UNIVERSITY OF CALIFORNIA

Los Angeles

Function and Regulation of Nucleotide Variants in RNA

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Bioengineering

by

Giovanni Quinones Valdez

2021

ABSTRACT OF THE DISSERTATION

Function and Regulation of Nucleotide Variants in RNA

by

Giovanni Quinones Valdez

Doctor of Philosophy in Bioengineering

University of California, Los Angeles, 2021

Professor Xinshu Xiao, Co-Chair

Professor Aaron S. Meyer, Co-Chair

RNA molecules harbor the information necessary for the synthesis of proteins and are essential to a wide variety of cellular processes. Variation of the RNA sequences results in significant phenotypic differences; however, the precise relationship between the two remains largely unknown. Thanks to the advent of high-throughput sequencing technologies, we now have the opportunity to study the transcriptome with unprecedented detail and characterize many different types of variants present in the RNA. In the present work, we developed novel computational approaches and performed in-depth analysis of RNA-sequencing (RNA-seq) data with the overarching goal of studying the function and regulations of nucleotide variants in RNA.

We first aimed to understand the factors that regulate the most prevalent type of non-genetic nucleotide variant in human RNAs which is Adenosine-to-Inosine (A-to-I) editing. We analyzed bulk RNA-seq data obtained following the knockdown of over two hundred RNA Binding Proteins (RBPs) individually. This allowed us to study their role in the regulation of A-to-I editing at the transcriptome-wide scale. We identified several RBPs including DROSHA, ILF2/3, TROVE2, and TARDBP that significantly alter editing levels through various mechanisms including directly targeting the expression of ADAR1, protein-protein interaction, and direct binding to edited regions.

Next, to study the effect of nucleotide variants, we made use of single-cell RNA-sequencing (scRNA-seq) data. This technology offers a unique glimpse of the transcriptome at the single cell-resolution. However, identification of nucleotide variants in scRNA-seq remains challenging and very few methods are available for this purpose. Here, we present scAllele, a novel method that detects both single nucleotide variants (SNVs) and microindels in scRNA-seq with high accuracy and sensitivity. In addition, scAllele identifies functional relationships between the identified variants and alternative RNA processing. We applied scAllele to scRNA-seq data derived from lung cancer patients (matched tumor and normal) and detected over 150 allele-specific splicing events that were unique to each condition or showed differential prevalence.

Based on scAllele, we further developed a new method, namely T-Allele, to identify nucleotide variants and their linkage patterns in third-generation RNA-seq data. We demonstrated that the precision of variant calls by T-Allele is robust despite the relatively high sequencing error rate of this type of data. Using T-Allele, we identified up to 44 haplotype-specific alternative splicing events in each of the 8 cell lines included in our study. We also showed T-allele’s ability to segregate alternative splicing events regulated by nucleotide variants from those whose regulation involved other factors.

The dissertation of Giovanni Quinones Valdez is approved.

Zhilin Qu

Desmond Smith

Wei Wang

Aaron S. Meyer, Committee Co-Chair

Xinshu Xiao, Committee Co-Chair

University of California, Los Angeles

2021

To my family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SUPPLEMENTAL FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to thank Dr. Xinshu (Grace) Xiao. Grace is not only an excellent scientist, but more importantly, she is the best advisor a student can ask for. She always makes time to meet with all her students. She spreads energy, motivation, and love for science to the entire group. I will always be thankful for her patient and compassionate mentoring. She has become the main role model for my professional career.

I would also like to thank all my committee members. Dr. Aaron Meyer, Dr. Wei Wang, Dr. Desmond Smith, and Dr. Zhiling Qu. I am very thankful for their feedback and their advice which was essential for the completion of my research projects.

Thanks to all the professors that throughout my undergraduate and graduate education have guided me. Special thanks to Dr. Eduardo Vallejos, Dr. Mehul Bhakta, Dr. Alfred S. Lewin and Dr. Manas Biswal who accepted me into their labs, trained me as a researcher and grave me lots of advice for my academic career. Thanks, in particular, to Dr. Manas Biswal, who was the person that taught me the most. I will always appreciate his teaching, work ethic and his friendship. Special thanks to Dr. Melanie Correll and Dr. Eric McLemore who were committee members for my undergraduate honors dissertation. Dr. Correll and Dr. McLemore looked after me as an undergraduate student and went above and beyond to help me get into graduate school.

I am also thankful and feel very privileged to work with such diverse and talented research group. Special thanks to Dr. Yun-Hua (Esther) Hsiao who mentored me when I first joined the Xiao lab and Dr. Ei-Wen Yang for being a close friend inside and outside of the lab. Thanks to the rest of my lab mates: Dr. Tracey Chan, Mudra Choudhury, Christina Burghard, Ting Fu, Carlos Gonzales-Figueroa, Jonathan Hervoso, Elaine Huang, Tadeo Spