variants

We extracted human variants from the ExAC[126] database and excluded indels using GATK SelectVariants tool[166]. Further, with a threshold of adjusted minor allele frequency (MAF) less than 0.01, we obtained 1,017,886 rare variants. Based on GENCODE[167] basic v24 annotation, we selected 54,959 SNPs that are annotated in the 3' UTR. To avoid unwanted enzyme digestion and amplification within the oligos, we filtered out sequences that contain restriction enzyme sites and subpool primer sequences (Table S3). We overlapped rare 3' UTR ExAC variants with a collection of clinically relevant variants reported in ClinVar[127], ClViC[128], COSMIC[129], and iGAP[130]. The resulting 1032 SNPs, which we referred to as clinically relevant

rare 3' UTR variants, were prioritized for final testing. In total, we designed 14,575 variants separated into two chips (chip1 & chip2) for synthesis by Twist Bioscience. Each chip contains 3-4 subpools with different 5' and 3' adaptors that can be amplified using subpool primers respectively (See supplemental protocol). Both reference and alternative alleles for each variant were included in the same subpool. Each oligo is 200nt in length, with 164nt being the flanking sequence centered around the variant of interest. The rest of the oligo contain forward subpool primer binding site (15nt), reverse subpool primer binding site (15nt), and restriction enzyme site (6nt) for cloning.

2.5.3 Generation of MapUTR master plasmid

We extracted human variants from the ExAC[126] database and excluded indels using GATK SelectVariants tool[166]. Further, with a threshold of adjusted minor allele frequency (MAF) less than 0.01, we obtained 1,017,886 rare variants. Based on GENCODE[167] basic v24 annotation, we selected 54,959 SNPs that are annotated in the 3' UTR. To avoid unwanted enzyme digestion and amplification within the oligos, we filtered out sequences that contain restriction enzyme sites and subpool primer sequences (Table S3). We overlapped rare 3' UTR ExAC variants with a collection of clinically relevant variants reported in ClinVar[127], ClViC[128], COSMIC[129], and iGAP[130]. The resulting 1032 SNPs, which we referred to as clinically relevant rare 3' UTR variants, were prioritized for final testing. In total, we designed 14,575 variants separated into two chips (chip1 & chip2) for synthesis by Twist Bioscience. Each chip contains 3-4 subpools with different 5' and 3' adaptors that can be amplified using subpool primers respectively (See supplemental protocol). Both reference and alternative alleles for each variant were included in the same subpool. Each oligo is 200nt in length, with 164nt being the flanking sequence centered around the variant of interest. The rest of the oligo contain forward subpool

primer binding site (15nt), reverse subpool primer binding site (15nt), and restriction enzyme site (6nt) for cloning.

2.5.4 Cloning of synthesized oligos into MapUTR master plasmids

We resuspended the chip oligos with ultrapure distilled water (Thermo Fisher Scientific, Cat# 10977015) at a final concentration of 1ng/µl. Each subpool was amplified using subpool-specific primers (See supplemental protocol). The reverse subpool primer contains a BamHI restriction enzyme site, which allows subsequent digestion and ligation into the master plasmid. To avoid potential bias due to over-amplification, we first assembled qPCR reactions with PowerUp™ SYBR® Green Master Mix (Thermo Fisher Scientific, Cat# A25742) with 1ng oligos as templates in a 50µl reaction. We determined the cycle number where the slope of the amplification curve began to decrease. We repeated the PCR with Q5 polymerase (NEB, Cat# M0492L) using the cycle number determined from the qPCR pre-run, which is usually 17-19 cycles. The PCR products were cleaned up using the DNA Clean & Concentrator kit (Zymo Research, Cat# D4004).

Next, DNA digestion reactions were set up for both PCR products (oligo inserts) and master plasmids using EcoRI-HF (NEB, Cat# R3101S) and BamHI-HF (NEB, Cat# R3136S), followed by incubation at 37 °C overnight. All digestion reactions were terminated with enzyme heat inactivation at 65 °C for 20min. For purification, the digested plasmids were resolved in 1% agarose gel and the desired band was gel purified using Zymoclean™ Gel DNA Recovery Kit (Zymo Research, Cat# D4002). The digested PCR products were directly cleaned up using the DNA Clean & Concentrator kit (Zymo Research, Cat# D4004). Cleaned-up PCR products and digested plasmids were ligated at a 10:1 molar ratio with T7 DNA ligase (NEB, Cat# M0318).

Ligation reactions were incubated at 25 °C on a thermal cycler for one hour, followed by a clean-up using the DNA Clean & Concentrator kit (Zymo Research, Cat# D4004), with water elution.

Finally, the purified ligation products were electroporated into 10-beta Electrocompetent *E. coli* (NEB, Cat# C3020K) using Gene Pulser Xcell Electroporation Systems (NEB, Cat# 1652660) at 2.0 kV, 200 Omega, and 25 μF. The transformed *E. coli* were spread onto 150mm selective plates at 37 °C overnight. Colonies with at least 100X coverage of the oligo library (e.g., 0.2M colonies for 2000 designed oligos) were harvested for plasmid isolation using ZymoPURE II Plasmid Midiprep Kit (Zymo Research, Cat# D4200). For the initial quality check, the isolated plasmid library was sent for Sanger sequencing with a sequencing primer (See supplemental protocol) complementary to the polyA signal region shared by all plasmids.

2.5.5 Cell culture and electroporation

HEK293 and HeLa cells were maintained in DMEM (Gibco, Cat# 11995065) with 10% FBS (Gibco, Cat# 26140079) and antibiotic-antimycotic reagent (Gibco, Cat# 15240062) at 37 °C with 5% CO_2 supply. Cells were passaged the day before electroporation to make sure they are actively dividing by the time of electroporation. Prior to electroporation, HEK293/HeLa cells were disassociated with Trypsin-EDTA (Gibco, Cat# 25300120), washed with growth media, and resuspended in OptiMEM (Gibco, Cat# 31985062) at a cell density of 10M/ml. For a typical subpool library with 0.2M colonies, 3ug plasmid library was electroporated into 15M HEK293/HeLa cells (See DNA/Cell ratio optimization below) for each biological replicate, for a total number of three biological replicates. Cell electroporation was done using the Gene Pulser Xcell Electroporation System (NEB, Cat# 1652660) with the following settings: square wave,

25msec, 220V, 0.4cm. After electroporation, the HEK293/HeLa cells were incubated in growth media at 37 °C for 24 hours.

2.5.6 mRNA isolation

24 hours after electroporation, HEK293/HeLa cells were lysed using TRIzol (Thermo Fisher Scientific, Cat# 15596026). Each 500 µl TRIzol-lysed solution was mixed with 100 µl chloroform (Fisher Chemical, Cat# C298-500) to allow phase separation. The upper aqueous phase was transferred and mixed with equal volume ethanol (200 proof, Fisher BioReagents). The mixture was loaded to the column supplied by the Direct-zol RNA Miniprep Plus kit (Zymo Research, Cat# R2072) to isolate total RNA following the manufacturer's protocol. PolyA selection was carried out to isolate mRNA from total RNA using Dynabeads™ Oligo$(dT)_{25}$ (Thermo Fisher Scientific, Cat# 61002). The concentration of mRNA in each sample was quantified using the Qubit RNA HS Assay Kit (Thermo Fisher Scientific, Cat# Q32852) with the Qubit Fluorometer (Thermo Fisher Scientific).

2.5.7 Generation of UMI-containing libraries

To measure mRNA abundance, we generated UMI-containing libraries from the plasmid library (DNA) before electroporation, as well as mRNA isolated after electroporation. The mRNA was reverse transcribed into cDNA with the SuperScript™ IV Reverse Transcriptase (Thermo Fisher Scientific, Cat# 18090010) using a gene-specific reverse transcription (RT) primer (MPP3) that contains a 15-mer unique molecular identifier (UMI), which was synthesized as -NNNNNNNNNNNNNNN- (See supplemental protocol for primer sequences). After RT, mRNA was removed via RNase H treatment.

Both cDNA and plasmid DNA underwent two rounds of PCR (Fig. S1). The first-round PCR (2-3 cycles) utilized primers (MPP2 & MPP3) to add UMIs to cDNA or plasmid DNA samples. We assumed that there is little PCR amplification bias in the initial 3-cycle UMI addition step. The first-round PCR products were cleaned up using the DNA Clean & Concentrator kit (Zymo Research, Cat# D4004). Next, a second-round PCR was performed using the purified first-round PCR products and primers (MPP2 & MPP4), which added sample indexes and Illumina sequencing adaptors (P5/7). To avoid over-amplification, a pilot reaction for the second-step PCR was performed using PowerUp™ SYBR® Green Master Mix (Thermo Fisher Scientific, Cat# A25742) and ran on a qPCR thermal cycler. An amplification curve was obtained for each sample to determine the cycle number before the plateau. The second-round PCR was then conducted using the cycle number (or less) determined from the qPCR pre-run (See supplemental protocol). All PCR steps for sequencing library generation were performed using the Q5 polymerase (NEB, Cat# M0492L).

PCR reactions for the same sample were pooled and purified using the DNA Clean & Concentrator kit (Zymo Research, Cat# D4004). Purified PCR products were resolved on 2% agarose gel and the band at the expected library size (377bp) was cut out and purified using the Zymoclean™ Gel DNA Recovery Kit (Zymo Research, Cat# D4002). UMI measurement libraries made from DNA/RNA were mixed and sequenced using custom sequencing primers (See supplemental protocol) on Hiseq3000 PE150 or Novaseq SP PE150 with 15% PhiX spike-in.

2.5.8 DNA/Cell ratio optimization

To optimize the DNA/Cell ratio during cell electroporation, different amounts of plasmid libraries (i.e., 2ng, 10ng, 62.5ng, 200ng, and 1ug) were electroporated into 5M HEK293 cells, respectively. Total RNA was isolated 24 hours post electroporation. UMI-containing libraries were generated from plasmid libraries before electroporation and mRNA isolated from electroporated cells (see details above). To check the allelic ratios for the three control SNPs (APP chr21:27253559 minus strand G>A, ABCB1 chr7:87133366 minus strand A>G, CYP2A7 chr19:41381398 minus strand G>A), each control gene was amplified by a gene-specific reverse primer and the common P5 forward primer (See Table S4). To avoid overamplification of the target genes in the libraries, a qPCR reaction was assembled for each condition to determine the cycle number before the plateau. Control genes were then amplified from each UMI measurement library with the cycle numbers determined by qPCR using the Q5 polymerase (NEB, Cat# M0492L). PCR amplicons were gel-purified and sent for Sanger sequencing. The allelic ratio for each SNP was estimated based on the peak signal for each base, which was quantified using 4Peaks. RNA/DNA ratios were calculated by dividing the allelic ratios in the RNA samples by the allelic ratios in the DNA samples.

2.5.9 Mismatch rate analysis for DNA and RNA reads

To assess the quality of sequencing data obtained from MapUTR, we examined the mismatch rate at any given position along the length of the design sequences. For each nucleotide position that was covered by sequencing reads, we calculated the percent mismatch as follows:

$$Mismatch\ rate\ (i) = \frac{Number\ of\ mismatches\ at\ position\ i}{Number\ of\ reads\ covering\ position\ i} \times 100$$

This calculation excludes the 15nt primer sequences on the ends of the design sequence.

2.5.10 Estimation of variant effect sizes

Paired-end reads of 150nt each were obtained for 3 replicates each of DNA and RNA libraries. Read 1 contains a UMI (15nt), reverse transcription (RT) primer (14nt), REC2 restriction enzyme site (6nt), subpool primer (15nt), and 100nt of the design sequence. Read 2, however, consists entirely of the design sequence. UMIs, together with the RT primer, REC2 restriction enzyme site, and subpool primer, were extracted from read 1 and added to the read name using UMItools[168]. The reads were then aligned to the reference sequences using Bowtie 2[169], allowing up to 1 mismatch per alignment. Since both reference and alternative alleles were designed in our library, it would be challenging to differentiate between reads with the designed SNV and those with sequencing errors at the same position as the designed SNV. To address this, we only used perfectly mapped reads and reads with 1 mismatch as long as the mismatch does not occur in the same position as the designed SNV.

Further, we removed PCR duplicates by only retaining one of the multiple reads with the same UMI that map to the same reference sequence. The UMIs were then counted in the DNA- and RNA-seq libraries, and the counts were quantile-normalized across the 3 replicates. For each tested allele, we calculated the regulatory activity as follows:

$$Activity\ Score = \frac{C_{RNA}}{C_{DNA}},$$

where C_{RNA} is the normalized RNA counts and C_{DNA} is the normalized DNA counts for the allele. To quantify the relative effects of rare variants on steady-state mRNA abundance, we modeled RNA counts as a function of DNA counts for both reference and alternative alleles using

MPRAnalyze[132]. To call functional variants, we required 10% increase or decrease (|lnFC| >= 0.1) in RNA abundance in the alternative allele relative to the reference and FDR <=0.1.

2.5.11 Motif discovery

The reference and variant sequences were separated for motif analysis depending on their observed effect in the initial MapUTR analysis. Briefly, we compared the mRNA abundance of the reference and alternative allele of each rare variant, and categorized the sequence containing the alternative allele as an upregulating sequence (and that with the reference allele as a downregulating sequence) if the alternative allele yielded higher expression than the reference, and *vice versa*. Subsequently, we obtained the sequence around the variant position by taking the 5 bases upstream and downstream of the variant, making a total of 11 bases, for the reference and alternative alleles respectively. In this way, the reference and alternative alleles of the rare variants were appropriately included in the search for overrepresented 6mers in the upregulating and downregulating sequences. The reference and variant sequences deemed upregulating were combined to make one large superset, and the same was done for the downregulating sequences. Then, we did a *de novo* motif search of the RNA sequences with HOMER[136], limiting the number of motifs to 25. Here, we used the downregulating sequences as background when identifying motifs in the upregulating set and *vice versa*.

2.5.12 Integrative analyses of RBPs and discovered motifs

We then examined whether functional variants from MapUTR are located in RBP binding sites as follows. First, we extracted predicted RBP binding sites by PrismNet[170], which overlapped 26.6% and 22.4% of the functional variants in HEK293 and HeLa cells, respectively. Second, we defined a functional variant to be located in an RBP binding site if the variant overlaps a

motif of the RBP (defined by RBNS) and resides in PrismNet-predicted binding site of the same RBP.

The HOMER-identified motifs (i.e., kmers) were matched with motifs (i.e., kmers) of each RBP tested in a previous RBNS experimental[138]. For each variant, we used the associated motifs identified by HOMER and the associated RBPs from RBNS to assess the effect of the variant on RBP binding. Specifically, we used the DeepRipe model[139] to calculate the predicted difference in RBP binding between the reference and the alternative alleles. As controls, an equal number of random rare dbSNPs were sampled per chromosome. For the controls, RBP overlaps were simulated as follows. For each RBP, N (the same number of true RBP-motif overlaps for functional variants) random control SNPs were chosen. Then the control SNPs were scored with DeepRiPe similarly as for functional variants. The distribution of absolute changes in binding (reference vs. alternative alleles) was compared between the functional variants and random control SNPs. These steps were repeated separately for upregulating and downregulating variants in HEK293 and HeLa respectively.

2.5.13 Analysis of functional 3' UTR SNPs in LD with GWAS SNPs

The GWAS catalog[18] was filtered to retain significant SNP-trait associations ($p < 5.0 \times 10^{-8}$). In addition, GWAS SNPs were assigned LD blocks according to the CEU population LD structure, requiring $R^2 \geq 0.9$ and $D' \geq 0.9$. Then the number of functional 3' UTR variants in LD with at least one GWAS SNP and within 200kb of the GWAS SNP was counted. As controls, we randomly sampled a similar number of common or rare SNPs from dbSNP and computed their overlaps with GWAS SNPs. Using Fisher's Exact test, the proportion of overlap was compared between the functional set and each of the control sets.

2.5.14 Gene ontology (GO) enrichment analysis

A union list of functional variants from HEK293 and HeLa cells was generated to select the top 500 variants ranked by their absolute values of fold changes. The resulting 339 genes were chosen as the query genes for GO enrichment analysis. For each query gene, a control gene was randomly chosen among the background genes (excluding the query genes) tested in MapUTR. In this way, a control set of genes was constructed that has the same number of genes as the query set. This process was repeated 10,000 times. The *p*-value of enrichment of each GO term in the query genes was calculated using a normal distribution fit to the occurrence of the GO term in the 10,000 sets of controls. To call significance, FDR < 0.05 and occurrence (number of genes associated with a term) >= 5 were used. For GO analysis in each cell line, a similar strategy was used that analyzed 447 query genes in HEK293 and 423 query genes in HeLa. For HEK293, FDR < 0.05 and occurrence >= 7 were used to call significance. For HeLa, FDR < 0.05 and occurrence >= 10 were used to call significance.

2.5.15 Cancer driver genes in MapUTR

A table of 568 annotated cancer driver genes was obtained from the IntOGen[143] database (https://www.intogen.org/download). Genes containing MapUTR functional variants were overlapped with the cancer driver genes. To plot the heatmap of the overlapped cancer driver genes, the variant with the largest absolute value of fold change was reported for each gene in each cell line.

2.5.16 Expression outlier detection in TCGA

Genotype data including both germline and somatic mutations in The Cancer Genome Atlas (TCGA) was obtained from the Pan-Cancer Analysis of Whole Genomes (PCAWG)[144] through the ICGC Data Portal (http://dcc.icgc.org/pcawg/). FPKM data was downloaded from the Genomic Data Commons (GDC) Data portal. For MapUTR variants that are present in TCGA, z-scores of gene FPKM values were calculated for individuals with reference alleles or variant alleles (usually heterozygous variant). An expression outlier is defined by a z-score > 2 or z-score < -2 dependent on the lnFC value reported in MapUTR. To test if expression outliers were significantly enriched in individuals with variant alleles, a Fisher's exact tests was performed for each MapUTR variant in an individual cancer type. To call significance, *p*-value < 0.05 was used.

2.5.17 Generation of single-cell clones containing MapUTR variants via prime editing

To introduce a MapUTR variant to the HEK293T genome using prime editing[153–155], the spacer and extension sequences for epegRNAs and nick gRNAs were designed using pegFinder[171]. A linker pattern was designed for each epegRNA using pegLIT[154]. For epegRNA constructs, the spacer, extension (contains a unique linker), and pegRNA scaffold sequences (See Table S4) were cloned into the pU6-tevopreq1-GG-acceptor (Addgene, Plasmid #174038) via Golden Gate assembly. Similarly, the spacer and nick sgRNA scaffold sequences (See Table S4) were cloned into the pU6-pegRNA-GG-acceptor (Addgene, Plasmid #132777) to generate nick gRNA expressing constructs.

HEK293T cells were maintained in DMEM (Gibco, Cat# 11995065) with 10% FBS (Gibco, Cat# 26140079) and the antibiotic-antimycotic reagent (Gibco, Cat# 15240062) at 37 °C with 5% CO_2 supply. HEK293T cells were seeded in 48-well plates to reach 50% confluency by the time of transfection. The enhanced prime editing system[155], consisting of plasmids expressing the epegRNA (250ng), nick gRNA (83ng), and prime editor (750ng), i.e., pCMV-PEmax-P2A-hMLH1dn (Addgene, Plasmid #174828), was used for cell transfection. For *MFN2*, only the epegRNA and prime editor were used for cell transfection due to a higher editing efficiency compared to the other strategy that includes an additional nick gRNA. Cell transfection was performed with Lipofectamine™ 3000 Transfection Reagent (Thermo Fisher Scientific, Cat# L3000015) according to the manufacturer's protocol. For genotyping, genomic DNA (gDNA) was extracted and amplified with primers specific to each candidate variant (See Table S4). PCR amplicons were purified and sent for Sanger sequencing with one of the PCR primers. Three days after cell transfection, the transfected cells were re-plated into 96-well plates by serial dilution to generate single-cell clones. Single-cell clones were then expanded and genotyped via Sanger sequencing.

2.5.18 Measurement of mRNA expression levels via qRT-PCR

Single-cell clones with MapUTR variants were maintained in DMEM (Gibco, Cat# 11995065) with 10% FBS (Gibco, Cat# 26140079) and antibiotic-antimycotic reagent (Gibco, Cat# 15240062) at 37 °C with 5% CO_2 supply. For RNA isolation, cells were washed with PBS (Gibco, Cat# 14190144) and lysed with TRIzol (Thermo Fisher Scientific, Cat# 15596026). Total RNA was isolated using the Direct-zol RNA Miniprep Plus kit (Zymo Research, Cat# R2072) following the manufacturer's protocol. 1~2 μg of total RNA was used for cDNA synthesis with SuperScript™ IV Reverse Transcriptase (Thermo Fisher Scientific, Cat# 18090010) using

random hexamers. To measure mRNA expression levels of genes containing MapUTR variants, 1 µl cDNA was used for qPCR reactions using PowerUp™ SYBR® Green Master Mix (Thermo Fisher Scientific). Primers used for qPCR (same as gDNA PCR primers) were listed in Table S4. The reaction was performed in the CFX96 Touch Real-Time PCR detection system (Bio-Rad) with the following settings: 50 °C for 10 min, 95 °C for 2 min, 95 °C for 15 s, 60 °C for 30 s, and with the last two steps repeated for 45-55 cycles. The expression of genes containing MapUTR variants (*MFN2*, *FOSL2*, and *IRAK1*) was normalized against the expression of *TBP*. For mRNA stability assays, single-cell clones with MapUTR variants were treated with 10 µg/ml actinomycin D (Sigma-Aldrich) in growth media. Cells were harvested at different time points (2 h, 8 h, and 24 h) post actinomycin D (ActD) treatment for RNA isolation and RT-qPCR. Two technical replicates were performed for each single-cell clone during RT-qPCR. For each gene, 4 to 6 single-cell clones were used for either reference or variant alleles. Samples of reference and variant alleles collected at the same time point were analyzed in one PCR plate to allow for proper comparisons. P-values were calculated using one-tailed Student's t-test. To call significance, p-value < 0.05 was used.

2.5.19 Cell proliferation assay

Single-cell clones with MapUTR variants were seeded at 3,000 cells per well in the 96-well plates. For each single-cell clone, five technical replicates (wells) were performed. After 24 h incubation at 37 °C, the plate was transferred to the Incucyte® S3 live-cell analysis system (Sartorius) to monitor cell proliferation. Images were taken every 2 h and analyzed for confluency. Data were analyzed and plotted using Graphpad Prism 7. P-values were calculated using one-tailed Student's t-test. To call significance, p-value < 0.05 was used.

2.6 Acknowledgements

We thank members of the Xiao laboratory for helpful discussions and comments on this work. The results published here are in part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. This work was supported in part by grants from the National Institutes of Health (U01HG009417, R01CA262686, and R01AG075206 to X.X.) and the Jonsson Comprehensive Cancer Center at UCLA. T.F. was supported by the UCLA Hyde Fellowship and Dissertation Year Fellowship. K.A. was supported by the University of California-Historically Black Colleges and Universities (UC-HBCU) Fellowship. T.W.C. was supported by the NIH T32LM012424. S.T. was supported by the NIH T32GM145388. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

2.7 Figures

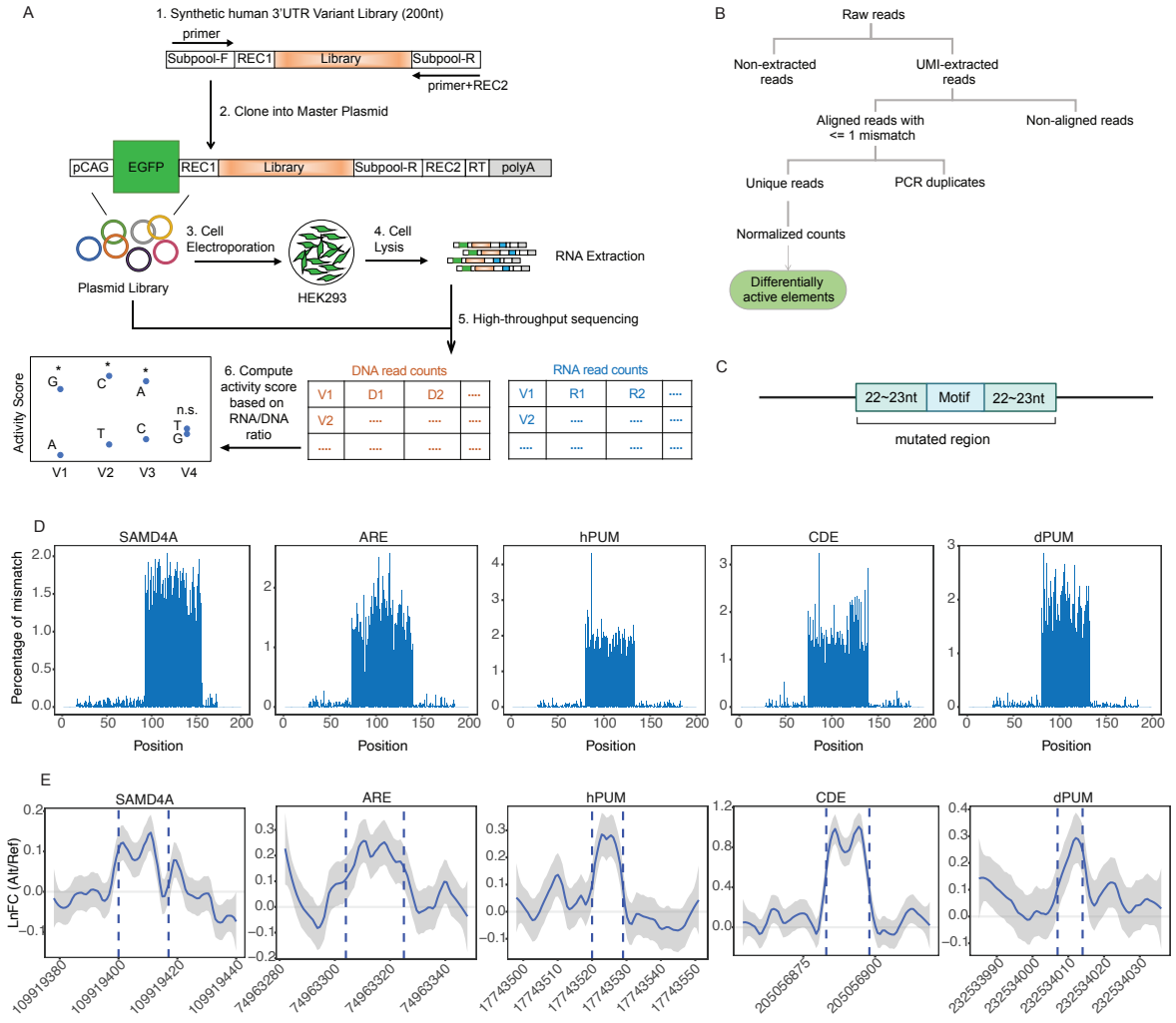


Figure 2.1 MapUTR captures functional 3' UTR variants in well-known motifs

(A) General workflow of MapUTR. See also Figure S1.

(B) Detailed Computational workflow diagram.

(C) Diagram of oligos with random mutations in the motif and its flanking regions (22~23nt). The Illustration was created with BioRender.com.

(D) Mismatch rate (%) per position along the length of DNA sequences harboring known motifs: SAMD4A (in gene *CHRDL1*), ARE (in gene *CXCL2*), hPUM (in gene *MYOD1*), CDE (in gene *RBBP5*), and dPUM (in gene *SIPA1L2*).

(E) Normalized changes in RNA abundance (i.e., activity score) as a result of alternative alleles in HEK293 cells. LnFC are averaged across the 3 tested alternative alleles per position.

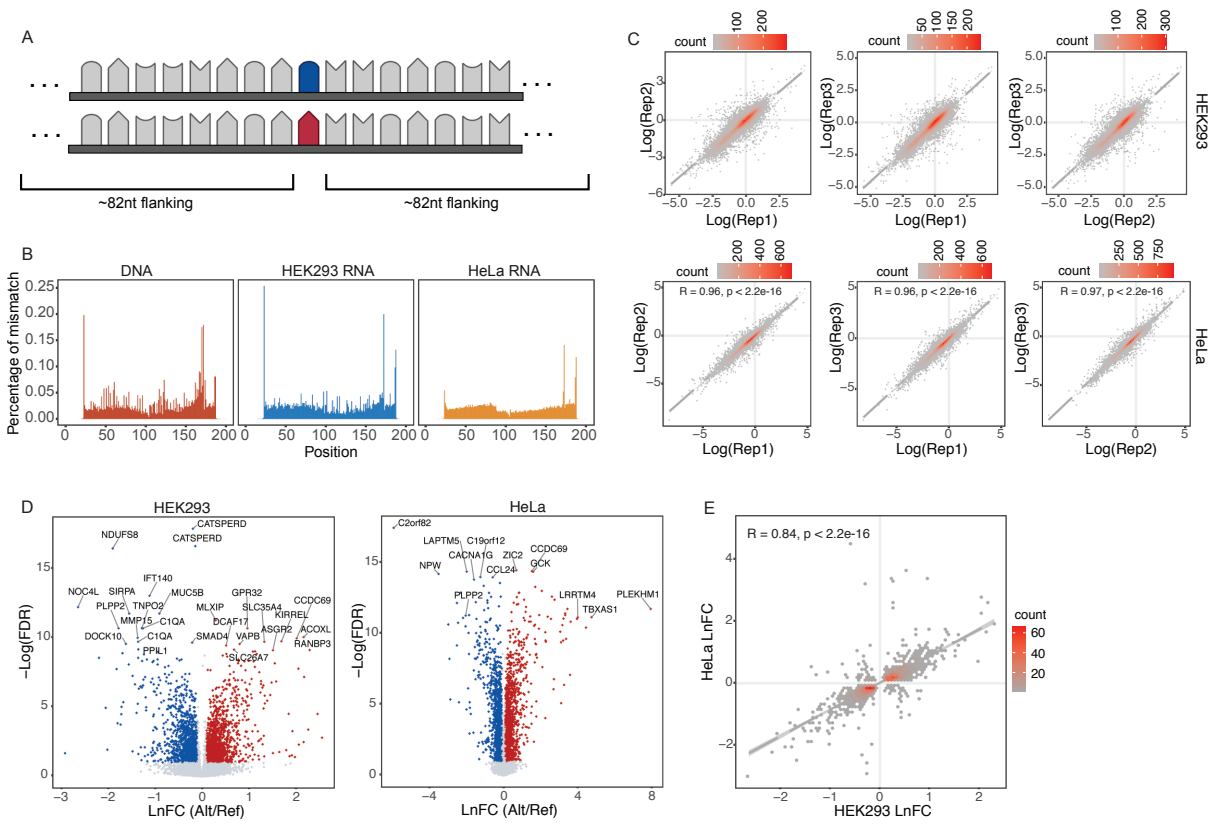


Figure 2.2 MapUTR identified functional rare 3' UTR variants regulating mRNA abundance

(A) Design of oligos containing rare variants from ExAC.

(B) Average percentage of read mismatches in DNA and RNA sequencing libraries.

(C) MapUTR reproducibility as measured by the activity score correlation between biological replicates. Three biological replicates are shown for each cell line.

(D) Functional activity of rare 3' UTR variants tested in HEK293 and HeLa. Volcano plots illustrate lnFCs of alternative alleles relative to the reference alleles.

(E) Correlation of lnFC values of significant functional variants shared in HEK293 and HeLa cells.

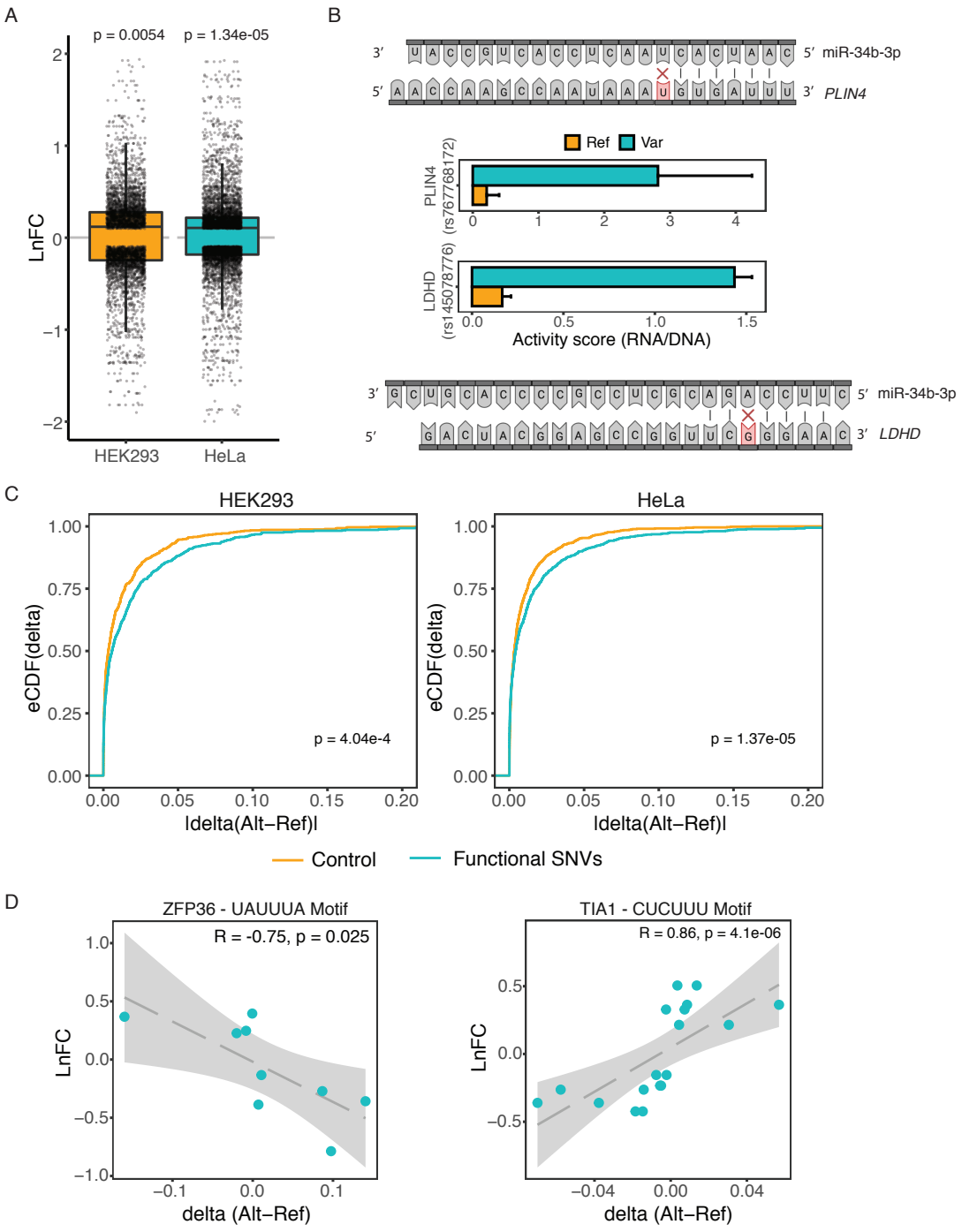


Figure 2.3 Mechanisms of RNA abundance regulation via 3' UTR

(A) Functional SNPs in miRNA targets exhibited a bias in effect in HEK293 and HeLa. P-values were obtained by a chi-squared test.

(B) Functional variants in *PLIN4* (HEK293) and *LDHD* (HeLa) disrupted the target sites for miR-34b-3p (top) and miR-3180-5p (bottom), respectively. MapUTR activity scores are shown in the middle.

(C) Functional MapUTR variants significantly altered RBP binding (See also Figure S3-4). X-axis shows the binding score difference of the reference and alternative alleles predicted by DeepRipe.

(D) Changes in mRNA abundance detected by MapUTR were corroborated by changes in RBP binding. X-axis similar as (C). Top: Functional variants that increased ZFP36 binding to AU-rich element (AUUUA) exhibited decreased gene expression. Bottom: Functional variants that increased TIA1 binding to CU-rich element (CUCUUU) led to upregulation of gene expression.

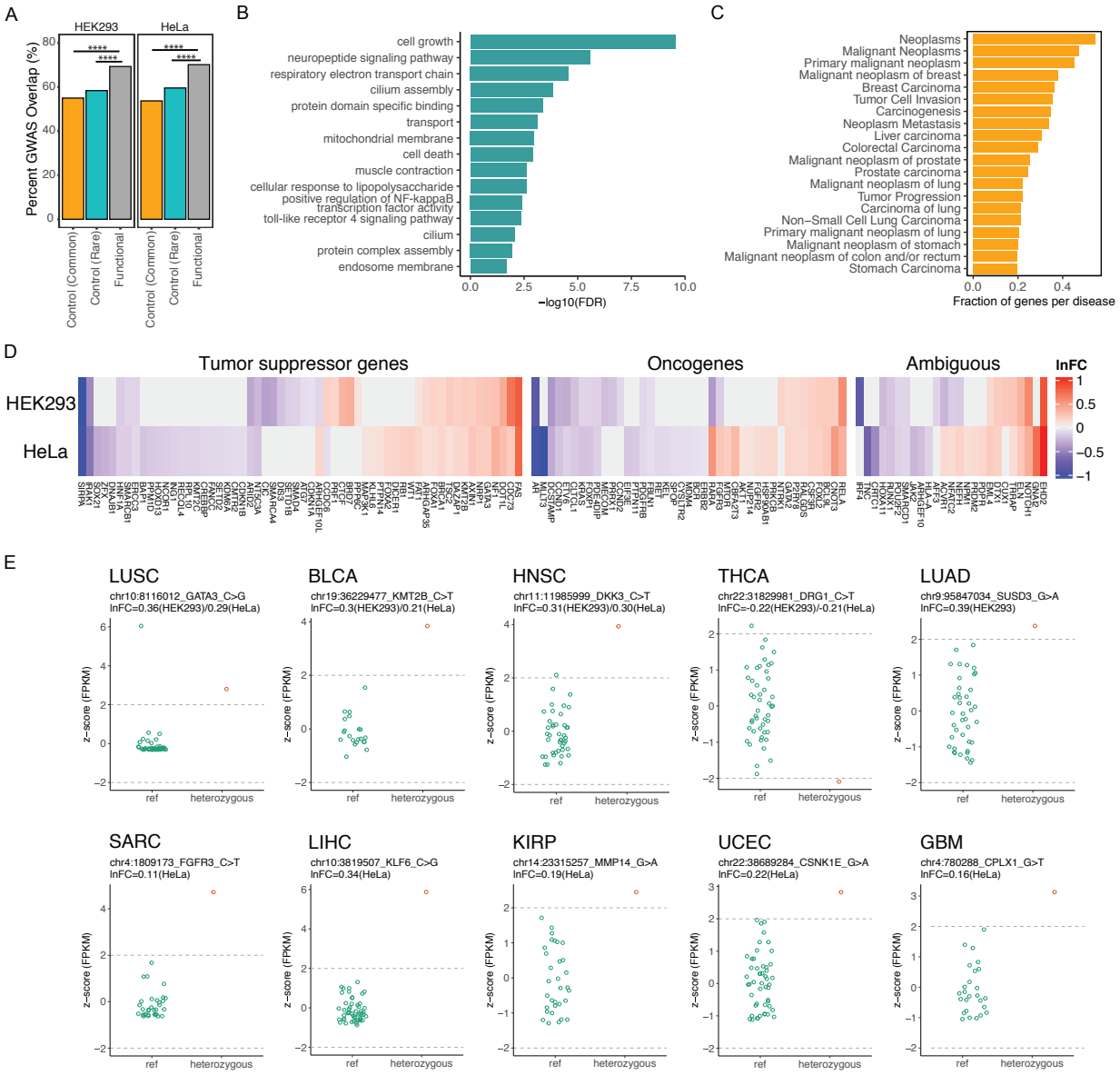


Figure 2.4 Functional relevance of significant variants in cancer

(A) Proportion of variants associated with GWAS SNPs among functional MapUTR rare variants, rare dbSNP controls, and common dbSNP controls.

(B) Gene ontology terms enriched in the genes with large-effect functional variants (top 500) in HEK293 and HeLa.

(C) Disease associations most represented by MapUTR functional variants. Top 20 diseases were plotted.

(D) Cancer driver genes containing MapUTR functional variants. For genes with multiple functional variants, the variant with the largest absolute value of lnFC was plotted in each cell line.

(E) MapUTR functional variants found in gene expression outliers in TCGA. See also Figure S5.

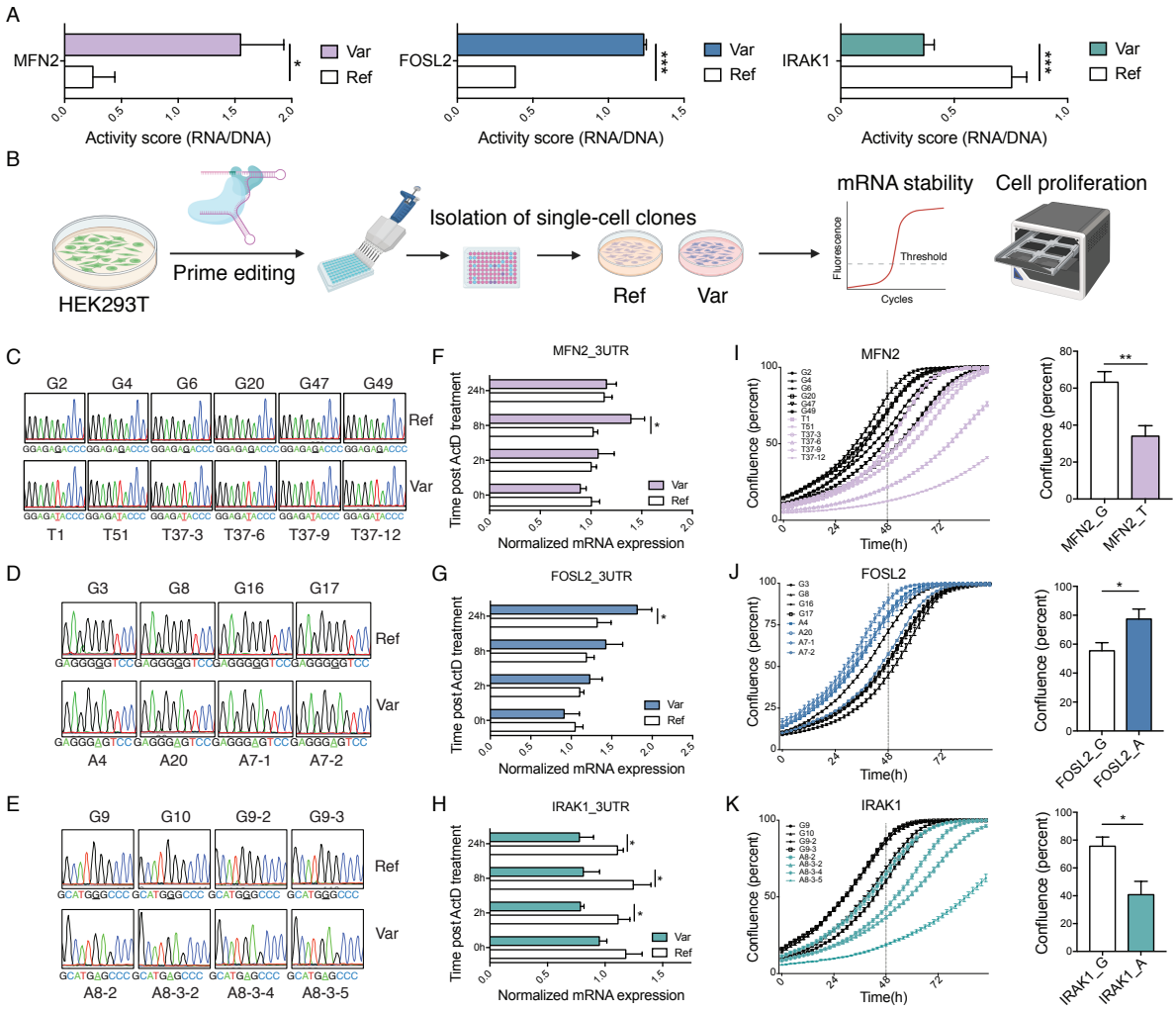


Figure 2.5 Functional rare 3' UTR variants regulate mRNA stability and cell proliferation in HEK293T cells

(A) MapUTR activity scores (RNA/DNA) of functional variants (*MFN2: rs777822288*, *FOSL2: rs11884725*, and *IRAK1: rs782486025*) measured in HEK293 cells. Data are plotted as mean +/- SEM. P-values were calculated using MPRAnalyze. *p<0.05, **p<0.001, ***p<0.001.

(B) Diagram of validation workflow. HEK293T cells were transfected with plasmids expressing PEmax enzymes and epegRNAs to introduce the functional variant of interest. Single-cell clones were isolated and genotyped as either reference (Ref) or variant (Var) clones. The single

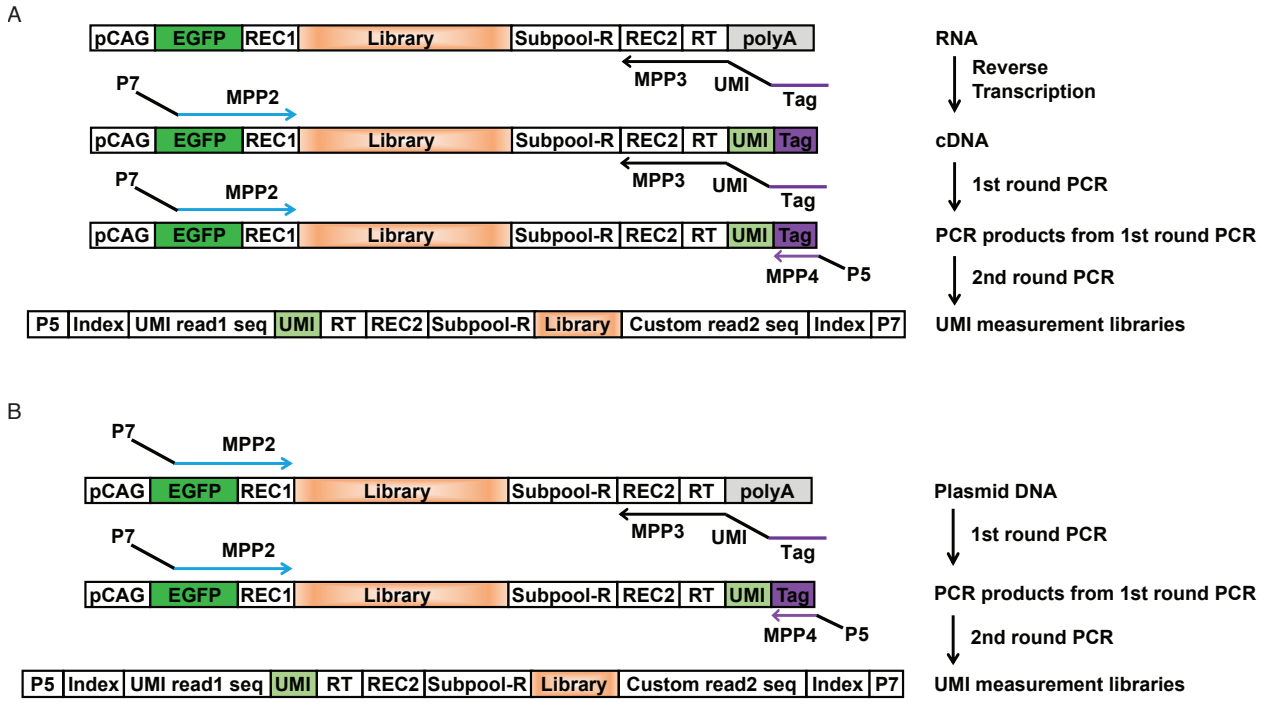
clones were used for downstream assays to test for mRNA expression/stability or cell proliferation. Created with BioRender.com.

(C-E) Sanger sequencing results confirming the genotype of single clones with reference (Ref) and variant (Var) alleles for *MFN2* (C), *FOSL2* (D), and *IRAK1* (E).

(F-H) Normalized mRNA expression level of reference (Ref) and variant (Var) alleles of *MFN2* (F), *FOSL2* (G) and *IRAK1* (H). Cells were treated with 10 μg/ml actinomycin D (ActD) and harvested at 2 h, 8 h, and 24 h post treatment to test for mRNA stability. For *MFN2*, six biological replicates (six clones) for each allele were included in the experiment. For *FOSL2* and *IRAK1*, four biological replicates (four clones) per allele were included in the experiment. Data are plotted as Mean +/- SEM. P-values were calculated using one-tailed Student's t-test. **p*<0.05.

(I-K) Cell proliferation assay of single clones with reference (Ref) and variant (Var) alleles for *MFN2* (I), *FOSL2* (J), and *IRAK1* (K). For *MFN2*, six biological replicates (six clones) for each allele were included in the experiment. For *FOSL2* and *IRAK1*, four biological replicates (four clones) per allele were included in the experiment. The dashed line indicates the cell confluence values at 48h, which are plotted on the right. Data are plotted as mean +/- SEM. P-values were calculated using one-tailed Student's t-test. **p*<0.05, ***p*<0.01.

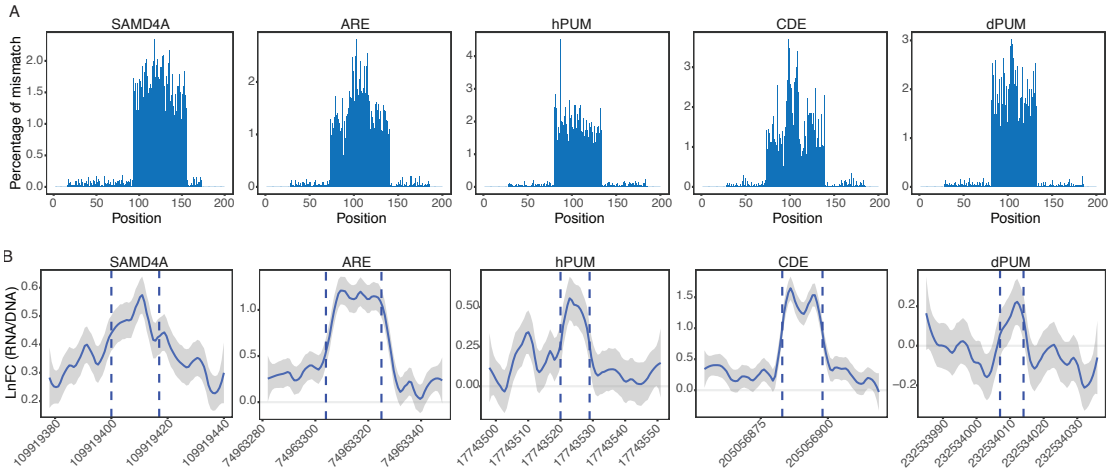
2.8 Supplementary Figures



Supplementary Figure 2.1 Generation of UMI measurement libraries

(A) RNA-seq library generation for mRNA isolated from HEK293/HeLa cells electroporated with plasmid libraries.

(B) DNA-seq library generation for plasmid libraries used for cell electroporation.



Supplementary Figure 2.2 MapUTR sequencing quality and accuracy

(A) Mismatch rate (%) per position along the length of RNA sequences harboring known motifs: SAMD4A (in gene *CHRDL1*), ARE (in gene *CXCL2*), hPUM (in gene *MYOD1*), CDE (in gene *RBBP5*), and dPUM (in gene *SIPA1L2*).

(B) Normalized changes in RNA abundance (i.e., activity score) as a result of alternative alleles in HeLa cells. LnFC were averaged across the 3 tested alternative alleles per position.

Supplementary Figure 2.3 Overrepresented motifs in functional sequences

(A) HEK293 downregulating motifs

(B) HEK293 upregulating motifs

(C) HeLa downregulating motifs

(D) HeLa upregulating motifs

A

RBP	Motif (Downregulating)
ELAVL4	
FUS	
HNRNPD	
HNRNPK	
IGF2BP2	
KHSRP	
MBNL1	
PCBP1	
PCBP2	
PUM2	
RBFOX2	
TAF15	
TARDBP	
TIA1	
ZFP36	

RBP	Motif (Upregulating)
EIF4G2	
FUS	
IGF2BP2	
MBNL1	
PCBP1	
PCBP2	
TAF15	
TARDBP	
TIA1	

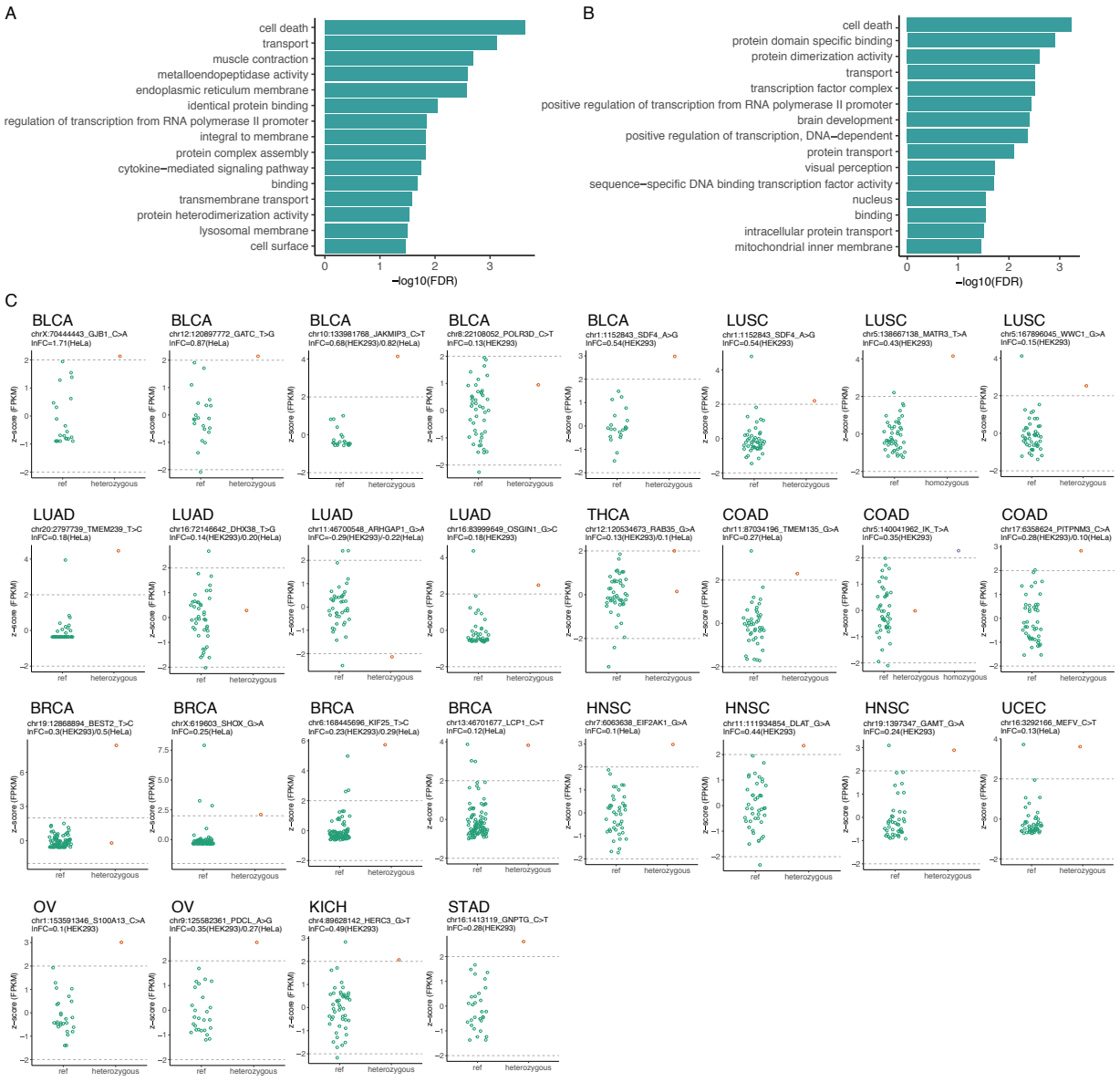
B

RBP	Motif (Downregulating)
EIF4G2	
FUS	
HNRNPK	
IGF2BP1	
IGF2BP2	
MBNL1	
PCBP1	
PCBP2	
TARDBP	
TIA1	

RBP	Motif (Upregulating)
EIF4G2	
ELAVL4	
FUS	
HNRNPC	
HNRNPD	
IGF2BP2	
KHSRP	
PCBP1	
TAF15	
TARDBP	
TIA1	

Supplementary Figure 2.4 Motifs bound by RBPs according to RBNS data

(A-B) RBPs motifs overrepresented in downregulating (left) and upregulating sequences (right) in HEK293 (A) and HeLa (B).



Supplementary Figure 2.5 Functional relevance of MPRA significant variants

(A) Gene ontology terms enriched in the genes with large-effect (top 500) functional variants found in HEK293 cells.

(B) Gene ontology terms enriched in the genes with large-effect (top 500) functional variants found in HeLa cells.

(C) MapUTR functional variants found in gene expression outliers in TCGA. Related to Figure 4E.

2.9 Supplementary Tables

	APP	ABCB1	CYP2A7
0.4ng/1M		0.77	
2ng/1M	0.22	0.70	
12.5ng/1M	1.55	0.74	
40ng/1M	0.24	0.33	0.76
200ng/1M	1.09	0.99	1.21
Expected Results	**>1**	**1**	**>1**

Supplementary Table 2.1 DNA/Cell ratio optimization using cell electroporation

Values in the table show RNA/DNA ratios tested in each condition.

Gene	Chrom	Start (hg19)	End	Strand	Motif	Sequence	Function
CXCL2	chr4	74963304	74963325	-	ARE	UAUUUAUUUAUUUAUUUAUUUAU	Destabilize
RBBP5	chr1	205056883	205056898	-	CDE	UCCUUUCUGUGAAAGG	Destabilize
CHRDL1	chrX	109919400	109919417	-	SAMD4A	AAGCUGCAGCUGGACUGC	Destabilize
MYOD1	chr11	17743520	17743529	+	hPUM	UGUAAAUAAG	Destabilize
SIPA1L2	chr1	232534007	232534014	-	dPUM	UGUACAGA	Destabilize

Supplementary Table 2.2 Well-known 3' UTR motifs tested in Figure 1E

Chip	Sequence	Filters
Chip1	Original sequence (200nt: F+rec1+lib+R)	• Only one occurrence of REC1 (EcorRI) • Does not contain REC2 (BamHI) • Only one occurrence of each of the corresponding subpool1 reverse amplification sequences (8 nt in the 3’ end of the primer) • Only one occurrence of corresponding subpool1 forward sequence (the entire 15 nt of the primer) • Only one occurrence of corresponding subpool1 reverse sequence (the entire 15 nt of the primer)
Chip2 Chip3	Original sequence (200nt: F+rec1+lib+R)	• Only one occurrence of REC1 (EcorRI) • Does not contain REC2 (BamHI) • Only one occurrence of each of the corresponding subpool1 forward and reverse amplification sequences (8 nt in the 3’ end of the primer) • Does not contain any pair of other subpools' (subpool 2,3,4) forward and reverse amplification sequences (8 nt in the 3’ end of the primer)
	Reverse complement of the original sequence	• Only one occurrence of REC1 (EcorRI) • Does not contain REC2 (BamHI) • Does not contain corresponding subpool1 forward amplification sequence (8 nt in the 3’ end of the primer) • Does not contain corresponding subpool1 reverse amplification sequence (8 nt in the 3’ end of the primer) • Does not contain any pair of other subpools' (subpool 2,3,4) forward and reverse amplification sequences (8 nt in the 3’ end of the primer)

Supplementary Table 2.3 Filtering Criteria for MPRA Library Design

Section 1: Primers used for master plasmid cloning and DNA/Cell ratio optimization.

Primer	Sequence
CMV_NdeI_F	GTGTATCATATGCCAAGTACGCCCCCTATTGACG
EGFP_EcoRI_R	GTAGAATTCTTACTTGTACAGCTCGTCCATGCCGAGAGTGATCCCGG
EGFP_SacI_R	CCCCGGTGAAGAGCTCCTCGC
EGFP_SacI_F	GCGAGGAGCTCTTCACCGGGG
BamHI_subpool2_HpaI_R	TGTTAACGTTCCGCAGCCAGGATCCCGGGCCCGCGGTACC
RT_Rd1_polyA_BamHI_F	GGGATCCAGATCGGAAGAGCGTCGTGTAGGGAAAAACTTGTTTATTGCAGCT TATAATGG
polyA_MluI_R	TTTACGCGTTAAGATACATTGATGAG
GFP_ACCTTA_AAA_R	GTAGAATTCTTATTTGTACAGCTCGTCCATGCCTAAGGTGATCCCGGCGGCGG TCAC
GFP_AAA_F	CACTCTCGGCATGGACGAGCTGTACAAATAAGAATTC
P5_F	AATGATACGGCGACCACCGAG
APP_R	TGGTTTGTGACCCAATTAAGTCCTAC
ABCB1_R	TTCCTCAGTCAAGTTCAGAGTCTTCAG
CYP2A7_R	CGTGGTGGCTAGAGGGAAGAG

Section 2: Primers used for prime editing and RT-qPCR.

Primer	Sequence
pegRNA_scaffold_f	5' [Phos] agagctagaaatagcaagttaaaataaggctagtccgttatcaacttgaaaaagtggCACCGAGTCG 3'
pegRNA_scaffold_r	5' [Phos] GCACcgactcggtGccacttttcaagttgataacggactagccttattttaacttgctatttctag 3'
nichsgRNA_scaffold_f	/5Phos/agagctagaaatagcaagttaaaataaggctagtccgttatcaacttgaaaaagtggCACCGAGTCG GTGC
nicksgRNA_scaffold_r	/5Phos/AAAAGCACcgactcggtGccacttttcaagttgataacggactagccttattttaacttgctatttctag
spacerF_MFN2_P13_RT18	caccGGCCATACTTCTTTCAGAAAgtttt
spacerR_MFN2_P13_RT18	ctctaaaacTTTCTGAAAGAAGTATGGCC
extF_MFN2_PBS13_RT18	gtgcCTGAGGGAGATACCCTTTCTGAAAGAAGTATAAATAATG
extR_MFN2_PBS13_RT18	cgcgCATTATTTATACTTCTTTCAGAAAGGGTATCTCCCTCAG
spacerF_FOSL2_p13_rt14	caccgTCCCTCCCCAGCTCCGGAGGgtttt
spacerR_FOSL2_p13_rt14	ctctaaaacCCTCCGGAGCTGGGGAGGGAc
extF_FOSL2_pbs13_rt14	gtgcGAGGAGGACTCCCTCCGGAGCTGGGGATTAAATGA
extR_FOSL2_pbs13_rt14	cgcgTCATTTAATCCCCAGCTCCGGAGGGAGTCCTCCTC
spacerF_FOSL2_3b_nick	caccGGAGCGAGGAGGACTCCCTCgtttt
spacerR_FOSL2_3b_nick	ctctaaaacGAGGGAGTCCTCCTCGCTCC
spacerF_IRAK1_p13_rt15	caccgCTTCTCTCCCCCGCGGGCATgtttt
spacerR_IRAK1_p13_rt15	ctctaaaacATGCCCGCGGGGGAGAGAAGc
extF_IRAK1_pbs13_rt15	gtgcGGGTGGGGGCTCATGCCCGCGGGGGAGAATATATAA
extR_IRAK1_pbs13_rt15	cgcgTTATATATTCTCCCCCGCGGGCATGAGCCCCCACCC
spacerF_IRAK1_3b_nick	caccGGGTGGGGGCTCATGCCCGCgtttt
spacerR_IRAK1_3b_nick	ctctaaaacGCGGGCATGAGCCCCCACCC
mfn2_gDNA_F	TGAATGGACAGGGGCCACTTC
mfn2_gDNA_R	CAGATTATAGTGGGAACTTCCCCAAAG

FOSL2_gDNA_F	ATCCTGCTCCAAGGCTC
FOSL2_gDNA_R	ACTGCTAAGTCCCACCTG
IRAK1_gDNA_F	AGCCTCCTCACTGGATG
IRAK1_gDNA_R	TGTGTTCACCTGGGCAG
MFN2_3UTR_qPCR_R	TCATTCATTCCCCAGGGGCTAC
qTBP-Fw	CAGCAACTTCCTCAATTCCTTG
qTBP-Rv	GCTGTTTAACTTCGCTTCCG

Supplementary Table 2.4 Additional list of primers

2.10 Supplementary Protocol

Part I. Plasmid library cloning

Materials

Reagents

- UltraPure DNase/RNase-Free Distilled Water (Invitrogen, Cat# 10977015)
- PowerUp™ SYBR® Green Master Mix (Thermo Fisher Scientific, Cat# A25743)
- Q5 Hot Start High-Fidelity 2x Master Mix (NEB, Cat# M0494L)
- Zymo DNA clean & Concentrator Kit (Zymo Research, Cat# D4004)
- Zymoclean™ Gel DNA Recovery Kit (Zymo Research, Cat# D4002)
- EcoRI-HF (NEB, Cat# R3101S)
- BamHI-HF (NEB, Cat# R3136S)
- T7 DNA Ligase (NEB, Cat# M0318)
- 10-beta Electrocompetent *E. coli* (NEB, Cat# C3020K)
- Electroporation Cuvettes, 0.1cm gap (Bio-Rad, Cat. # 1652089)
- Fisherbrand™ Petri Dishes with Clear Lid, 150mm x 15mm (Fisher Scientific, Cat# FB0875714)
- ZymoPURE II Plasmid Midiprep Kit (Zymo Research, Cat# D4200)

Primers

Name	Sequence
Subpool1_F	GGTCGAGCCGGAACT
subpool2_F	CGATCGCCCTTGGTG
subpool3_F	GGGTCACGCGTAGGA
Subpool4_F	CGCGTCGAGTAGGGT
Subpool1_BamHI_R	TTACGTGGATCCGGATGCGCACCCAGA
subpool2_BamHI_R	TTACGTGGATCCGGTTTAGCCGGCGTG
subpool3_BamHI_R	TTACGTGGATCCGTTCCGCAGCCACAC
Subpool4_BamHI_R	TTACGTGGATCCGCCGTGTGAAGCTGG
polyA_MluI_R	TTTACGCGTTAAGATACATTGATGAG

*All primers listed above were synthesized by IDT with standard desalting purification.

Procedure

1. Resuspend oligo library (Twist Biosciences) in Ultrapure distilled water at a final concentration of 1ng/µl.
2. Assemble qPCR reaction for each subpool (below showing the reaction for chip1.subpool1, same for other subpools)

Reagent	Volume(µl)
PowerUp SYBR 2x master mix	25
Subpool1_F (10µM)	2.5
Subpool1_BamHI_R (10µM)	2.5
Resuspended oligos	1
dH_2O	19

Load 20µl of the mixed reaction to the qPCR 96-well plate (2 wells per subpool)

Run qPCR as follows:

UDG activation	50 °C	10min
Initial denaturation	95 °C	2min
45 cycles	95 °C	15s
	60 °C	30s

3. Repeat PCR by replacing the SYBR master mix with the Q5 hot start HF master mix, set up three 50µl reactions for each subpool, use the cycle number where the slope begins to decrease in the qPCR pre-run (usually 17-19 cycles).

Reagent	Volume(µl)
Q5 2x master mix	75
Subpool1_F (10µM)	7.5
Subpool1_BamHI_R (10µM)	7.5
Resuspended oligos	3
dH_2O	57

Distribute the mixed reaction to 3 PCR tubes (50µleach).

Run PCR as follows:

Initial denaturation	98 °C	30s
17-19 cycles	98 °C	10s
	60 °C	30s
	72 °C	30s
Final extension	72 °C	2min

4. Run 20µl of PCR products on a 2% agarose gel to check the band size; save the rest PCR products for direct PCR clean up using the Zymo DNA clean & concentrator kit, elute with distilled water.
5. Quantitate DNA concentration of purified PCR products with BioDrop Fluorometer.
6. Digest 2µg of the master plasmids and 100ng of the purified PCR products with EcoRI-HF and BamHI-HF overnight at 37 °C.
7. Heat inactivation of restriction enzymes at 65 °C for 20min.
8. Load the digested master plasmid reaction on a 1% agarose gel and gel purify the band at 5.7kb.
9. Clean up the digested PCR products directly with the Zymo DNA clean & concentrator kit.
10. Set up ligation with freshly cut vector and inserts at 1:10 molar ratio.

Reagent	Amount
2x T7 ligase buffer	20µl
Digested master plasmid	100ng
Digested inserts	36.6ng
T7 DNA ligase	2µl
dH_2O	to 40µl

*Set up one ligation reaction per subpool; set up a ligation reaction without inserts (inserts replaced with water) for background calculation.

Incubate the reaction at 25°C for 1hour and then keep on ice.

11. Clean up the ligation reaction using the Zymo DNA clean & concentrator kit, elute with 8µl distilled water.
12. Mix 1µl purified ligation products with 25µl 10-beta electrocompetent *E. coli*. Transfer the mixture the a prechilled 0.1cm electroporation cuvette, and perform electroporation following the manufactory's protocol. Immediately add 750µl pre-warmed 10-beta outgrowth medium into the cuvette and transfer the mixture to a 1.5ml microcentrifuge tube.
13. Recover transformed *E. coli* at 37 °C for an hour.
14. Make a serial dilution of the transformed *E. coli* (1:1, 1:10, 1:100, 1:1000). Plate 250µl transformed *E. coli* per 150mm Kanamycin-selective plate.
15. Grow plates at 37 °C overnight
16. Count colonies on the serial dilution plates. The number of colonies represents the plasmid complexity per subpool library. For a subpool with 2000 variants, to ensure 100x coverage, harvest 0.2M colonies (harvest 0.4M colonies to account for the loss during plasmid isolation). One 25µl transformation typically yields colonies ranging from 0.4M~4.5M. Set up multiple electroporation reactions to make sure to get enough colonies for a given subpool.

17. Add 5ml LB media to the selective plate and gently scrape off the colonies. Combine all the colonies for a subpool in a 50ml tube. Mix well.
18. Measure the OD600 of the colony suspension. Make 1-to-10 or 1-to-100 dilutions as necessary for accurate measurement.
19. The plasmid library can be directly extracted from the harvested colonies. Pellet 5ml colony suspension (OD>10) for a library with 0.6M colonies; scale the amount of colony suspension according to the library coverage. Extract plasmids from the pellet using the ZymoPURE II Plasmid Midiprep Kit.
20. Alternatively, for a subpool with 0.2M colonies, seed 22M *E. coli* (1 OD= 80M *E. coli*/ml) in 50ml LB media (Kan-selective). Grow the culture overnight at 37 °C and extract plasmids using the ZymoPURE II Plasmid Midiprep Kit.
21. Send the plasmid library to Sanger sequencing using the "polyA_MluI_R" primer.

Part II. Cell Electroporation and isolation of mRNA

Materials

Reagents

- Electroporation Cuvettes, 0.4cm gap (Bio-Rad, Cat# 1652086)
- OptiMEM (Gibco, Cat# 31985062)
- Growth Media
    - DMEM (Gibco, Cat# 11995065)
    - 10% FBS (Gibco, Cat# 26140079)
    - Antibiotic-Antimycotic reagent (Gibco, Cat# 15240062)
- Trypsin-EDTA (Gibco, Cat# 25300120)

- TRIzol (Thermo Fisher Scientific, Cat# 15596026)
- Direct-zol RNA Miniprep Plus kit (Zymo Research, Cat# R2072)
- Dynabeads™ Oligo$(dT)_{25}$ (Thermo Fisher Scientific, Cat# 61002)
- NEBNext® Poly(A) mRNA Magnetic Isolation Module (NEB)
- Qubit RNA HS Assay Kit (Thermo Fisher Scientific, Cat# Q32852)

Procedure

*The following numbers are designed for the subpool with 2000 variants (ref+alt) and 100x coverage (i.e., 0.2M colonies, see table below for scale up recommendations)

1. On day 0, seed >45M HEK293 cells in 150mm dishes and make sure they are less than 80% confluent (should be actively dividing cells) by the time of electroporation.
2. One day 1, trypsinize the HEK293 cells, resuspend with growth media, and count cell numbers.
3. Spin down the 293 cells and resuspend with ice-cold OptiMEM at a cell density of 10M/ml.
4. For each electroporation, mix 750µl (7.5M) cells with 1.5µg plasmid libraries in a pre-chilled microcentrifuge tube, transfer the mixture to a pre-chilled 0.4cm electroporation cuvette, perform electroporation (square wave, 25msec, 220V, 0.4cm).
5. Immediately add 1ml warm growth media to the cuvette and transfer the cells to a 150mm petri dish.
6. Combine 2 x 7.5M cell transformants in one 150mm petri dish for one replicate. Perform three replicates for each subpool.
7. Incubate cells at 37 °C for 24h.

8. After 24h, wash the cells in one 150mm petri dish with 10ml pre-warmed PBS, then add 5ml TRIzol to each plate. Lyse the cells at RT for 10min, then transfer the mixture to 1.5ml microcentrifuge tubes. Distribute 500µl lysed mixture per tube. Add 100µl chloroform to each tube and then mix well. Incubate at RT for 5min. Centrifuge at >13,000 g, 4° C for 15min. Carefully transfer the aqueous upper phase into a new 1.5ml tube. Add equal volume of 100% ethanol and mix well. Load the mixture to six columns supplied by Direct-zol RNA Miniprep Plus kit (Zymo Research, Cat# R2072) to isolate total RNA following the manufacturer's protocol.
9. Isolate the mRNA using the Dynabeads™ Oligo$(dT)_{25}$ (Thermo Fisher Scientific, Cat# 61002). Use 400µg total RNA for one replicate. Quantitate the mRNA yield with Qubit Fluorometer. Store the mRNA at -80°C.

Table. DNA and cells used during electroporation for plasmid libraries with different complexity

Variants	Coverage	Colonies	DNA/replicate	Cells/replicate
2000	100x	0.2M	3µg	15M
6000	100x	0.6M	6µg	30M

Part III. Generation of UMI measurement libraries

Materials

Reagents

- SuperScript™ IV First-Strand Synthesis System (Thermo Fisher Scientific, Cat# 18091050)
- Q5 Hot Start High-Fidelity 2x Master Mix (NEB, Cat# M0494L)

- UltraPure DNase/RNase-Free Distilled Water (Invitrogen, Cat# 10977015)
- PowerUp™ SYBR® Green Master Mix (Thermo Fisher Scientific, Cat# A25743)
- Zymo DNA clean & Concentrator Kit (Zymo Research, Cat# D4004)
- Zymoclean™ Gel DNA Recovery Kit (Zymo Research, Cat# D4002)

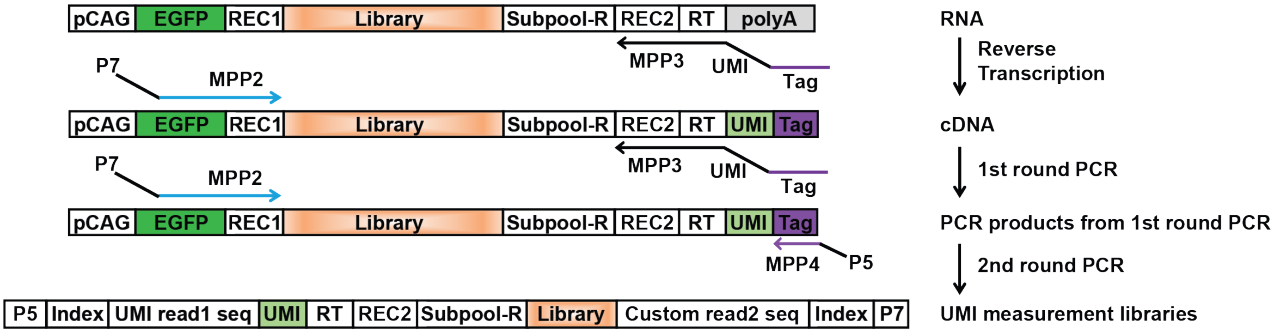
Primers

Name	Sequence
MPP3	GTGATTGGAGTTCAGACGTGTGTTCTGCTGACGNNNNNNNNNNNNNNNNCGCTCTTCCGATCTGGATCC
MPP2_352	CAAGCAGAAGACGGCATACGAGATTTGGACTTCACCTTAGGCATGGACGAGCTGTACAAATAAGAATTC
MPP2_361	CAAGCAGAAGACGGCATACGAGATCCTCGGTACACCTTAGGCATGGACGAGCTGTACAAATAAGAATTC
MPP2_362	CAAGCAGAAGACGGCATACGAGATAGACTTGGCACCTTAGGCATGGACGAGCTGTACAAATAAGAATTC
MPP2_363	CAAGCAGAAGACGGCATACGAGATATGAGGCTCACCTTAGGCATGGACGAGCTGTACAAATAAGAATTC
MPP2_364	CAAGCAGAAGACGGCATACGAGATGCAGAATCCACCTTAGGCATGGACGAGCTGTACAAATAAGAATTC
MPP2_365	CAAGCAGAAGACGGCATACGAGATGTTGTCCGCACCTTAGGCATGGACGAGCTGTACAAATAAGAATTC
MPP2_366	CAAGCAGAAGACGGCATACGAGATCATGCCATCACCTTAGGCATGGACGAGCTGTACAAATAAGAATTC
MPP2_367	CAAGCAGAAGACGGCATACGAGATTCTCATTCCACCTTAGGCATGGACGAGCTGTACAAATAAGAATTC
MPP2_368	CAAGCAGAAGACGGCATACGAGATGCGCCTGTCACCTTAGGCATGGACGAGCTGTACAAATAAGAATTC
MPP4	AATGATACGGCGACCACCGAGATCTACACTACTCATAGTGATTGGAGTTCAGACGTGTGTTCTGCTGAC*G
MPP1	AATGATACGGCGACCACCGAGATCTACACTATGAGTATTTCCCTACACGACGCTCTTCCG
MPP2	CAAGCAGAAGACGGCATACGAGATTTGGACTTCACCTTAGGCATGGACGAGCTGTAC
UMI.R1.seq	GTGATTGGAGTTCAGACGTGTGTTCTGCTGACG
Read2.seq	CACCTTAGGCATGGACGAGCTGTACAAATAAGAATTC
Index7.seq	GAATTCTTATTTGTACAGCTCGTCCATGCCTAAGGTG

*All primers listed above were synthesized by IDT with PAGE purification, red highlighted sequences are the indexes for sample pooling purposes.

Procedure

a. UMI addition for mRNA



1. Anneal RT primers to the template mRNA

Reagent	Amount
MPP3 (2μM)	1μl
dNTP mix (10mM)	1μl
mRNA	500ng
Nuclease-free water	To 13μl

Heat on a thermal cycler at 65°C for 5min

Promptly remove the samples and put them on ice for 2min

*Set-up one no-RT control

**Use all mRNA(~7μg) isolated from 400μg total RNA for a subpool with 0.2M colonies. Determine the mRNA input by a trial run (e.g., 1 RT reaction) of this protocol with input standards to estimate the complexity of the libraries. The complexity of the libraries should match the downstream sequencing read coverage. (e.g., 20M complexity for 20M reads) Do multiple RT reactions when making the real libraries.

2. Prepare RT reaction mix

Reagent	Volume(μl)
5x SSIV Buffer	4
DTT (100mM)	1
RNase Inhibitor	1
SuperScriptIV RT	1

Mix and centrifuge

3. Combine RT reaction mix with annealed RNA by pipetting up and down
4. Incubation reactions

RT incubation	50 °C	30min
Inactivation	80 °C	10min

5. Add 1μl RNase H to the 20μl RT reaction and incubate at 37 °C for 20min.

6. Add the UMI with 3-cycle of PCR (first-round)

Reagent	Volume(μl)
Q5 2x master mix	20
MPP2_352 (10μM)	2
MPP3 (10μM)	2
cDNA	10
dH_2O	6

*Use MPP2 primers with different indexes for different samples

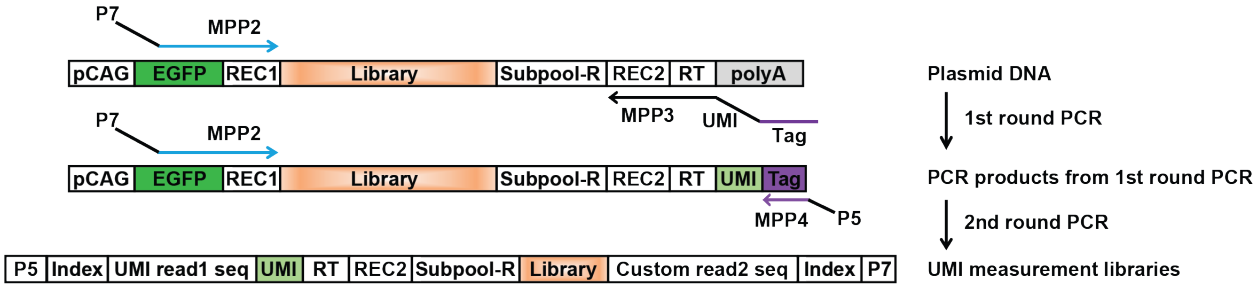
**Assemble more PCR reactions as needed to use all the cDNA samples when making the libraries.

Run PCR as follows:

Initial denaturation	98 °C	1min
3 cycles	98 °C	15s
	50 °C	30s
	72 °C	1min
Final extension	72 °C	10min

7. Pool the PCR reactions with the same index in step 6. Directly purify the PCR reactions with the Zymo DNA Clean & Concentrator Kit. Use multiple columns as needed. Elute the samples with distilled water (5μl for one RT reaction).

b. UMI addition for plasmid DNA



8. Add the UMI to the plasmid DNA using Q5 polymerase (first-round)

Reagent	Amount
2x Q5 master mix	20µl
MPP2_366 (10µM)	2µl
MPP3 (10µM)	2µl
DNA	200ng
dH_2O	to 40µl

*Use 400ng DNA for a subpool with 0.2M colonies. Set up two 40µl reactions. For subpools with 0.6M colonies, use 800ng DNA.

**Determine the DNA input by a trial run of this protocol with standards to estimate the complexity of the libraries. The complexity of the libraries should match the downstream sequencing read coverage. (e.g., 20M complexity for 20M reads)

Run PCR as follows:

Initial denaturation	98 °C	1min
3 cycles	98 °C	15s
	50 °C	30s
	72 °C	1min
Final extension	72 °C	10min

9. Pool the PCR reactions with the same index in step 8. Directly purify the PCR reactions with the Zymo DNA Clean & Concentrator Kit. Elute the samples with distilled water.

c. Second round library amplification and complexity estimation for mRNA or plasmid DNA

10. Assemble qPCR reactions using part of the eluents from step 7 or step 9 to find the maximum PCR cycles numbers:

Reagent	Volume(µl)
PowerUp SYBR master mix	10
MPP2_$X^{\#}$ (10µM)	1
MPP4 (10µM)	1

Eluent from step7	5
dH_2O	3

Load 20µl of the mixed reaction to the qPCR 96-well plate.

#MPP2_X represents the same MPP2 indexed primer used in step 6 or step 8.

Run qPCR as follows:

UDG activation	50 °C	10min
Initial denaturation	95 °C	2min
25 cycles	95 °C	15s
	60 °C	30s

For each sample, check the amplification curve to determine the cycle number before the plateau. Use this cycle number as a cap for the second-round PCR cycles.

11. Set up PCR reaction with DNA standards for library complexity calculation. Make dilutions of the plasmid library to generate DNA standards with the following concentrations: 0.2ng/µl, 0.1ng/µl, 0.05ng/µl, 0.02ng/µl, 0.01ng/µl, 0.005ng/µl, 0.002ng/µl.

Reagent	**Volume(µl)**
Q5 2x master mix	10
MPP1 (10µM)	1
MPP2 (10µM)	1
DNA standards	5
dH_2O	3

12. Assemble the second-round PCR run to generate the UMI measurement libraries.

Reagent	**Volume(µl)**
Q5 2x master mix	10
MPP2_X (10µM)	1
MPP4 (10µM)	1
Eluent from step 6 or 8	5

dH_2O	3

Set up multiple reactions for the same sample until all the eluents from step 7 or 9 are used.

For both library amplification reactions and the standard reactions from step 11, run PCR as follows:

Initial denaturation	98 °C	30s
7-11 cycles	98 °C	10s
	60 °C	30s
	72 °C	30s
Final extension	72 °C	2min

*PCR cycles should be determined by running different PCR cycles with the real library material and find the lowest cycles with visible library band on the agarose gel. PCR cycle numbers should be less than the cycle number determined in step 10. For a subpool of 0.2M colonies, the optimized second-round PCR cycle numbers are 11 cycles for mRNA (step 7) and 8 cycles for plasmid DNA (step 9). For a subpool of 0.6M colonies, the PCR cycles for DNA can be lowered to 7 cycles.

13. Mix 20µl PCR reactions of libraries or standards with 4µl 6x loading dye. Run 20µl of each mixture on a 2% agarose gel. Estimate the library complexity by comparing the band intensity of the libraries (377bp) with the band intensity of the amplicon (342bp) generated from the DNA standards. Use ImageJ for band intensity quantification.
14. Pool the remaining second-round PCR reactions of UMI measurement libraries for the same sample (same index). Purify the reactions with the Zymo DNA Clean & Concentrator Kit. Elute with 20ul distilled water. Resolve the eluents on a 2% agarose gel. Gel purify the band of libraries (377bp) with the Zymoclean™ Gel DNA Recovery Kit. Elute the UMI measurement libraries with 20µl distilled water.

15. Quantitate the libraries with Qubit Fluorometer
16. Mix UMI measurement libraries (generated from both mRNA and plasmid DNA) equally and sequence on Hiseq3000 PE150 or Novaseq SP PE150 with 15% PhiX spike-in and custom sequencing primers (UMI.R1.seq, Read2.seq, and Index7.seq).

CHAPTER 3

RNA editing in cancer impacts mRNA abundance in immune response pathways

3.1 Abstract

RNA editing generates modifications to the RNA sequences, thereby increasing protein diversity and shaping various layers of gene regulation. Recent studies have revealed global shifts in editing levels across many cancer types, as well as a few specific mechanisms implicating individual sites in tumorigenesis or metastasis. However, most tumor-associated sites, predominantly in noncoding regions, have unknown functional relevance. Here, we carry out integrative analysis of RNA editing profiles between epithelial (E) and mesenchymal (M) tumors, since epithelial-mesenchymal transition (EMT) is a key paradigm for metastasis. We identify distinct editing patterns between E and M tumors in seven cancer types using TCGA data, an observation further supported by single-cell RNA-seq data and ADAR perturbation experiments in cell culture. Through computational analyses and experimental validations, we show that differential editing sites between E and M phenotypes function by regulating mRNA abundance of their respective genes. Our analysis of >120 RNA-binding proteins revealed ILF3 as a potential regulator of this process, supported by experimental validations. Consistent with the known roles of ILF3 in immune response, E-M differential editing sites are enriched in genes involved in immune and viral processes. The strongest target of editing-dependent ILF3 regulation is the transcript encoding PKR, a crucial player in immune and viral response.
Our study reports widespread differences in RNA editing between epithelial and mesenchymal tumors and a novel mechanism of editing-dependent regulation of mRNA abundance. It reveals

the broad impact of RNA editing in cancer and its relevance to cancer-related immune pathways.

3.2 Introduction

RNA editing, the modification of specific nucleotides in RNA sequences, expands diversity in proteins and gene regulatory mechanisms[172,173]. The most frequent type of RNA editing in human cells is A-to-I editing, which refers to the deamination of adenosine (A) to inosine (I) and is catalyzed by the Adenosine Deaminases Acting on RNA (ADAR) family of enzymes[174]. Three ADAR genes are encoded in the human genome, namely ADAR1, ADAR2 and ADAR3. Catalytically active ADAR1 and ADAR2 are widely expressed across tissues. In contrast, ADAR3 is exclusively expressed in certain brain regions and is catalytically inactive[175]. As inosine is recognized as guanosine (G) in translation and sequencing, A-to-I editing is also referred to as A-to-G editing. Though millions of editing events have been revealed across the human transcriptome, only a small proportion of editing events have been functionally characterized. The effects of most editing sites, primarily within non-coding regions, have yet to be discovered[58,176].

Increasing evidence has established the importance of RNA editing dysregulation in cancer. A number of studies have delineated mechanisms through which individual RNA editing sites, mostly causing recoding events (i.e., amino acid changes), promote or suppress tumor development[82,89,177,178]. Besides functioning in tumorigenesis, edited RNA transcripts can be translated into edited peptides, which may be recognized as cancer antigens and activate an anti-tumor immune response[87,88]. Furthermore, across various cancer types, genome-wide aberrations in RNA editing were observed and associated with clinical features[77,78,179]. Within

each cancer type, editing levels generally increased or decreased in tumors, compared to matched normal samples. Editing levels of specific sites were correlated with tumor stage, subtype, and patient survival, and for a smaller subset of nonsynonymous sites, editing altered cell proliferation and drug sensitivity in cell line experiments[77]. As RNA editing has the potential to inform development of improved cancer diagnostics and patient-specific treatments, thorough investigation of the precise functions and regulatory mechanisms of the many cancer-type-specific RNA editing changes is needed[178].

In cancer progression, activation of epithelial-mesenchymal transition (EMT) facilitates metastasis by enabling tumor cells to gain an invasive phenotype, infiltrate the circulatory and lymphatic systems, and reach distant sites for colonization[180–182]. A few RNA editing sites have been associated with this process so far. Specifically, editing events in SLC22A3, FAK, COPA, RHOQ, and miR-200b were demonstrated to accelerate metastasis[80,87,183–186]. While miR-200b normally targets ZEB1 and ZEB2, which are key EMT-inducing transcription factors, editing alters its targets and enhances cell invasiveness and motility[186]. The SLC22A3 recoding event also promoted EMT, causing expression changes in EMT marker genes[183]. In contrast, a recoding event in GABRA3 inhibited metastasis and was present only in non-invasive cell lines and non-metastatic tumors[185]. These studies highlight the functional relevance of RNA editing in metastasis and the merit of a more comprehensive investigation.

Here, we present a global analysis and comparison of RNA editing profiles between epithelial (E) and mesenchymal (M) phenotypes of primary tumors across multiple cancer types. Using RNA-seq data derived from bulk tumors and single cells, we observed distinct editing patterns between phenotypes, with editing differences often enriched among immune response pathway genes. Supported by experimental evidence, we show that differential editing sites affect host gene mRNA abundance and identify a novel mechanism of editing-dependent

stabilization of mRNAs by ILF3. One of the target genes of ILF3 is EIF2AK2, coding for Protein Kinase R (PKR), a key player in immune and viral response.

3.3 Results

3.3.1 Altered RNA editing profiles between epithelial and mesenchymal tumors

EMT is known to be accompanied by substantial transcriptome remodeling[181,187–191]. Given the previously reported functional relevance of RNA editing in EMT[183,186,192], we hypothesized that epithelial and mesenchymal tumors possess different transcriptome-wide RNA editing profiles. Thus, we analyzed RNA-seq datasets of primary tumors from The Cancer Genome Atlas (TCGA). We focused on seven cancer types that have been previously studied in the context of EMT and have relatively large sample sizes available from TCGA (Fig. 1A). To classify tumors into epithelial (E) and mesenchymal (M) phenotypes, we utilized a well-established EMT scoring system, where scoring and categorization of tumors into these E and M phenotypes enabled systematic identification of cancer-specific differences in treatment response between phenotypes, as well as associations with survival[193]. Of all categorized tumors for each cancer type, we further refined the subset of tumors such that metadata were matched between the two groups (Supplementary Table 1).

Applying our previously published methods [74,172,194], we quantified editing levels at over 4 million editing sites recorded in the REDIportal database[14]. We then identified sites that were differentially edited between E and M tumors in each cancer type. To control for false discoveries, we filtered out predicted differential editing sites that repeatedly exhibited differences in editing when phenotype labels were shuffled randomly. Principal components

analysis on differential editing levels showed that E and M tumors could be well separated by the first two principal components of editing (Fig. 1A). These first two principal components did not appear to be confounded by sample metadata and suggest that most of the variation in editing is explained by the distinction of E and M phenotypes (Supplementary Fig. 1).

Based on the differential editing sites, most cancer types, including LUAD, LUSC, PRAD, KIRC and HNSC, demonstrated a hyperediting trend in the M phenotype (Fig. 1B). In contrast, two cancer types, BRCA and OV, had a trend of hypoediting in the M samples. The majority of differential editing sites in all cancer types were located in the 3' untranslated regions (UTRs) or introns (Fig. 1C). The above results suggest that distinct RNA editing profiles exist between E and M phenotypes.

3.3.2 Editing patterns are shared among cancer types and distinct from differential expression

Given dominant trends of hyperediting or hypoediting that distinguished E and M phenotypes in an individual cancer type, we asked whether genes with differential editing patterns were shared or distinct across cancer types. We examined the statistical significance of overlap in differentially edited genes between pairs of cancer types by Rank-rank Hypergeometric Overlap (RRHO). Extending Gene Set Enrichment Analysis (GSEA) to two dimensions, RRHO tests the significance of the intersection of gene lists, ranked by a metric of differential expression, across two genome-wide datasets[195]. We applied RRHO to RNA editing here by ranking genes according to the significance of tested editing differences between E and M and the direction of editing differences (Methods). In addition to shared directionality of global editing trends, we found significant overlap in genes with editing changes among multiple cancer types (Fig. 2A).

Within pairs of cancer types, most significant overlaps were enriched at the bottom left or top right corners, where genes were hyperedited or hypoedited in both cancer types, respectively. These significant overlaps in genes based on differential editing suggest that editing changes in EMT may affect common pathways across cancer types.

It should be noted that differentially edited genes do not overlap with differentially expressed genes (Fig. 2B). This observation indicates that gene expression changes in EMT did not confound the RNA editing differences observed. Thus, altered editing potentially represents a distinct layer of molecular changes in EMT.

3.3.3 Differential editing occurs in genes of immune relevance

Next, we examined the gene ontologies enriched among genes with differential editing in EMT. In this analysis, background control genes were chosen randomly from those that did not have differential editing sites but had similar gene length and GC content as the differentially edited genes (Methods). Across multiple cancer types, differentially edited genes were enriched with viral-host interaction features, interferon (IFN) and other immune response pathways, metabolic processes, and translational regulation (Fig. 2C, Supplementary Fig. 2).

The observation of immune-relevant categories is of particular interest. RNA editing has been described as a mechanism to label endogenous double-stranded RNAs and consequently prevent IFN induction[61,196–199]. However, the roles of editing events in genes directly associated with immune response, such as those in the IFN response pathways, have not been well characterized. Our observation indicates that RNA editing may directly affect immune response genes in EMT.

3.3.4 Contribution of cell types to differential editing

Given the observed enrichment of differential editing in immune-relevant genes, we asked whether our identified differential editing events primarily occur in cancer cells or in other cell types in the tumor microenvironment. To address this question, we analyzed single cell (sc) RNA-seq data from three non-small cell lung cancer (NSCLC) patients, each with three tumor samples from the tumor edge, core, and in-between[200]. Following quality control measures, we clustered the cells in two rounds and labeled cell types based on marker genes to obtain T cells, B cells, myeloid cells, endothelial cells (EC), fibroblasts (Fibro), epithelial cells (Epi), mast cells, alveolar cells, and cancer cells (Supplementary Fig. 3A-C, Methods). Supporting the accuracy of this clustering, expression of marker genes was generally highest in their expected cell types when RPKM was calculated from pooled cells and when a signature gene expression matrix was predicted by CIBERSORTx[201] (Supplementary Fig. 3D).

To gauge the contribution of individual cell types to bulk tumor differential editing, we examined gene expression and editing profiles of each cell type. Specifically, we pooled cells of each type and calculated the percent of differentially edited genes from the bulk tumor analysis that were expressed in each cell type. Cancer cells expressed the highest proportion of genes that were differentially edited (Fig. 3A). We then measured the extent of editing in each cell type by calculating the percent of bulk tumor differential editing sites that were edited. Consistent with the expression analysis, the highest proportion of differential sites were edited in cancer cells (Fig. 3B). Therefore, the editing differences observed among bulk tumors may be mainly attributable to the cancer cells.

We next separated cancer cells to epithelial and mesenchymal cell clusters (Fig. 3C, Methods). Sampling epithelial cells to match mesenchymal cells in terms of cell number (200 cells) and metadata, we pooled cells within each phenotype together and detected RNA editing events (Supplementary Fig. 4). Although the scRNA-seq primarily sequences the 3' ends of mRNAs, a relatively small number of RNA editing events were still captured. We identified nine editing sites with significant differences between E and M (Fig. 3D). All nine differential sites exhibited higher editing levels in the M phenotype, which is consistent with the hyperediting trend in M observed in bulk LUAD and LUSC tumors (Fig. 1B). Two sites overlapped with differentially edited sites in LUAD or LUSC and both had hyperediting in M cells, consistent with the direction in bulk tumors (Supplementary Fig. 5). This small overlap likely reflects the low coverage on editing sites in the single cell data, and/or the possibility that more differential editing sites, which were not identified in our study due to limits in power, exist in the bulk tumors.

Notable differentially edited genes include RHOA, which is active in cell migration and is associated with metastasis in multiple cancer types[202–204], and ARL16, a reported negative regulator of RIG-I activity[205], consistent with the observed enrichment of immune-relevant genes that were differentially edited in bulk tumors. Overall, the findings from single cell data support the hypothesis that editing differences between bulk E and M tumors mainly reflect changes occurring in cancer cells.

3.3.5 ADAR1 or ADAR2 knockdown induced EMT

Given the differential editing profiles between E and M tumors, an important question is whether the editing changes are functionally relevant to EMT. To address this question, we first

examined if changes in ADAR expression affect EMT. Using cell culture systems commonly employed in EMT studies, we carried out knockdown (KD) experiments of ADAR1 or ADAR2 in two cell lines, A549 and MCF10A, via siRNAs. Upon ADAR1 KD, A549 cells showed elongated spindle-like mesenchymal morphology (Fig. 4A). We also confirmed the loss of epithelial markers (E-cadherin and γ-Catenin) and gain of mesenchymal marker (Vimentin) in ADAR1 KD A549 cells (Fig. 4B). Similar results were observed upon ADAR2 KD in A549 cells (Fig. 4C-D) and reproducible in MCF10A cells (Fig. 4E-F). These findings suggest that loss of either catalytically active ADAR enabled EMT in the two cell lines. The phenotypic changes following ADAR2 KD are consistent with a previous report that ADAR2-deficiency can induce EMT in SW480 cells[192]. Together, these results indicate that knockdown of ADARs promotes EMT.

As expected, ADAR KD induced significant editing changes measured by RNA-seq in A549 cells (Supplementary Fig. 6A-B), with ADAR1 KD affecting a large number of editing sites but ADAR2 having fewer targets. A minority of ADAR2-responding sites had increased editing upon ADAR2 KD, reflecting the likely compensation by ADAR1. The reverse, compensation of ADAR1 loss by ADAR2, was not observed. Among the lung cancer E-M differential editing sites that were testable in the above A549 RNA-seq data, the vast majority responded to KD of either ADAR or double KD (Supplementary Fig. 6C). These results confirm the impairment of RNA editing at genome scale upon the loss of ADARs.

We next examined mRNA expression of ADARs in the bulk E and M tumors across cancer types. In several cancer types with a hyperediting trend in M, higher mRNA expression of ADAR1 or ADAR2 likely contributed to increased editing levels in M tumors (Supplementary Fig. 7). However, ADAR expression was not consistent with RNA editing differences for some cancer types. Thus, although ADAR KD caused EMT in cell culture models, ADAR expression alone may not sufficiently explain the global editing trends observed in bulk tumors.

3.3.6 Impact of RNA editing on mRNA abundance

Given ADAR's primary role in RNA editing, we next asked how RNA editing may affect genes relevant to EMT, especially those related to immune response (Fig. 2C). Since a relatively large fraction of differential editing sites is located in 3' UTRs, we examined the hypothesis that these sites may affect mRNA abundance of their respective genes. Thus, we first calculated the correlation between editing levels and mRNA abundance for differentially edited sites observed in the E-M comparison. Using a regression model accounting for confounding factors including age, gender and race, we observed a total of 127 genes whose editing sites are significantly correlated with mRNA abundance (FDR<10%) in at least one type of cancer (Fig. 5A). In addition, among these genes, 77% (94 of 122 testable genes) demonstrated a significant correlation in at least one human tissue type based on a similar analysis of GTEx data, 78% (73/94 genes) of which showed the same direction of correlation between cancer and at least one GTEx tissue.

To further evaluate the regulatory role of RNA editing on mRNA abundance, we next examined the change in mRNA expression levels upon ADAR1 KD. We used ADAR1 KD RNA-Seq data from 5 cell lines: U87, HepG2, K562, HeLa and B cells[172,206,207], respectively. Out of the 127 edited genes identified above, 126 of them were detectable at an expression level of at least 1 FPKM (and edited) in at least one cell line (control or ADAR1 KD condition). Among them, 71% (89 genes, red dots, Fig. 5B) showed inverse correlation between ADAR1 KD and editing level coefficient in at least one cell line (Fig. 5B). These genes showed an enrichment of negative expression changes upon ADAR1 KD, indicating a likely stabilizing effect imposed by RNA editing ($p = 2.7e-4$, binomial test). Among expression-correlated editing sites in the 89

genes, 64% are located in 3' UTRs, a percentage that's significantly higher than that of E-M differential editing sites in general (p = 2.4e-4, Fig. 5C). We thus refer to the 89 genes as putative target genes whose expression is modulated by RNA editing (Supplementary Table 2).

Next, we experimentally validated the regulation of mRNA abundance by six editing sites within three genes: RNF24, RHOA, and MRPS16. We used a minigene reporter with bi-directional promoters for mCherry and eYFP[37] and cloned edited and unedited versions of each editing site and its surrounding 3' UTR region into the 3' UTR of mCherry. Using expression of eYFP as an internal control, we compared mCherry expression between cells carrying the edited and unedited versions for each editing site. All six editing sites induced significant expression differences in the direction consistent with the editing-expression correlations observed in primary tumors (Fig. 5D, Supplementary Table 3). While positive editing associations were dominant among predicted target genes, there also exist negative associations between editing and expression levels. We tested one example of the latter category (RHOA).

3.3.7 ILF3 as an editing-dependent regulator of mRNA abundance

Since mRNA stability is closely regulated by RNA-binding proteins (RBPs)[208–211], we next asked whether RBPs are involved in the modulation of mRNA abundance by RNA editing sites. To this end, we analyzed enhanced ultraviolet crosslinking and immunoprecipitation (eCLIP) datasets of 126 RBPs in two cell lines (HepG2 and K562) from ENCODE[206,212]. We asked whether RBP binding signals are enriched significantly closer to editing sites in the 89 potential target genes than expected by chance. This analysis identified ILF3 as a top protein with significantly short distances to the editing sites in both cell lines (Supplementary Fig. 8A). To validate this finding

and test this relationship in a different cell type, we performed eCLIP-seq of ILF3 in A549 cells. The same observation was made via this data set (Fig. 6A). As observed in HepG2 and K562 cells, differential editing sites within predicted target genes were significantly closer to ILF3 binding regions in A549 cells than random gene-matched control sites. Furthermore, 75 (84%) of the 89 genes showed a significant correlation between their gene expression and the expression of ILF3 (FDR<10%), 37 of which had an absolute correlation coefficient of at least 0.2 (Fig. 6B). Importantly, the majority of the significant correlations were positive, consistent with the known roles of ILF3 in stabilizing its target mRNAs[213–215].

3.3.8 Impact of ILF3 on immune-relevant genes

ILF3 promotes an antiviral response through its binding to RNAs[216–218]. Given the fact that immune-relevant genes are differentially edited in E-M (Fig. 2C), we next asked whether ILF3 regulates the mRNA abundance of these EMT-associated differentially edited, immune-relevant genes. Among the 89 genes whose expression was affected by RNA editing, 20 genes fall into the immune or viral GO categories. Interestingly, the ILF3 binding sites were significantly closer to the differential editing sites of these 20 genes than differential sites in immune-related genes without editing-expression associations (Fig. 6C). Together, these results suggest that ILF3 binds close to the editing sites of immune-related genes.

Since we observed that differential editing between bulk E and M tumors mainly reflected changes occurring in cancer cells (Fig. 3A-B), we next asked whether the above regulatory relationship between ILF3 and immune-related genes also occurs in cancer cells. To this end, we analyzed gene expression of individual cell types identified in the NSCLC scRNA-seq dataset. Within each cell type, we correlated ILF3 expression with expression of the 20

immune-related target genes. In cancer cells, all 20 genes had expression levels positively correlated with ILF3 expression at 10% FDR (Fig. 6D). Though significant correlations were also observed in other cell types, only cancer cells showed correlation coefficients of at least 0.2 in magnitude. This result suggests that the mRNA stabilizing function of ILF3 is prominent in cancer cells, in line with our observation that E-M differential editing primarily occurs in cancer cells.

3.3.9 PKR expression is affected by 3' UTR editing through ILF3 regulation

Among the 20 immune-related genes putatively regulated by ILF3, the gene EIF2AK2, coding for Protein Kinase R (PKR), had most significant expression-editing correlation (Supplemental Table 2) and expression correlation with ILF3 (Fig. 6D). Activated by dsRNA, PKR suppresses translation and promotes apoptosis through its phosphorylation activity[219,220]. PKR also regulates various signaling pathways, such as NF-κB and p38 MAPK, in response to cellular stress[219]. Using the editing minigene reporter, we examined the individual effects of seven 3' UTR editing sites on PKR mRNA abundance in A549 cells. Five of the seven editing sites showed significantly higher normalized mCherry expression compared to their unedited counterparts (Fig. 6E, Supplementary Fig. 8B). To assess the collective impact of multiple RNA editing sites on PKR mRNA abundance, we measured endogenous PKR expression in A549 cells upon ADAR1 or ADAR2 KD. We first confirmed that the 3' UTR editing sites in PKR were edited endogenously in A549 cells. Importantly, these editing sites are mainly regulated by ADAR1 instead of ADAR2 (Supplementary Fig. 8C). Upon ADAR1 KD, PKR expression level was significantly reduced by about 40% (Fig. 6F). In contrast, PKR expression did not change upon ADAR2 KD, as expected. These results suggest that the editing sites enhanced PKR mRNA abundance, consistent with the positive editing-expression correlation in primary tumors.

Based on the eCLIP data, the five editing sites that individually promoted PKR mRNA abundance are located within ILF3 binding sites (Fig. 6G, Supplementary Fig. 8D-E). To test the hypothesis that ILF3 regulates PKR mRNA abundance in an editing-dependent manner, we generated ILF3 KD A549 cells (Fig. 7A). The edited and unedited reporters, demonstrating differential expression in control cells, no longer produced different expression levels upon ILF3 KD (Fig. 7B). Together, our data suggest that ILF3 promotes PKR mRNA expression in an editing-dependent manner by binding to the PKR mRNA.

3.3.10 ILF3 knockdown induced EMT in A549 cells

Since ILF3 was found to stabilize transcripts that were differentially edited between E and M tumors, we next asked if ILF3 regulates the EMT process. We carried out ILF3 KD experiments via two different siRNAs in A549 cells. Upon ILF3 KD, cell morphology changed from tightly connected, round cells towards more dispersed, spindle-shaped cells (Fig. 7C), consistent with expected EMT phenotypes. Additionally, we observed reduced expression of the epithelial marker E-cadherin along with increased expression of the mesenchymal marker N-cadherin in the ILF3 KD cells (Fig. 7D, E for protein and RNA levels, respectively). Thus, these data show that ILF3 deficiency induces EMT in A549 cells, supporting a significant role of ILF3 in regulating EMT.

3.4 Discussion

As most cancer patient deaths are due to metastasis, thorough understanding of the molecular mechanisms underlying metastasis is crucial to developing effective preventative measures[221].

EMT plasticity is thought to underlie cell dissemination and metastatic formation in many cancer types[182]. Supported by studies on primary tumors and various model systems, features of EMT have been associated with metastasis[180,182,222,223]. For instance, higher expression of mesenchymal markers, with preserved epithelial markers in the absence of nearly all canonical EMT transcription factors, was detected in cells located at the leading edge of primary human HNSC tumors[223]. Furthermore, this partial EMT program was correlated with multiple metastatic characteristics, including abundance of lymph node metastases, lymphovascular invasion, and tumor grade[223]. While mutations are understood to drive primary tumorigenesis and are often found in reported oncogenes and tumor suppressor genes, the existence of recurrently mutated genes specific to metastasis is not clear[182]. Accordingly, mechanisms regulating cell invasiveness beyond genetic variation need to be more thoroughly investigated. Our study is the first to report a systematic characterization of RNA editing in EMT phenotypes across several cancer types. Through a combination of experimental and computational analyses, we observed many editing differences in EMT-relevant genes, especially those related to immune and viral response, with the potential of affecting mRNA abundance of these genes. We also show that higher expression levels of these edited transcripts may be due to stabilization by ILF3.

Located in noncoding regions, most editing sites have unknown function. To assess the contribution of differential editing to altered cell phenotypes in cancer, we focused on the capacity of editing to regulate host gene mRNA abundance. To our knowledge, very few studies have examined this question on the transcriptome-wide scale[224,225]. Previously, several studies demonstrated this regulatory role for a handful of editing sites through alteration of miRNA binding sequences or mRNA secondary structure or otherwise unknown mechanisms[58,80,93,226–230]. Expanding on these previous studies, we incorporated tissue-rich data from GTEx and ADAR KD expression changes from five cell lines to computationally support associations of editing with mRNA abundance. We also validated the effects of specific editing sites and

explored the involvement of RBPs in this regulatory mechanism. It should be noted that we were able to detect associations between editing and mRNA abundance levels, even though differentially expressed genes did not significantly overlap differentially edited genes. These findings do not contradict each other because editing levels are relatively low. Consequently, inosine may affect mRNA abundance, but when present at low levels, may not necessarily lead to significant expression differences.

Considering tumor heterogeneity and the roles of stromal and immune cells in EMT, it is important to examine the contributions of different cell types to differential editing observed in the E-M comparisons. Our results using single-cell data supported that cancer cells are a main cell type underlying differential editing between E and M phenotypes in lung cancer, although contributions by other cell types cannot be excluded. Furthermore, cancer cells demonstrated the strongest expression correlation between ILF3 and immune-relevant differentially edited genes among all cell types considered in lung cancer. These findings suggest that RNA editing is likely an important aspect of transcriptome remodeling of cancer cells in EMT, at least in lung cancer. Single-cell analysis of RNA editing in other cancer types should be conducted in the future.

Our cell line experiments showed EMT induction upon KD of either ADAR1 or ADAR2 in lung and breast cell lines. In contrast, we observed hyperediting in M tumors of most cancer types. The seemingly opposite trends may reflect the complexity of tumor biology that is not effectively recapitulated by cell culture models. Although the cell culture models can support the likely importance of RNA editing in EMT, the exact mechanisms and related regulation can only be investigated using *in vivo* models in the future. In addition, we did not observe large differences in ADAR expression levels that are consistent with observed editing differences between E and M tumors for all cancer types. Other proteins that directly or indirectly affect

ADAR function likely contribute to the regulation of E-M RNA editing differences, which remains to be investigated.

RNA editing is known to be important to innate immunity by preventing viral dsRNA sensors, such as MDA5 and RIG-I, from sensing host dsRNA[196,199,231]. In this study, we provided multiple lines of evidence to support that RNA editing differences in EMT may affect immune response genes directly, adding a new dimension to the relationships between RNA editing and innate immunity. Interestingly, a major RBP that mediates this relationship is ILF3. ILF3 was identified as a PKR substrate and serves as a negative regulator of viral replication upon phosphorylation[216,232]. Upon viral infection and sensing of viral dsRNA, PKR activates, suppresses translation, and promotes apoptosis of affected cells[220]. Importantly, this mechanism has been targeted in oncolytic virotherapy for cancer. Cancer cells that have low PKR expression are sensitive to oncolytic viruses[233–235]. Our study showed that ILF3 mediates the RNA editing-dependent regulation of PKR expression. We also observed that ILF3 KD induced EMT in A549 cells. These data reveal novel insights into the reciprocal regulation between PKR and ILF3 and their potential contributions to EMT. Additional studies on their interaction during viral infection or cancer treatment will also be informative for therapeutic development. Previously, ADAR1 loss has been shown to render tumor cells sensitive to immunotherapy through enhanced inflammatory response[70,236]. Our findings on the regulation of immune response genes by RNA editing may add additional mechanisms in this process that will need further investigation.

The functional roles of RNA editing in cancer have been increasingly recognized in recent years. Highlighting the extensive editing differences between EMT phenotypes and their impact on mRNA abundance, especially for genes involved in the immune response, our work

extends the basis for future studies on the contribution of editing to metastasis and patient outcomes.

3.5 Methods

3.5.1 Plasmid construction

For bi-directional reporters, full length or partial 3' UTR regions (1~2kb) of candidate genes were cloned from the genomic DNA extracted from HMLE or A549 cells. Edited versions of 3' UTR inserts were generated using overlap-extension PCR (Supplementary table 3). Edited and unedited versions of 3' UTR regions were then cloned into the pTRE-BI-red/yellow vector via ClaI and SalI-HF enzyme sites[37]. To obtain a lentiviral vector expressing ILF3 shRNA, oligos containing the target sequence (GGTCTTCCTAGAGCGTATAAA, TRCN0000329788) were ordered from Integrated DNA Technologies (IDT) and cloned into pLKO.1 via EcoRI and AgeI enzyme sites.

3.5.2 Cell culture and transfection

A549, Hela and HEK293T cells were maintained in DMEM with 10% FBS and Antibiotic-Antimycotic reagent (Gibco). MCF10A cells were maintained in DMEM/F12, supplemented with 5% Horse serum, 20ng/ml human EGF (PeproTech), 0.5mg/ml Hydrocortisone (Sigma), 100ng/ml Cholera Toxin (Sigma), 10ug/ml Insulin (Sigma), and Antibiotic-Antimycotic reagent (Gibco). For siRNA treatment, A549 or MCF10A cells were seeded at 1×10^5 cells per well in 6-well plates. After 24 hours, siRNAs (Supplementary table 3) were introduced at the final

concentration of 10nM~100nM using lipofectamine RNAiMAX (Invitrogen) according to the manufacturer's protocol. Media were changed 24 hours post-transfection, and cells were harvested 72 hours post-transfection. For transfection of bi-directional reporters, Hela and HEK293T cells were seeded in 12-well plates to reach 90% confluency by the time of transfection. A549 cells were seeded at $0.15x10^5$ cells per well in 12-well plates 24 hours before transfection. Reporter plasmids were transfected at 200ng per 12-well with lipofectamine 3000 (Invitrogen), following the manufacturer's protocol. Cells were harvested 16 hours post-transfection.

3.5.3 Western blot

Cells were lysed with RIPA buffer containing protease inhibitor (EDTA-free, Thermo Fisher Scientific) at 4°C for 30 minutes. The whole cell lysates were then centrifuged at 12,000g, 4°C for 15 minutes. The supernatants were collected for protein concentration measurement using Bradford assay (Pierce™ Detergent Compatible Bradford Assay Kit, Thermo Fisher Scientific). Protein samples were prepared by mixing protein lysates with 4x SDS protein loading dye at 3:1 ratio. The mixture was boiled for 5 minutes. 10 ug of each protein samples were loaded on SDS-PAGE gels and transferred to nitrocellulose membranes for antibody incubations. Antibodies used were as follows: ADAR1 antibody (Santa Cruz Biotechnology, sc-73408, 1:200), ADAR2 antibody (Santa Cruz Biotechnology, sc-73409, 1:200), E-cadherin antibody (Cell Signaling Technology, #3195, 1:1000), γ-Catenin antibody (BD Transduction Laboratories, 610253, 1:8000), N-cadherin antibody (BD Transduction Laboratories, 610920, 1:500), Vimentin antibody (Cell Signaling Technology, 5741, 1:1000), NF90(ILF3) antibody (BETHYL Laboratories, A303-651A, 1:1000), β-actin-HRP antibody(Santa Cruz Biotechnology, sc-47778, 1:2000), goat anti-rabbit IgG-HRP(Santa Cruz Biotechnology, sc-2004, 1:2000), goat anti-

mouse IgG-HRP(Santa Cruz Biotechnology, sc-2005, 1:2000). Membrane blots were incubated with SuperSignal West Pico PLUS Chemiluminescent Substrate (Thermo Fisher Scientific) and visualized under the imager (Syngene PXi).

3.5.4 RNA isolation and real-time qPCR

Cells were lysed using TRIzol (Thermo Fisher Scientific). Total RNA was isolated using Direct-zol RNA Miniprep Plus kit (Zymo Research) following the manufacturer's protocol. 2 ug of total RNA was used for cDNA synthesis with SuperScript IV (Thermo Fisher Scientific). The real-time qPCR reaction was assembled using the PowerUp™ SYBR® Green Master Mix (Thermo Fisher Scientific). Primers used for qPCR are listed in Supplementary Table 3. The reaction was performed in the CFX96 Touch Real-Time PCR detection system (Bio-Rad) with the following settings: 50°C for 10 minutes, 95°C for 2 minutes, 95°C for 15 seconds, 60°C for 30 seconds, and with the last two steps repeated for 45 cycles. For bi-directional reporter assays, mCherry expression was normalized against eYFP expression within the same sample. ILF3 expression was normalized against the expression of internal control gene *TBP*. For qPCR validating the eCLIP peaks, the final libraries were diluted to the same concentration at 0.01ng/ul. 5ul of diluted libraries were used in each qPCR reaction. Around 80 bp upstream each EIF2AK2 editing site was amplified. The expression of each EIF2AK2 region was normalized against the expression of 18s.

3.5.5 Quantification of RNA editing levels by Sanger sequencing

Regions of interest were amplified from cDNA using Thermo Scientific™ DreamTaq™ Green PCR Master Mix (2X). Primers used for PCR are listed in Supplementary Table 3. The

amplicons were gel extracted and premixed with the reverse primer for Sanger sequencing. The peak signals of A and G nucleotides were measured by 4Peaks for editing level calculation (G/(A+G)).

3.5.6 Categorization of tumors as epithelial and mesenchymal

We downloaded fragments per kilobase million (FPKM) data of primary tumors from patients across seven cancer types in TCGA: BRCA, LUAD, LUSC, PRAD, OV, KIRC, and HNSC, from the Genomic Data Commons (GDC) Data Portal[237]. To assess E and M phenotypes of the tumors of each cancer type, we quantified the enrichment of E and M gene sets by applying gene set variation analysis (GSVA)[238]. We obtained pan-cancer E and M gene sets from a 2014 publication by Tan and colleagues (Table S1A from their publication)[193]. Tumors with high E scores and low M scores were considered to have an E phenotype, while tumors with low E and high M scores were classified as M. Subsets of E and M tumors were selected for each cancer type to minimize confounding of E and M distinction by patient and sample metadata.

3.5.7 Quantification and comparison of RNA editing levels in TCGA tumors

We downloaded RNA-seq fastq files of categorized tumors from the GDC Legacy Archive. We mapped reads to hg19 with HISAT2, using default parameters. Dense clusters of editing sites, or hyperedited regions, can lead to many mismatches in reads. Consequently, these reads may be left unmapped and hinder accurate detection of editing in these regions. To rescue reads that were originally unmapped due to high density of editing activity, we applied a hyperediting pipeline and combined the recovered reads with uniquely mapped reads for downstream analyses[74,239]. To analyze editing sites of high confidence, we downloaded the REDIportal

database, comprising over 4 million editing sites identified across 55 tissues of 150 healthy humans from GTEx [14,240]. We applied methods used in our previous studies to detect editing at REDIportal sites in the tumor samples. We filtered out editing sites found in dbSNP (version 147) and COSMIC (version 81), except for reported cancer-related editing sites[77,81,177,183,241–243], since editing sites have been shown to be mistakenly recorded as SNPs[5,244]. Within each sample, we also filtered out editing events that overlapped with sample-specific somatic mutations and copy number variants. Somatic variants were obtained from the publicly released MC3 MAF[245], and copy number variants were obtained from copy number segment data downloaded from the GDC data portal.

Differential editing sites were defined as editing sites with significantly different editing levels between E and M phenotypes. To identify such sites, we used an adaptive coverage approach[74]. For an individual editing site, we determined the highest read coverage threshold that was satisfied in at least five samples of both phenotypes, among twenty, fifteen, and ten reads. If none of these thresholds was satisfied and fewer than ten samples in each phenotype had at least five reads covering the site, we did not test the site for differential editing. Using the highest coverage determined, we calculated the mean editing levels among samples of each phenotype separately. We then consecutively lowered the read coverage threshold by 5 reads and compared the new mean editing levels of each phenotype, when including additional samples, to the original high-coverage-only editing means. If the differences in mean editing levels were less than 0.03, we used the lower read coverage threshold to delineate which samples to include for the differential test. Editing levels between E and M samples were compared by a Wilcoxon rank-sum test. Editing differences were considered significant if the Wilcoxon p-value < 0.05 and the magnitude of the difference ≥ 0.05. To account for false positives, we shuffled phenotype labels and retested for significant differences for each

differential editing site, 100 times. If a site showed significant differences for shuffled labels over ten times, it was filtered out and no longer considered a differential editing site.

3.5.8 Identification of differentially expressed genes

HTSeq-Count data were downloaded from the GDC data portal. We identified genes with significantly different mRNA expression levels between E and M tumors of each cancer type, using limma-voom[246]. Metadata significantly correlated with the top two principal components of expression were included as covariates in the linear models. Expression differences were considered significant if log2-fold change was at least 1 and adjusted p-value was less than 0.05.

3.5.9 Rank-rank hypergeometric overlap

To measure the similarity in patterns of editing changes across cancer types, we ranked genes based on differential editing between E and M phenotypes for each cancer type. More specifically, the ranking metric was the statistical significance of the differential editing test (-log10(Wilcoxon p-value)), multiplied by the sign of the editing difference (mean of M editing levels – mean of E editing levels). Accordingly, genes at the top of the ranked list had the highest increases in editing in M, while genes at the bottom had the largest decreases in editing in M. For each gene with multiple editing sites tested, the site with the most significant change in editing levels was used to represent the gene. We used the RRHO package within Bioconductor in R to test for significance of overlap between ranked gene lists, with a step size of 30 genes between each rank[247].

We also ran RRHO between gene rankings by differential editing and differential gene expression for each cancer type. To order genes based on differential gene expression, genes were ranked according to the signed statistical significance of differential expression tests (signed by the direction of expression change in M). As a result, genes at the top of the list were more highly expressed in M and genes at the bottom, more lowly expressed in M.

To make RRHO maps comparable across cancer types and across overlaps based on differential editing and differential expression, we scaled the log-transformed p-values to account for different lengths of gene lists and then applied the Benjamini-Yekutieli correction for multiple testing[195].

3.5.10 Gene ontology enrichment analysis

To evaluate whether an individual GO term was enriched in differential editing in one cancer type, we compared the occurrence of the term among query genes – genes containing differential editing sites – to its occurrences within 10,000 sets of control genes. In each set, one control gene for each query gene was randomly selected among non-differentially edited genes that matched the query gene based on gene length and GC content (within 10%). Query genes that did not have at least ten matched control genes were excluded. We calculated the p-value of the term's enrichment among query genes from the normal distribution fit to occurrences of the term among control gene sets. We repeated this assessment of GO term enrichment separately for lists of differential hyperedited and hypoedited genes in each cancer type.

Likewise, we tested the occurrence of each GO term represented among differentially expressed genes to its occurrences among 10,000 sets of non-differentially expressed control

genes, randomly selected to match the differentially expressed query genes for gene length and GC content.

3.5.11 scRNA-seq dataset analysis

We downloaded fastq files from 15 tumor samples of five NSCLC patients[248] and ran CellRanger (version 3.0.2) to map reads and obtain count matrices. We excluded the tumor samples from three LUSC patients exhibiting low percentages of valid barcodes and mapped reads. For the remaining samples, we loaded the filtered feature-barcode matrices from CellRanger and merged the datasets into a single Seurat object with the R package Seurat[249] (version 3.0.2). Next, we filtered out cells that did not meet the following criteria: 101-6000 expressed genes, over 200 UMIs, and less than 10% UMIs corresponding to the mitochondrial genome. Following normalization by sctransform[250] (version 0.2.0), we performed dimensional reduction with PCA. Based on an elbow plot, we decided to consider the first ten PCs for downstream clustering and TSNE embedding. To assign cell identity labels to clusters, we matched differentially expressed genes of clusters to reported marker genes. One cluster had differentially expressed markers of multiple cell types, so we subclustered its cells. To assess the accuracy of our final labeling of nine cell types, we examined expression of marker genes across the cell types in two approaches. In one approach, we used CIBERSORTx[251] to generate a gene expression signature matrix, which is a matrix of expression signatures characterizing cell types. To create this matrix from expression profiles of single cells labeled by cell type, CIBERSORTx identified differentially expressed genes. In the second approach, we pooled reads of each cell type together and calculated RPKM. These RPKM values calculated from pooled cells were also used to correlate ILF3 expression with expression of editing-correlated genes.

To identify cancer cells with E and M phenotypes, we subclustered the cancer cells. To this end, we first ran sctransform and PCA on only the cancer cells. Using the first twelve PCs, we clustered the cells and performed non-linear dimension reduction by UMAP. As a cluster of 200 M cells was identified, we sampled 200 E cells with similar numbers of features, numbers of UMIs, and percentages of reads mapped to the mitochondrial genome. For each phenotype, we compiled reads of cells together and detected editing levels at REDIportal sites. For each testable editing site, E and M editing levels were compared by a Fisher's Exact test. An editing site was considered differential if the difference in editing levels was at least 0.05 and the Fisher's Exact p-value < 0.05.

3.5.12 RNA-seq generation for ADAR KD A549 cells

A549 cells were seeded at 1×10^5 cells per well in 6-well plates 24 hours before siRNA transfection. siRNAs (Supplementary Table 3) were introduced at the final concentration of 22nM using lipofectamine RNAiMAX (Invitrogen), according to the manufacturer's protocol. For individual KD of ADAR1 or ADAR2, 11nM siRNA of ADAR1 or ADAR2 were mixed with 11nM control siRNAs. For double KD of ADAR1 and ADAR2, 11nM siRNA of ADAR1 and 11nM siRNA of ADAR2 were mixed. Media were changed 24 hours post-transfection. The transfected cells were harvested 48 hours post-transfection. Total RNA was extracted for RNA-seq library generation for three biological replicates of each condition. RNA sequencing libraries were generated using NEBNext Ultra II Directional RNA library Prep kit and NEBNext multiplex oligos for Illumina according to the manufacturer's instructions (New England Biolabs, E7760S). Library concentrations were measured by Qubit fluorometric assay (Life Technologies), and libraries were sequenced on an Illumina HiSeq-4000 with 150-bp paired-end reads.

3.5.13 A549 ADAR KD RNA-seq analysis

Following mapping of RNA-seq reads with HISAT2 and a hyperediting pipeline[74], we detected editing events at REDIportal sites as we did for the TCGA tumor samples. We then removed dbSNP variants while retaining previously reported cancer editing sites. To identify differential editing sites between each ADAR KD condition and control or between each individual ADAR KD and double KD, we used REDIT-LLR on sites that were edited in the control condition (editing level ≥ 0.05)[72]. A site was considered differentially edited if the difference in mean editing levels between conditions was at least 0.05 and REDIT-LLR p-value < 0.05.

3.5.14 Regression analysis

For each differential editing site, association between editing level and host gene mRNA abundance was tested by fitting a linear model of log-transformed gene FPKM against editing level and potentially confounding covariates (using the lm function in R). For associations in GTEx data, we included age, gender, and race as covariates. For associations in TCGA data, we included metadata that were significantly correlated with the top two principal components of expression, as in the differential expression analysis.

3.5.15 eCLIP-seq generation

Following a published protocol[212], we performed an eCLIP experiment comprising three libraries from two ILF3-immunoprecipitated biological replicates and one control. The antibody used for this experiment is: ILF3/NF90 antibody (Bethyl Laboratories, A303-651A).

For each sample, 10M A549 cells were ultraviolet (UV) crosslinked at 254 nm (800 mJ cm^{-2}). We then performed cell lysis, RNA fragmentation, immunoprecipitation, adapter ligation, and other library preparation steps on UV crosslinked samples, as described[212]. For the size-matched input control (SMInput), we prepared a library from sampling 2% of one pre-immunoprecipitation UV crosslinked sample. This control is used to normalize binding signal, given biases that may be introduced through various experimental steps.

3.5.16 eCLIP-seq peak calling and distance analysis

We obtained eCLIP peak data for 96 RBPs in K562, 83 RBPs in HepG2, and ILF3 in A549 cells, as described previously[206]. Briefly, after demultiplexing and trimming adapters, we aligned reads in multiple rounds with STAR. First, reads aligning to rRNA sequences were discarded, and then the unmapped reads were aligned to Alu sequences, permitting a maximum of 100 alignments for an individual read. In the final alignment step, the remaining unmapped reads were uniquely aligned to the hg19 genome. Then read enrichment within a sliding window, considering both genome and Alu-aligned reads, was tested for significance by a Poisson model in order to call eCLIP peaks[206,252].

To assess the proximity of a single RBP's binding to differential editing sites compared to random controls, we calculated the distance from each differential editing site or control to the closest eCLIP peak in the same gene. Control sites consisted of adenosines within genes containing differential editing sites[74]. We then calculated the area under the curve (AUC) of the cumulative distribution of distances from differential editing sites to the closest eCLIP peaks. Given our interest in close binding, we considered distances up to 10,000 bases only for AUC calculation. Similarly, we calculated the AUC of the distribution of closest distances between

eCLIP peaks and controls, for each of 10,000 sets of random controls. We computed the p-value of the AUC for differential editing sites from the normal distribution fit to the AUC values of control sets[74].

3.6 Acknowledgements

We thank members of the Xiao and Cheng laboratories for helpful discussions and comments on this work. The results published here are in part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. We thank the GTEx consortium for generating the RNA-seq data. We thank the ENCODE Project Consortium (specifically the groups of Dr. Gene Yeo and Dr. Brenton Graveley) for generating the eCLIP-seq and RNA-seq data sets used in this study. We appreciate the helpful discussions with Dr. Eric Van Nostrand and Dr. Gene Yeo on the ILF3 eCLIP-seq experiments.

This work was supported in part by grants from the National Institutes of Health (U01HG009417, R01AG056476 to X.X. and R35GM131876 to C.C.) and the Jonsson Comprehensive Cancer Center at UCLA. C.C. is a CPRIT Scholar in Cancer Research (RR160009). T.W.C. was supported by the NIH-NCI National Cancer Institute T32LM012424.

3.7 Data Availability

eCLIP-seq and RNA-seq data sets are available at the Gene Expression Omnibus (GEO): https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147487. Other data analyzed in this study are from the GDC data portal at https://portal.gdc.cancer.gov/, the GTEx portal at https://gtexportal.org/home/, the ENCODE project at http://www.encodeproject.org, the

REDIportal database at http://srv00.recas.ba.infn.it/atlas/, ArrayExpress at https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6149/, and GEO under accession numbers GSE28040 and GSE38233.

3.8 Figures

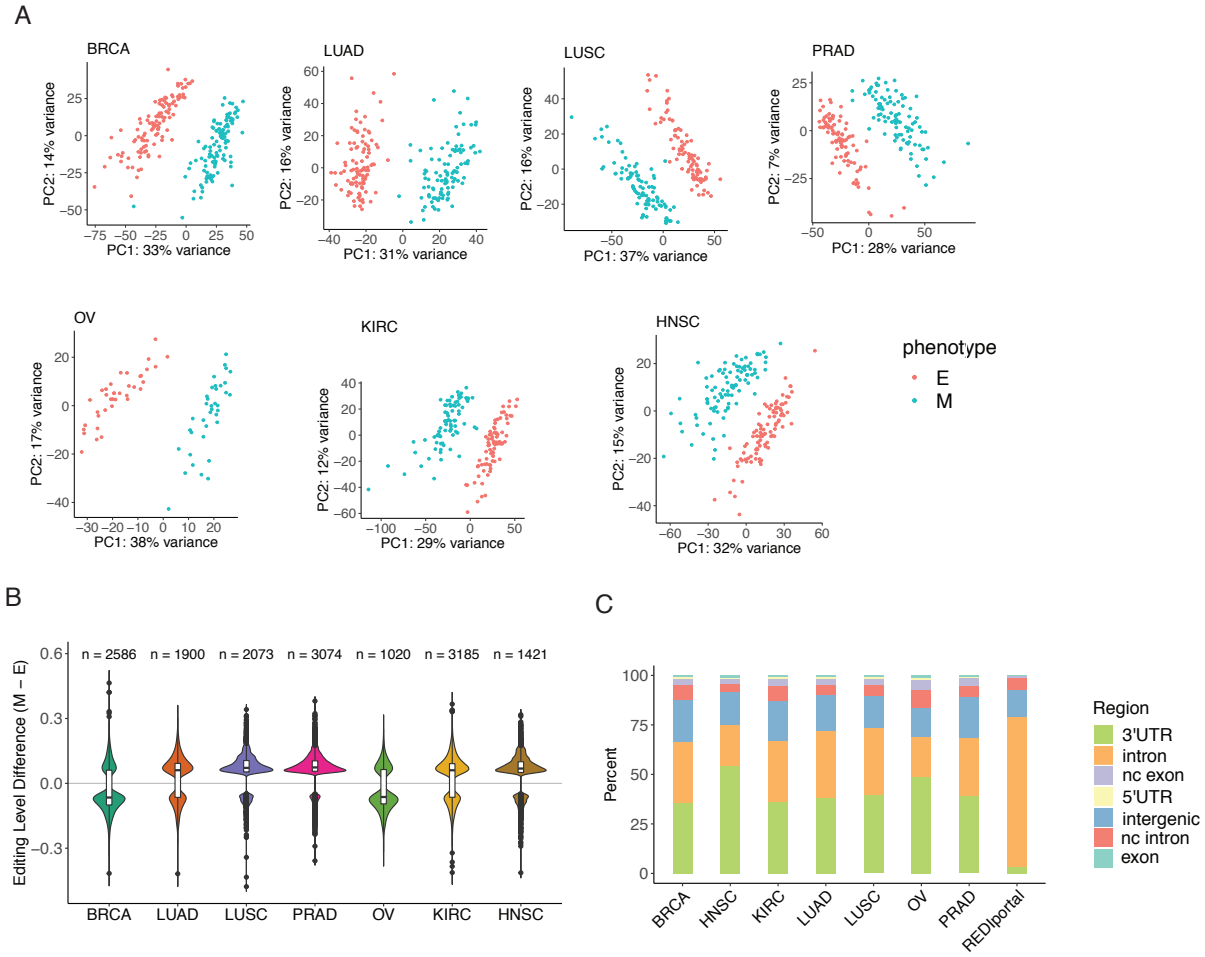


Figure 3.1 Overview of differential editing in cancer EMT

The following cancer types were studied: breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), prostate adenocarcinoma

(PRAD), ovarian serous cystadenocarcinoma (OV), kidney renal clear cell carcinoma (KIRC), head and neck squamous cell carcinoma (HNSC). **A** First two principal components of differential editing profiles separate tumor samples into epithelial (E) and mesenchymal (M) phenotypes across cancer types. **B** Distributions of differences in mean editing levels between E and M tumors in each cancer type. The number of differential editing sites is listed on top of each distribution. **C** Differential editing sites are mostly found in 3′ UTR and intronic regions in all cancer types, with higher proportions of 3′ UTR sites compared to that of all editing sites from the REDIportal database.

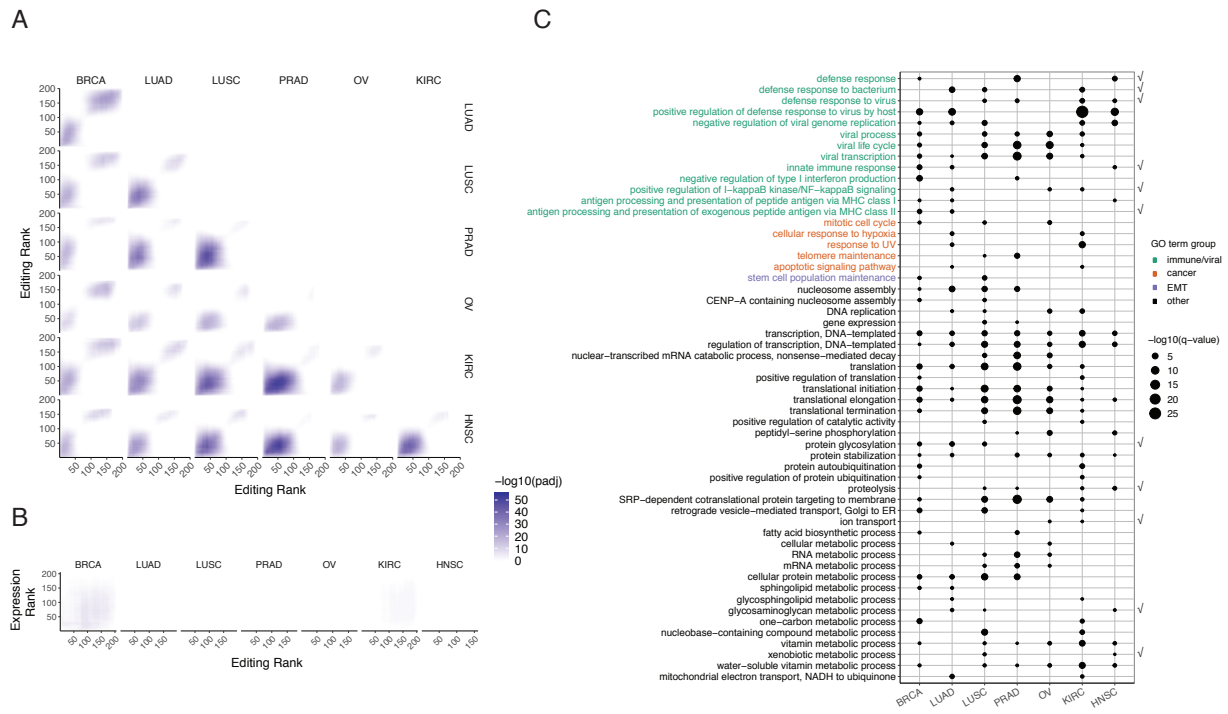


Figure 3.2 Differential editing patterns are shared among cancer types yet distinct from differential gene expression

A Rank-rank hypergeometric overlap (RRHO) map of RNA editing across pairs of cancer types. Each heatmap (for two cancer types) represents the matrix of log10- transformed adjusted p values that indicate the extent of overlap in two gene lists at each possible pair of ranks. For an individual cancer type, genes were ranked by the signed significance of RNA editing differences (M-E). Genes with higher editing in the M phenotype are at lower ranks, while those with higher editing levels in E tumors are at higher ranks. Higher pixel darkness indicates stronger enrichment of overlapping genes within the rank thresholds given by the x and y coordinates. The step size between ranks was 30 genes. **B** RRHO map of editing and gene expression within each cancer type. Each heatmap contains log10-transformed adjusted p values of hypergeometric overlap between genes ranked by editing differences (x-axis) and genes ranked by expression differences (y-axis) in a single cancer type. Similar to ranking genes by

differential editing, genes were ranked by the signed significance of expression differences, such that genes at lower ranks have higher expression in M tumors, while genes at higher ranks have higher expression in the E phenotype. The step size between ranks was 30 genes. **C** Significance of enrichment of gene ontology (GO) terms in differentially edited genes of each cancer type represented by point size (log10-transformed adjusted p value). Terms significantly enriched in at least two cancer types are shown. Check mark on the right indicates terms that were also significantly enriched in differentially expressed genes in at least two cancer types. Text color indicates category of biological relevance.

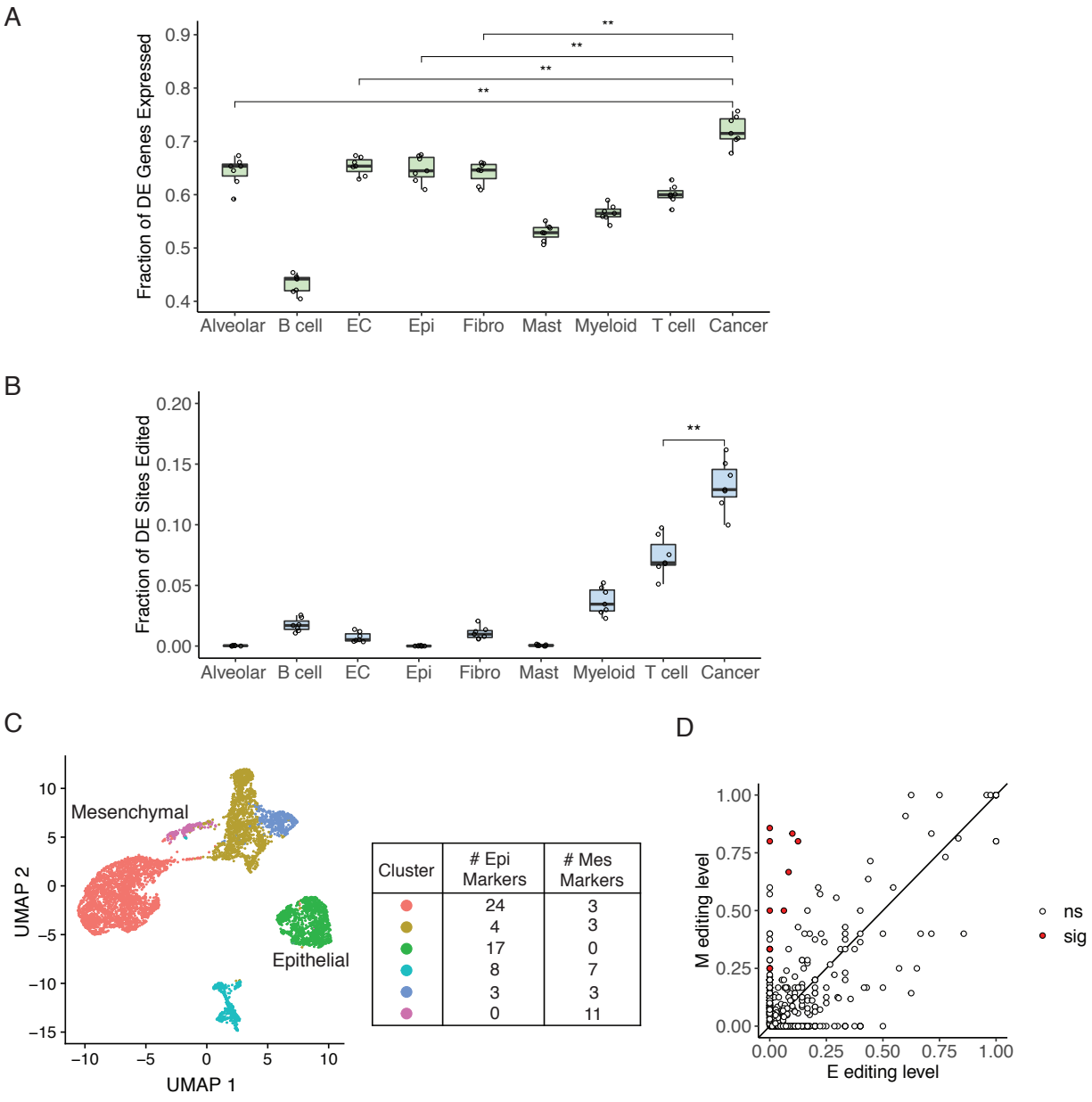


Figure 3.3 Contribution of cell types to differential editing

A Proportions of differentially edited (DE) genes from bulk tumor analysis that were expressed in cell types identified in lung cancer single-cell RNA-seq data. Each point represents the proportion of genes from one cancer type. A gene was considered as expressed in a cell type if its expression ≥ 1 RPKM. RPKM values were calculated within each cell type by pooling reads

of the same cell type together. Proportions were compared for top cell types by Mann Whitney U test, with significance of p values shown. **p ≤ 0.01. EC stands for endothelial cells. **B** Proportion of differential editing sites from bulk tumor analysis that were edited in individual cell types. A site was considered as edited in a cell type if the site was covered by at least 5 reads and editing was supported by at least 2 reads. Each point represents the proportion of sites from one cancer type. Proportions for top cell types were compared by Mann Whitney U test, with p value significance shown. **p ≤ 0.01. **C** UMAP projection of 6526 tumor cells based on expression profiles, colored by cluster assignment (scatterplot, left). By differential expression of epithelial or mesenchymal markers (table, right), green and purple clusters were labeled as epithelial and mesenchymal, respectively. **D** Scatterplot of editing levels of pooled E and M cells, with y = x line. Editing sites exhibiting significant differences between E and M were labeled in red. Differences were considered significant if the difference between editing levels ≥ 0.05 and Fisher's exact p value < 0.05.

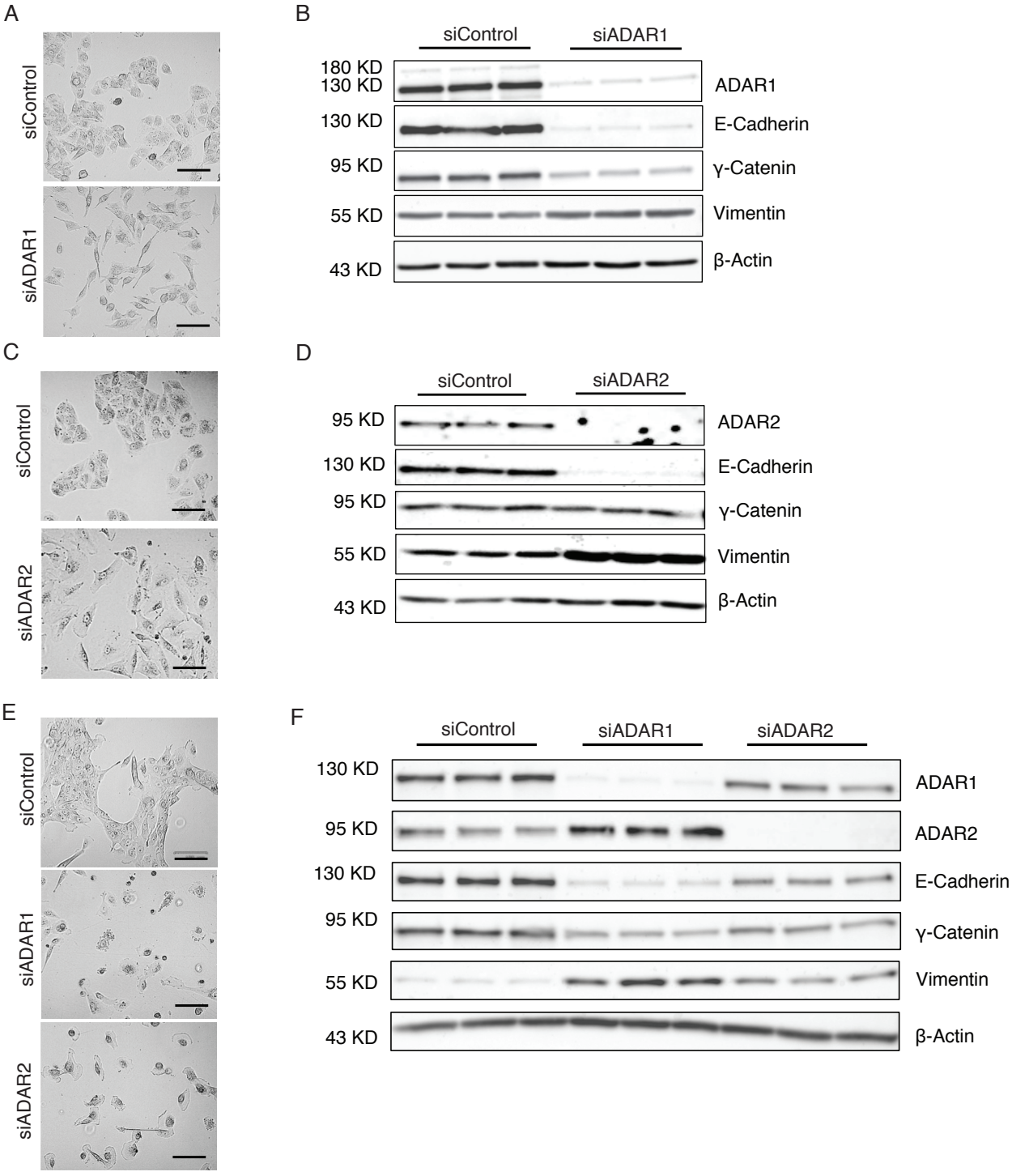


Figure 3.4 ADAR1 or ADAR2 knockdown induced EMT

A Images of A549 cells transfected with siRNAs for ADAR1 knockdown (KD) (siADAR1) or control siRNAs (siControl). Scale bars, 100 μm. **B** Loss of epithelial markers (E- cadherin and γ-Catenin) and induction of mesenchymal marker (Vimentin) in A549 cells upon ADAR1 KD. Cells were treated with 100 nM siRNA for 72 h. Three biological replicates were used in each condition. **C** Images of A549 cells transfected with siRNAs for ADAR2 KD (siADAR2) or control siRNAs (siControl). Scale bars, 100 μm. **D** Loss of epithelial markers (E-cadherin and γ-Catenin) and induction of mesenchymal marker (Vimentin) in A549 cells upon ADAR2 KD. Cells were treated with 11 nM siRNA for 72 h. Three biological replicates were used in each condition. **E** Images of MCF10A cells with ADAR1 or ADAR2 KD or control siRNAs. Scale bars, 100 μm. **F** Loss of epithelial markers (E-cadherin and γ-Catenin) and induction of mesenchymal markers (Vimentin) in MCF10A cells upon ADAR1 KD or ADAR2 KD. Cells were treated with 11 nM siRNA for 72 h. Three biological replicates were used in each condition.

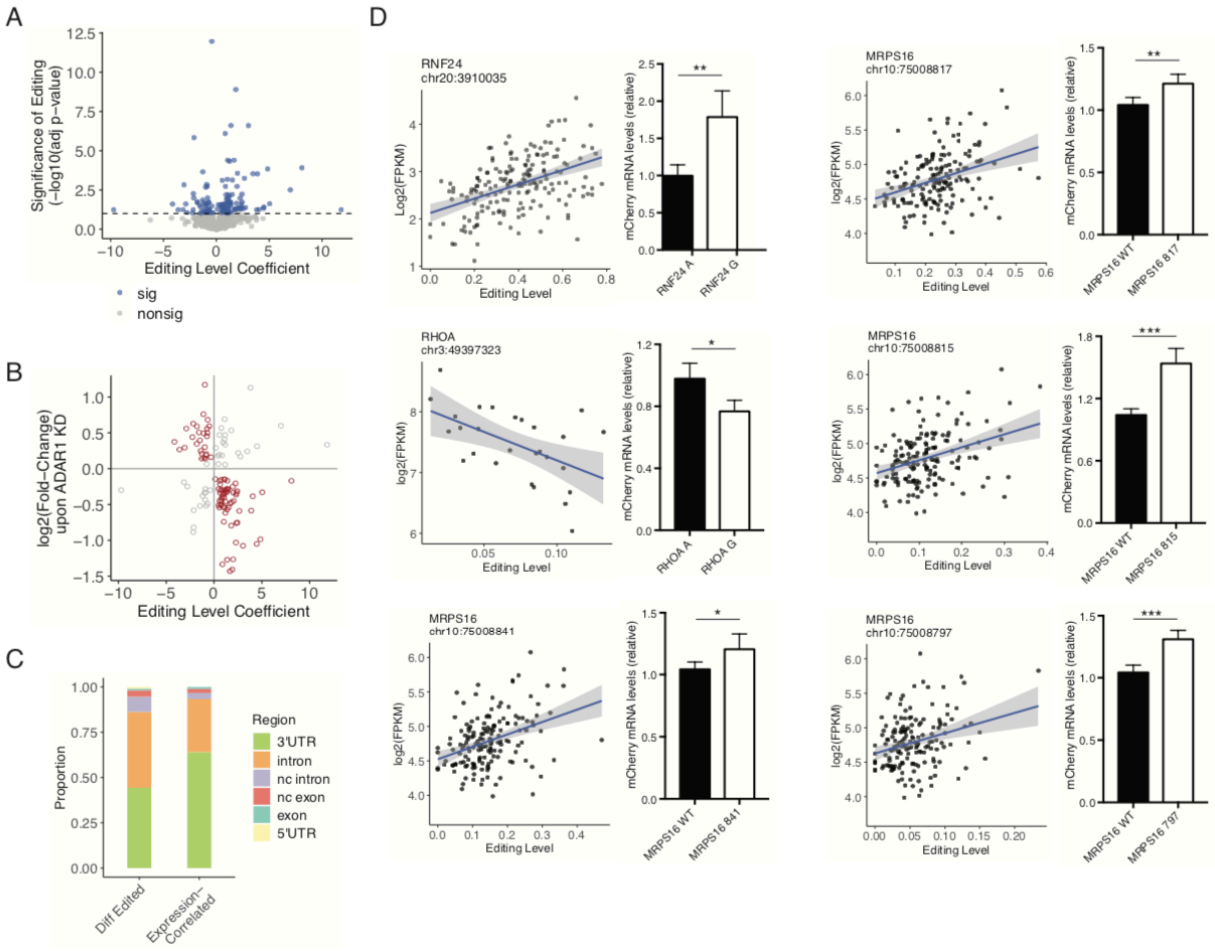


Figure 3.5 Effects of editing on mRNA abundance

A Scatterplot of coefficient estimate and statistical significance (log10-transformed adjusted p value) of editing level as a predictor of host mRNA expression in linear regression, accounting for potential confounding variables. For genes with multiple editing sites associated with expression, the most significantly associated site was used. Dashed line indicates significance threshold based on 10% false discovery rate (FDR). **B** Scatterplot of editing level coefficient estimate from multiple linear regression models used in A and log2-transformed fold change of the corresponding gene observed in ADAR1 KD cells. Red points indicate expression changes in the direction consistent with the sign of the editing association, in contrast to the gray points.

C Editing sites associated with host expression (Expression-Correlated) are more often found in 3′ UTR regions, compared to all differential editing sites (Diff Edited, not including intergenic sites). **D** Validation of six editing sites affecting host mRNA abundance. For each site, a scatterplot of editing level and log2-transformed mRNA expression in the TCGA data is shown. On the right of each scatterplot is mCherry expression, normalized by eYFP expression, of minigenes with A or G, corresponding to nonedited or edited versions of the sites in the 3′UTR of each gene. All minigenes were tested in Hela cells with five biological replicates. Normalized expression values (mean ± SD) were compared between edited and nonedited versions by two-sided t-test. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. Note that RHOA and MRPS16 editing sites were identified as differential sites in the single-cell RNA-seq analysis (Fig. 3c).

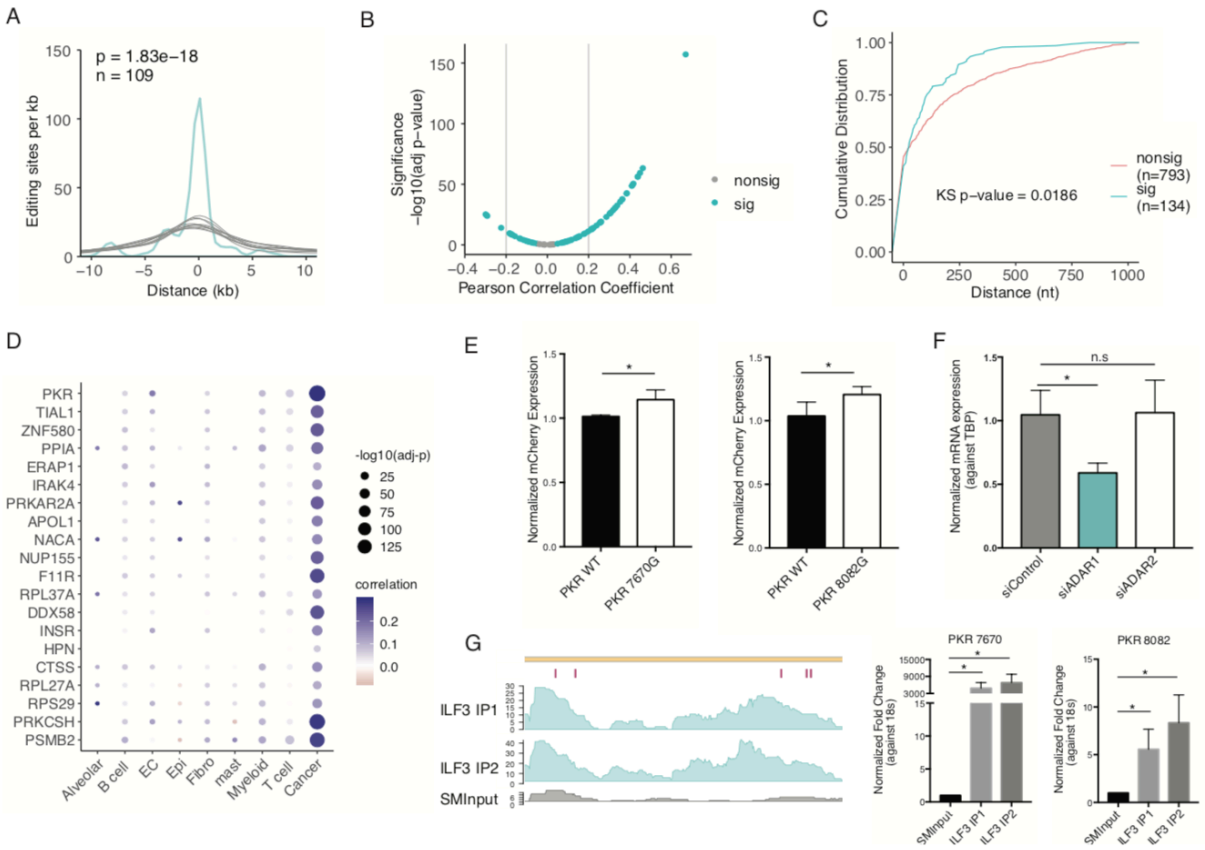


Figure 3.6 ILF3 binds closely to the differential editing sites in editing-expression-correlated genes

A Histogram of distances between differential editing sites in editing-correlated genes and the closest ILF3 eCLIP peaks in A549 cells (turquoise), up to 10 kb. Gray curves represent distances for 10 sets of randomly picked A's in the same genes as differential editing sites. Number of differential editing sites is given by n. p value was calculated by comparing the area under the curve (AUC) of the distance distribution for differential editing sites to a normal distribution fit to the AUC values of 10,000 sets of random gene-matched A's. **B** Scatterplot of Pearson correlation coefficient and significance (log10-transformed adjusted p value) of correlation between ILF3 mRNA expression and mRNA expression of editing-correlated genes.

Genes passing 10% FDR are labeled as significant (sig, turquoise), others as nonsig. **C** Cumulative distributions of distances between ILF3 eCLIP peaks and differential editing sites within editing- expression-associated genes (sig) or differential editing sites in genes without editing- expression associations (nonsig), up to 1 kb. Only genes associated with immune and viral related GO terms were included. p value calculated by the Kolmogorov-Smirnov test. **D** For each cell type in the lung cancer scRNA-seq dataset, ILF3 mRNA expression was correlated with mRNA expression of editing-expression-correlated genes (identified in the TCGA data) by Pearson correlation. Genes associated with any immune or viral-related GO term are shown. The size of each point indicates significance of correlation and color corresponds to values of the correlation coefficient. **E** Normalized mCherry expression (mean ± SD) for nonedited or edited versions of sites in the 3′UTR of PKR in A549 cells. Five biological replicates were performed. p value calculated by two-sided t-test (same below), *p < 0.05. **F** Normalized mRNA expression (mean ± SD) of endogenous PKR in siControl, siADAR1, and siADAR2 A549 cells. Three biological replicates were performed. *p < 0.05. n.s., not significant. **G** Read coverage of ILF3 eCLIP-seq in A549 cells for two biological replicates (ILF3 IP1 and ILF3 IP2, turquoise) and size-matched input (SMInput, gray). The five validated 3′ UTR editing sites affecting PKR mRNA abundance in A549 cells are labeled in magenta (left). Right: Validation of PKR eCLIP signal overlapping two editing sites. PKR expression (mean ± SD) was measured by qRT-PCR in the IP or SMInput samples and normalized against the expression of 18s rRNA, *p < 0.05. (n = 3).

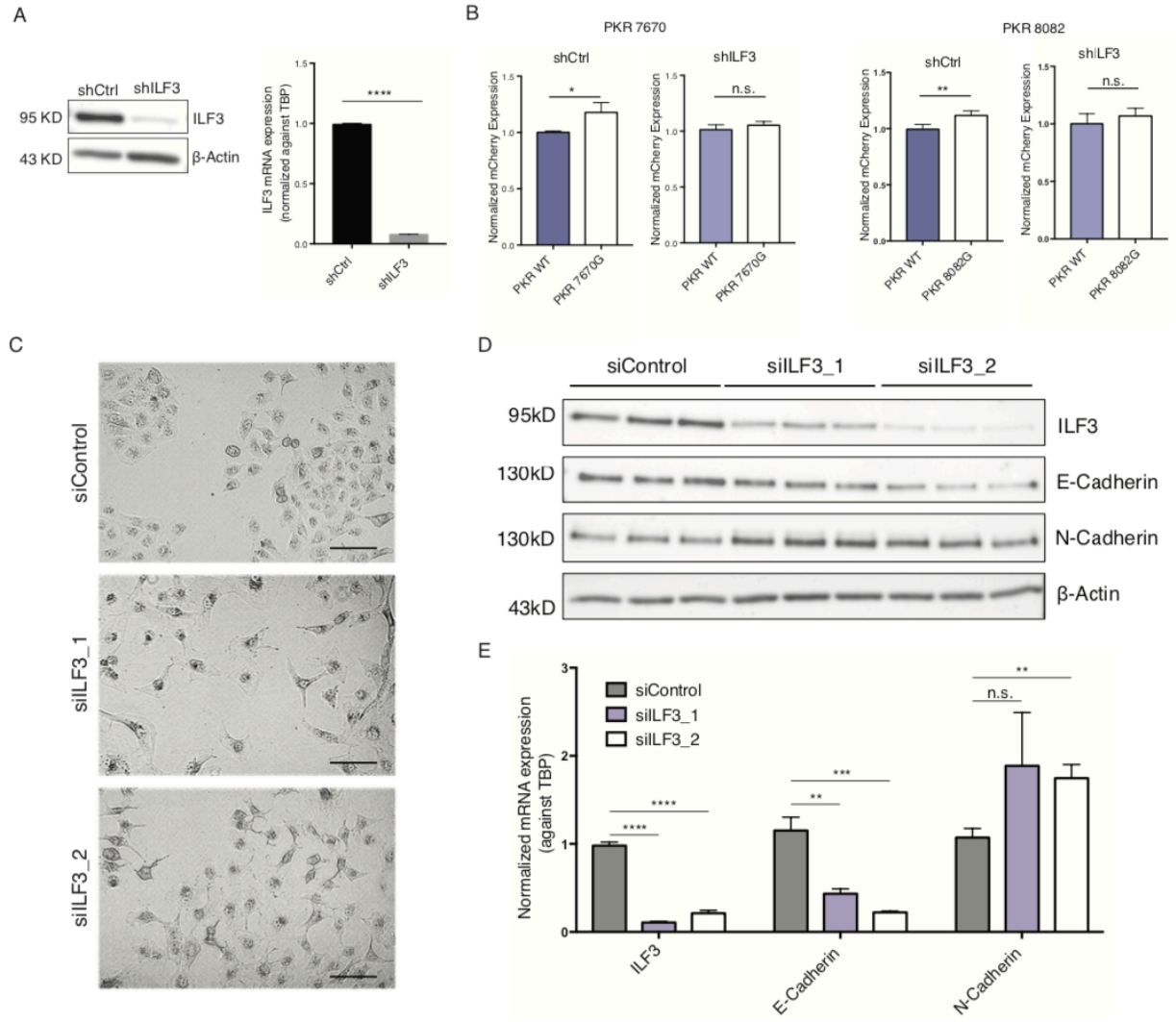
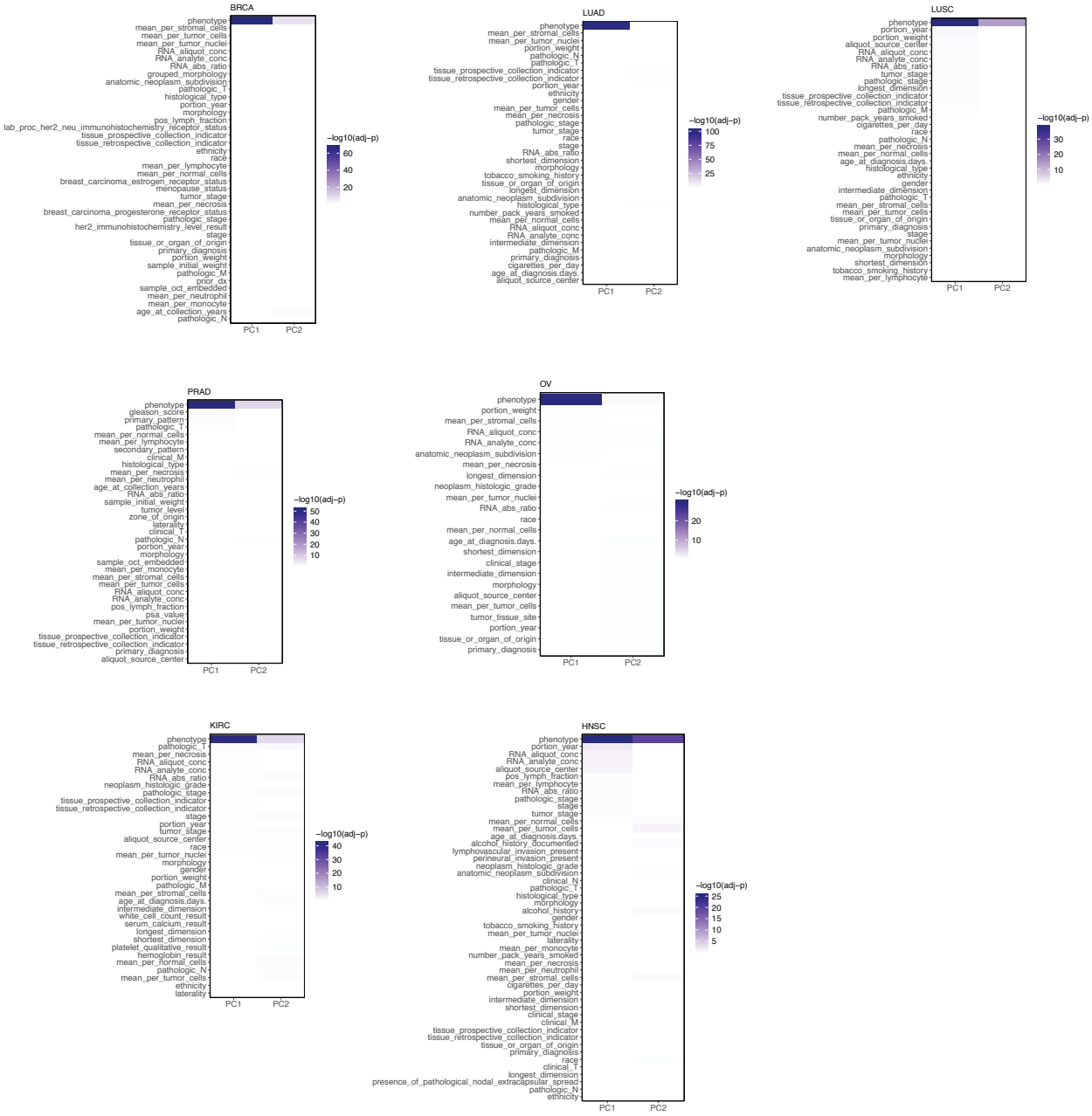


Figure 3.7 ILF3 regulates PKR mRNA abundance and EMT in A549 cells

A Western blot confirming shRNAmediated ILF3 KD in A549 cells (left). ILF3 mRNA levels (mean ± SD) were quantified in A549 shCtrl and ILF3 KD cells by qRT-PCR (right). ILF3 mRNA expression was normalized against gene TBP mRNA expression. Three biological replicates were performed. p value calculated via t-test, ****p < 0.0001. **B** Normalized mCherry expression (mean ± SD) for nonedited or edited versions of sites in the 3′ UTR of PKR in shCtrl or ILF3 KD A549 cells. Five biological replicates were performed. Normalized expression values were

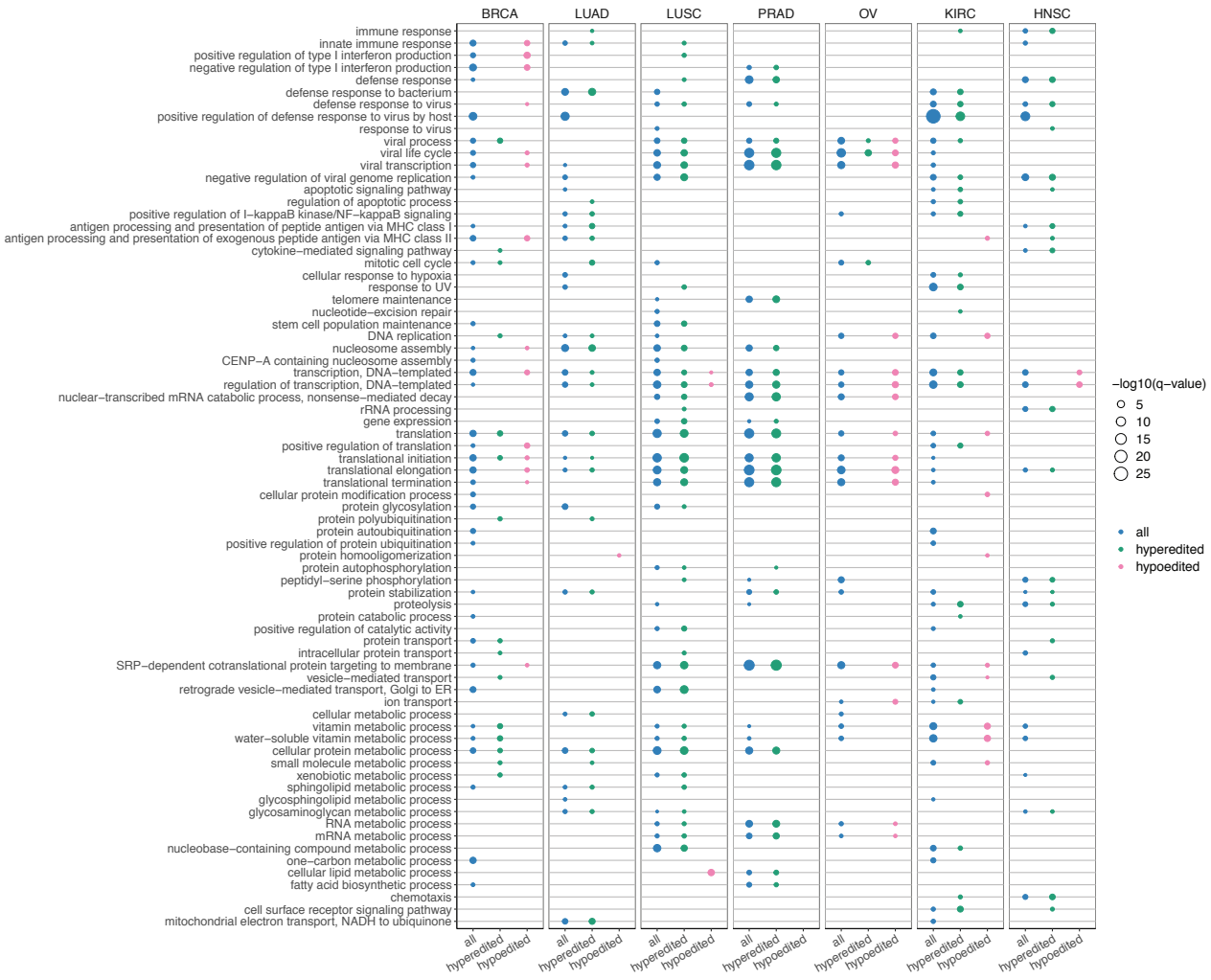
compared between edited and nonedited versions by two-sided t-test. *p < 0.05, **p < 0.01, n.s., not significant. **C** Images of A549 cells transfected with siRNAs targeting ILF3 (two different siRNAs were used to KD ILF3, siILF3_1, and siILF3_2) or control siRNAs (siControl). Scale bars: 100 μm. **D** Western blot detecting protein levels of ILF3, E-Cadherin, N-Cadherin, and internal control β-Actin in the siControl, siILF3_1, and siILF3_2 A549 cells. Three biological replicates were carried out for each experiment. **E** Normalized mRNA expression levels (mean ± SD) for ILF3, E-Cadherin, and N-Cadherin in the siControl, siILF3_1, and siILF3_2 A549 cells. Three biological replicates were carried out for each experiment. The expression values were compared between siILF3 and siControl via t- test. **p < 0.01, ***p < 0.001, ****p < 0.0001, n.s., not significant.

3.9 Supplementary Figures



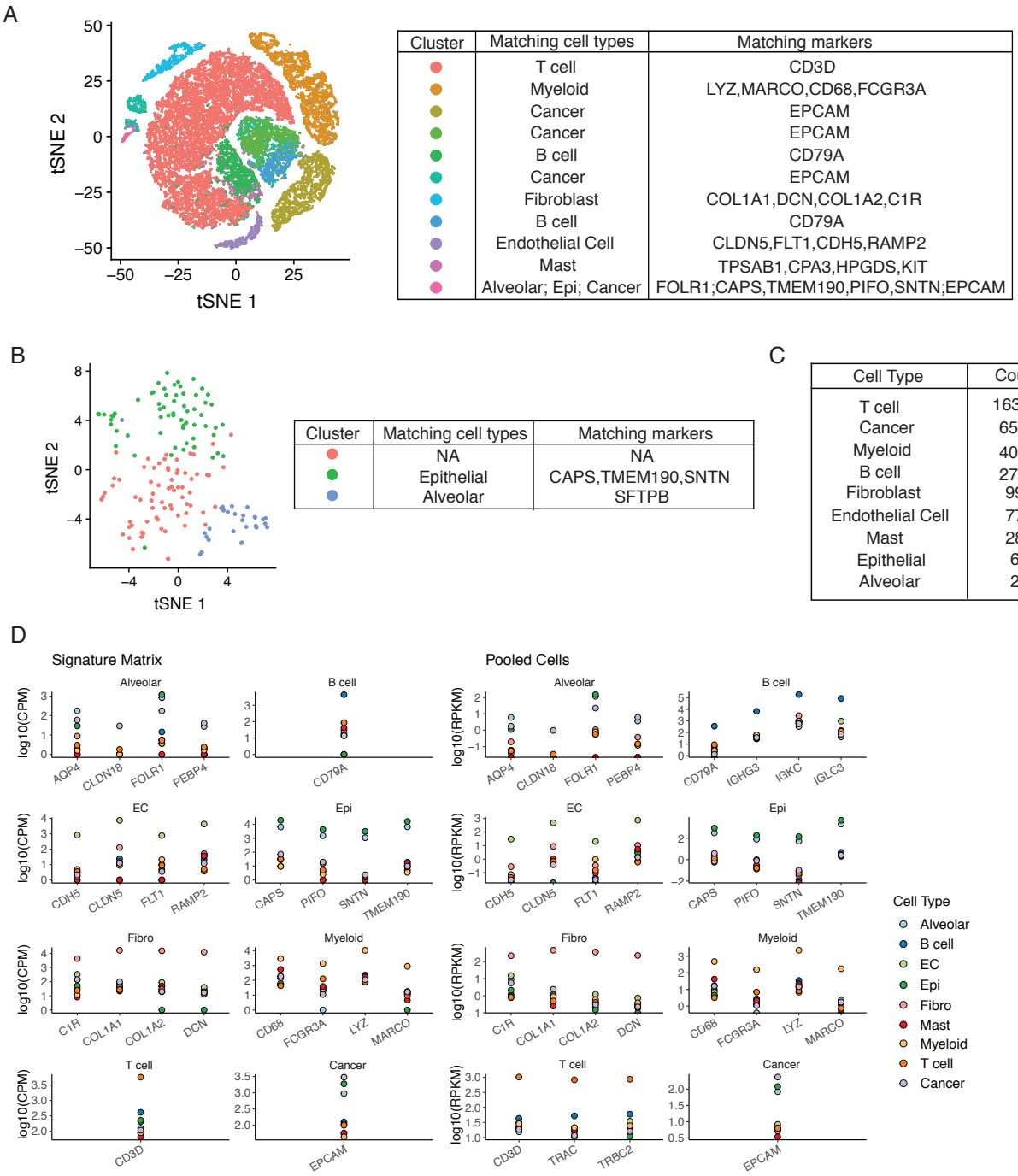
Supplementary Figure 3.1 Differential editing not confounded by metadata

Heatmaps of significance (log10-transformed adjusted p-values) of correlations between the top two principal components and E/M phenotype among metadata fields in each cancer type. Darker color indicates smaller p-value and stronger association.



Supplementary Figure 3.2 Gene ontology enrichment among differentially edited genes

Significance of enrichment of gene ontology (GO) terms among all differentially edited genes (blue), only hyperedited genes (green) or only hypoedited genes (pink) of each cancer type. Point size represents the statistical significance of enrichment (log10-transformed adjusted p-value). Terms significantly enriched in at least two cancer types are shown. For cancer types with a global hyperediting trend in M tumors, GO enrichment among hyperedited genes is similar to that among all differentially edited genes. Likewise, for cancer types with a hypoediting trend (BRCA and OV), enrichment among hypoedited genes is similar to that among all differentially edited genes.



Cluster	Matching cell types	Matching markers
●	T cell	CD3D
●	Myeloid	LYZ,MARCO,CD68,FCGR3A
●	Cancer	EPCAM
●	Cancer	EPCAM
●	B cell	CD79A
●	Cancer	EPCAM
●	Fibroblast	COL1A1,DCN,COL1A2,C1R
●	B cell	CD79A
●	Endothelial Cell	CLDN5,FLT1,CDH5,RAMP2
●	Mast	TPSAB1,CPA3,HPGDS,KIT
●	Alveolar; Epi; Cancer	FOLR1;CAPS,TMEM190,PIFO,SNTN;EPCAM

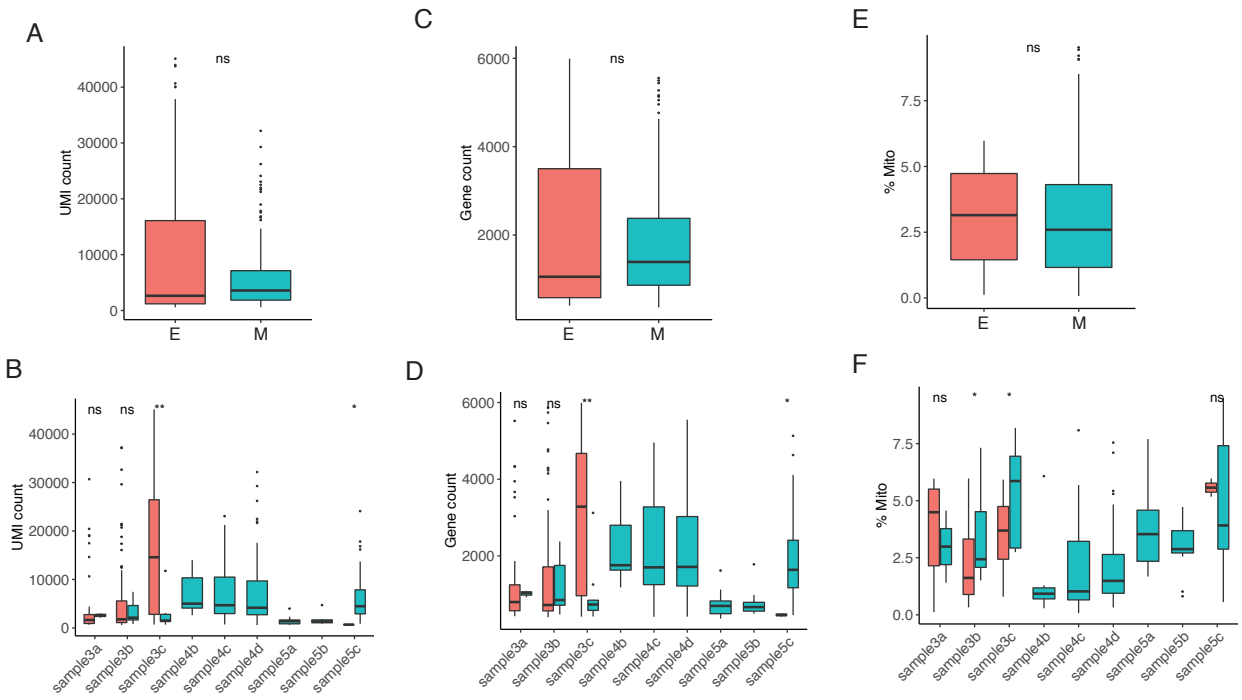
Cluster	Matching cell types	Matching markers
●	NA	NA
●	Epithelial	CAPS,TMEM190,SNTN
●	Alveolar	SFTPB

Cell Type	Count
T cell	16354
Cancer	6526
Myeloid	4063
B cell	2770
Fibroblast	998
Endothelial Cell	771
Mast	285
Epithelial	62
Alveolar	29

Supplementary Figure 3.3 Clustering of single cells from three lung cancer tumors

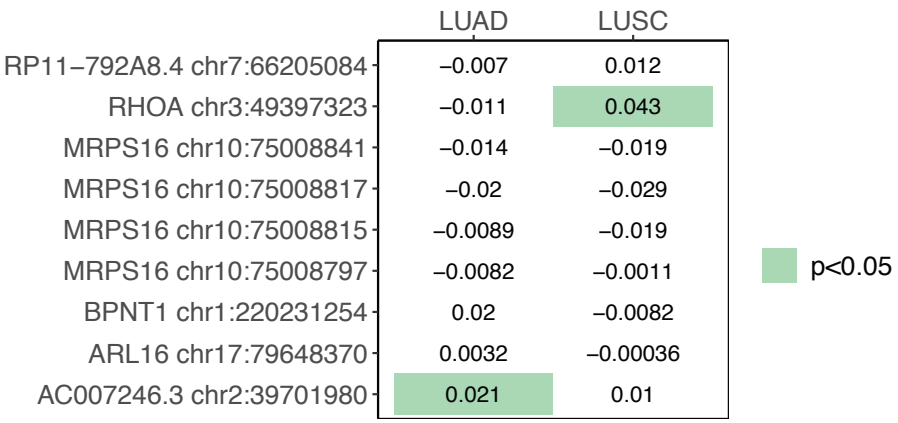
A TSNE projection of cells based on expression profiles, with color indicating cluster identity

(left). Cell types were assigned to clusters by matching differentially expressed genes of clusters to known cell type markers (right). **B** TSNE projection of only cells from cluster 10 to further refine cell type assignment (left). Similar to A, cell types were labeled using differentially expressed genes that matched cell type markers (right). **C** Counts of cells for each cell type after 2 rounds of clustering and cell type assignment (A and B). **D** Log2-transformed expression values of marker genes across cell types. Signature matrix on the left indicates expression values assigned for each cell type by CIBERSORTx. On the right, Pooled Cells indicate that expression values were calculated from pooling reads from cells of the same type together.



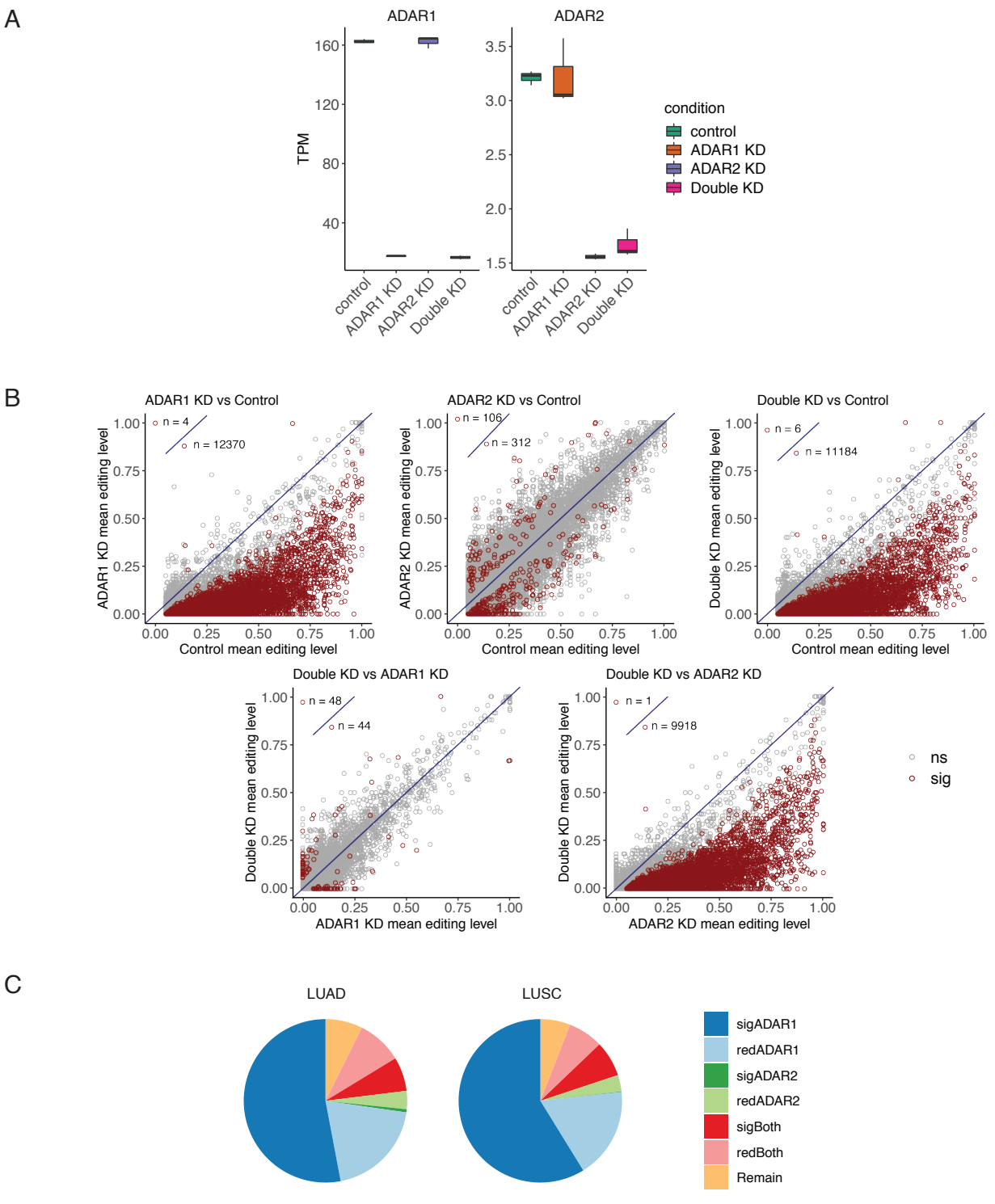
Supplementary Figure 3.4 E and M assignment of single cells not confounded by metadata

Comparison between E and M cells altogether (top) and within each tumor sample (bottom) of metadata fields: UMI count (A-B), gene count (C-D), and percent of reads mapping to the mitochondrial genome (E-F). Metadata values were compared by Mann Whitney U tests, and significance of p-values are shown. ns: $p > 0.05$, * $p \leq 0.05$, ** $p \leq 0.01$.



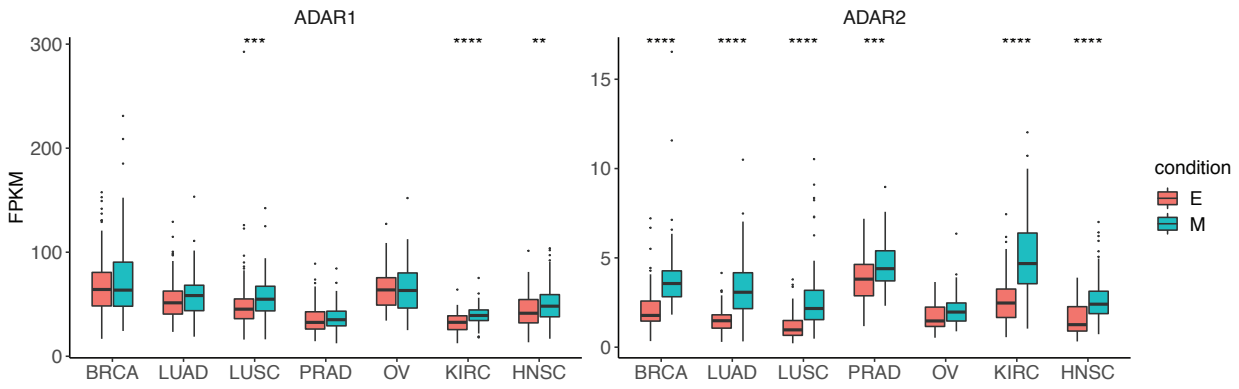
Supplementary Figure 3.5 LUAD and LUSC tumor editing differences of differential

For each editing site, the difference in mean editing levels between M and E tumors (M - E) in each cancer type is listed. Green highlight indicates Wilcoxon p-value < 0.05.



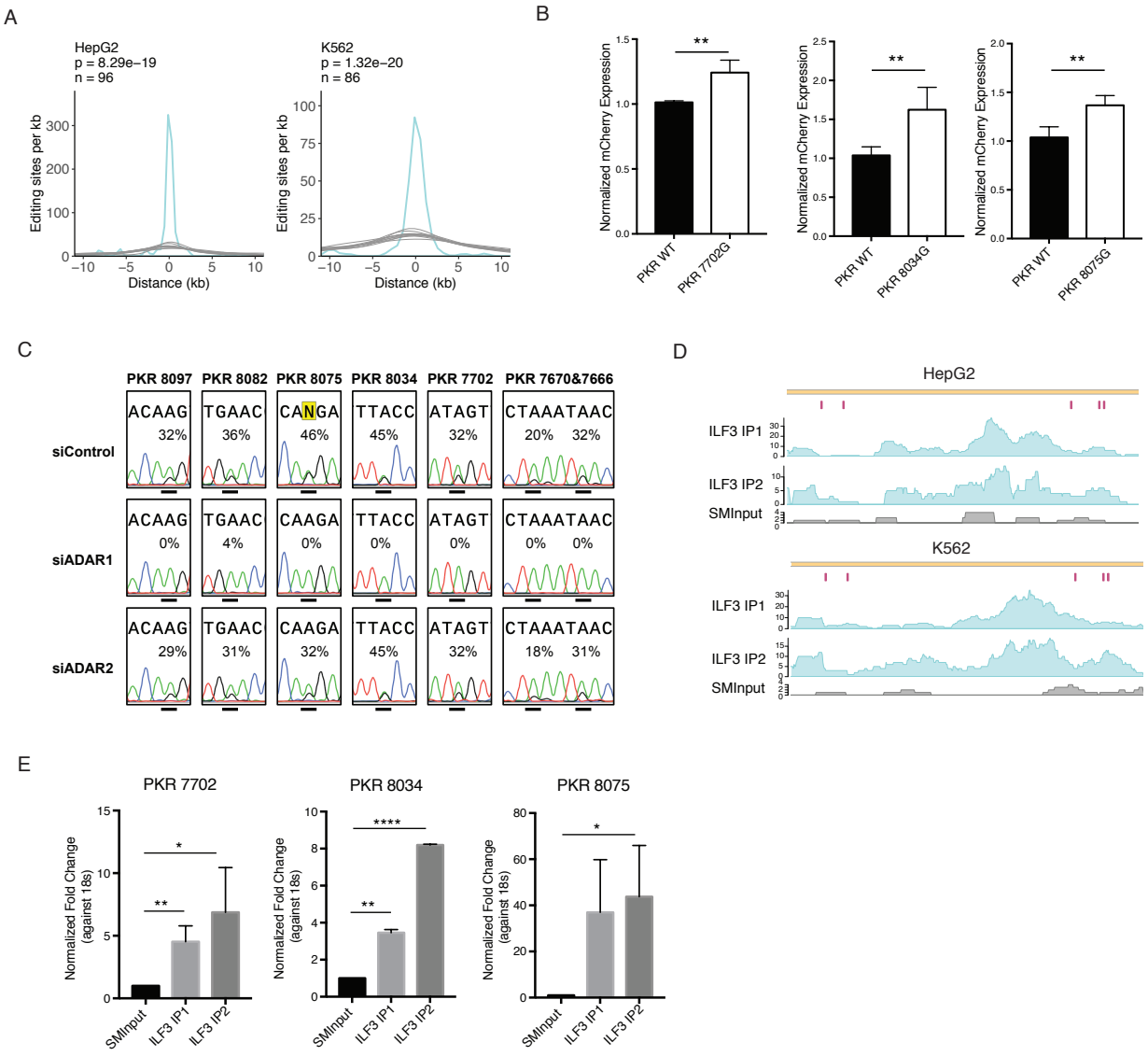
Supplementary Figure 3.6 Altered editing upon knockdown of ADAR1, ADAR2, or both

A Distributions of mRNA expression of ADAR1 and ADAR2 under ADAR KD and control conditions. Expression levels were quantified as transcripts per million (TPM). **B** Mean editing levels of testable sites in five comparisons between ADAR KD conditions or control experiment. Sites with significant editing differences between conditions are colored red, while gray represents nondifferential sites. Y=x line shown in blue. **C** Proportions of lung cancer E-M differential sites that were also differential in ADAR KD conditions (compared to controls). sigADAR1: sites that were differential only in ADAR1 KD. sigADAR2: sites that were differential only in ADAR2 KD. sigBoth: sites that were differential in both ADAR1 KD and ADAR2 KD, or in double KD. The prefix 'red' indicates reduced editing level by at least 0.05 upon KD from control, but did not pass the statistical significance requirement. 'Remain': editing sites that were not significantly different or reduced across any comparison.



Supplementary Figure 3.7 Expression of ADARs in E and M tumors

Distributions of mRNA expression of ADAR1 (left) and ADAR2 (right) in E and M tumors across cancer types. Expression values, measured as Fragments Per Kilobase per Million mapped reads (FPKM), were compared by Mann Whitney U tests, and significance of p-values are shown. ** $p \leq 0.01$; *** $p \leq 0.001$; **** $p \leq 0.0001$.



Supplementary Figure 3.8 ILF3 binds closely to the differential editing sites in editing-expression correlated genes

A Histogram of distances between differential editing sites in editing-correlated genes and the closest ILF3 eCLIP peaks in HepG2 and K562 cells (turquoise), up to 10 kb. Gray curves represent distances for 10 sets of randomly picked A's in the same genes as differential editing sites. Number of differential editing sites is given by n for each cell line. P-value was calculated

by comparing the area under the curve (AUC) of the distance distribution for differential editing sites to a normal distribution fit to the AUC values of 10,000 sets of random gene-matched A's. **B** Normalized mCherry expression for nonedited or edited versions of sites in the 3'UTR of PKR in A549 cells. Five biological replicates were performed. Normalized expression values were compared between edited and nonedited versions by two-sided t-test. **p<0.01. **C** Editing levels of PKR 3'UTR editing sites in siControl, siADAR1 and siADAR2 A549 cells measured by Sanger sequencing. The peak signals of A and G nucleotides were measured by 4Peaks for editing level calculation (G/(A+G)). The editing level of each editing site (underlined) is shown in the graph. **D** Read coverage of ILF3 eCLIP-seq in HepG2 and K562 cells for two biological replicates (ILF3 IP1 and ILF3 IP2, turquoise) and size-matched input (SMInput, gray) in each cell line. The five validated 3' UTR editing sites affecting PKR mRNA abundance in A549 cells are labeled in magenta. **E** Validation of PKR eCLIP signal overlapping three editing sites. PKR expression was measured by qRT-PCR in the IP or SMInput samples and normalized against the expression of 18s rRNA. Three technical replicates were performed (other than two replicates for 8034). P-value calculated by t-test. *p<0.05, **p<0.01, ****p<0.0001.

3.10 Supplementary Tables

Primary_Site	Cancer_Type	Abbreviation	E	M
Breast	Breast Invasive Carcinoma	BRCA	110	110
Lung	Lung Adenocarcinoma	LUAD	100	100
Lung	Lung Squamous Cell Carcinoma	LUSC	93	93
Prostate	Prostate Adenocarcinoma	PRAD	97	97
Ovary	Ovarian Serous Cystadenocarcinoma	OV	37	37
Kidney	Kidney Renal Clear Cell Carcinoma	KIRC	78	78
Head and Neck	Head and Neck Squamous Cell Carcinoma	HNSC	94	94

Supplementary Table 3.1 Primary tumor samples used in this study

Cancer types and the corresponding numbers of categorized E and M tumor samples analyzed in this study.

gene	rna_region	editing_site	editlevel_est	adj_edit_pvalue	consis_cell	consis_logfoldchange
ACOX1	3UTR	chr17:73940077	0.675590107	0.057558713	K562	-0.3
ALDH6A1	3UTR	chr14:74526975	2.642817316	0.000566336	HepG2,K562	-0.18,-0.59
APOL1	3UTR	chr22:36662697	5.043197874	0.023870074	Hela_cyto	-0.33
APOL6	3UTR	chr22:36057354	1.325809518	0.019982168	U87	-0.284
APOOL	3UTR	chrX:84346877	0.881841353	0.024752325	U87,K562	-0.3,-0.39
ARHGAP29	intron	chr1:94634050	2.330463483	0.03925702	HepG2	-0.76
ARPIN	3UTR	chr15:90444654	0.92754555	0.022276919	U87	-0.535
ARRDC1	intron	chr9:140504867	-1.806802676	0.019982168	Hela_w,Hela_cyto	0.347,0.229
ASB16-AS1	noncodingIntron	chr17:42255926	0.596184402	0.074699173	HepG2	-0.17
C19orf48	intron	chr19:51306695	2.354947542	0.025715263	U87	-0.339
C19orf71	intron	chr19:3539995	2.291666376	0.046870966	K562,Hela_w,Hela_cyto	-0.15,-0.461,-1.03
CCNYL1	3UTR	chr2:208619479	1.084244552	3.98E-05	U87	-0.307
CEP104	3UTR	chr1:3729872	-0.807298896	0.084946682	Hela_cyto	0.183
CERS2	noncodingIntron	chr1:150933884	2.051475252	0.044422316	Hela_cyto	-0.451
CINP	intron	chr14:102813733	0.72775171	0.03712208	U87,HepG2,K562,Hela_w,Hela_cyto	-0.211,-0.2,-0.2,-0.511,-0.387
CPM	intron	chr12:69243317	3.885443958	0.000297212	U87,Hela_w,Hela_cyto	-0.257,-0.22,-0.586
CTSS	3UTR	chr1:150704508	4.848763834	0.000143564	U87	-0.985
CXorf56	3UTR	chrX:118673269	0.492247834	0.085175369	U87,K562,Hela_w	-0.265,-0.51,-0.216
CYP20A1	3UTR	chr2:204170105	-0.32963639	0.085175369	Bcell	0.16
DDX58	3UTR	chr9:32456380	1.228575069	0.007953617	U87,HepG2	-0.218,-0.3
EIF2AK2	3UTR	chr2:37328097	1.382670474	2.46E-07	U87	-0.167
EMC1	3UTR	chr1:195426609	1.373657011	0.022662878	U87,Hela_w,Hela_cyto	-0.534,-0.148,-1.27
ERAP1	3UTR	chr5:96110544	1.356393229	0.033976694	U87	-0.286
F11R	3UTR	chr1:160967890	1.02385868	0.094957352	HepG2,K562,Hela_cyto	-0.21,-0.32,-0.741

FAM129A	3UTR	chr1:184761312	-1.577706747	0.028593802	Hela_w	0.506
FAM20B	3UTR	chr1:179042544	1.059798513	0.05628566	HepG2	-0.15
FCF1	3UTR	chr14:75202544	0.791115041	0.03925702	U87,K562	-0.458,-0.25
FGD5-AS1	noncodingExon	chr3:149866104	-0.664122639	0.0638645	Hela_w,Hela_cyto	0.473,0.683
FLNA	intron	chrX:153579411	1.999356154	0.000270927	U87,Hela_w,Hela_cyto	-0.274,-0.301,-1.27
FOXRED2	3UTR	chr22:36884057	2.434689666	0.018654036	U87,Hela_w,Hela_cyto	-0.753,-0.152,-0.19
GFOD2	3UTR	chr16:67716119	-1.381194392	0.072767736	Hela_w,Hela_cyto	0.25,0.21
GGCX	3UTR	chr2:85773390	1.080678262	0.057558713	HepG2,K562,Hela_cyto	-0.35,-0.27,-0.232
H2AFV	intron	chr7:44872505	0.47460002	0.03897698	U87	-0.285
H6PD	3UTR	chr1:9328120	1.348878853	0.018654036	HepG2	-0.57
HPN	intron	chr19:35538724	-1.792221114	0.01799247	HepG2	0.59
HSPB11	intron	chr1:54387885	0.737673592	0.025715263	U87,K562	-0.22,-0.42
IGFBP7	exon	chr4:57976234	3.027681172	2.48E-07	Hela_w,Hela_cyto	-0.17,-1.08
INSR	intron	chr19:7146479	-1.472096252	0.038996066	U87,HepG2	0.395,0.19
IRAK4	3UTR	chr12:44181739	-1.084182315	0.085721376	Hela_w,Hela_cyto	0.436,0.494
KNOP1	3UTR	chr16:19714150	1.382159872	0.077693483	U87,HepG2,K562	-0.267,-0.32,-0.23
LPP	3UTR	chr3:188598857	2.422937272	0.005835479	U87,HepG2	-0.314,-0.21
MDM4	3UTR	chr1:204526595	1.113953173	0.022276919	K562	-0.54
METTL7A	3UTR	chr12:51324639	-1.111493276	0.057558713	HepG2,Hela_w	0.76,0.509
MFSD12	intron	chr19:35440230	1.90702215	0.022276919	HepG2,Hela_w,Hela_cyto	-0.27,-1.06,-1.41
MREG	3UTR	chr2:216808321	-2.891316786	0.024460223	U87,HepG2,Hela_w,Hela_cyto	0.286,0.28,0.56,0.509
MRTO4	3UTR	chr1:19586458	0.585692697	0.025465398	U87,K562,Hela_w	-0.452,-0.27,-0.168
MYO19	intron	chr17:34853684	1.518375774	0.000735249	U87,HepG2	-0.541,-0.26

NACA	intron	chr12:57125246	1.121990286	0.085084237	U87,K562	-0.15,-0.15
NBPF10	3UTR	chr1:145369556	0.127875665	0.051780731	U87	-0.145
NOP14	3UTR	chr4:2940462	1.239264708	4.72E-05	U87	-0.33
NPLOC4	3UTR	chr17:79530071	2.025722247	0.049162216	U87,HepG2,Hela_cyto	-0.336,-0.22,-0.141
NUP155	3UTR	chr5:37291446	1.183902296	0.081405156	U87,HepG2,K562	-0.298,-0.18,-0.46
PAICS	3UTR	chr4:57326875	-2.171471286	0.005405145	Hela_w,Hela_cyto	0.614,0.629
PCCB	intron	chr3:136050150	4.046404225	0.055133335	U87,HepG2,K562	-0.385,-0.36,-0.35
PDLIM5	intron	chr4:95542079	-4.144319345	0.024752325	Hela_w,Hela_cyto	0.374,0.241
PINK1-AS	noncodingExon	chr1:20976109	1.182220784	0.088652742	Hela_w,Hela_cyto	-0.307,-0.616
PLBD2	3UTR	chr12:113827916	1.666217711	3.98E-05	U87,Hela_cyto	-1.43,-0.735
PPIA	3UTR	chr7:44841857	-1.281380018	0.000220889	HepG2,Hela_cyto	0.21,0.181
PRKAR2A	3UTR	chr3:48787856	-0.736600952	0.018612605	Hela_w,Hela_cyto	0.362,0.452
PRKCSH	intron	chr19:11561241	-0.71592712	0.019855 95	U87,HepG2,K562	0.314,0.56,0.24
PSMB2	3UTR	chr1:36067752	1.044145565	0.000659271	U87,HepG2,K562,Hela_w,Hela_cyto	-0.325,-0.19,-0.24,-0.403,-0.263
PXMP4	3UTR	chr20:32293069	-0.70072022	0.058943984	Hela_w,Hela_cyto	0.498,0.446
RBBP9	3UTR	chr20:18468339	0.915381444	0.006388843	U87,Bcell	-0.358,-0.188
RBM8A	3UTR	chr1:145513094	0.636371761	0.041058 97	U87	-0.217
RDH13	intron	chr19:55550840	-0.902703462	0.054168949	HepG2	0.14
RNF24	3UTR	chr20:3910035	1.843212681	1.26E-09	U87,HepG2	-0.469,-0.4
RPL27A	3UTR	chr11:8708835	2.246128853	0.010482013	U87,Hela_w	-0.177,-0.206
RPL37A	3UTR	chr2:217366903	0.876470214	0.046870966	Hela_w	-0.295
RPL7L1	3UTR	chr6:428556051	-0.826267054	0.010067513	Hela_w,Hela_cyto	0.184,0.205

RPS29	3UTR	chr14:50039685	-2.260993657	0.091869511	HepG2,Bcell	0.44,0.166
SERBP1	3UTR	chr1:67874859	-0.506580781	0.085175369	Hela_w,Hela_cyto	0.594,0.545
SERINC2	intron	chr1:31895247	4.465208772	0.045831535	HepG2,Hela_w,Hela_cyto	-1.05,-0.685,-0.803
SLC25A16	intron	chr10:70279509	-3.60902466	0.057558713	Hela_cyto	0.266
SLC35F5	noncodingIntron	chr2:114465433	0.751160766	0.057338066	HepG2,K562,Hela_cyto	-0.46,-0.49,-0.503
SMPD4	intron	chr2:130915581	-1.292145797	0.013733024	HepG2	0.14
TIAL1	3UTR	chr10:121331855	-3.068906487	0.001484171	Hela_w,Hela_cyto	0.294,0.236
TIMM50	intron	chr19:39982388	-1.057991191	0.02951378	HepG2	0.15
TMEM120B	3UTR	chr12:122216342	0.64131591	0.049727225	HepG2,K562,Hela_w	-0.34,-0.27,-0.229
TMEM59	3UTR	chr1:54496708	1.802146018	0.048523878	HepG2,K562,Hela_cyto	-0.19,-0.29,-0.794
TTC9C	intron	chr11:62507302	0.553395068	0.000767773	Hela_w,Hela_cyto	-0.542,-0.369
TXNDC15	3UTR	chr5:134236752	0.872473286	0.001864405	HepG2,Hela_w,Hela_cyto	-0.44,-0.681,-1.33
UBA1	intron	chrX:47067783	1.955430639	0.001864405	U87,Hela_cyto	-0.19,-0.252
VPS41	3UTR	chr7:38764332	0.754541392	0.081323723	HepG2,Hela_cyto	-0.15,-0.154
ZDHHC20	3UTR	chr13:21949165	1.493075791	0.006085932	Hela_cyto	-0.288
ZNF432	3UTR	chr19:52535007	8.096825832	0.000121773	HepG2,K562	-0.17,-0.14
ZNF552	intron	chr19:58321472	2.305572887	0.052306474	U87	-0.363
ZNF580	intron	chr19:56149705	-0.96540403	0.081405156	U87,K562,Hela_cyto	0.226,0.3,0.29
ZNF587B	3UTR	chr19:58357228	-0.957656315	0.024752325	Hela_w,Hela_cyto	0.872,1.17
ZSWIM1	3UTR	chr20:44513183	1.686372763	0.05628566	U87,Hela_w	-0.485,-0.249

Supplementary Table 3.2 List of editing sites predicted to regulate host gene mRNA abundance

Editing-expression associations (editlevel_est represents editing level regression coefficient and adj_edit_pvalue is the adjusted p-value of the coefficient) were supported by consistent expression changes upon ADAR KD in at least one cell line.

Cloning for minigenes	
Name	**Sequences**
RNF24 Fw	GTACCATCGATAGTCTCATGTGGATATGCCTG
RNF24 Rv	ACGCGTCGACGTTGGGGTAATTTCTGTTGTC
RNF24 G R	GTAACAAGACCCCGTCTCAACAACAAC
RNF24 G F	GTTGTTGTTGAGACGGGGTCTTGTTAC
RhoA_ClaI_F	GTACCATCGATAACCTTGCTGCAAGCACAG
RhoA_SalI_R	ACGCGTCGACGGATACAGGAAGTTTAGAAAACTGCCTTTATTC
RhoA_G_F	GTTGGTAACTTTTGTGAATTGGGCTGTAACTAC
RhoA_G_R	GTAGTTACAGCCCAATTCACAAAAGTTACCAAC
MRPS16_ClaI_F	GTACCATCGATATGAGCTGACTTTAGTGAGCATAG
MRPS16_SalI_R	ACGCGTCGACGGAAAATTGAAATCGCACACTGAAATATC
MRPS16_841_R	CAACACCACAGCCAGCCAATTTTTTAAG
MRPS16_841_F	CTTAAAAAATTGGCTGGCTGTGGTGTTG
MRPS16_817_R	CGAGTAGCTGGAACtACGGGTGAG
MRPS16_817_F	CTCACCCGTaGTTCCAGCTACTCG
MRPS16_815_R	CGAGTAGCTGGAACCATGGGTGAG
MRPS16_815_F	CTCACCCATGGTTCCAGCTACTCG
MRPS16_797_R	ACTgCCTCAGCCTCCCGAGTAGCTGGA
MRPS16_797_F	TCCAGCTACTCGGGAGGCTGAGGcAGT
PKR_ClaI_F	GTACCATCGATTGTTACATCATTGCACTTGTAAGTAC
PKR_SalI_R	ACGCGTCGACAATGTCTAGCATGGGCAAATC
PKR_8097_F	CAAGTAAATACAGGTCTCAGTCAGATG
PKR_8097_R	CATCTGACTGAGACCTGTATTTACTTG
PKR_8082_F	CAGTCAGATGGACCCCAAGAGCCAC
PKR_8082_R	TGGCTCTTGGGGTCCATCTGACTGAG
PKR_8075_F	GATGAACCCCAGGAGCCACATG
PKR_8075_R	CATGTGGCTCCTGGGGTTCATC
PKR_8034_F	TCTCACACTTTTGCCTGTTACATGG
PKR_8034_R	CCATGTAACAGGCAAAAGTGTGAGA
PKR_7702_F	GAATCACAGTTGATGGTTATATGGTGAC
PKR_7702_R	GTCACCATATAACCATCAACTGTGATTC
PKR_7670_F	GTGGCTTAAATTCTGAATAACTAGAAACTG
PKR_7670_R	CAGTTTCTAGTTATTCAGAATTTAAGCCAC
PKR_7666_F	GGCTTAAATTCTAAATGACTAGAAACTGTATAATAG
PKR_7666_R	GCCTATTATACAGTTTCTAGTCATTTAGAATTTAAG

PKR_p2_seq	AGGAGTTGGCAACTAATTG
Tre-F-seq	ACTACACCATCGTGGAACAG
Tre-R-seq	GATTATGATCCTCTGGAG
Generating ILF3 KD constructs	
Name	**Sequences**
lLF3_sh_F	CCGGGGTCTTCCTAGAGCGTCTAAACTCGAGTTTAGACGCTCTAGGAAGACCTTTTTG
lLF3_sh_R	AATTCAAAAAGGTCTTCCTAGAGCGTCTAAACTCGAGTTTAGACGCTCTAGGAAGACC
qPCR primers	
Name	**Sequences**
mCherry qPCR F	ACTACGACGCTGAGGTCAA
mCherry randomR	CGTTCGTACTGTTCCACGATG
eYFP qPCR F	AAGATCCGCCACAACATCGA
eYFP randomR	ACTCCAGCAGGACCATGTG
qTBP-Fw	CAGCAACTTCCTCAATTCCTTG
qTBP-Rv	GCTGTTTAACTTCGCTTCCG
NF90(ILF3) qPCR F	AACCATGGAGGCTACATGAAT
NF90(ILF3) qPCR R	CGCTCTAGGAAGACCCAAAATC
18S_F	CTCTTAGCTGAGTGTCCCGC
18S_R	CTGATCGTCTTCGAACCTCC
7670 qPCR F	CAGGTCCAAATCAAATTAACCCCATAAG
7670 qPCR R	CTGTATAATAGGCAAAACTGTGAGGC
7702 qPCR F	TTGCCTCACAGTTTTGCCTATTATAC
7702 qPCR R	GTTGATAGTTATATGGTGACATTAGTGGC
8034 qPCR F	AGATGTACAGTCGCCCCAC
8034 qPCR R	TTGTCTCACACTTTTACCTGTTACATG
8075 qPCR F	GCCCCACTACTGGCTTAAATTC
8075 qPCR R	GAGCCACATGTATTTGAGGGGTAC
8082 qPCR F	CTGAAAACCATGTAACAGGTAAAAGTG
8082 qPCR R	CAGTCAGATGAACCCCAAGAG
PKR qPCR F	CCTGTCCTCTGGTTCTTTTGCT
PKR qPCR R	GATGATTCAGAAGCGAGTGTGC
Ecadherin qPCR F	GCCCTTGGAGCCGCAG

Ecadherin qPCR R	TCAAAATTCACTCTGCCCAGGA
ZO-1 qPCR F	CAACATACAGTGACGCTTCACA
ZO-1 qPCR R	CACTATTGACGTTTCCCCACTC
Ncadherin qPCR F	TCAGGCGTCTGTAGAGGCTT
Ncadherin qPCR R	ATGCACATCCTTCGATAAGACTG
Vimentin qPCR F	CGAGGAGAGCAGGATTTCTC
Vimentin qPCR R	GGTATCAACCAGAGGGAGTGA
MMP9 qPCR F	TTCTACGGCCACTACTGTGCCT
MMP9 qPCR R	AATCGCCAGTACTTCCCATCCT
siRNAs	
Name	**Sequences**
ADAR1	CGCAGAGUUCCUCACCUGUAUU (Thermo Scienctific Dharmacon)
ADAR2	GCCUGGUUUGCAGUACACTT (Ambion cat# AM51331)
Non-targeting siRNA #2	N/A (Thermo Scienctific Dharmacon)
DsiRNAs	
Name	**Sequences**
hs.Ri.ILF3.13.1_F	rGrGrArArCrUrCrUrArUrCrArCrArArUrUrUrGrArArArAGA
hs.Ri.ILF3.13.1_R	rUrCrUrUrUrUrCrArArArUrUrGrUrGrArUrArGrArGrUrUrCrCrUrU
hs.Ri.ILF3.13.3_F	rGrCrArArArGrCrArUrUrCrUrUrCrCrGrUrUrUrArUrCrCAA
hs.Ri.ILF3.13.3_R	rUrUrGrGrArUrArArArCrGrGrArArGrArArUrGrCrUrUrUrGrCrCrA
Negative Control DsiRNA	N/A (IDT)
Checking endogenous PKR editing levels	
Name	**Sequences**
PKR_editing_F	GGGATTAAGGAAAGGTAAGCATCAAAG
PKR_editing_R	CAGGTCCAAATCAAATTAACCCCATAAG

Supplementary Table 3.3 List of primers and siRNAs used in Chapter 3

CHAPTER 4

Multifaceted role of RNA editing in promoting loss-of-function of PODXL in cancer

4.1 Abstract

PODXL, a protein that is dysregulated in multiple cancers, plays an important role in promoting cancer metastasis. Previous studies showed that RNA editing in PODXL may alter its cellular function to promote cell invasion and tumorigenesis. However, the underlying mechanisms of this editing-mediated functional change of PODXL remain unknown. In this study, we report that RNA editing in *PODXL* affects the gene function by both mediating alternative splicing and creating an amino acid change (i.e., recoding). Specifically, RNA editing promotes the inclusion of a *PODXL* alternative exon, resulting in the formation of the PODXL long isoform that contains a recoding editing site. Using cells stably overexpressing the three PODXL isoforms (short isoform, unedited long isoform, and edited long isoform), we show that the edited PODXL long isoform is more prone to protease digestion and has the strongest effects on reducing cell migration and cisplatin chemoresistance. We discovered that the editing level of the *PODXL* recoding site and the inclusion level of the *PODXL* alternative exon are strongly associated with overall patient survival and progression-free interval in Kidney Renal Clear Cell Carcinoma (KIRC). In general, we observed a significant enrichment of exonic RNA editing sites in alternatively spliced exons. These findings suggest that exonic RNA editing sites may enhance proteomic diversity through alternative splicing, in addition to amino acid changes, a previously under-appreciated aspect of RNA editing function.

4.2 Introduction

RNA editing is a fundamental process in gene expression that introduces deletions, insertions, and base substitutions in the RNA transcripts. This process can happen as soon as the nascent RNA arises, thus potentially impacting transcriptome diversity by altering splicing[50,105,253], modification[254,255], localization[256,257], abundance[57,58,258], and translation[259,260] of RNA molecules. The most common type of RNA editing is adenosine-to-inosine (A-to-I) editing, which is also called A-to-G editing because inosine is interpreted as guanosine by the cellular machineries. A-to-I editing is carried out by the adenosine deaminase acting on RNA (ADAR) enzymes[10,261]. In human, all ADAR proteins (ADAR1, ADAR2, and ADAR3) share the dsRNA-binding domains as well as the deaminase domain that exert the catalytic function[174].

A-to-I RNA editing is dysregulated in cancer[64,262]. Analysis of RNA-seq data of human tumor samples revealed substantial changes in RNA editing in multiple cancer types[77,78]. Much attention has been given to editing sites located in the protein-coding regions, considering their potential in altering amino acid sequences (i.e., recoding). Indeed, a few functional recoding editing sites have been reported in regulating tumorigenesis[79,81–83,87,183–185]. One such example is the recoding editing event (H241R) in the gene *PODXL* (podocalyxin-like), first studied in gastric cancer[79]. This editing site confers a loss of function of PODXL, which results in decreased cell invasion and tumor growth compared to the wild-type protein when overexpressed in the MKN28 cells[79]. Yet it remains unclear how RNA editing alters the function of PODXL.

PODXL is a type I transmembrane protein expressed in various tissues including kidney podocytes, hematopoietic progenitor cells, vascular endothelia, and a subset of neurons[263]. The

extracellular domain of PODXL is highly glycosylated and sialylated, contributing to the negatively charged cell surface coat of the glomerular epithelium that maintains proper structure of the glomeruli in the kidney[264–266]. PODXL is also found to be abnormally upregulated in more than 10 types of cancer[266–272]. Association studies suggested that PODXL is a potential biomarker for cancer diagnosis and prognosis in multiple cancers[272–277]. In addition, PODXL plays important roles in cancer metastasis through promoting cell migration[269,272,277,278], cell invasion[79,269,270,277], cell extravasation[279], immune evasion[280], and chemoresistance[267,272,281,282]. Therefore, PODXL is also a valuable therapeutic target for cancer metastasis[268,283].

Most previous studies on *PODXL* editing pursued the assumption that the RNA recoding site functions mainly through amino acid changes in the protein products. Nonetheless, it is known that some exonic editing sites rely on intronic editing complementary sequences to form double-stranded RNA substrates for ADARs[50,52,105]. These structures (and their interactions with ADAR) may impact downstream RNA processing steps. Since the recoding site of *PODXL* is in the alternatively spliced exon, we hypothesized that it may also be involved in RNA splicing regulation. We observed that the two exonic RNA editing sites of *PODXL* synergistically enhance alternative exon inclusion. Through alternative splicing and editing-mediated amino acid changes, the *PODXL* gene can give rise to three isoforms that are functionally distinct in protease digestion patterns, cell migration, and cisplatin chemoresistance.

4.3 Results

4.3.1 RNA editing can potentially affect *PODXL* alternative splicing

PODXL encodes for two transcript isoforms. The third exon of *PODXL* can be alternatively skipped, leading to an in-frame short isoform (Fig. 1A). Both isoforms are endogenously expressed in A549 and Hela cell lines with the shorter isoform being dominant (Fig. 1B). Interestingly, the alternative exon of *PODXL* forms dsRNA structures with its upstream intronic sequences (Fig. 1C). A previous study reported two A-to-I RNA editing sites (A722G and A714G) in the alternative exon of *PODXL*[79], one of which induces an amino acid change (namely, recoding site, H241R resulting from A722G).

Since the two RNA editing sites are relatively close to the 3' splice site of the alternative exon (+8, +16, Fig. 1C), we hypothesized that RNA editing may affect splicing of *PODXL*. To test this hypothesis, we sub-cloned the alternative exon and its flanking intronic regions (~500bp on each side, encompassing the predicted dsRNA structure) into a splicing reporter that was developed previously[35,50,284], with modifications (Fig. S1A, see Methods). This reporter contains two exons (together encoding the *GFP*) that are upstream and downstream of the tested alternative exon, respectively. We generated four constructs representing the four possible combinations of the editing status of the two RNA editing sites (AA, AG, GA, and GG, A and G representing the unedited and edited versions, respectively). These splicing reporters were then transfected into Hela cells individually to test for exon inclusion rate. Hela cells were used here because they are easy to transfect and show alternative splicing pattern for the endogenous *PODXL* (indicating presence of *trans*-factors for its alternative splicing). Note that the minigenes were not edited in the presence of endogenous ADAR proteins (Fig. S1B, C). We observed a general increase in exon inclusion rate when the G allele was introduced (Fig. 1D). The GA version had a larger effect on splicing compared to the AG version, and the combination of two editing sites (GG) gave the largest increase in *PODXL* exon inclusion (Fig. 1D). These results support the hypothesis that RNA editing events on the *PODXL* alternative exon affect its splicing.

4.3.2 ADAR2-dependent *PODXL* alternative splicing

Next, we asked whether splicing of the *PODXL* minigene depends on ADAR expression. According to previous literature, both A722G and A714G are regulated by ADAR2, whereas ADAR1 only affects A714G[79]. Upon co-transfection of an ADAR1 overexpression vector and the *PODXL* minigene (AA version), we did not observe changes in editing levels at either editing site (Fig. S1B, C). However, the A722G editing responded to ADAR2 overexpression. Thus, we focused on the impact of ADAR2 on *PODXL* splicing.

We generated three ADAR2 mutants including the binding mutant EAA[285], the editing-enhanced mutant E488Q[286,287], and the editing-deficient mutant E396A[285]. We co-transfected each ADAR2 mutant or the wild-type (WT) ADAR2 and the *PODXL* splicing reporter (AA version) into Hela cells (Fig. 1E). First, we confirmed that the *PODXL* reporter is not edited when co-transfected with an empty backbone vector or the editing-deficient mutant E396A. In contrast, co-expression of the WT ADAR2 enhanced the editing of the A722G site, and co-expression of the editing-enhanced mutant E488Q greatly increased the editing levels of both A714G and A722G sites in the reporter (Fig. 1F). Interestingly, co-expression of the RNA binding mutant EAA also enhanced the editing level of the A722G site, indicating that this ADAR2 mutant can induce editing of the *PODXL* transcripts without the RNA binding domains. This editing activity may be enabled by RNA binding through endogenous ADAR2 in complex with the ADAR2 mutant. Alternatively, the deaminase domain of the ADAR2 mutant may facilitate RNA binding, as previously reported[288,289].

Next, we quantified the splicing level of the *PODXL* minigene in the co-transfection assays. The editing-enhanced mutant E488Q of ADAR2 showed an increase in exon inclusion compared to empty control or the editing-deficient mutant E396A (Fig. 1G). This result is consistent with the above finding where the GG version of the minigene had the highest exon inclusion level (Fig. 1D). Curiously, overexpression of ADAR2 WT led to a slightly reduced exon inclusion rate compared to the empty control or its binding mutant EAA. We hypothesize that RNA binding by ADAR2 may inhibit the exon inclusion of *PODXL*, which counteracts the effect of increased RNA editing levels.

4.3.3 Edited PODXL long isoform is more prone to protease digestion

In addition to the regulation of *PODXL* splicing, one of the RNA editing sites (A722G) also introduces a recoding event (H241R) to the PODXL protein (Fig. 1A). PODXL is a transmembrane protein that plays important roles in maintaining the filtration slit in the glomerulus[263,264]. Previous studies showed PODXL enrichment in the leading edges of A549 cells during cell migration [290] and PODXL overexpression increased cell migration in A549 cells [277]. Interestingly, the PODXL alternative exon containing the recoding site is located in the extracellular domain. Thus, we first examined whether the alternative PODXL isoforms are located on the cell membrane. We generated stable A549 cell lines overexpressing different PODXL isoforms (short isoform where the alternative exon is skipped, wild-type long isoform, long isoform with the H241R recoding event) as well as the control vectors (empty). We performed cell fractionation on the PODXL-overexpressing (PODXL-OE) A549 cells and examined the protein expression levels of different PODXL isoforms in the cytoplasmic and membrane fractions. As expected, PODXL showed robust expression in the membrane fraction

and no substantial difference was observed in the cellular localization of different PODXL isoforms (Fig. S2).

Next, we asked if the presence of the alternative exon with or without the recoding event could change the protein conformation on the cell membrane. To test this, we treated the live PODXL-overexpressing A549 cells with the protease Trypsin and proteinase K. We reasoned that if different PODXL isoforms differ in their extracellular regions, they may possess different digestion patterns after the treatment. Strikingly, we observed a distinct trypsin digestion pattern for the PODXL long isoform with the H241R recoding event. Almost all the upper band of the PODXL long isoform with H241R was digested while the upper band of the other two isoforms remained relatively intact (Fig. 2A).

Upon examination of the amino acid sequences of each isoform, we observed that the recoding event H241R on the PODXL long isoform indeed creates a novel trypsin digestion site (Fig. 2B). However, it should be noted that there are 38 trypsin digestion sites already present in the PODXL short and long isoforms without the H241R event. Thus, it is unlikely that the additional trypsin site led to a dramatic digestion change of the H241R isoform. Rather, we hypothesize that the PODXL long isoform with H241R editing has a different protein conformation that renders it more accessible to protease digestion. To test this hypothesis, we treated the live PODXL-overexpressing A549 cells with proteinase K, which does not have differential digestion sites between different PODXL isoforms. Interestingly, cells expressing the PODXL long isoform with the H241R recoding event showed higher sensitivity to protease K digestion (Fig. 2C). These findings suggest that the H241R recoding event on the PODXL long isoform may alter protein conformation and thereby render them more prone to protease digestions.

4.3.4 PODXL isoforms regulate cell migration

PODXL was reported to regulate cell migration and invasion in various cell line models, including A549 cells[267,277]. Thus, we next asked whether PODXL isoforms impact cell migration and cell invasion abilities differently. We performed scratch wound cell migration assays using A549 cells with either stable PODXL overexpression or KD (Fig. S3). Wounds were introduced to each well of confluent cells in 96-well plates, which were then imaged using the Incucyte® S3 live-cell analysis system at 2-hour intervals (Fig. 3A, showing a subset of time points to save space). We used the relative wound density as a measurement for the cell migration ability. Starting from 10h post-wound generation, we observed increased cell migration in cells overexpressing PODXL and a decrease in cell migration in cells with PODXL KD (Fig. 3B), which is consistent with previous literature[277]. Interestingly, the PODXL short isoform was associated with the highest cell migration ability, whereas cells overexpressing the long isoform with H241R had the lowest migration ability among the three isoforms (Fig. 3C). To exclude the possible confounding effect of cell proliferation differences associated with the three isoforms, we seeded these cells at equal cell density and monitored their proliferation (Fig. S3D). We observed no significant differences in cell proliferation associated with the three isoforms as measured by confluence at 10h post the wound generation (Fig. 3D). Overall, these findings suggest that the PODXL isoforms have functional differences in regulating cell migration.

To test the effects of PODXL isoforms on cell invasion, we performed the 3D 96-well scratch wound invasion assay. Similar to the cell migration assay, we introduced wounds to confluent cells in 96-wells and then added Matrigel on top of the cells to create a 3D matrix mimicking the extracellular matrix. However, we did not find significant differences in cell

invasion associated with different PODXL isoforms (Fig. S3E). While we observed a decrease in cell invasion for cells with PODXL KD (Fig. S3E), this may reflect the reduced cell proliferation upon PODXL KD (Fig. S3D).

4.3.5 PODXL isoforms regulate cisplatin chemoresistance

Cisplatin is a wide-spectrum anti-tumor drug that has been applied to treat various human solid tumors. Previous studies reported that PODXL promotes cell resistance to cisplatin in oral tongue squamous cell carcinoma and osteosarcoma (OS)[281,282]. To further understand the functional differences of PODXL isoforms, we examined cisplatin chemoresistance of human OS cell line U2OS overexpressing different PODXL isoforms as well as U2OS cells with PODXL KD (Fig. S4). We conducted the cell cytotoxicity assay with cisplatin at a concentration of 30 μM, which confers around 50% death rate for the U2OS cells overexpressing PODXL isoforms after 48h treatment (Fig. 4A). To remove bias in cell proliferation, we calculated cytotoxic indexes by normalizing the dead cell object counts against the total number of DNA-containing object counts. We observed significantly higher cytotoxic index values (~1) in U2OS cells with PODXL KD compared to those treated with scrambled controls (shctrl) (Fig. 4A), confirming a lower level of cisplatin chemoresistance in KD cells. We also found that cells overexpressing all three PODXL isoforms showed significantly lower cytotoxic index values compared to cells expressing the empty backbone (Fig. 4A). These results are consistent with previously reported function of PODXL in promoting cisplatin chemoresistance[282]. Interestingly, cells overexpressing the PODXL long isoform with the H241R recoding event showed a slightly higher cytotoxic index compared to cells overexpressing the unedited version (Fig. 4A). Thus, the H241R recoding event dampended the function of PODXL in promoting cisplatin chemoresistance.

To further compare the cisplatin chemoresistance associated with different PODXL isoforms, we performed the cytotoxicity assay under various concentrations of cisplatin. Figure 4B-C show the dose-response curves and the half maximal effective concentration (EC_{50}) of cisplatin. A higher EC_{50} value reflects higher cisplatin chemoresistance. The wild type U2OS cells and U2OS cells overexpressing empty controls had mean EC_{50} values of 20.6 μM and 19.6 μM, respectively (Fig. 4C). In contrast, U2OS cells overexpressing PODXL wild type isoforms showed increased mean EC_{50} values (24.9 μM for the short isoform and 24.4 μM for the long isoform) (Fig. 4C). Cells overexpressing the PODXL long isoform with the H241R recoding event had a mean EC_{50} value of 23.5 μM, which was slightly lower than its wild-type counterpart (Fig. 4C). In addition, the mean EC_{50} values for U2OS cells with PODXL KDs were 7.5 μM (shPODXL1) and 9.9 μM (shPODXL2), both of which were lower than the U2OS shctrl cells (14.1 μM) (Fig. 4C). The above results again confirm the role of PODXL in promoting cisplatin chemoresistance. In addition, the edited PODXL long isoform showed reduced capacity in this role compared to the short or unedited isoform.

Upon the addition of cisplatin to the U2OS cells, we observed the loss of cell-to-cell contact and cell shrinkage, indicating the induction of cell apoptosis. To directly measure the effect of PODXL isoforms on cisplatin-induced cell apoptosis, we performed the cell apoptosis assay in U2OS cells at a fixed cisplatin concentration of 30 μM. Following cisplatin treatment, we calculated the normalized apoptotic index, defined as the number of caspase-3/7 positive objects divided by the total number of DNA-containing objects. Consistent with the cytotoxic index results, KD of PODXL increased the cell sensitivity to cisplatin as reflected by the increased apoptotic indexes in U2OS shPODXL1/2 cells (Fig. 4D). We also observed decreased apoptotic index in U2OS cells overexpressing PODXL isoforms than the U2OS empty control cells (Fig. 4E). Among the three PODXL isoforms, the PODXL long isoform with H241R showed the highest apoptotic index, whereas the short PODXL isoform had the lowest apoptotic index

values (Fig. 4E). These results suggest that the PODXL short isoform has the strongest resistance against cisplatin, followed by the PODXL long isoform, then the PODXL long isoform with the H241R recoding site, respectively.

4.3.6 *PODXL* editing and splicing are clinically informative

Given the distinct effects of PODXL isoforms on cancer cell migration and sensitivity to cisplatin, we investigated the clinical relevance of the *PODXL* alternative exon and its editing levels. For this analysis, we quantified editing (A722G and A714G) and inclusion of the alternative exon in Kidney Renal Clear Cell Carcinoma (KIRC) tumors from The Cancer Genome Atlas (TCGA), as PODXL is closely involved in kidney function[263–266]. Comparisons across clinical stages revealed significantly reduced editing levels of A722G in patients with advanced disease (Fig. 5A). Parallel with editing levels of this recoding site, the inclusion of the alternative exon significantly decreased over the progression of tumor stages (Fig. 5B). These coordinated changes across tumor stages are consistent with the effect of editing on alternative splicing (Fig. 1D) and regulation of cell migration by PODXL isoforms (Fig. 3). Considering next the prognostic value of these *PODXL* alternative exon features, we observed that low editing (A722G) and low alternative exon inclusion were each significantly associated with worse overall survival (Fig. 5C-D). Low inclusion of the alternative exon, regulated by diminished editing levels, corresponds to a relatively higher abundance of the short isoform. Consequently, the presence of this short isoform may contribute to enhanced migratory capacity, which may lead to poorer patient prognosis.

4.3.7 Known exonic editing sites are enriched in alternative exons

Our data suggest a coupling between RNA editing and splicing in *PODXL*, both of which are functionally relevant. Similar to *PODXL*, the RNA recoding sites in the glutamate receptor subunit B (*GluR-B*) pre-mRNA modulate its splicing[105]. RNA editing also plays important roles in the exonization of *Alu*-exons[56]. Thus, we hypothesize that the coupling between RNA editing and alternative splicing is a widespread phenomenon. To support this hypothesis, we examined all known exonic editing sites cataloged by the REDIportal database[291]. Strikingly, alternative exons encompassed a significantly higher number of known exonic editing sites compared to random control exons ($p = 8.6e-10$, Fig. 6A). Similar enrichment was also found if only recoding editing sites were included (from REDIportal) ($p = 2.1e-9$, Fig. 6B).

This enrichment of editing sites in alternatively spliced exons may reflect an apparent relationship resulting from the prevalence of RNA editing in *Alus* and enrichment of alternative splicing among *Alu*-exons. To assess this possibility, we checked the overlap between the editing-harboring alternative exons and *Alu* annotations. Only ~30% of these alternative exons showed overlap with the Alu elements (Fig. 6C), indicating that the observation of enriched exonic editing sites in alternative exons is not completely associated with *Alu* exonizations. Rather, it may also reflect the functional roles of RNA editing in regulating alternative splicing (Fig. 6D). These findings support the coupling between RNA editing and alternative splicing. Although the exact mechanisms remain unknown, it is possible that a large number of exonic editing sites, in addition to inducing potential codon changes, may also enhance proteomic diversity by regulating RNA splicing. The alternative exons overlapping with exonic editing sites are enriched in gene ontology terms such as platelet activation, metal ion binding, and response to DNA damage (Fig. S5). Dysregulated editing and alternative splicing may thus further contribute to molecular phenotypes characteristic of human diseases.

4.4 Discussion

Among the vast number of RNA editing sites in human transcriptomes, recoding sites, i.e., those that alter amino acid sequences, have been the focus of many studies due to their readily appreciated impact on protein sequences[79,81–83,87,183–185]. Here we report a novel mechanism in which the recoding site in *PODXL*, a gene abnormally expressed in cancer, promotes PODXL loss-of-function via both alternative splicing and protein-recoding. PODXL is a transmembrane protein expressed in various tissues including the kidney podocytes[263]. Abnormally expressed in multiple types of cancer[266–272], PODXL is a potential biomarker for cancer diagnosis and prognosis[272–277], and a therapeutic target for cancer metastasis in multiple cancers[268,283].

We showed that the recoding editing site, residing in an alternatively skipped exon of *PODXL*, promotes the inclusion of this exon. The long isoform of PODXL resulting from exon inclusion, and further, the edited version of the long isoform, reduce the protein's function in promoting cell migration and cisplatin chemoresistance. Consistently, higher editing and higher exon inclusion were associated with better patient survival in KIRC. Although the function of *PODXL* and its recoding site has been reported previously[79], our study affords an in-depth functional comparison of the alternative isoforms and edited versions of PODXL, an aspect that had not been fully appreciated and may have clinical relevance.

RNA editing may affect alternative splicing through multiple mechanisms such as by creating splice site sequences, altering exonic splicing enhancers or silencers, or via a kinetic competition between the splicing machineries and ADARs[50,56,105,292,293]. For the two RNA editing sites of *PODXL*, using the Mutation Analysis tool provided by Human Spicing Finder[294], we did not detect any significant alterations in splicing signals that would explain our observations. Alternatively, the impact of RNA editing on *PODXL* splicing may be explained by a two-facet model. First, ADARs and the spliceosome compete for the dsRNA substrate formed between

the alternative exon and the flanking intron, which explains the inhibitory effect of ADAR2 binding on the *PODXL* alternative exon inclusion (Fig 1G, Fig. 6D). Second, once edited by the ADARs, the dsRNA structure is more accessible to the spliceosome, which enhances exon inclusion (Fig 6D). In this model, the two aspects of ADAR function, RNA binding and RNA editing, have opposing impacts on alternative splicing. Future work is needed to examine the kinetic and binding properties of ADAR and its relationship to the spliceosome.

Our study highlighted a previously under-appreciated aspect that RNA editing sites promote proteomic diversity not only through amino acid changes but also through alternative splicing. Interestingly, we found a strong enrichment of known exonic editing sites in alternatively spliced exons (Fig 6A). This observation could be the consequence of *Alu* exonization, which gives rise to alternative exons that are prone to RNA editing due to the dsRNA structure of *Alu*. However, only ~30% of the editing-containing alternative exons overlap with *Alu* elements (Fig. 6C), indicating that the enrichment cannot be fully explained by *Alu* exonization. Rather, it may reflect the general roles of exonic editing sites and ADARs in regulating alternative splicing.

Interestingly, we found that the PODXL long isoform with the H241R recoding event is more prone to protease digestion (Fig 2). The H241R recoding site directly creates a new trypsin digestion site (Fig 2B). It is possible that the long isoform with the H241R recoding site undergoes conformational changes that further expose the digestion sites to the proteases, explaining its higher protease sensitivity compared to the other isoforms (Fig 2A, 2C). This observation may have close relevance to cancer metastasis. The tumor microenvironment exhibits abnormal protease activities that modulate tumor invasion and metastasis[295]. While digestion of the extracellular matrix facilitates cell invasion, cleavage of the extracellular domain of transmembrane receptors also modulates the associated intracellular signaling pathways that

are important in tumor cell survival and drug resistance[295,296]. It is possible that the PODXL long isoform with the H241R recoding event is subject to higher proteolytic pressure, thus altering the intracellular binding complex (NHERF1/2-Ezrin-Actin)[297,298] or downstream signaling pathways (PI3K/Akt[282,298] and Bmi1/FAK[281], etc.) that regulate its function in cell migration and drug resistance.

We showed that higher editing and higher exon inclusion in *PODXL* were associated with better patient survival in KIRC. The cell-based assays in this study yielded converging results using multiple cell lines. In addition, previous literature supports the ubiquitous involvement of PODXL in multiple cancer types[267]. Thus, we hypothesize that *PODXL* editing and splicing levels may have clinical relevance in multiple cancer types. However, significant association of *PODXL* editing and splicing with patient survival was not observed in other cancer types (e.g., lung adenocarcinoma, Fig. S6). This may be due to lack of accurate measures of editing and splicing levels that demand high coverage of *PODXL* transcripts in the RNA-seq data (Fig. S6E), since the other tissues (e.g., lung) had much lower *PODXL* expression than kidney[299]. In the future, the relevance of *PODXL* editing and splicing to a wide range of cancer types needs to be further examined.

In summary, our study highlights the functional importance of *PODXL* editing in multiple cancer-related processes. We showed that RNA editing in *PODXL* affects the protein function by inducing changes in both alternative splicing and protein sequences. Such multifaceted roles of RNA editing were previously under-appreciated and may exist for many editing sites in the coding regions.

4.5 Methods

4.5.1 Cell culture

A549 (male), HeLa (female), U2OS (female), and HEK293T (female) cells were maintained in DMEM (Gibco) with 10% FBS (Gibco) and antibiotic-antimycotic reagent (Gibco) at 37 °C with 5% CO_2 supply. Cell lines have not been authenticated.

4.5.2 PODXL overexpression and knockdown

To generate the PODXL overexpression stable cell lines, the coding region of each PODXL isoforms was cloned into the pLJM1-EGFP vector (Addgene) using the restriction enzymes AgeI-HF (NEB) and EcoRI-HF (NEB). The primers used for cloning are listed in Table S1. For empty control (pLJM1-Empty), the *EGFP* coding sequences in the pLJM1-EGFP construct were replaced with a short fragment of multi-cloning sites that do not express any proteins. The resulting pLJM1 vectors (pLJM1-Empty, pLJM1-PODXL-short-isoform, pLJM1-PODXL-long-isoform-A, pLJM1-PODXL-long-isoform-G) were separately co-transfected with the dR8.91 and VSVG plasmids into the HEK293T cells using lipofectamine3000 (Invitrogen) according to the manufacturer's protocol. The lentivirus-containing media were collected every 24h for a total of 72 hours. The lentivirus-containing media were then filtered through 0.2 μm PES filters (VWR). Prior to lentiviral transduction, the A549 cells or the U2OS cells were seeded at 0.1M per well in 6-well plates. 500 μl of virus-containing media were added to each well with the addition of polybrene (Santa Cruz Biotechnology) at a final concentration of 8 μg/ml. 24h post cell transduction, fresh media were added to the cells with puromycin (Fisher BioReagents) at a final concentration of 1 μg/ml. The transduced cells were maintained in the puromycin-containing media for at least 7 days prior to any experiments. To generate PODXL knockdown stable cells,

two PODXL-targeting shRNA constructs (TRCN0000296029, targets 3'UTR; TRCN0000310117, targets CDS; primers provided in Table S1) were cloned into the pLKO.1-TRC cloning vector. The two PODXL shRNA constructs (pLKO.1-shPODXL1 and pLKO.1-shPODXL2), together with the pLKO.1-scramble shRNA control, were co-transfected with the dR8.91 and VSVG plasmids into HKE293T cells for lentiviral packaging, individually. The viruses were then used to transduce A549 or U20S cells to create PODXL knockdown stable cells using similar procedures as described above. For all plasmid constructions described above, NEB Stable Competent *E.coli* were used.

4.5.3 RNA isolation and cDNA generation

Cells were washed with PBS (Gibco) and lysed with TRIzol (Thermo Fisher Scientific). Each 500 µl TRIzol-lysed solution was mixed with 100 µl chloroform (Fisher Chemical) to allow phase separation. The upper aqueous phase was transferred and mixed with equal volume ethanol (200 proof, Fisher BioReagents). The mixture was loaded to the column supplied by the Direct-zol RNA Miniprep Plus kit (Zymo Research) to isolate total RNA following the manufacturer's protocol. 1~2 µg of total RNA was used for cDNA synthesis with SuperScript IV (Thermo Fisher Scientific) using random hexamers.

4.5.4 Detection of *PODXL* isoforms via PCR

Primers used for *PODXL* isoform detection are listed in Table S1. For endogenous *PODXL* isoform detection, 1 µl of the cDNA was used for PCR using DreamTaq™ Green PCR Master Mix (2X) (Thermo Fisher Scientific). The PCR reaction was carried out with an annealing temperature of 55 °C for 28 cycles. The PCR products were resolved in a 1% agarose gel with

Ethidium bromide (Sigma) staining and visualized under the imager (Syngene PXi). For *PODXL* isoform detection in splicing reporters, 1 µl of the cDNA was used for PCR using DreamTaq™ Green PCR Master Mix (2X) (Thermo Fisher Scientific), which underwent 28 cycles with an annealing temperature of 60 °C. The PCR products were resolved in a 6% PAGE gel, stained with SYBR Safe DNA Gel Stain (Thermo Fisher Scientific), and visualized under the imager (Syngene PXi). Images were analyzed using ImageJ to quantify band intensity for both short and long isoforms. The exon inclusion rate was calculated by (short/(short+long)). To measure the expression of *PODXL* (all isoforms) or just the long isoforms via qPCR, 1 µl of cDNA was used, together with the PowerUp™ SYBR® Green Master Mix (Thermo Fisher Scientific). The reaction was performed in the CFX96 Touch Real-Time PCR detection system (Bio-Rad) with the following settings: 50 °C for 10 min, 95 °C for 2 min, 95 °C for 15 s, 60 °C for 30 s, and with the last two steps repeated for 45 cycles. For the *PODXL* overexpression samples, the expression of *PODXL* was normalized against the expression of *18S* or *TBP*. For the *PODXL* knockdown samples, the expression of *PODXL* was normalized against the expression of *TBP*.

4.5.5 RNA structure predictions

The sequences of the PODXL alternative exon and its flanking introns (~500bp each) were folded using mFold[300] with default settings.

4.5.6 Construction of splicing minigene reporters

The *PODXL* alternative exon and its flanking intronic sequences (~500bp upstream and downstream of the exon) were cloned into the pZW1-GFP splicing reporter that was described previously[35,50,284]. To generate an in-frame transcript when *PODXL* alternative exon is included,

we introduced two insertions (+1c, +ag) at the splice sites of the two *GFP* sub-exons. A T-to-G mutation (named as 3ssTg) was also introduced to the *PODXL* 3'splice site to increase the basal exon inclusion rate so that it approximately matches the endogenous exon inclusion level. Primers used for making modifications to the splicing reporter and introducing the *PODXL* editing sites are listed in Table S1. NEB 5-alpha Competent *E.coli* were used for plasmid construction.

4.5.7 ADAR overexpressing constructs

The coding sequences for the ADAR1 p110 isoform and ADAR1 p150 isoform were amplified from the constructs previously generated in our lab[252] and cloned into the pcDNA4-TO-FLAG-myc-His vector (Invitrogen) using restriction enzymes NotI-HF (NEB) and BstBI (NEB). ADAR2 mutant constructs (EAA, E396A, and E488Q) were generated by introducing the recoding mutations to the pcDNA4-ADAR2-WT construct previously generated in our lab[74]. In general, the ADAR2 coding sequences were reamplified to introduce mutations using overlap extension PCR, followed by digestion and ligation into the pcDNA4-TO-FLAG-myc-His vector via the restriction enzymes NotI-HF (NEB) and XbaI (NEB). All PCR reactions were performed using the Q5® Hot Start High-Fidelity 2X Master Mix (NEB). The primers used for PCR reactions are listed in Table S1. NEB 10-beta Competent *E.coli* were used for plasmid construction.

4.5.8 Western blot

The cells were washed with cold 1x PBS and then lysed with cold RIPA buffer with Pierce™ Protease Inhibitor Tablets (EDTA-free, Thermo Fisher Scientific) freshly added. After incubation at 4 °C for 30 min, the whole cell lysate was collected and centrifuged for 30 min at 12,000 g,

4 °C. The supernatants were transferred to a new tube for protein concentration measurement using the Pierce™ Detergent Compatible Bradford Assay Kit (Thermo Fisher Scientific). After mixing with the 4× SDS protein loading dye, the protein samples were boiled at 95 °C for 5 min and loaded onto SDS-PAGE gels with the PageRuler™ Prestained Protein Ladder, 10 to 180 kDa (Thermo Fisher Scientific), followed by protein transfer to nitrocellulose membranes (GE Healthcare) and antibody incubations. Antibodies used were as follows: ADAR1 antibody (Santa Cruz Biotechnology, sc-73408, 1:200), ADAR2 antibody (Santa Cruz Biotechnology, sc-73409, 1:200), FLAG antibody (Sigma, F1804, 1:1000), PODXL antibody (Santa Cruz Biotechnology, sc-23904, 1:500), HSP 90α/β antibody(Santa Cruz Biotechnology, sc-13119, 1: 500), EGFR antibody (Cell Signaling, #4267, 1:1000), β-actin-HRP antibody (Santa Cruz Biotechnology, sc-47778, 1: 2000), goat anti-rabbit IgG-HRP (Santa Cruz Biotechnology, sc-2004, 1:2000), and goat anti-mouse IgG-HRP (Santa Cruz Biotechnology, sc-2005, 1:2000). To visualize, the membrane blots were incubated with SuperSignal West Pico PLUS Chemiluminescent Substrate (Thermo Fisher Scientific) and then imaged using the Syngene Pxi imager.

4.5.9 Splicing reporter assay

Hela cells were seeded in 12-well plates to reach 90% confluency by the time of cell transfection. 375ng of reporter plasmids were transfected into each well with lipofectamine 3000 (Invitrogen). For ADAR co-transfection experiments, 1,250ng of ADAR-overexpressing plasmids and 375ng of reporter plasmids were transfected into each well with lipofectamine 3000. The total RNA was harvested 24h post cell transfection and processed to detect *PODXL* isoforms generated from the splicing reporters, as described above.

4.5.10 Quantification of RNA editing levels

The *PODXL* alternative exon was amplified from the cDNA of Hela cells transfected with the splicing reporters. The primers used are listed in Table S1. PCR reactions were performed with DreamTaq™ Green PCR Master Mix (2X) (Thermo Fisher Scientific) for 28 cycles. PCR products were resolved in 1% agarose gel and then purified using Zymoclean™ Gel DNA Recovery Kit (Zymo Research). The reverse primer was mixed with the amplicons and sent for Sanger sequencing. To quantify the RNA editing levels, the peak signals of both A alleles and G alleles were measured using 4Peaks, followed by editing level calculation (G/(A+G)).

4.5.11 Cell fractionation and protease digestion

Cell fractionation was performed using the Subcellular Protein Fractionation Kit for Cultured Cells (Thermo Fisher Scientific). The PODXL-OE A549 cells were washed with cold 1x PBS and directly lysed with the Cytoplasmic Extraction Buffer provided by the kit. For protease digestion assay, the PODXL-OE A549 cells were washed with warm 1x PBS, and then treated with either TrypLE™ Express (Gibco) or protease K (100 μg/ml, Zymo Research) at 37 °C for 5 min. The digested cells were then collected with growth media and centrifuged at 500 g for 5 min. The cell pellet was further processed using the Subcellular Protein Fractionation Kit for Cultured Cells (Thermo Fisher Scientific) following the manufacturer's protocol.

4.5.12 Prediction of protease cleavage sites

The protease cleavage sites of the PODXL long isoforms (WT and H241R) were predicted using PeptideCutter[301].

4.5.13 Cell proliferation assay

The PODXL-OE A549 cells were seeded at 3,000 cells per well in the 96-well plates. After 24 h incubation at 37 °C, the plate was transferred to the Incucyte® S3 live-cell analysis system (Sartorius) to monitor cell proliferation. Images were taken every 2 h and analyzed for confluency.

4.5.14 Cell migration assay

The PODXL-OE A549 cells were seeded at 30,000 cells per well in the Incucyte® Imagelock 96-well plate (Sartorius) to reach 100% confluency after 24 h incubation. A scratch wound was created on each well using the WoundMaker™ (Sartorius) followed by fresh media change. The 96-well plate was monitored for wound closure by imaging every 2 h. The images were analyzed to calculate the relative wound density, a measure of the density of the wound region relative to the density of the cell region, as recommended by the Incucyte® manual for 96-well scratch wound cell migration assay.

4.5.15 Cell invasion assay

Prior to cell seeding, the Incucyte® Imagelock 96-well plate (Sartorius) was pre-coated with 50 µl of 100 µg/ml Matrigel (Corning) in each well at 37 °C for at least 2 h. The PODXL-OE A549 cells were seeded at 30,000 cells per well in the coated plate and incubated at 37 °C for 24h to reach 100% confluency. After scratch wound introduction using WoundMaker™ (Sartorius), 50 µl of 8mg/ml Matrigel were added to each well and incubated at 37 °C for 20 min until the Matrigel is solidified. Additional 250 µl of cell growth media were added to each well

before the plate was transferred to the Incucyte® S3 live-cell analysis system. Images were taken every 2 h to monitor cell invasion activities. The images were analyzed to calculate the relative wound density as described above.

4.5.16 Cell cytotoxicity assay

The U2OS cells with PODXL-OE or PODXL-KD were seeded at 3,000 cells per well in the 96-well plate and incubated at 37 °C for 24 h. The cells were then treated with 100 µl cell growth media containing 30 µM cisplatin (Selleck Chemical LLC) and the Incucyte® Cytotox Red Dye for counting dead cells. Both phase-contrast and red-fluorescence images were taken for each well every 2 h under the Incucyte® S3 live-cell analysis system. At 48 h after treatment, the assay was terminated by adding 20 µl of 12 µM (diluted in 1x PBS) Vybrant™ DyeCycle™ Green Stain (Invitrogen™) directly to each well (final dye conc. at 2 µM) and imaged using the Incucyte® S3 live-cell analysis system for phase-contrast and green-fluorescence images. The cytotoxic index was calculated by dividing the dead cell numbers (red-fluorescence object counts) by the total number of DNA-containing cells (green-fluorescence object counts).

4.5.17 Determination of the EC_{50} value of cisplatin

The cell cytotoxicity assay was performed under a range of concentrations (0.1 µM, 1 µM, 3 µM, 7 µM, 10 µM, 15 µM, 30 µM, and 200 µM) of cisplatin (Selleck Chemical LLC). The cytotoxic index at 48h of cisplatin treatment was calculated and plotted against the cisplatin concentrations in GraphPad Prism. The EC_{50} value was calculated for each cell line based on the dose-response curve (nonlinear regression) using the “find ECanything” function in GraphPad Prism.

4.5.18 Cell apoptosis assay

The U2OS cells with PODXL-OE or PODXL-KD were seeded at 3,000 cells per well in the 96-well plate. After 24 h incubation in the cell incubator, each well was replaced with 100ul cell growth media containing 30 μM cisplatin (Selleck Chemical LLC) and the Incucyte® Caspase 3/7 Green Dye for apoptosis (Sartorius). Images in phase-contrast and green-fluorescence were taken every 2 h using the Incucyte® S3 live-cell analysis system. At either 26 h or 40 h post cisplatin addition, the assay was ended by adding 20 μl of 12 μM (diluted in 1x PBS) Vybrant™ DyeCycle™ Green Stain (Invitrogen™) directly to each well and imaged again. The apoptotic index was calculated by dividing the number of apoptotic objects (green-fluorescence object counts before the addition of Vybrant™ DyeCycle™ Green Stain) by the total number of DNA-containing objects (green-fluorescence object counts after the addition of Vybrant™ DyeCycle™ Green Stain).

4.5.19 Statistics for cell-based assays

Data were analyzed and plotted using Graphpad Prism 7. Statistical details for each experiment can be found in the figure legends. P-value < 0.05 was used to call significance.

4.5.20 Quantification of editing and PSI in TCGA

Using the BAM slicing functionality of the Genomic Data Commons (GDC) Application Programming Interface (API), we downloaded *PODXL*-overlapping bam files for tumors of Kidney Renal Clear Cell Carcinoma (KIRC) and Lung Adenocarcinoma (LUAD) patients in The

Cancer Genome Atlas (TCGA). After retaining only uniquely mapped reads, we applied our previously published methods[74,172,194,258] to calculate editing ratios at two A-to-I RNA editing sites in the alternative exon of *PODXL* (A722G and A714G). To quantify inclusion levels of the *PODXL* alternative exon, we calculated its percent spliced in (PSI) using previously described methods[302]. Comparison between tumor stages were done using Wilcoxon rank sum test and *p*-value <= 0.05 was used to determine significance.

4.5.21 Survival associations

We downloaded the TCGA Pan-Cancer Clinical Data Resource [303] to obtain survival times for KIRC and LUAD patients. High and low editing groups of patients were defined based on tertiles of *PODXL* A722G editing levels in each cancer type. Similarly, patients were categorized into high and low alternative exon inclusion groups in each cancer type, based on PSI tertiles. With the R package survival, we used the log-rank test to compare overall survival between high and low *PODXL* A722G editing groups, as well as between high and low alternative exon inclusion groups in KIRC and LUAD. P-value < 0.05 was used to call significance. We visualized the Kaplan Meier survival curves using the R package survminer.

4.5.22 Enrichment of editing in alternative exons

To test whether recoding sites were enriched in alternatively spliced exons, we first annotated recoding sites within the REDIportal V2 database[291] by running ANNOVAR[304]. After defining alternative and constitutive exons based on gene annotations from the Consensus Coding Sequence (CCDS) project, we counted the number of recoding sites within alternative exons. We also counted overlapping recoding sites within 1000 sets of control exons, which were

randomly selected from the same genes containing alternative exons. An enrichment p-value for recoding events within alternative exons was calculated from a normal distribution fit to the recoding site counts within control exons. We similarly tested the enrichment of editing sites (not limited to recoding sites) within alternative exons by using counts of editing sites overlapping alternative exons and 1000 sets of control exons. P-value < 0.05 was used to call significance. The alternative exons were overlapped with the *Alu* annotations downloaded from the UCSC Genome Browser.

4.5.23 Gene ontology (GO) enrichment analysis

For each query gene that contains AS exons overlapping with RNA editing sites, a control gene was randomly chosen among the background genes (excluding the query genes) in the CCDS database, 10,000 times. The p value of the enrichment of each GO term in the query genes was calculated using the normal distribution fit to the occurrence of the GO term in the 10,000 sets of control genes. To call significance, FDR < 0.05 and occurrence >= 10 were used.

4.6 Acknowledgements

We thank members of the Xiao laboratory for helpful discussions and comments on this work. The results published here are in part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. This work was supported in part by grants from the National Institutes of Health (U01HG009417, R01CA262686 to XX) and the Jonsson Comprehensive Cancer Center at UCLA.

4.7 Figures

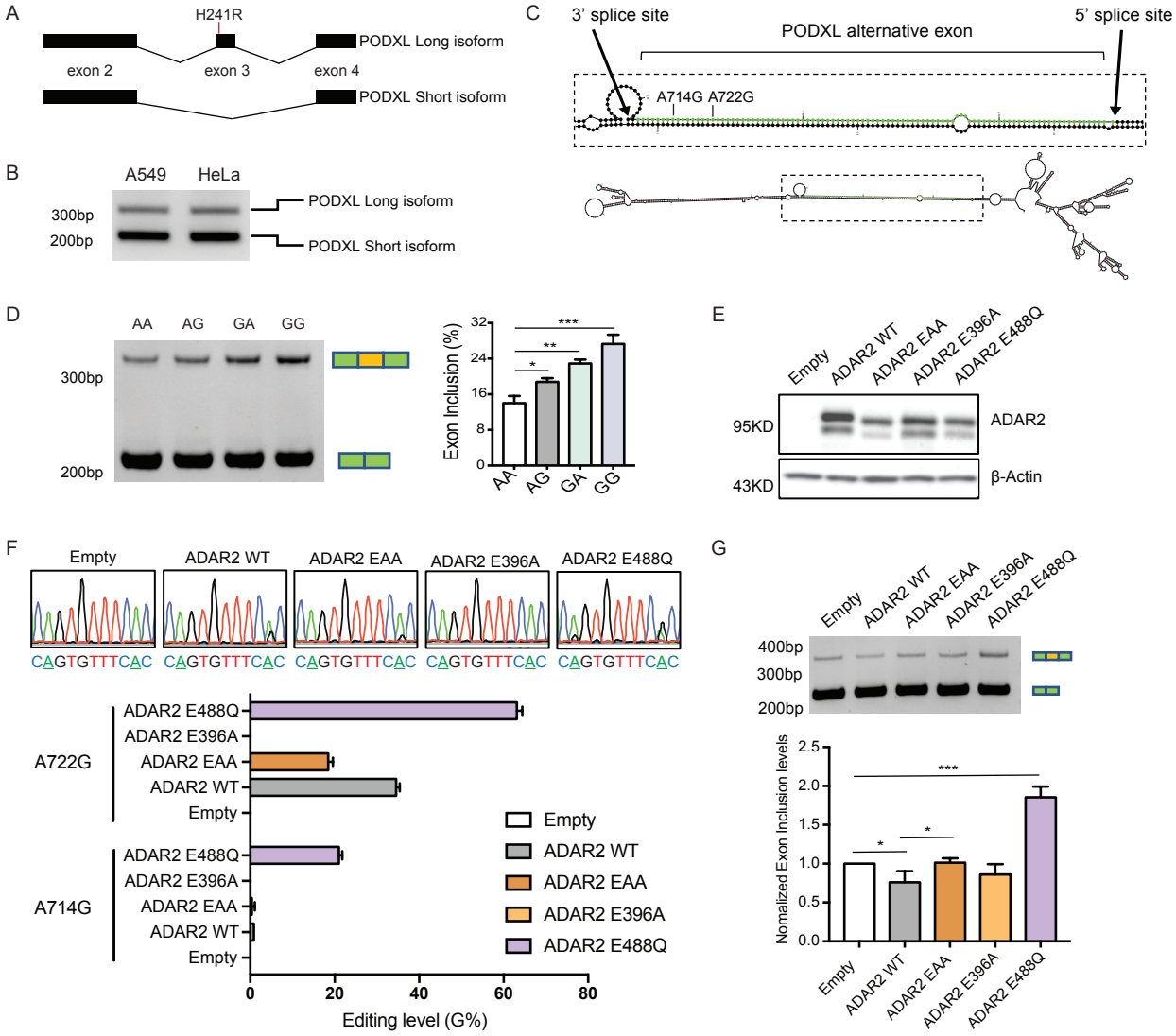


Figure 4.1 RNA editing and ADAR2 regulate *PODXL* alternative splicing

(A) The long and short isoforms of *PODXL*. Partial gene structures (exon 2 ~ 4) are shown. The H241R recoding event is labeled in the alternative exon (exon 3) of the long isoform.

(B) Agarose gel image of the endogenous *PODXL* PCR products amplified from the cDNA of A549 and HeLa cells, respectively.

(C) Predicted RNA structure of the *PODXL* alternative exon (green) with its flanking introns (black). Locations of the two RNA editing sites are labeled.

(D) Left: PAGE gel resolving the amplicons of transcripts derived from the *PODXL* splicing reporters in HeLa cells with four combinations of the A714G and A722G editing events (AA, AG, GA, GG, G represents edited and A, unedited). In the AG reporter, G was introduced at the A722G site. In the GA reporter, G was introduced at the A714G site. Right: Quantification of the *PODXL* alternative exon inclusion rate for each reporter based on the PAGE gel result (measured by ImageJ). Three biological replicates were included. Data are plotted as mean ± SEM. The *p*-values were calculated using Student's t-test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). See also Figure S1A.

(E) Western blot showing overexpression of ADAR2 and its mutants in HeLa cells. The upper bands represent the FLAG-ADAR2 fusion proteins (see also Figure S1B). The lower bands represent ADAR2 or mutant proteins without FLAG tagging, which may result from alternative translation start sites in the overexpression constructs.

(F) Top: Sanger sequencing to detect RNA editing of the A714G and A722G editing sites (underlined As) in the minigene reporters after co-transfection with ADAR2 overexpression vectors or the empty control in HeLa cells. Bottom: Quantification of the RNA editing levels based on the Sanger sequencing peaks (measured using 4Peaks). Three biological replicates were included. Data are plotted as mean ± SEM. See also Figure S1C.

(G) Top: PAGE gel image of the amplicons of *PODXL* transcripts derived from the splicing reporters co-transfected with the ADAR2 overexpression vectors or the empty control in HeLa cells. Bottom: Normalized exon inclusion levels based on the PAGE gel band intensity (measured by ImageJ). Three biological replicates were included. For each replicate, the exon inclusion levels were normalized against the empty control. Data are represented as mean ± SEM. The *p*-values were calculated using Student's t-test (*$p < 0.05$, ***$p < 0.001$).

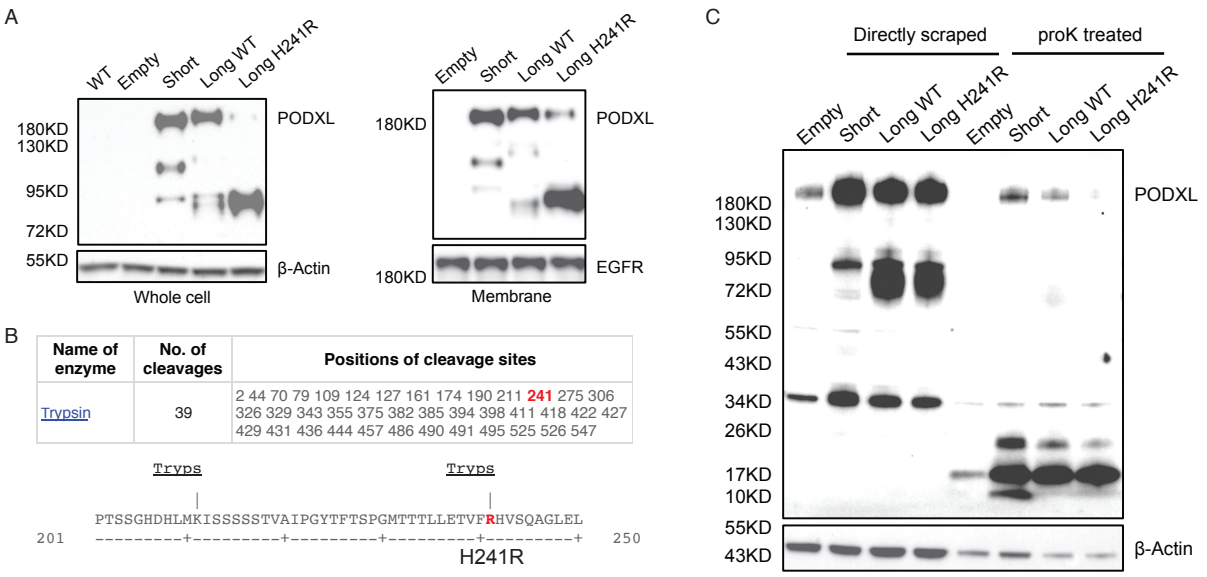


Figure 4.2 PODXL long isoform with the H241R recoding event is more prone to protease digestion than other isoforms

(A) Western blot of trypsinized A549 cells overexpressing PODXL isoforms. Left: whole cell lysates. Right: cell lysates from the cell membrane fraction. The upper bands (~180KD) represent intact PODXL proteins with posttranslational modifications. The middle and lower bands represent truncated PODXL proteins due to trypsin digestion. See also Figure S2.

(B) Recoding RNA editing event H241R creates a novel trypsin digestion site on the PODXL long isoform. Top: Predicted trypsin cleavage sites on the edited PODXL long isoform. Bottom: Amino acid sequences of around the H241R site with trypsin digestion sites labeled.

(C) Western blot of protease K treated A549 cells overexpressing PODXL isoforms. Whole cell lysates were used for Western blot. The upper bands (~180KD) represent intact PODXL proteins with posttranslational modifications. The bands at 72KD ~ 95KD are likely truncated PODXL proteins due to protein degradation. The bands at ~34KD are likely PODXL proteins lacking posttranslational modifications. The10KD to 26KD bands are digested PODXL proteins due to protease K treatment.

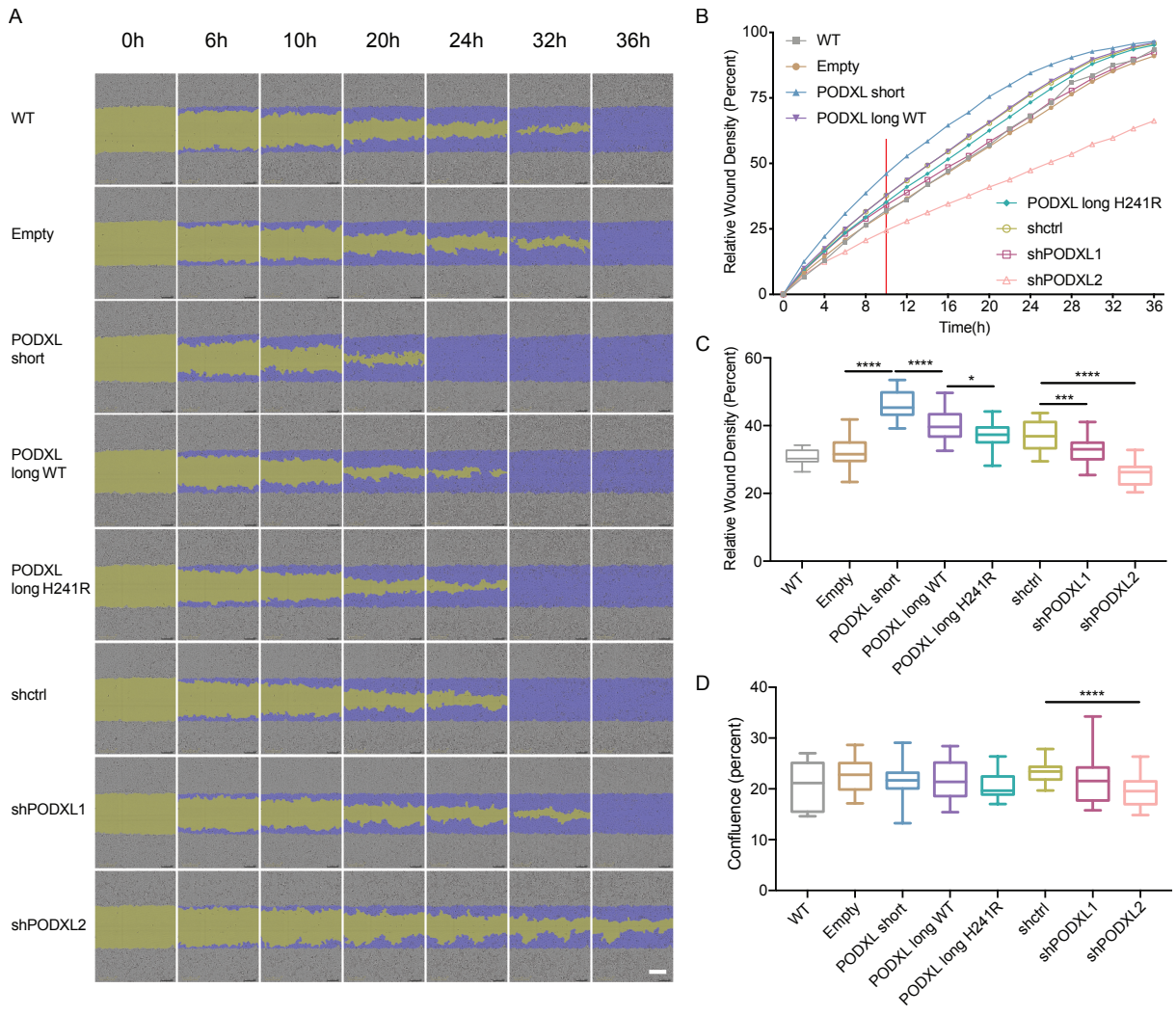


Figure 4.3 PODXL isoforms regulate cell migration to different degrees

(A) Cell migration assay of A549 cells overexpressing PODXL isoforms or with PODXL knockdown (PODXL-OE/KD A549 cells). WT: wild-type A549 cells. Empty: A549 cells overexpressing the empty backbone. shctrl: A549 cells with scrambled control shRNA. Two alternative shRNAs for PODXL were used (shPODXL1, shPODXL2). Phase-contrast images at different time points were shown (purple: initial scratch wound mask, yellow: wound region). Scale bar: white line at the bottom right, 300 μm. See also Figure S3.

(B) Cell migration curve of the PODXL-OE/KD A549 cells described in (A). The plot shows one set of experiments performed with three biological replicates. Red line highlights data at 10 h post wound creation.

(C) Quantification of cell migration with relative wound density for the PODXL-OE/KD A549 cells. Data at 10 h post wound creation are shown, which was the earliest time point when we observed significant differences in cell migration between PODXL isoforms. Two independent sets of experiments were performed with three biological replicates included in each experiment. Data are plotted as mean ± SEM. The p-values were calculated using Student's t-test (*$p < 0.05$, ***$p < 0.001$, ****$p < 0.0001$).

(D) Quantification of cell proliferation using cell confluence levels for the PODXL OE/KD A549 cells. Data at 10 h post wound creation are shown, to exclude the possible effect of cell proliferation on cell migration differences shown in C. Two independent sets of experiments were performed with three biological replicates included in each experiment. Data are plotted as mean ± SEM. The p-values were calculated using Student's t-test (****$p < 0.0001$). See also Figure S3.

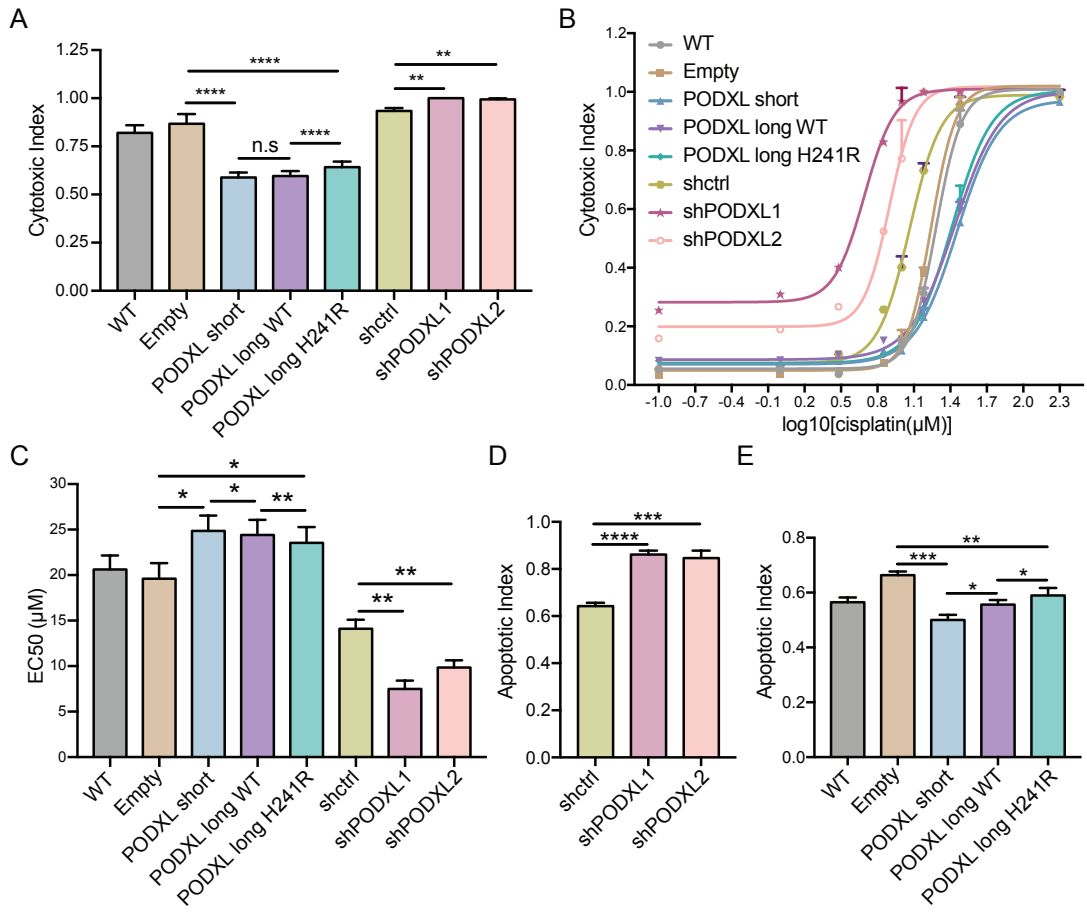


Figure 4.4 PODXL isoforms regulate cell sensitivity to cisplatin to different degrees

(A) Cytotoxic index values of U2OS WT cells, U2OS cells overexpressing empty (control) backbone, the PODXL isoforms, scrambled control shRNA (shctrl), or PODXL shRNAs (shPODXL1, shPODXL2) were generated (See also Figure S4). Cells were treated with 30 μM cisplatin for 48h. Cytotoxic index was calculated by normalizing the dead cell object counts against the total number of DNA-containing object counts. Three independent sets of experiments were performed with three biological replicates included in each experiment. Data are plotted as mean ± SEM. The *p*-values were calculated using Student's t-test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$).

(B) Dose-response curves for the U2OS cells with PODXL overexpression or KD, and controls (WT, Empty, shctrl). The plot shows one set of experiment performed with two biological replicates.

(C) EC50 of U2OS cells with PODXL overexpression or KD, and control cell lines. Three independent sets of experiments were performed with two biological replicates included in each experiment. Data are plotted as mean ± SEM. The *p*-values were calculated using Student's t-test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$).

(D) Apoptotic Index values of U2OS cells with PODXL KD. The assay was ended at 26 h after cisplatin treatment (30 μM). Two independent sets of experiments were performed with three biological replicates included in each experiment. Data are plotted as mean ± SEM. The *p*-values were calculated using Student's t-test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$).

(E) Similar as (D), for PODXL OE cells. The assay was ended at 40 h after cisplatin treatment (30 μM).

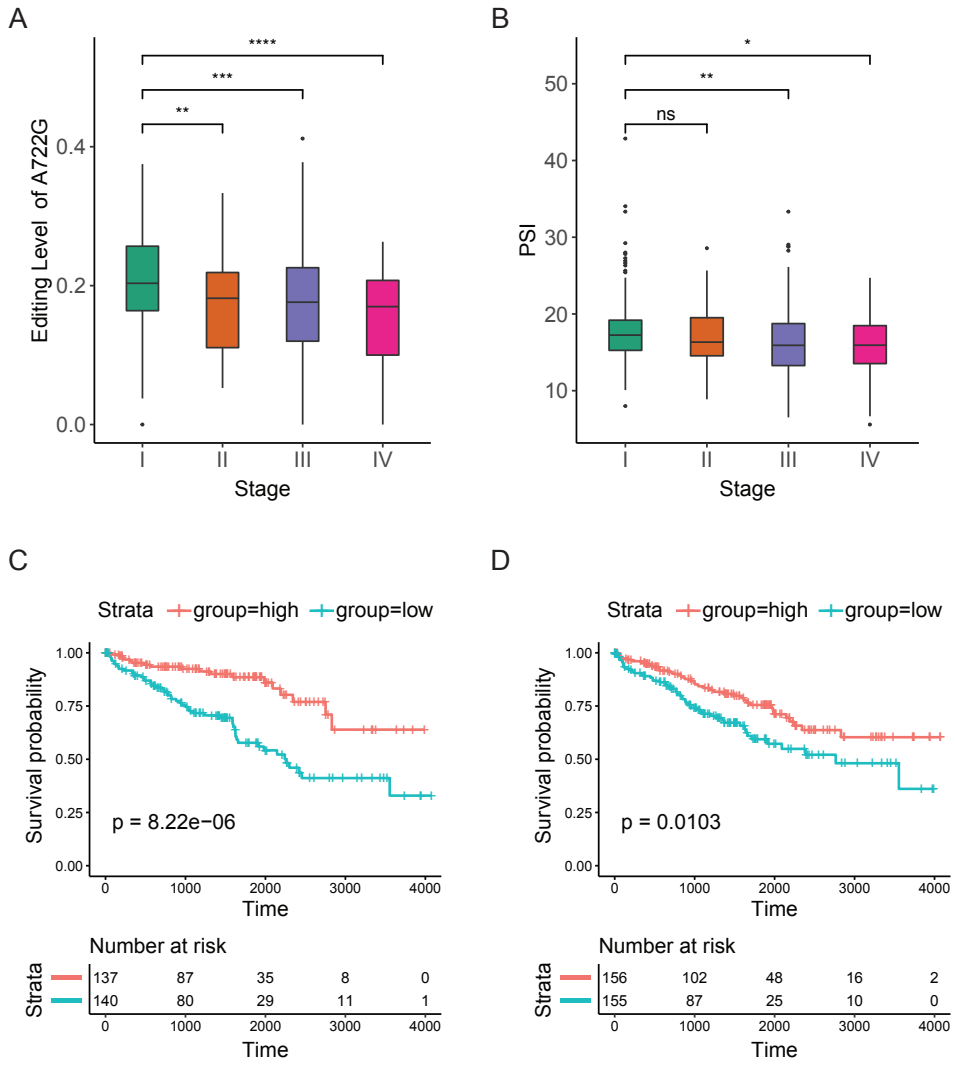


Figure 4.5 Clinical relevance of *PODXL* editing and splicing in KIRC

(A) Significant decrease in editing level of the A722G site over stage progression of KIRC. The *p*-values were calculated using Wilcoxon rank sum test (**p <= 0.01, ***p <= 0.001, ****p <= 0.0001).

(B) Significant decrease in *PODXL* alternative exon inclusion (measured by PSI) over stage progression of KIRC. The *p*-values were calculated using Wilcoxon rank sum test (*p <= 0.05, **p <= 0.01, ns p > 0.05).

(C) Lower editing levels of the A722G site associated with worse overall survival in KIRC. Patients were grouped into high (red) and low (blue) groups by editing level tertiles. The *p*-value was calculated by the log-rank test.

(D) Lower *PODXL* alternative exon inclusion associated with overall survival in KIRC. Patients were grouped into high (red) and low (blue) groups by PSI tertiles. The *p*-value was calculated by the log-rank test. See also Figure S6.

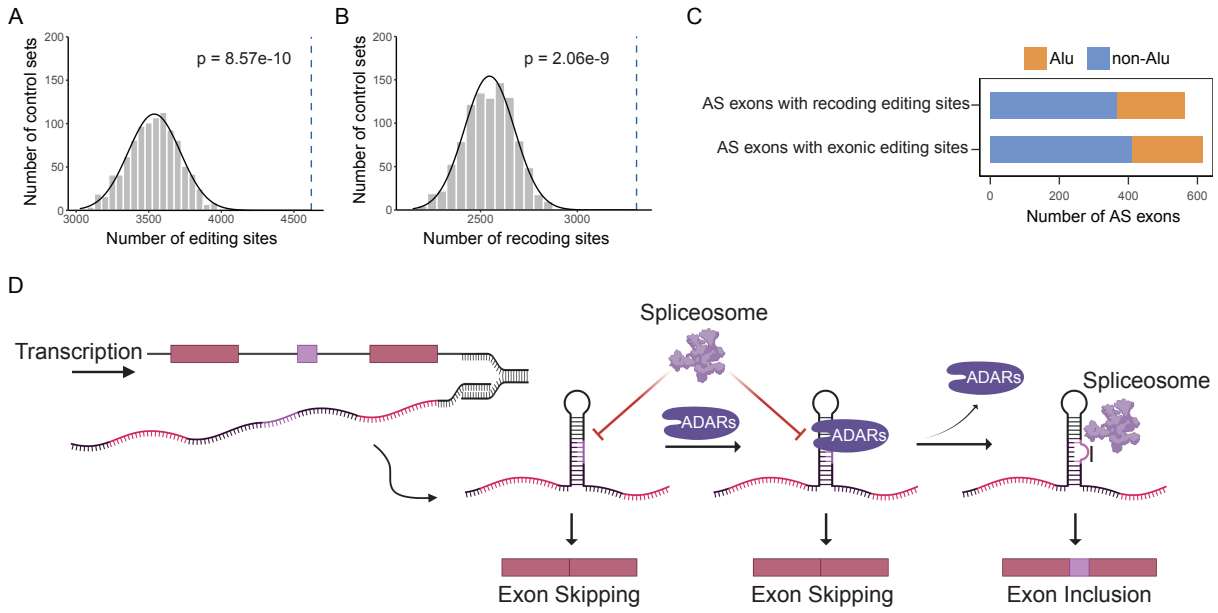


Figure 4.6 Exonic editing sites are enriched in alternative spliced exons

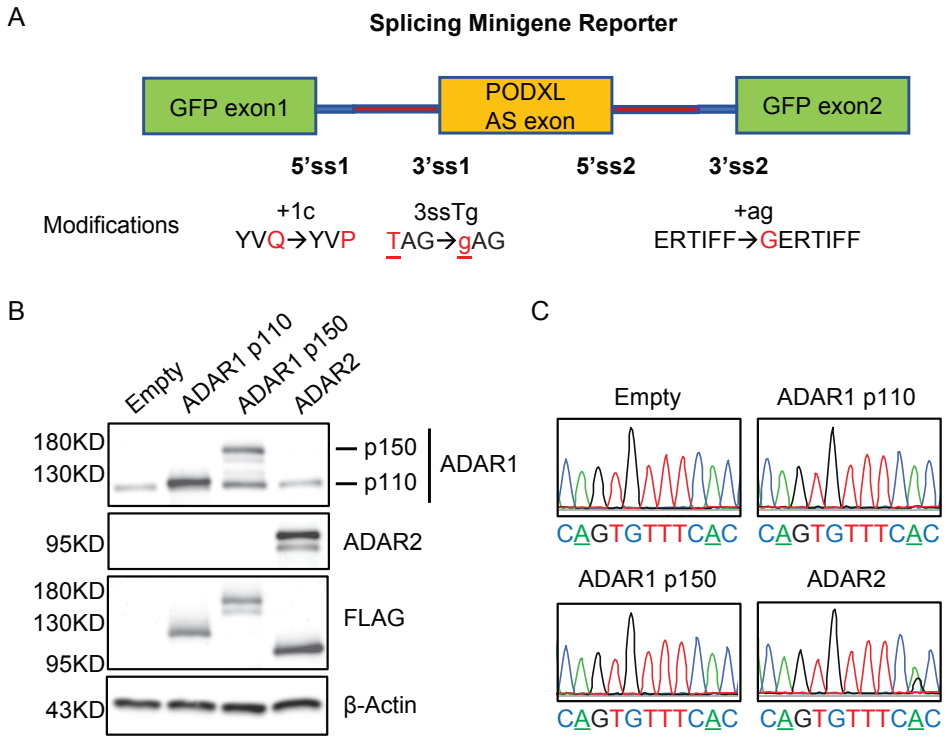
(A) Number of exonic editing sites from REDIportal overlapping alternative exons (blue dashed line), compared to the numbers of exonic editing sites in 1000 sets of random control exons (gray histogram). Black curve represents the normal distribution fit to the histogram. P value was calculated using the normal fit.

(B) Similar to A, but for recoding exonic editing sites only.

(C) Number of editing-containing alternative exons overlapping with Alu elements. See also Figure S5.

(D) The multi-facet model of the impact of ADAR/RNA editing on alternative splicing. In the pre-mRNA transcript, the alternative exon forms dsRNA structures with the flanking introns. ADARs bind to the dsRNA structure, which may compete with the spliceosome and prevent splicing. On the other hand, ADARs introduce RNA editing sites to the transcript that destabilize the dsRNA structure. The edited pre-mRNA thus allows access of the spliceosome to the splice sites of the alternative exon, promoting exon inclusion. The illustration was created using BioRender.com.

4.8 Supplementary Figures

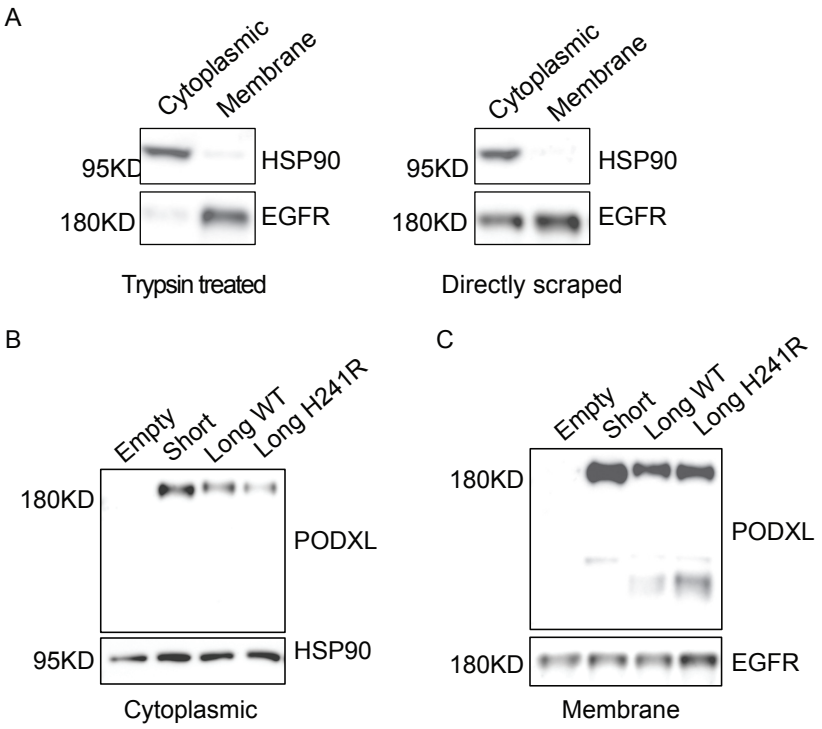


Supplementary Figure 4.1 Co-transfection of ADARs with *PODXL* splicing reporters in HeLa cells

(A) Illustration of different modifications (+1c, +ag, 3ssTg) made to the *PODXL* splicing reporters. This reporter contains two *GFP* split-exons that are upstream and downstream of the tested alternative exon. Two insertion modifications (+1c, +ag) were made to the splicing reporter to generate an in-frame transcript when *PODXL* alternative exon is included. The amino acid changes were indicated for each insertion modification. A T-to-G mutation was introduced to the 3' splice site of the *PODXL* alternative exon to increase the basal inclusion rate so that it approximately matches the endogenous exon inclusion level.

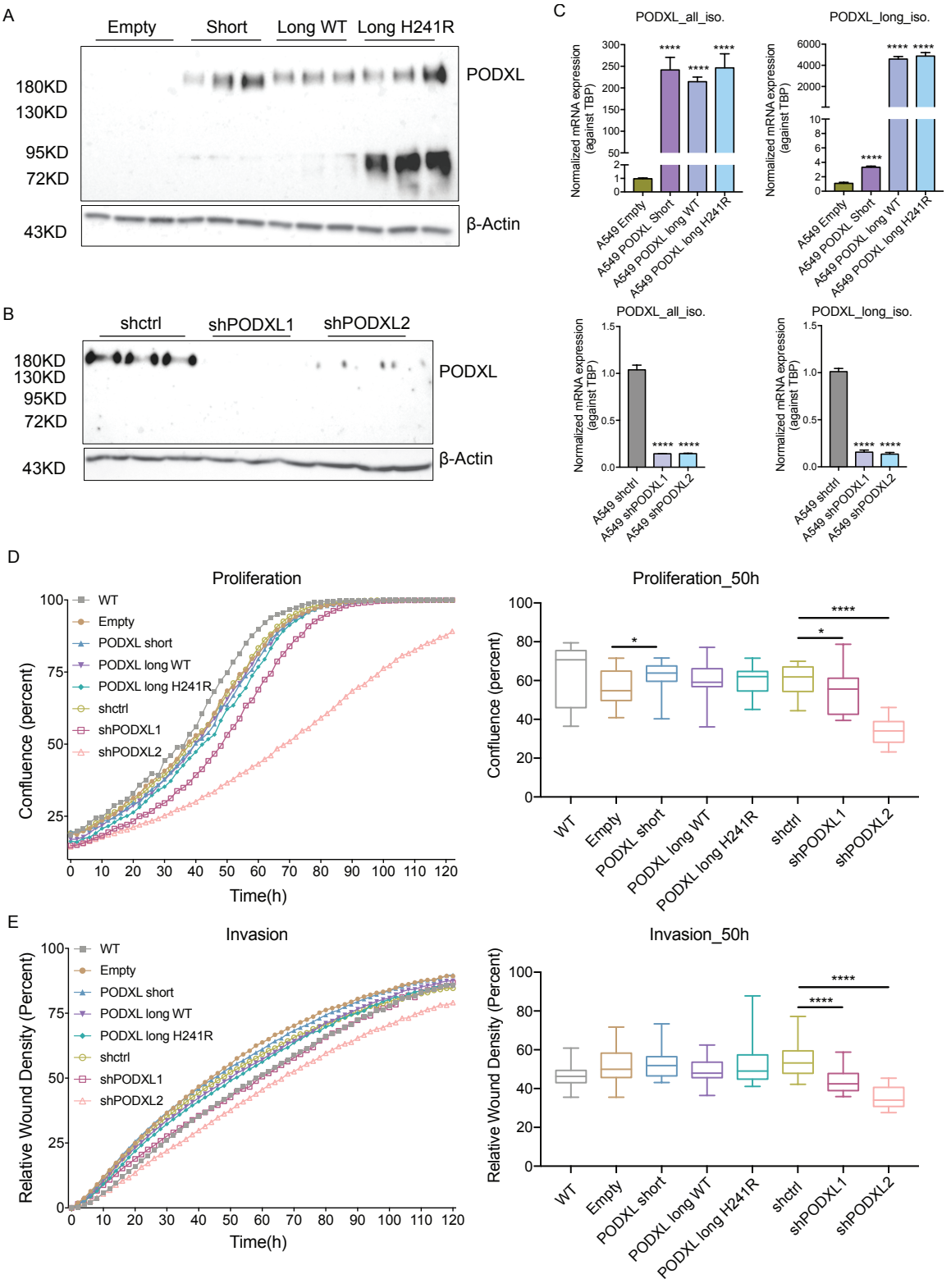
(B) Western blot showing the overexpression of ADARs in Hela cells. All ADARs are FLAG tagged. For ADAR1 p150 overexpression, the minor bands between p110 and p150 likely represent truncated proteins due to alternative translation initiation. For ADAR2 overexpression, the upper bands represent the FLAG-ADAR2 fusion proteins (see FLAG Western). The lower bands represent ADAR2 proteins without FLAG tagging, which may result from alternative translation start sites in the overexpression constructs.

(C) Sanger sequencing traces to detect the A714G and A722G editing sites (underlined As) on the reporters after co-transfection with the ADARs and the empty control in Hela cells.



Supplementary Figure 4.2 Cellular localizations of PODXL isoforms

(A) Western blot detecting marker genes for cytoplasmic (HSP90) and membrane (EGFR) fractions of wild type A549 cells. Cell fractionations were performed with or without trypsin digestion (see Methods). Trypsin treated: cells treated with trypsin before cell fractionation. Directly scraped: cells directly lysed and scraped from cell culture plates for cell fractionation. (B-C) Western blot detecting PODXL expression in the cytoplasmic (B) and membrane (C) fractions of A549 cells overexpressing different PODXL isoforms. Cells are directly scraped for cell fractionation.



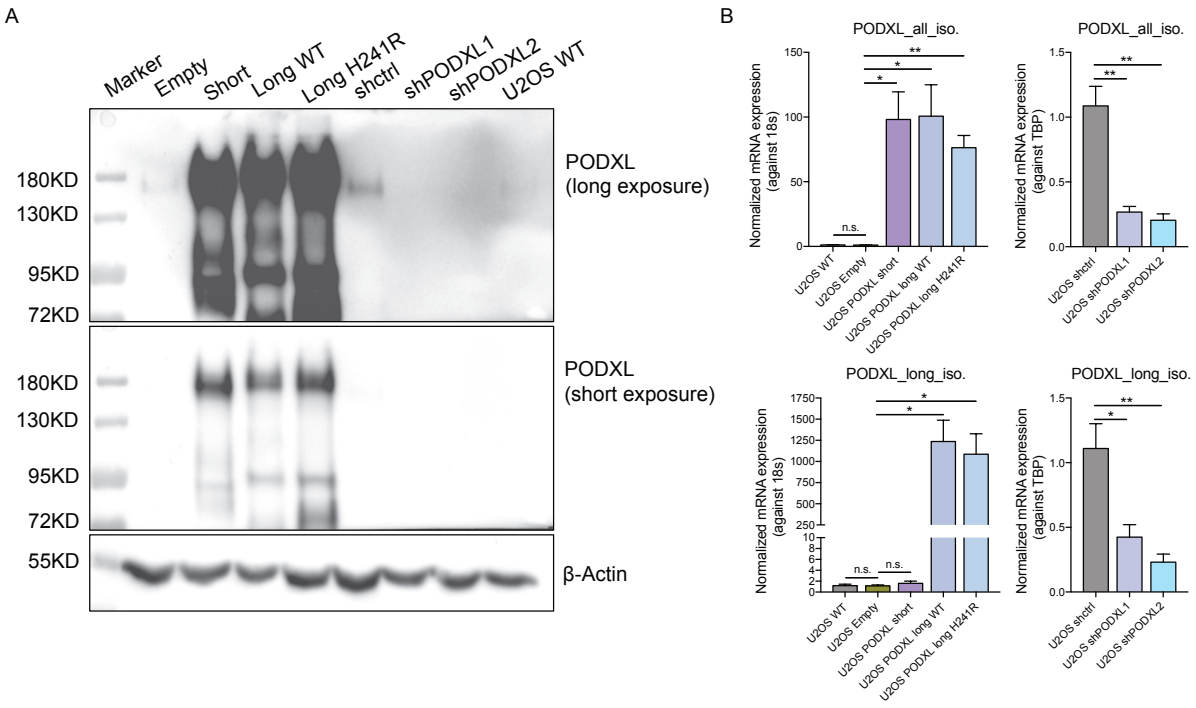
Supplementary Figure 4.3 Cell proliferation and invasion assay of A549 cells with PODXL overexpression and knockdown

(A-B) Western blot detecting PODXL overexpression (A) and knockdown (B) in A549 cells. Three biological replicates are shown.

(C) Normalized mRNA expression levels of all PODXL isoforms (PODXL_all_iso.) and the PODXL long isoform (PODXL_long_iso.) in A549 cells with PODXL overexpression or KD, and controls (WT, Empty, shctrl). Three biological replicates are included. Data are plotted as mean ± SEM. The *p*-values were calculated for each cell line compared to the corresponding controls (Empty or shctrl) using Student's t-test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$).

(D) Left: Cell proliferation curve of the A549 cells with PODXL overexpression or KD, and controls (WT, Empty, shctrl). The plot shows one set of experiment performed with three biological replicates. Right: Quantification of cell proliferation using cell confluence. Data at 50 h post wound creation are shown to examine the possible effect of cell proliferation on cell invasion shown in B. Two independent sets of experiments were performed with three biological replicates included in each experiment. Data are plotted as mean ± SEM. The *p*-values were calculated using Student's t-test (*$p < 0.05$, ***$p < 0.001$, ****$p < 0.0001$).

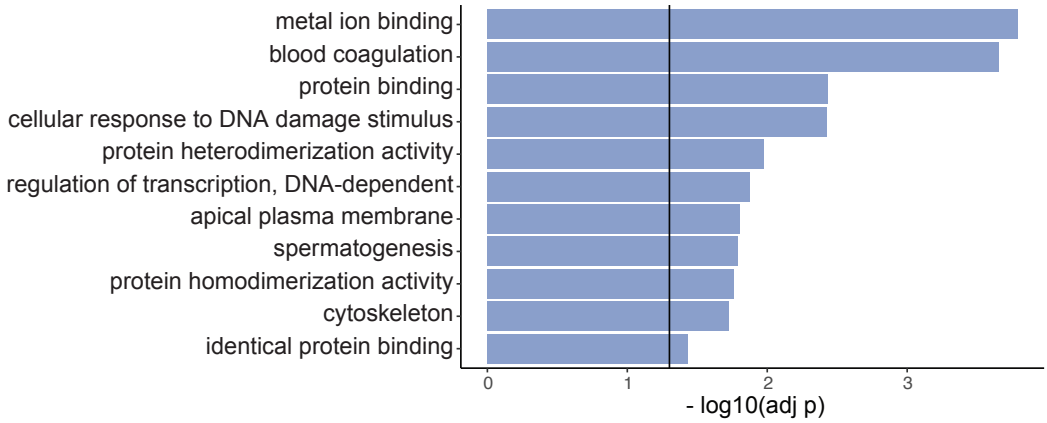
(E) Left: Cell invasion curve of the A549 cells with PODXL overexpression or KD, and controls (WT, Empty, shctrl). The plot shows one set of experiment performed with three biological replicates. Right: Quantification of cell invasion with relative wound density. Data at 50 h post wound creation are shown, when most cell lines reached around 50% relative wound density. Two independent sets of experiments were performed with three biological replicates included in each experiment. Data are plotted as mean ± SEM. The *p*-values were calculated using Student's t-test (*$p < 0.05$, ***$p < 0.001$, ****$p < 0.0001$).



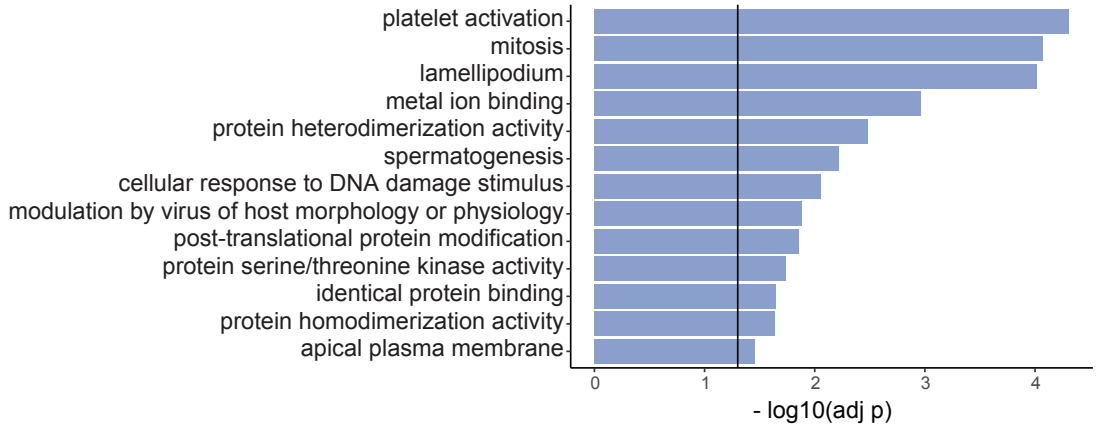
Supplementary Figure 4.4 PODXL overexpression and knockdown in U2OS cells

(A) Western blot detecting PODXL overexpression (A) and knockdown (B) in U2OS cells. (B) Normalized mRNA expression levels of all PODXL isoforms (PODXL_all_iso.) and the PODXL long isoforms (PODXL_long_iso.) in U2OS cells with PODXL overexpression or KD, and controls (WT, Empty, shctrl). Three biological replicates are included. Data are plotted as mean ± SEM. The *p*-values were calculated using Student's t-test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$, n.s., not significant).

A



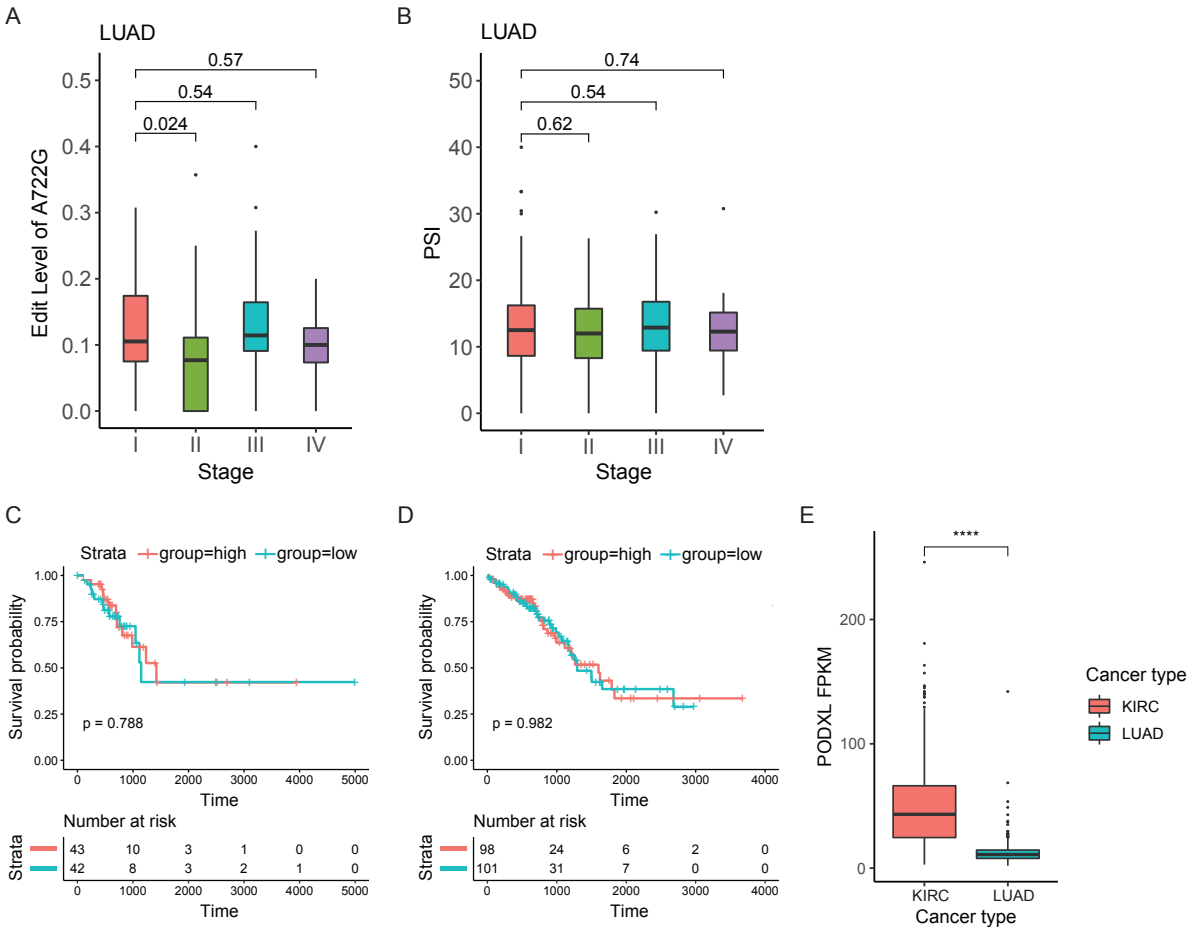
B



Supplementary Figure 4.5 Gene ontology terms enriched in the genes with alternative exons containing RNA editing sites

(A) Exon harboring recoding sites from REDIportal.

(B) Exons harboring any editing sites from REDIportal.



Supplementary Figure 4.6 Clinical relevance of *PODXL* editing and splicing in LUAD

(A) Editing level of the A722G site over stage progression of LUAD. The *p*-values were calculated using Wilcoxon rank sum test and annotated on the plot between each comparison. (B) *PODXL* alternative exon inclusion (measured by PSI) over stage progression of LUAD. The *p*-values were calculated using Wilcoxon rank sum test and annotated on the plot between each comparison.

(C) Overall survival of LUAD patients separated by editing levels of the A722G site. Patients were grouped into high (red) and low (blue) groups by editing level tertiles. The *p*-value was calculated by the log-rank test.

(D) Overall survival of LUAD patients separated by *PODXL* alternative exon inclusion. Patients were grouped into high (red) and low (blue) groups by PSI tertiles. The *p*-value was calculated by the log-rank test.

(E) PODXL expression level in primary tumors of KIRC and LUAD in TCGA. The *p*-values were calculated using Wilcoxon rank sum test (****$p <= 0.0001$).

4.9 Supplementary Tables

Primers used for PODXL overexpression constructs	
name	**sequences**
PODXL_kozac_AgeI_F	ctaccggtcgccaccATGCGCTGCGCGCTGGCGC
PODXL-EcoRI-R	tggcgaattcTTACTAGAGGTGTGTGTCTTC
PODXL-A722G-F	ACAGTGTTTCGCCATGTCAGCC
PODXL-A722G-R	GCTGACATGGCGAAACACTGTCTCTAGT
pLJM1-seq-R	gtggatctctgctgtccctg
Primers used for PODXL shRNA constructs	
name	**sequences**
PODXL_sh1_F (TRCN0000296029)	CCGGAGCCACGTAAGGGACTTTATACTCGAGTATAAAGTCCCTTACGTGGCTTTTTTG
PODXL_sh1_R (TRCN0000296029)	AATTCAAAAAAGCCACGTAAGGGACTTTATACTCGAGTATAAAGTCCCTTACGTGGCT
PODXL_sh2_F (TRCN0000310117)	CCGGACGAGCGGCTGAAGGACAAATCTCGAGATTTGTCCTTCAGCCGCTCGTTTTTTG
PODXL_sh2_R (TRCN0000310117)	AATTCAAAAAACGAGCGGCTGAAGGACAAATCTCGAGATTTGTCCTTCAGCCGCTCGT
Primers for endogenous PODXL isoform detection	
name	**sequences**
PODXL exonb F	CACTTCGACGCATCCTGTG
PODXL exond R	GTAGAGCTGGCTGGCATC
Primers for PODXL splicing minigene constructs	
name	**sequences**
pzw_AgeI_F	tccgctagcgctaccggtc
pzw_HindIII_R	CGCCTGGCaagctttTAAGAC
pzw_5ss1_+1c_F	cgaaggctacgtcccaggtaagtctcgaCGAAACaag
pzw_5ss1_+1c_R	cttGTTTCGtcgagacttacctgggacgtagccttcg
PODXL_HindIII_F	gagaagcttGCCAGGCGTGATGGCTCTG
PODXL_SacII_R	tatccgcggCCAGTGGAATAACCCGGCAAAG
PODXL_doubleA_F	AGAGACAGTGTTTCACCATGTCAGCC
pzw_podxl_3ss1_g9_65_DoubleA_R	CTGACATGGTGAAACACTGTCTCTcCTTGA
pzw_podxl_3ss1_g9_65_A714G_F	TTTCAAGgAGAGACGGTGTTTC

pzw_podxl_3ss1_g9_65_A714G_R	CTGACATGGTGAAACACCGTCTCTcCTTGA
PODXL-A722G-F	ACAGTGTTTCGCCATGTCAGCC
pzw_podxl_3ss1_g9_65_A722G_R	CTGACATGGCGAAACACTGTCTCTcCTTGA
PODXL_doubleG_F	AGAGACGGTGTTTCGCCATGTCAGCC
pzw_podxl_3ss1_g9_65_DoubleG_R	CTGACATGGCGAAACACCGTCTCTcCTTGA
PODXL Alu seqF	TAGCTGGGACTACAGGTGTG
PODXL Alu seqR	ACTTTGGGAGGCCAAGGTG
pzw_3ss2_+ag_SacII_F	TGGccgcggtctctttcttccaggagagcgcaccatcttcttc
pzw_BamHI_R	tccggtggatccttacttgtacagctcgtccatgc
Primers for PODXL isoform detection in splicing minigene	
name	**sequences**
Gexon F1 (gfp)	AGTGCTTCAGCCGCTACCC
Gexon Rv (gfp)	GTTGTACTCCAGCTTGTGCC
Primers for detecting PODXL isoforms via qPCR	
name	**sequences**
PODXL_longiso_qPCR_F	CACTTCGACGCATCCTGTG
PODXL_longiso_qPCR_R	ACTTTGGGAGGCCAAGGTG
PODXL_qPCR_both_F	TGCAGACACCACTACAGTTGC
PODXL_qPCR_both_R	ATGGTCATGTCCCGAGCTTG
18S_qPCR_F	CTCTTAGCTGAGTGTCCCGC
18S_qPCR_R	CTGATCGTCTTCGAACCTCC
TBP_qPCR_F	CAGCAACTTCCTCAATTCCTTG
TBP_qPCR_R	GCTGTTTAACTTCGCTTCCG
Primers for ADAR overexpression constructs	
name	**sequences**
Flag_Fw	CATCGACTACAAGGATGACG
p110 NotI F	AAGGAAAAAAGCGGCCGCAAGCCGAGATCAAGGAGAAAATCTG
ADAR1_BstBI_stop_R	atactgttcgaaCTATACTGGGCAGAGATAAAAGTTCTTTTCCTC
ADAR2_XbaI_R	CCCTCTAGACCGGGCG
ADAR2_EAA1_F	GGCTCTGGTCCCACAGAGGCAAAGGCAGCACTCCATGCTGCTGAGAAGG
ADAR2_EAA1_R	CCTTCTCAGCAGCATGGAGTGCTGCCTTTGCCTCTGTGGGACCAGAGCC
ADAR2_EAA2_F	GGCTCGGGGAGAAACGAGGCGCTTGCCGCGGCCCGGGCTGCGC

ADAR2_EAA2_R	GCGCAGCCCGGGCCGCGGCAAGCGCCTCGTTTCTCCCCGAGCC
ADAR2_E396A_F	CATTAAATGACTGCCATGCAGCAATAATATCTCGGAGATCCTT
ADAR2_E396A_R	AAGGATCTCCGAGATATTATTGCTGCATGGCAGTCATTTAATG
ADAR2_E488Q_F	GACCAAAATAGAGTCTGGTCAGGGGACGATTCCAGTGCG
ADAR2_E488Q_R	CGCACTGGAATCGTCCCCTGACCAGACTCTATTTTGGTC
Primers for PODXL minigene editing detection	
name	**sequences**
EGFP_SacI_F	GCGAGGAGCTCTTCACCGGGG
PODXL_EGFP_R	tggtgcgctcCTGTAATCCCAG

Supplementary Table 4.1 Oligonucleotides used in Chapter 4

CHAPTER 5

Concluding Remarks

A growing number of RNA single-nucleotide variants (SNVs), either derived from genetic variants or RNA editing, has been associated with complex human traits. Despite a significant progress in identifying SNVs of interest, functional studies connecting SNVs to phenotypes are lagging. In this work, we developed a massively parallel reporter assay (MPRA), enabling the high-throughput screen of rare 3' UTR variants regulating mRNA abundance. We also examined the function of differentially edited sites in cancer, revealing an editing-dependent stabilization mechanism. Further, we characterized two exonic RNA editing sites in *PODXL*, and demonstrated a previously underestimated role of exonic RNA editing in regulating alternative splicing.

In Chapter 2, we tested 14,575 rare variants using the MPRA platform and identified 5,437 functional 3' UTR variants regulating mRNA abundance. We showed that many functional variants had a close relevance to human diseases, especially cancer. Further, we uncovered 181 functional variants in cancer driver genes and 37 functional variants present in TCGA gene expression outliers, which we nominated as causal variants in cancer. Specifically, we characterized three variants in cancer-associated genes, revealing their functional roles in regulation mRNA stability and cell proliferation in their native genomic context.

Our massively parallel screen uncovered many functional 3' UTR variants that regulate mRNA abundance post-transcriptionally, i.e., mRNA stability. As shown in the control experiment on five well-known destabilizing motifs, our assay was able to capture the stabilizing

effects of random mutations, which disrupted the motifs. Yet, the assay does have some limitations. For example, other regulatory mechanisms, such as transcription, alternative polyadenylation, and RNA localization, may still play a role and affect the final readout of this assay. This ambiguity in the readout may impact studies of downstream mechanisms, for example, motif enrichment analysis. To elucidate the mechanism, it will be beneficial to link each functional variant to its potential *trans*-factors. However, our knowledge of *trans*-factor binding has limited specificity, often with multiple *trans*-factors predicted for a single variant, making it hard to pinpoint the exact *trans*-factor and the underlying mechanisms. Future efforts in mapping the *cis*-regulatory elements to *trans*-acting factors will be helpful to solve this problem.

Like many MPRAs, our method suffers from the intrinsic limitations of episomal MPRAs: 1) a limited length of test sequences (200nt in our assay) unable to capture functional variants with structural or distal effects; 2) a lack of genomic context; 3) transient transfection artifacts (e.g., elevated immune responses[305]). Improvement in DNA oligo synthesis, combined with genome-integrated reporter assays may help to overcome some of these limitations. Yet, genome-integrated methods are still limited in their flexibility to test different cell lines[41]. Alternatively, CRISPR perturbation screens combined with the scRNA-seq readout would be ideal to measure the functional effects of SNVs. Currently, it remains costly to do such screens, partly due to the high cost of single-cell sequencing[46]. Besides, although great advances have been made in the genome editing field, base editing still has limited efficacy. In addition, some genomic regions, such as highly repetitive regions, or sequences far away from the protospacer adjacent motif (PAM) sequences, are hard to modify. Future development of genome editing technology will likely increase the usability of CRISPR screens.

In Chapter 3, we reported the global dysregulation of RNA editing between epithelial and mesenchymal tumors across seven cancer types in TCGA. A similar alteration of RNA editing was also observed in scRNA-seq of lung cancer, with the majority of differential editing sites detected in cancer cells. Importantly, knockdown of ADAR1 or ADAR2 induced EMT in human cell lines. Supported by correlation analysis and experimental validations, we showed that 3' UTR editing sites differential between epithelial and mesenchymal tumors regulate mRNA abundance. Further, we uncovered an RBP, ILF3, which binds in proximity to many differential editing sites correlated with gene expression, accounting for editing-mediated stabilization of PKR, a key player in innate immune responses.

Our analysis of RNA editing in epithelial and mesenchymal tumors and ADAR KD experiments showed a strong relevance of RNA editing to EMT. However, it remains unknown if the observed RNA editing alteration is a cause or a consequence of EMT. While both ADAR1 and ADAR2 knockdown induced EMT, we only observed widespread downregulation of RNA editing in ADAR1 KD cells, but not ADAR2 KD, indicating that a large fraction of altered RNA editing is not necessary for EMT induction. Indeed, RNA editing alteration in a specific target was reported to induce EMT[192]. Previous studies observed that ADAR2 editing is important in regulating the steady-state levels of miR-200, the loss of which induced EMT in colorectal cancer[192]. Another interesting target of ADAR2 editing is *SLC22A3*, a suppressor for EMT processes, whose editing led to reduced gene expression, thus promoting tumor invasion and metastasis in esophageal cancer[183]. Nonetheless, the association between altered RNA editing profiles and EMT makes these editing sites valuable as potential biomarkers for EMT. Moreover, even if they do not directly contribute to EMT, they may have other functional roles in cancer. Further characterization of these differential RNA editing sites, such as via MPRAs on mRNA abundance and splicing, will be helpful to elucidate their roles in cancer.

Our study of RNA editing in EMT also raised many other interesting questions. First, we showed that ADAR1 and ADAR2 KD induced EMT in A549 and MCF10A cells. Yet, the underlying mechanisms remain undetermined. Both ADAR1 and ADAR2 have RNA binding domains and RNA deaminase domains responsible for RNA binding and catalysis of RNA editing, respectively[12]. Further experiments using mutant ADARs will be helpful to examine whether these proteins regulate EMT through RNA binding or RNA editing (or both). Another interesting question is how ILF3 KD induced EMT in A549 cells. We observed that ILF3 potentially regulates editing-dependent stabilization of immune-relevant genes. Yet, this mechanism alone is unlikely to explain EMT changes upon ILF3 KD. Instead of detecting a downregulation of immune-relevant genes after ILF3 KD, we observed an induction of many ISGs (data not shown), including PKR (which we validated for the editing-dependent stabilization model). The above observations may have resulted from the diverse, multifaceted function of ILF3. In addition to regulating mRNA stability, ILF3 also plays important roles in modulating transcription, mRNA localization, translation, and miRNA biogenesis[217]. Interestingly, an antisense long non-coding RNA of *ILF3*, ILF3 divergent transcript (*ILF3-AS1*), is reported to promote EMT in hepatocellular carcinoma[306] but inhibit EMT in cervical cancer[307]. Both studies reported the functional roles of *ILF3-AS1* in regulating miRNAs[306,307]. Further studies on ILF3, especially its relationship with *ILF3-AS1* and miRNAs, will be helpful to understand the mechanism of ILF3-mediated EMT in A549 cells.

In Chapter 4, we discovered an unexpected role of exonic RNA editing in promoting alternative splicing of *PODXL*, resulting in three PODXL isoforms functionally distinct in protease digestion patterns, cell migration, and cisplatin chemoresistance. Consistent with our cell-based studies, we found that lower editing levels of the *PODXL* recoding site and lower inclusion of the *PODXL* alterative exon correlated with worse overall survival in kidney cancer (KIRC). Further, we showed that, in general, exonic RNA editing sites were enriched in

alternatively spliced exons, indicating a potentially prevalent role of exonic RNA editing in regulating alternative splicing.

Our findings of different protease digestion patterns among PODXL isoforms motivate further structural studies on the PODXL extracellular domain, which contains the *PODXL* recoding site. In our proposed model, the recoding event on PODXL may alter its protein conformation, which is more prone to protease digestion. Currently the protein structure of PODXL is not solved, making it hard to model structural alterations upon the amino acid change caused by RNA editing. We attempted to predict the protein structure of PODXL using AlphaFold[308]. A small disruption of the alpha helix, formed by the *PODXL* alternative exon, was observed for the edited PODXL isoform (data not shown). Future experimental studies are needed to assess the protein conformational change caused by the recoding event in PODXL, and its impact on the interaction of PODXL with other binding factors, either extracellularly or intracellularly.

Chapter 4 also revealed a more generalized role of exonic RNA editing in regulating alternative splicing, which will need further investigation. Direct impact of exonic RNA editing on alternative splicing should be addressed at a larger scale, potentially using MPRAs. Yet, as mentioned above, MPRAs are limited in the length of the tested sequences, which is a bottleneck in this effort, since most RNA editing sites are associated with double-stranded RNA (dsRNA) structures formed by long sequences. Alternatively, high-content CRISPR screens can be used to avoid the limitations of MPRAs. However, most editing sites are located in repetitive regions, such as *Alus*, which are hard to target specifically using CRISPR. Future efforts should be dedicated to study SNVs located in these long and repetitive regions.

In summary, our studies enabled improved understanding of the functional roles of SNVs, both genetic variants and RNA editing sites, in post-transcriptional regulation. The MPRA platform established in this work can be applied to decipher the functions of many SNVs in different cellular contexts. The insights generated in our work builds a foundation for future functional and translational discoveries.

References

1. Acuna-Hidalgo, R., Veltman, J. A. & Hoischen, A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol. 2016 171* **17**, 1–19 (2016).
2. Goldmann, J. M. *et al.* Parent-of-origin-specific signatures of de novo mutations. *Nat. Genet. 2016 488* **48**, 935–939 (2016).
3. Auton, A. *et al.* A global reference for human genetic variation. *Nat. 2015 5267571* **526**, 68–74 (2015).
4. Sherry, S. T., Ward, M. & Sirotkin, K. dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. *Genome Res.* **9**, 677–679 (1999).
5. Eisenberg, E. *et al.* Identification of RNA editing sites in the SNP database. *Nucleic Acids Res.* **33**, 4612–4617 (2005).
6. Picardi, E. & Pesole, G. REDItools: high-throughput RNA editing detection made easy. *Bioinformatics* **29**, 1813–1814 (2013).
7. Ramaswami, G. *et al.* Identifying RNA editing sites using RNA sequencing data alone. *Nat. Methods 2013 102* **10**, 128–132 (2013).
8. Zhu, S., Xiang, J. F., Chen, T., Chen, L. L. & Yang, L. Prediction of constitutive A-to-I editing sites from human transcriptomes in the absence of genomic sequences. *BMC Genomics* **14**, 1–16 (2013).
9. Zhang, Q. & Xiao, X. Genome sequence–independent identification of RNA editing sites. *Nat. Methods 2015 124* **12**, 347–350 (2015).
10. Nishikura, K. Functions and Regulation of RNA Editing by ADAR Deaminases. *Annu. Rev.*

Biochem. **79**, 321–349 (2010).

11. Kim, D. D. Y. *et al.* Widespread RNA Editing of Embedded Alu Elements in the Human Transcriptome. *Genome Res.* **14**, 1719–1725 (2004).
12. Bass, B. L. RNA Editing by Adenosine Deaminases That Act on RNA. *Annu. Rev. Biochem.* **71**, 817–846 (2002).
13. Bazak, L. *et al.* A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res.* **24**, 365–76 (2014).
14. Picardi, E., D'Erchia, A. M., Lo Giudice, C. & Pesole, G. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res.* **45**, D750–D757 (2017).
15. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179 (2020).
16. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–D798 (2015).
17. Bush, W. S. & Moore, J. H. Chapter 11: Genome-Wide Association Studies. *PLOS Comput. Biol.* **8**, e1002822 (2012).
18. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
19. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491 (2018).
20. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. *Hum. Mol. Genet.*

24, R111–R119 (2015).

21. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9362–9367 (2009).

22. Shameer, K., Tripathi, L. P., Kalari, K. R., Dudley, J. T. & Sowdhamini, R. Interpreting functional effects of coding variants: challenges in proteome-scale prediction, annotation and assessment. *Brief. Bioinform.* **17**, 841–862 (2016).

23. Anna, A. & Monika, G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J. Appl. Genet.* **59**, 253 (2018).

24. Steri1, M. M. L. I. | M. B. W. | V. O. Genetic variants in mRNA untranslated regions. *WIREs RNA* (2017).

25. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet. 2013 456* **45**, 580–585 (2013).

26. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).

27. Kang, E. Y. *et al.* Discovering single nucleotide polymorphisms regulating human gene expression using allele specific expression from RNA-seq data. *Genetics* **204**, 1057–1064 (2016).

28. Li, G. *et al.* Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Res.* **40**, e104–e104 (2012).

29. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods 2015 1210* **12**, 931–934 (2015).

30. Xiong, H. Y. *et al.* The human splicing code reveals new insights into the genetic determinants of disease. *Science (80-.).* **347**, (2015).

31. Ramaswami, G. *et al.* Genetic mapping uncovers cis-regulatory landscape of RNA editing. *Nat. Commun. 2015 61* **6**, 1–9 (2015).

32. Mo, X. B., Zhang, Y. H. & Lei, S. F. Genome-wide identification of m6A-associated SNPs as potential functional variants for bone mineral density. *Osteoporos. Int.* **29**, 2029–2039 (2018).

33. Yekta, S., Shih, I. H. & Bartel, D. P. MicroRNA-Directed Cleavage of HOXB8 mRNA. *Science (80-.).* **304**, 594–596 (2004).

34. Sheives, P. & Lynch, K. W. Identification of cells deficient in signaling-induced alternative splicing by use of somatic cell genetics. *RNA* **8**, 1473 (2002).

35. Wang, Z. *et al.* Systematic Identification and Analysis of Exonic Splicing Silencers. *Cell* **119**, 831–845 (2004).

36. Ipe, J. *et al.* PASSPORT-seq: A novel high-throughput bioassay to functionally test polymorphisms in micro-RNA target sites. *Front. Genet.* **9**, 1–10 (2018).

37. Yang, E.-W. *et al.* Allele-specific binding of RNA-binding proteins reveals functional genetic variants in the RNA. *Nat. Commun.* **10**, 1338 (2019).

38. Gasperini, M., Starita, L. & Shendure, J. The power of multiplexed functional analysis of genetic variants. **11**, (2016).

39. Oikonomou, P., Goodarzi, H. & Tavazoie, S. Systematic identification of regulatory elements in conserved 3’ UTRs of human transcripts. *Cell Rep.* **7**, 281–292 (2014).

40. Vainberg Slutskin, I., Weingarten-Gabbay, S., Nir, R., Weinberger, A. & Segal, E.

Unraveling the determinants of microRNA mediated regulation using a massively parallel reporter assay. *Nat. Commun.* **9**, (2018).

41. Cheung, R. *et al.* A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions. *Mol. Cell* **73**, 183-194.e8 (2019).

42. Zhao, W. *et al.* Massively parallel functional annotation of 3' untranslated regions. *Nat. Biotechnol.* **32**, 387–391 (2014).

43. Rabani, M., Pieper, L., Chew, G. L. & Schier, A. F. Erratum: A Massively Parallel Reporter Assay of 3' UTR Sequences Identifies In Vivo Rules for mRNA Degradation (Molecular Cell (2017) 68(6) (1083–1094.e5) (S1097276517308730) (10.1016/j.molcel.2017.11.014)). *Mol. Cell* **70**, 565 (2018).

44. Adamson, S. I., Zhan, L. & Graveley, B. R. Vex-seq: High-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. *Genome Biol.* **19**, 1–12 (2018).

45. Lubelsky, Y. & Ulitsky, I. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature* **555**, 107–111 (2018).

46. Bock, C. *et al.* High-content CRISPR screening. *Nat. Rev. Methods Prim. 2022 21* **2**, 1–23 (2022).

47. Kreimer, A. *et al.* Predicting gene expression in massively parallel reporter assays: A comparative study. *Hum. Mutat.* **38**, 1240–1250 (2017).

48. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–

D894 (2019).

49. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nat. 2017 5507675* **550**, 239–243 (2017).

50. Hsiao, Y.-H. E. *et al.* RNA editing in nascent RNA affects pre-mRNA splicing. *Genome Res.* **28**, 812–823 (2018).

51. Ferris, Z. E., Li, Q. & Germann, M. W. Substituting Inosine for Guanosine in DNA: Structural and Dynamic Consequences: *https://doi.org/10.1177/1934578X19850032* **14**, 1–7 (2019).

52. Higuchi, M. *et al.* RNA editing of AMPA receptor subunit GluR-B: A base-paired intron-exon structure determines position and efficiency. *Cell* **75**, 1361–1370 (1993).

53. Brusa, R. *et al.* Early-Onset Epilepsy and Postnatal Lethality Associated with an Editing-Deficient GluR-B Allele in Mice. *Science (80-.).* **270**, 1677–1680 (1995).

54. Jain, M. *et al.* RNA editing of Filamin A pre-mRNA regulates vascular contraction and diastolic blood pressure. *EMBO J.* **37**, e94813 (2018).

55. Picardi, E. *et al.* Profiling RNA editing in human tissues: towards the inosinome Atlas. *Sci. Reports 2015 51* **5**, 1–17 (2015).

56. Lev-Maor, G. *et al.* RNA-editing-mediated exon evolution. *Genome Biol.* **8**, 1–12 (2007).

57. Kawahara, Y. *et al.* Frequency and fate of microRNA editing in human brain. *Nucleic Acids Res.* **36**, 5270 (2008).

58. Brümmer, A., Yang, Y., Chan, T. W. & Xiao, X. Structure-mediated modulation of mRNA abundance by A-to-I editing. *Nat. Commun.* **8**, 1–12 (2017).

59. Reich, D. P. & Bass, B. L. Mapping the dsRNA World. *Cold Spring Harb. Perspect. Biol.* **11**,

a035352 (2019).

60. Liddicoat, B. J. *et al.* RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. *Science (80-.).* **349**, 1115–1120 (2015).

61. Chung, H. *et al.* Human ADAR1 Prevents Endogenous RNA from Triggering Translational Shutdown. *Cell* **172**, 811-824.e14 (2018).

62. Chen, Y. G. & Hur, S. Cellular origins of dsRNA, their recognition and consequences. *Nat. Rev. Mol. Cell Biol. 2021 234* **23**, 286–301 (2021).

63. Williams, B. R. G. PKR; a sentinel kinase for cellular stress. *Oncogene* **18**, 6112–6120 (1999).

64. Jain, M., Jantsch, M. F. & Licht, K. The Editor's I on Disease Development. *Trends Genet.* **35**, 903–913 (2019).

65. Hartner, J. C., Walkley, C. R., Lu, J. & Orkin, S. H. ADAR1 is essential for the maintenance of hematopoiesis and suppression of interferon signaling. *Nat. Immunol.* **10**, 109–115 (2009).

66. Ben-Shoshan, S. O. *et al.* ADAR1 deletion induces NFκB and interferon signaling dependent liver inflammation and fibrosis. *RNA Biol.* **14**, 587–602 (2017).

67. Wang, H. *et al.* ADAR1 Suppresses the Activation of Cytosolic RNA-Sensing Signaling Pathways to Protect the Liver from Ischemia/Reperfusion Injury. *Sci. Rep.* (2016). doi:10.1038/srep20248

68. Wang, Q. *et al.* Stress-induced Apoptosis Associated with Null Mutation of ADAR1 RNA Editing Deaminase Gene. *J. Biol. Chem.* (2004). doi:10.1074/jbc.M310162200

69. Hartner, J. C. *et al.* Liver Disintegration in the Mouse Embryo Caused by Deficiency in the

RNA-editing Enzyme ADAR1. *J. Biol. Chem.* (2004). doi:10.1074/jbc.M311347200

70. Ishizuka, J. J. *et al.* Loss of ADAR1 in tumours overcomes resistance to immune checkpoint blockade. *Nature* 1 (2018). doi:10.1038/s41586-018-0768-9

71. Eisenberg, E. & Levanon, E. Y. A-to-I RNA editing — immune protector and transcriptome diversifier. *Nat. Rev. Genet.* **19**, 473–490 (2018).

72. Tran, S. S. *et al.* Statistical inference of differential RNA-editing sites from RNA-sequencing data by hierarchical modeling. *Bioinformatics* **36**, 2796–2804 (2020).

73. Khermesh, K. *et al.* Reduced levels of protein recoding by A-to-I RNA editing in Alzheimer's disease. *RNA* **22**, 290 (2016).

74. Tran, S. S. *et al.* Widespread RNA editing dysregulation in brains from autistic individuals. *Nat. Neurosci.* **22**, 25–36 (2019).

75. Breen, M. S. *et al.* Global landscape and genetic regulation of RNA editing in cortical samples from individuals with schizophrenia. *Nat. Neurosci. 2019 229* **22**, 1402–1412 (2019).

76. Ma, Y. *et al.* Atlas of RNA editing events affecting protein expression in aged and Alzheimer's disease human brain tissue. *Nat. Commun. 2021 121* **12**, 1–16 (2021).

77. Han, L. *et al.* The Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in Human Cancers. *Cancer Cell* **28**, 515–528 (2015).

78. Paz-Yaacov, N. *et al.* Elevated RNA Editing Activity Is a Major Contributor to Transcriptomic Diversity in Tumors. *Cell Rep.* **13**, 267–276 (2015).

79. Chan, T. H. M. *et al.* ADAR-Mediated RNA Editing Predicts Progression and Prognosis of Gastric Cancer. *Gastroenterology* **151**, 637-650.e10 (2016).

80. Amin, E. M. *et al.* The RNA-editing enzyme ADAR promotes lung adenocarcinoma migration and invasion by stabilizing FAK. *Sci. Signal.* **10**, (2017).

81. Chen, Y.-B. *et al.* ADAR2 functions as a tumor suppressor via editing IGFBP7 in esophageal squamous cell carcinoma. *Int. J. Oncol.* **50**, 622–630 (2017).

82. Chen, L. *et al.* Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma. *Nat. Med.* **19**, 209–216 (2013).

83. Takeda, S. *et al.* Activation of AZIN1 RNA editing is a novel mechanism that promotes invasive potential of cancer-associated fibroblasts in colorectal cancer. *Cancer Lett.* **444**, 127–135 (2019).

84. Qin, Y. R. *et al.* Adenosine-to-inosine RNA editing mediated by ADARs in esophageal squamous cell carcinoma. *Cancer Res.* **74**, 840–851 (2014).

85. Chan, T. H. M. *et al.* A disrupted RNA editing balance mediated by ADARs (Adenosine DeAminases that act on RNA) in human hepatocellular carcinoma. *Gut* **63**, 832–843 (2014).

86. Christofi, T. & Zaravinos, A. RNA editing in the forefront of epitranscriptomics and human health. *J. Transl. Med. 2019 171* **17**, 1–15 (2019).

87. Peng, X. *et al.* A-to-I RNA Editing Contributes to Proteomic Diversity in Cancer. *Cancer Cell* **33**, 1–12 (2018).

88. Zhang, M. *et al.* RNA editing derived epitopes function as cancer antigens to elicit immune responses. *Nat. Commun.* **9**, 3919 (2018).

89. Kung, C.-P., Maggi, L. B. & Weber, J. D. The Role of RNA Editing in Cancer Development and Metabolic Disorders. *Front. Endocrinol. (Lausanne).* **9**, 762 (2018).

90. Kawahara, Y. *et al.* RNA editing and death of motor neurons. *Nat. 2004 4276977* **427**, 801–801 (2004).

91. Silberberg, G., Lundin, D., Navon, R. & Öhman, M. Deregulation of the A-to-I RNA editing mechanism in psychiatric disorders. *Hum. Mol. Genet.* **21**, 311–321 (2012).

92. Srivastava, P. K. *et al.* Genome-wide analysis of differential RNA editing in epilepsy. *Genome Res.* **27**, 440–450 (2017).

93. Stellos, K. *et al.* Adenosine-to-inosine RNA editing controls cathepsin S expression in atherosclerosis by enabling HuR-mediated post-transcriptional regulation. *Nat. Med.* **22**, (2016).

94. Gal-Mark, N. *et al.* Abnormalities in A-to-I RNA editing patterns in CNS injuries correlate with dynamic changes in cell type composition. *Sci. Reports 2017 71* **7**, 1–12 (2017).

95. Filippini, A. *et al.* Absence of the Fragile X Mental Retardation Protein results in defects of RNA editing of neuronal mRNAs in mouse. *RNA Biol.* **14**, 1580–1591 (2017).

96. Mulvey, B., Lagunas, T. & Dougherty, J. D. Massively Parallel Reporter Assays: Defining Functional Psychiatric Genetic Variants Across Biological Contexts. *Biol. Psychiatry* **89**, 76–89 (2021).

97. Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519–1529 (2016).

98. Arnold, C. D. *et al.* Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science (80-.).* **339**, 1074–1077 (2013).

99. Klein, J. C. *et al.* Functional testing of thousands of osteoarthritis-associated variants for regulatory activity. *Nat. Commun.* **10**, 1–9 (2019).

100. Myint, L. *et al.* A screen of 1,049 schizophrenia and 30 Alzheimer's-associated variants for regulatory potential. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **183**, 61 (2020).

101. Choi, J. *et al.* Massively parallel reporter assays of melanoma risk variants identify MX2 as a gene promoting melanoma. *Nat. Commun.* **11**, 1–16 (2020).

102. Griesemer, D. *et al.* Genome-wide functional screen of 3'UTR variants uncovers causal variants for human disease and evolution. *Cell* **184**, 5247-5260.e19 (2021).

103. Goswami, C., Chattopadhyay, A. & Chuang, E. Y. Rare variants: data types and analysis strategies. *Ann. Transl. Med.* **9**, 961–961 (2021).

104. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet. 2018 505* **50**, 746–753 (2018).

105. Schoft, V. K., Schopoff, S. & Jantsch, M. F. Regulation of glutamate receptor B pre-mRNA splicing by RNA editing. *Nucleic Acids Res.* **35**, 3723 (2007).

106. Moore, M. J. From birth to death: The complex lives of eukaryotic mRNAs. *Science (80-.).* **309**, 1514–1518 (2005).

107. Matoulkova, E., Michalova, E., Vojtesek, B. & Hrstka, R. The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells. *RNA Biol.* **9**, 563–576 (2012).

108. Mazumder, B., Seshadri, V. & Fox, P. L. Translational control by the 3'-UTR: the ends specify the means. *Trends Biochem. Sci.* **28**, 91–98 (2003).

109. Mayr, C. What Are 3' UTRs Doing? *Cold Spring Harb. Perspect. Biol.* a034728 (2018). doi:10.1101/cshperspect.a034728

110. Theil, K., Herzog, M. & Rajewsky, N. Post-transcriptional Regulation by 3' UTRs Can Be

Masked by Regulatory Elements in 5′ UTRs. *Cell Rep.* **22**, 3217–3226 (2018).

111. Wanke, K. A., Devanna, P. & Vernes, S. C. Understanding Neurodevelopmental Disorders: The Promise of Regulatory Variation in the 3′UTRome. *Biological Psychiatry* **83**, 548–557 (2018).

112. Litterman, A. J. *et al.* A massively parallel 3′ UTR reporter assay reveals relationships between nucleotide content, sequence conservation, and mRNA destabilization. *Genome Res.* **29**, 896–906 (2019).

113. Lubelsky, Y. & Ulitsky, I. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature* **555**, 107–111 (2018).

114. Cottrell, K. A., Chaudhari, H. G., Cohen, B. A. & Djuranovic, S. PTRE-seq reveals mechanism and interactions of RNA binding proteins and miRNAs. *Nat. Commun.* **9**, 1–13 (2018).

115. Liu, S. *et al.* Systematic identification of regulatory variants associated with cancer risk. *Genome Biol.* **18**, 194 (2017).

116. Kalita, C. A. *et al.* High-throughput characterization of genetic effects on DNA-protein binding and gene transcription. *Genome Res.* **28**, 1701–1708 (2018).

117. Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173 (2009).

118. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265–270 (2012).

119. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).

120. Gordon, M. G. *et al.* lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nat. Protoc.* 1–28 (2020). doi:10.1038/s41596-020-0333-5

121. Alexopoulou, A. N., Couchman, J. R. & Whiteford, J. R. The CMV early enhancer/chicken β actin (CAG) promoter can be used to drive transgene expression during the differentiation of murine embryonic stem cells into vascular progenitors. *BMC Cell Biol.* **9**, 1–11 (2008).

122. Delay, C., Calon, F., Mathews, P. & Hébert, S. S. Alzheimer-specific variants in the 3'UTR of Amyloid precursor protein affect microRNA function. *Mol. Neurodegener.* **6**, 70 (2011).

123. Gow, J. M., Chinn, L. W. & Kroetz, D. L. The Effects of ABCB1 3'-Untranslated Region Variants on mRNA Stability. *Drug Metab. Dispos.* **36**, 10–15 (2008).

124. Wang, J., Pitarque, M. & Ingelman-Sundberg, M. 3'-UTR polymorphism in the human CYP2A6 gene affects mRNA stability and enzyme expression. *Biochem. Biophys. Res. Commun.* **340**, 491–497 (2006).

125. van Arensbergen, J. *et al.* High-throughput identification of human SNPs affecting regulatory element activity. *Nat. Genet.* 1 (2019). doi:10.1038/s41588-019-0455-2

126. Karczewski, K. J. *et al.* The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* **45**, D840 (2017).

127. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).

128. Griffith, M. *et al.* CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat. Genet. 2017 492* **49**, 170–174 (2017).

129. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).

130. Ruiz, A. *et al.* Follow-up of loci from the International Genomics of Alzheimer's Disease Project identifies TRIP4 as a novel susceptibility gene. *Transl. Psychiatry* **4**, e358 (2014).

131. Sleep, J. A., Schreiber, A. W. & Baumann, U. Sequencing error correction without a reference genome. *BMC Bioinformatics* **14**, 367 (2013).

132. Ashuach, T. *et al.* MPRAnalyze: Statistical framework for massively parallel reporter assays. *Genome Biol.* **20**, 1–17 (2019).

133. Agarwal, V., Bell, G. W., Nam, J. W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4**, (2015).

134. Pourteymour, S. *et al.* Perilipin 4 in human skeletal muscle: localization and effect of physical activity. *Physiol. Rep.* **3**, (2015).

135. Wang, Y., Li, G., Wan, F., Dai, B. & Ye, D. Prognostic value of D-lactate dehydrogenase in patients with clear cell renal cell carcinoma. *Oncol. Lett.* **16**, 866–874 (2018).

136. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

137. Lee, J. E., Lee, J. Y., Wilusz, J., Tian, B. & Wilusz, C. J. Systematic analysis of cis-elements in unstable mRNAs demonstrates that CUGBP1 is a key regulator of mRNA decay in muscle cells. *PLoS One* **5**, (2010).

138. Lambert, N. J., Robertson, A. D. & Burge, C. B. RNA Bind-n-Seq: Measuring the Binding Affinity Landscape of RNA-Binding Proteins. *Methods Enzymol.* **558**, 465–493 (2015).

139. Ghanbari, M. & Ohler, U. Deep neural networks for interpreting RNA-binding protein target preferences. *Genome Res.* **30**, 214–226 (2020).

140. Mukherjee, N. *et al.* Global target mRNA specification and regulation by the RNA-binding protein ZFP36. *Genome Biol.* **15**, (2014).

141. Meyer, C. *et al.* The TIA1 RNA-Binding Protein Family Regulates EIF2AK2-Mediated Stress Response and Cell Cycle Progression. *Mol. Cell* **69**, 622-635.e6 (2018).

142. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45**, D833–D839 (2017).

143. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. *Nat. Rev. Cancer 2020 2010* **20**, 555–572 (2020).

144. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nat. 2020 5787793* **578**, 82–93 (2020).

145. Dorn, G. W. Mitofusin 2 Dysfunction and Disease in Mice and Men. *Front. Physiol.* **11**, 782 (2020).

146. Li, Y. *et al.* The anti-tumor effects of Mfn2 in breast cancer are dependent on promoter DNA methylation, the P21 Ras motif and PKA phosphorylation site. *Oncol. Lett.* **15**, 8011–8018 (2018).

147. Xu, K. *et al.* MFN2 suppresses cancer progression through inhibition of mTORC2/Akt signaling. *Sci. Reports 2017 71* **7**, 1–13 (2017).

148. Liu, X. *et al.* Mfn2 inhibits proliferation and cell-cycle in Hela cells via Ras-NF-κB signal pathway. *Cancer Cell Int.* **19**, 1–9 (2019).

149. Tulchinsky, E. Fos family members: regulation, structure and role in oncogenic

transformation. *Histol. Histopathol.* **15**, 921–928 (2000).

150. He, J. *et al.* miR-597 inhibits breast cancer cell proliferation, migration and invasion through FOSL2. *Oncol. Rep.* **37**, 2672–2678 (2017).

151. Li, J., Zhou, L., Jiang, H., Lin, L. & Li, Y. Inhibition of FOSL2 aggravates the apoptosis of ovarian cancer cells by promoting the formation of inflammasomes. *Genes and Genomics* **44**, 29–38 (2022).

152. Singer, J. W. *et al.* Inhibition of interleukin-1 receptor-associated kinase 1 (IRAK1) as a therapeutic strategy. *Oncotarget* **9**, 33416–33439 (2018).

153. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).

154. Nelson, J. W. *et al.* Engineered pegRNAs improve prime editing efficiency. *Nat. Biotechnol. 2021* 1–9 (2021). doi:10.1038/s41587-021-01039-7

155. Chen, P. J. *et al.* Enhanced prime editing systems by manipulating cellular determinants of editing outcomes. *Cell* **0**, (2021).

156. Sobell, H. M. Actinomycin and DNA transcription. *Proc. Natl. Acad. Sci. U. S. A.* **82**, 5328–5331 (1985).

157. Mazzoni, C. & Falcone, C. mRNA stability and control of cell proliferation. *Biochem. Soc. Trans.* **39**, 1461–1465 (2011).

158. Johnson, E. L., Robinson, D. G. & Coller, H. A. Widespread changes in mRNA stability contribute to quiescence-specific gene expression patterns in a fibroblast model of quiescence. *BMC Genomics* **18**, 1–9 (2017).

159. Ferraro, N. M. *et al.* Diverse transcriptomic signatures across human tissues identify

functional rare genetic variation. *bioRxiv* 786053 (2019). doi:10.1101/786053

160. Griesemer, D. *et al.* Genome-wide functional screen of 3'UTR variants uncovers causal variants for human disease and evolution. *Cell* **184**, 5247-5260.e19 (2021).

161. Hernandez, R. D. *et al.* Ultrarare variants drive substantial cis heritability of human gene expression. *Nat. Genet. 2019 519* **51**, 1349–1355 (2019).

162. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet. 2011 1211* **12**, 745–755 (2011).

163. Benjamin, D. & Moroni, C. mRNA stability and cancer: an emerging link? *https://doi.org/10.1517/14712598.7.10.1515* **7**, 1515–1529 (2007).

164. Calabrese, C. *et al.* Genomic basis for RNA alterations in cancer. *Nature* **578**, 129–136 (2020).

165. Huang, K. lin *et al.* Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* **173**, 355-370.e14 (2018).

166. Van der Auwera, G., O'Connor, B. & Safari, an O. M. C. Using Docker, GATK, and WDL in Terra. *Genomics in the Cloud* 300 (2020).

167. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

168. Smith, T., Heger, A. & Sudbery, I. UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).

169. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods 2012 94* **9**, 357–359 (2012).

170. Sun, L. *et al.* Predicting dynamic cellular protein–RNA interactions by deep learning using in vivo RNA structures. *Cell Res. 2021 315* **31**, 495–516 (2021).

171. Chow, R. D., Chen, J. S., Shen, J. & Chen, S. A web tool for the design of prime-editing guide RNAs. *Nat. Biomed. Eng. 2020 52* **5**, 190–194 (2020).

172. Bahn, J. H. J. H. *et al.* Accurate Identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res.* **22**, 142–150 (2012).

173. Tan, M. H. *et al.* Dynamic landscape and regulation of RNA editing in mammals. *Nature* **550**, 249–254 (2017).

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

175. Chen, C. X. *et al.* A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains. *RNA* **6**, 755–767 (2000).

176. Walkley, C. R. & Li, J. B. Rewriting the transcriptome: adenosine-to- inosine RNA editing by ADARs. *Genome Biol.* **18**, 1–13 (2017).

177. Xu, L.-D., Öhman, M., Xu, L.-D. & Öhman, M. ADAR1 Editing and its Role in Cancer. *Genes (Basel).* **10**, (2019).

178. Xu, X., Wang, Y. & Liang, H. *The role of A-to-I RNA editing in cancer development*. *Current Opinion in Genetics and Development* **48**, 51–56 (Elsevier Current Trends, 2018).

179. Fumagalli, D. *et al.* Principles Governing A-to-I RNA Editing in the Breast Cancer Transcriptome. *Cell Rep.* **13**, 277–289 (2015).

180. Heerboth, S. *et al.* EMT and tumor metastasis. *Clin. Transl. Med.* **4**, 6 (2015).

181. Lamouille, S., Xu, J. & Derynck, R. Molecular mechanisms of epithelial–mesenchymal transition. *Nat. Rev. Mol. Cell Biol.* **15**, 178–196 (2014).

182. Dongre, A. & Weinberg, R. A. New insights into the mechanisms of epithelial–mesenchymal transition and implications for cancer. *Nat. Rev. Mol. Cell Biol.* **20**, 69–84 (2019).

183. Fu, L. *et al.* RNA editing of *SLC22A3* drives early tumor invasion and metastasis in familial esophageal cancer. *Proc. Natl. Acad. Sci.* **114**, E4631–E4640 (2017).

184. Han, S. *et al.* RNA editing in RHOQ promotes invasion potential in colorectal cancer. *J. Exp. Med.* **211**, 613–621 (2014).

185. Gumireddy, K. *et al.* The mRNA-edited form of GABRA3 suppresses GABRA3-mediated Akt activation and breast cancer metastasis. *Nat. Commun. 2016 71* **7**, 1–9 (2016).

186. Wang, Y. *et al.* Systematic characterization of A-to-I RNA editing hotspots in microRNAs across human cancers. (2017). doi:10.1101/gr.219741.116

187. Harvey, S. E. *et al.* Coregulation of alternative splicing by hnRNPM and ESRP1 during EMT. *RNA* **24**, 1326–1338 (2018).

188. Warzecha, C. C. & Carstens, R. P. Complex changes in alternative pre-mRNA splicing play a central role in the epithelial-to-mesenchymal transition (EMT). *Semin. Cancer Biol.* **22**, 417–427 (2012).

189. Brown, R. L. *et al.* CD44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression. *J. Clin. Invest.* **121**, 1064–1074 (2011).

190. Xu, Y. *et al.* Cell type-restricted activity of hnRNPM promotes breast cancer metastasis

via regulating alternative splicing. *Genes Dev.* **28**, 1191–1203 (2014).

191. Hu, X. *et al.* The RNA-binding protein AKAP8 suppresses tumor metastasis by antagonizing EMT-associated alternative splicing. *Nat. Commun.* **11**, 486 (2020).

192. Shelton, P. M. *et al.* The Secretion of miR-200s by a PKCζ/ADAR2 Signaling Axis Promotes Liver Metastasis in Colorectal Cancer. *Cell Rep.* 1178–1191 (2018). doi:10.1016/j.celrep.2018.03.118

193. Tan, T. Z. *et al.* Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol. Med.* **6**, 1279–93 (2014).

194. Lee, J. H., Ang, J. K. & Xiao, X. Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants. *RNA* **19**, 725–732 (2013).

195. Plaisier, S. B., Taschereau, R., Wong, J. A. & Graeber, T. G. Rank-rank hypergeometric overlap: Identification of statistically significant overlap between gene-expression signatures. *Nucleic Acids Res.* **38**, 1–17 (2010).

196. Liddicoat, B. J. *et al.* RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. *Science* **349**, 1115–20 (2015).

197. George, C. X., Ramaswami, G., Li, J. B. & Samuel, C. E. Editing of Cellular Self-RNAs by Adenosine Deaminase ADAR1 Suppresses Innate Immune Stress Responses. *J. Biol. Chem.* **291**, 6158–68 (2016).

198. Savva, Y. A., Rezaei, A., Laurent, G. S. & Reenan, R. A. Reprogramming, Circular Reasoning and Self versus Non-self: One-Stop Shopping with RNA Editing. *Front. Genet.* **7**, 1–8 (2016).

199. Mannion, N. M. M. *et al.* The RNA-Editing Enzyme ADAR1 Controls Innate Immune Responses to RNA. *Cell Rep.* **9**, 1482–1494 (2014).

200. Lambrechts, D. *et al.* Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat. Med.* **24**, 1277–1289 (2018).

201. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).

202. Zaoui, K. *et al.* Ran promotes membrane targeting and stabilization of RhoA to orchestrate ovarian cancer cell invasion. *Nat. Commun.* **10**, 2666 (2019).

203. Ridley, A. Rho GTPases and cell migration. *J. Cell Sci.* 2713–2722 (2001).

204. Yu, X. *et al.* CXCL12/CXCR4 promotes inflammation-driven colorectal cancer progression through activation of RhoA signaling by sponging miR-133a-3p. *J. Exp. Clin. Cancer Res.* **38**, 32 (2019).

205. Yang, Y.-K. *et al.* ARF-like protein 16 (ARL16) inhibits RIG-I by binding with its C-terminal domain in a GTP-dependent manner. *J. Biol. Chem.* **286**, 10568–80 (2011).

206. Quinones-Valdez, G. *et al.* Regulation of RNA editing by RNA-binding proteins in human cells. *Commun. Biol.* **2**, 19 (2019).

207. Wang, I. X. *et al.* ADAR Regulates RNA Editing , Transcript Stability , and Gene Expression. *Cell Rep.* **5**, 849–860 (2013).

208. Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Letters* **582**, 1977–1986 (2008).

209. Mayr, C. What are 3’ utrs doing? *Cold Spring Harb. Perspect. Biol.* **11**, (2019).

210. Dassi, E. Handshakes and Fights: The Regulatory Interplay of RNA-Binding Proteins. *Front.*

Mol. Biosci. **4**, 67 (2017).

211. Turner, M. & DÍaz-Muñoz, M. D. RNA-binding proteins control gene expression and cell fate in the immune system. *Nature Immunology* **19**, 120–129 (2018).

212. Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **13**, 508–514 (2016).

213. Shim, J., Lim, H., R.Yates, J. & Karin, M. Nuclear Export of NF90 Is Required for Interleukin-2 mRNA Stabilization. *Mol. Cell* **10**, 1331–1344 (2002).

214. Vumbaca, F., Phoenix, K. N., Rodriguez-Pinto, D., Han, D. K. & Claffey, K. P. Double-stranded RNA-binding protein regulates vascular endothelial growth factor mRNA stability, translation, and breast cancer angiogenesis. *Mol. Cell. Biol.* **28**, 772–83 (2008).

215. Kuwano, Y. *et al.* MKP-1 mRNA stabilization and translational control by RNA-binding proteins HuR and NF90. *Mol. Cell. Biol.* **28**, 4562–75 (2008).

216. Harashima, A., Guettouche, T. & Barber, G. N. Phosphorylation of the NFAR proteins by the dsRNA-dependent protein kinase PKR constitutes a novel mechanism of translational regulation and cellular defense. *Genes Dev.* **24**, 2640–53 (2010).

217. Castella, S., Bernard, R., Corno, M., Fradin, A. & Larcher, J.-C. Ilf3 and NF90 functions in RNA biology. *Wiley Interdiscip. Rev. RNA* **6**, 243–256 (2015).

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

219. Garcia, M. A. *et al.* Impact of Protein Kinase PKR in Cell Biology: from Antiviral to Antiproliferative Action. *Microbiol. Mol. Biol. Rev.* **70**, 1032–1060 (2006).

220. Gal-Ben-Ari, S., Barrera, I., Ehrlich, M. & Rosenblum, K. PKR: A Kinase to Remember.

Front. Mol. Neurosci. **11**, (2019).

221. Lambert, A. W., Pattabiraman, D. R. & Weinberg, R. A. Emerging Biological Principles of Metastasis. *Cell* **168**, 670–691 (2017).

222. Aiello, N. M. *et al.* EMT Subtype Influences Epithelial Plasticity and Mode of Cell Migration. *Dev. Cell* **45**, 681–695 (2018).

223. Puram, S. V *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611–1624 (2017).

224. Sharpnack, M. F. *et al.* Global Transcriptome Analysis of RNA Abundance Regulation by ADAR in Lung Adenocarcinoma. *EBioMedicine* **27**, 167–175 (2018).

225. Gu, T., Fu, A. Q., Bolt, M. J. & White, K. P. Clinical Relevance of Noncoding Adenosine-to-Inosine RNA Editing in Multiple Human Cancers. *JCO Clin. cancer informatics* **3**, 1–8 (2019).

226. Borchert, G. M. *et al.* Adenosine deamination in human transcripts generates novel microRNA binding sites. *Hum. Mol. Genet.* **18**, 4801–4807 (2009).

227. Wang, Q. *et al.* ADAR1 regulates ARHGAP26 gene expression through RNA editing by disrupting miR-30b-3p and miR-573 binding. *RNA* **19**, 1525–1536 (2013).

228. Zhang, L., Yang, C., Varelas, X. & Monti, S. Altered RNA editing in 3 ' UTR perturbs microRNA-mediated regulation of oncogenes and tumor-suppressors. *Sci. Rep.* 1–13 (2016). doi:10.1038/srep23226

229. Ma, C. *et al.* ADAR1 promotes robust hypoxia signaling via distinct regulation of multiple HIF-1α-inhibiting factors. *EMBO Rep.* **20**, (2019).

230. Sagredo, E. A. *et al.* ADAR1-mediated RNA-editing of 3'UTRs in breast cancer. *Biol. Res.*

51, 36 (2018).

231. Pestal, K. *et al.* Isoforms of RNA-Editing Enzyme ADAR1 Independently Control Nucleic Acid Sensor MDA5-Driven Autoimmunity and Multi-organ Development. *Immunity* **43**, 933–944 (2015).

232. Saunders, L. R. *et al.* Characterization of Two Evolutionarily Conserved, Alternatively Spliced Nuclear Phosphoproteins, NFAR-1 and -2, that Function in mRNA Processing and Interact with the Double-stranded RNA-dependent Protein Kinase, PKR. *J. Biol. Chem.* **276**, 32300–32312 (2001).

233. Strong, J. E., Coffey, M. C., Tang, D., Sabinin, P. & Lee, P. W. K. The molecular basis of viral oncolysis: Usurpation of the Ras signaling pathway by reovirus. *EMBO J.* (1998). doi:10.1093/emboj/17.12.3351

234. Stojdl, D. F. *et al.* Exploiting tumor-specific defects in the interferon pathway with a previously unknown oncolytic virus. *Nat. Med.* (2000). doi:10.1038/77558

235. Danziger, O., Shai, B., Sabo, Y., Bacharach, E. & Ehrlich, M. Combined genetic and epigenetic interferences with interferon signaling expose prostate cancer cells to viral infection. *Oncotarget* (2016). doi:10.18632/oncotarget.10313

236. Liu, H. *et al.* Tumor-derived IFN triggers chronic pathway agonism and sensitivity to ADAR loss. *Nat. Med.* 1 (2018). doi:10.1038/s41591-018-0302-5

237. Genomic Data Commons.

238. Hanzelmann, S., Castelo, R. & Guinney, J. GSVA : gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* **14**, (2013).

239. Porath, H. T., Carmi, S. & Levanon, E. Y. A genome-wide map of hyper-edited RNA reveals

numerous new sites. *Nat. Commun.* **5**, 4726 (2014).

240. REDIportal.

241. Dong, X. *et al.* CDK13 RNA Over-Editing Mediated by ADAR1 Associates with Poor Prognosis of Hepatocellular Carcinoma Patients. *Cell. Physiol. Biochem.* **47**, 2602–2612 (2018).

242. Jiang, Q. *et al.* Hyper-Editing of Cell-Cycle Regulatory and Tumor Suppressor RNA Promotes Malignant Progenitor Propagation. *Cancer Cell* **35**, 81-94.e7 (2019).

243. Beghini, A. *et al.* RNA hyperediting and alternative splicing of hematopoietic cell phosphatase (PTPN6) gene in acute myeloid leukemia. *Hum. Mol. Genet.* **9**, 2297–2304 (2000).

244. Levanon, E. Y. *et al.* Evolutionarily conserved human targets of adenosine to inosine RNA editing. *Nucleic Acids Res.* **33**, 1162–1168 (2005).

245. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines | NCI Genomic Data Commons.

246. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, 1–17 (2014).

247. Rosenblatt, J. & Stein, J. RRHO: Test overlap using the Rank-Rank Hypergeometric test. R package version 1.26.0. (2014). doi:10.18129/B9.bioc.RRHO

248. E-MTAB-6149 < Browse < ArrayExpress < EMBL-EBI.

249. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

250. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-

seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).

251. CIBERSORTx.

252. Bahn, J. H. *et al.* Genomic analysis of ADAR1 binding and its involvement in multiple RNA processing pathways. *Nat. Commun.* **6**, 1–13 (2015).

253. Rueter, S. M., Dawson, T. R. & Emeson, R. B. Regulation of alternative splicing by RNA editing. *Nature* **399**, 75–80 (1999).

254. Xiang, J. F. *et al.* N 6-Methyladenosines Modulate A-to-I RNA Editing. *Mol. Cell* **69**, 126-135.e6 (2018).

255. Rengaraj, P. *et al.* Interplays of different types of epitranscriptomic mRNA modifications. *https://doi.org/10.1080/15476286.2021.1969113* **18**, 19–30 (2021).

256. Zhang, Z. & Carmichael, G. G. The fate of dsRNA in the nucleus: a p54(nrb)-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. *Cell* **106**, 465–476 (2001).

257. Prasanth, K. V. *et al.* Regulating gene expression through RNA nuclear retention. *Cell* **123**, 249–263 (2005).

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

259. Higuchi, M. *et al.* Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nat. 2000 4066791* **406**, 78–81 (2000).

260. Licht, K. *et al.* Inosine induces context-dependent recoding and translational stalling. *Nucleic Acids Res.* **47**, 3–14 (2019).

261. Bass, B. L. & Weintraub, H. An unwinding activity that covalently modifies its double-

stranded RNA substrate. *Cell* **55**, 1089–1098 (1988).

262. Han, L. & Liang, H. RNA editing in cancer: Mechanistic, prognostic, and therapeutic implications. *Mol. Cell. Oncol.* **3**, (2016).

263. Nielsen, J. S. & McNagny, K. M. The role of podocalyxin in health and disease. *Journal of the American Society of Nephrology* **20**, 1669–1676 (2009).

264. Kerjaschki, D., Sharkey, D. J. & Farquhar, M. G. Identification and characterization of podocalyxin--the major sialoprotein of the renal glomerular epithelial cell. *J. Cell Biol.* **98**, 1591–1596 (1984).

265. Doyonnas, R. *et al.* Anuria, Omphalocele, and Perinatal Lethality in Mice Lacking the Cd34-Related Protein Podocalyxin. *J. Exp. Med.* **194**, 13 (2001).

266. Le Tran, N., Wang, Y. & Nie, G. Podocalyxin in Normal Tissue and Epithelial Cancer. *Cancers 2021, Vol. 13, Page 2863* **13**, 2863 (2021).

267. Tamayo-Orbegozo, E. *et al.* Emerging Role of Podocalyxin in the Progression of Mature B-Cell Non-Hodgkin Lymphoma. *Cancers (Basel).* **12**, 396 (2020).

268. Snyder, K. A. *et al.* Podocalyxin enhances breast tumor growth and metastasis and is a target for monoclonal antibody therapy. *Breast Cancer Res.* **17**, 46 (2015).

269. LIN, C.-W., SUN, M.-S. & WU, H.-C. Podocalyxin-like 1 is associated with tumor aggressiveness and metastatic gene expression in human oral squamous cell carcinoma. *Int. J. Oncol.* **45**, 710–718 (2014).

270. Wu, H. *et al.* Podocalyxin regulates astrocytoma cell invasion and survival against temozolomide. *Exp. Ther. Med.* **5**, 1025–1029 (2013).

271. Itai, S. *et al.* Podocalyxin is crucial for the growth of oral squamous cell carcinoma cell

line HSC-2. *Biochem. Biophys. Reports* **15**, 93–96 (2018).

272. Lee, W. Y. *et al.* Podocalyxin-like protein 1 regulates TAZ signaling and stemness properties in colon cancer. *International Journal of Molecular Sciences* **18**, 2047 (2017).

273. He, S., Du, W., Li, M., Yan, M. & Zheng, F. PODXL might be a new prognostic biomarker in various cancers: a meta-analysis and sequential verification with TCGA datasets. *BMC Cancer* **20**, 620 (2020).

274. Taniuchi, K., Tsuboi, M., Sakaguchi, M. & Saibara, T. Measurement of serum PODXL concentration for detection of pancreatic cancer. *Onco. Targets. Ther.* **Volume 11**, 1433–1445 (2018).

275. Zhang, J. *et al.* PODXL, negatively regulated by KLF4, promotes the EMT and metastasis and serves as a novel prognostic indicator of gastric cancer. *Gastric Cancer* **22**, 48–59 (2019).

276. Borg, D. *et al.* Expression of podocalyxin-like protein is an independent prognostic biomarker in resected esophageal and gastric adenocarcinoma. *BMC Clin. Pathol.* **16**, 13 (2016).

277. Kusumoto, H. *et al.* Podocalyxin influences malignant potential by controlling epithelial–mesenchymal transition in lung adenocarcinoma. *Cancer Sci.* **108**, 528–535 (2017).

278. Lee, H., Kong, J. S., Lee, S. S. & Kim, A. Radiation-induced overexpression of TGFβ and PODXL contributes to colorectal cancer cell radioresistance through enhanced motility. *Cells* **10**, 2087 (2021).

279. Fröse, J. *et al.* Epithelial-Mesenchymal Transition Induces Podocalyxin to Promote Extravasation via Ezrin Signaling. *Cell Rep.* **24**, 962–972 (2018).

280. Amo, L. *et al.* Podocalyxin-like protein 1 functions as an immunomodulatory molecule in breast cancer cells. *Cancer Lett.* **368**, 26–35 (2015).

281. Zhou, Y. *et al.* Bmi1 Essentially Mediates Podocalyxin-Enhanced Cisplatin Chemoresistance in Oral Tongue Squamous Cell Carcinoma. *PLoS One* **10**, e0123208 (2015).

282. Huang, Z., Huang, Y., He, H. & Ni, J. Podocalyxin promotes cisplatin chemoresistance in osteosarcoma cells through phosphatidylinositide 3-kinase signaling. *Mol. Med. Rep.* **12**, 3916–3922 (2015).

283. CHIJIIWA, Y. *et al.* Overexpression of microRNA-5100 decreases the aggressive phenotype of pancreatic cancer cells by targeting PODXL. *Int. J. Oncol.* **48**, 1688–1700 (2016).

284. Xiao, X. *et al.* Splice site strength-dependent activity and genetic buffering by poly-G runs. *Nat. Struct. Mol. Biol.* **16**, 1094–1100 (2009).

285. Valente, L. & Nishikura, K. RNA Binding-independent Dimerization of Adenosine Deaminases Acting on RNA and Dominant Negative Effects of Nonfunctional Subunits on Dimer Functions. *J. Biol. Chem.* **282**, 16054–16061 (2007).

286. Kuttan, A. & Bass, B. L. Mechanistic insights into editing-site specificity of ADARs. *Proc. Natl. Acad. Sci. U. S. A.* **109**, (2012).

287. Phelps, K. J. *et al.* Recognition of duplex RNA by the deaminase domain of the RNA editing enzyme ADAR2. *Nucleic Acids Res.* **43**, 1123–1132 (2015).

288. Matthews, M. M. *et al.* Structures of human ADAR2 bound to dsRNA reveal base-flipping mechanism and basis for site selectivity. *Nat. Struct. Mol. Biol.* **23**, 426–433 (2016).

289. Wang, Y., Park, S. H. & Beal, P. A. Selective Recognition of RNA Substrates by ADAR Deaminase Domains. *Biochemistry* **57**, 1640–1651 (2018).

290. Meng, X., Ezzati, P. & Wilkins, J. A. Requirement of podocalyxin in TGF-beta induced epithelial mesenchymal transition. *PLoS One* **6**, 1–10 (2011).

291. Mansi, L. *et al.* REDIportal: millions of novel A-to-I RNA editing events from thousands of RNAseq experiments. *Nucleic Acids Res.* **49**, D1012–D1019 (2021).

292. Licht, K., Kapoor, U., Mayrhofer, E. & Jantsch, M. F. Adenosine to Inosine editing frequency controlled by splicing efficiency. *Nucleic Acids Res.* **44**, 6398–6408 (2016).

293. Licht, K. *et al.* A high resolution A-to-I editing map in the mouse identifies editing events controlled by pre-mRNA splicing. *Genome Res.* **29**, 1453–1463 (2019).

294. Desmet, F. O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* **37**, e67 (2009).

295. Park, K. C., Dharmasivam, M. & Richardson, D. R. The Role of Extracellular Proteases in Tumor Progression and the Development of Innovative Metal Ion Chelators That Inhibit Their Activity. *Int. J. Mol. Sci.* **21**, 1–22 (2020).

296. Miller, M. A., Sullivan, R. J. & Lauffenburger, D. A. Molecular Pathways: Receptor Ectodomain Shedding in Treatment, Resistance, and Monitoring of Cancer. *Clin. Cancer Res.* **23**, 623–629 (2017).

297. Nielsen, J. S. & McNagny, K. M. Novel functions of the CD34 family. *J. Cell Sci.* **121**, 3683–92 (2008).

298. Sizemore, S., Cicek, M., Sizemore, N., Kwok, P. N. & Casey, G. Podocalyxin increases the aggressive phenotype of breast and prostate cancer cells in vitro through its interaction

with ezrin. *Cancer Res.* **67**, 6183–6191 (2007).

299. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science (80-.).* **347**, (2015).

300. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).

301. Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., B. A. Protein Identification and Analysis Tools on the ExPASy Server. in *The Proteomics Protocols Handbook* (ed. John M. Walker) 571–607 (Humana Press, 2005).

302. Barbosa-Morais, N. L. *et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587–1593 (2012).

303. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Article Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. 400–416 (2018). doi:10.1016/j.cell.2018.02.052

304. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164–e164 (2010).

305. Hagen, L., Sharma, A., Aas, P. A. & Slupphaug, G. Off-target responses in the HeLa proteome subsequent to transient plasmid-mediated transfection. *Biochim. Biophys. Acta - Proteins Proteomics* **1854**, 84–90 (2015).

306. Yan, G. *et al.* LncRNA ILF3-AS1 promotes cell migration, invasion and EMT process in hepatocellular carcinoma via the miR-628-5p/MEIS2 axis to activate the Notch pathway. *Dig. Liver Dis.* **54**, 125–135 (2022).

307. Linmei, Z. H. U. *et al.* Mechanism underlying long non-coding RNA ILF3-AS1-mediated inhibition of cervical cancer cell proliferation, invasion and migration, and promotion of

apoptosis. *Mol. Med. Rep.* **24**, 1–12 (2021).

308. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nat. 2021 5967873* **596**, 583–589 (2021).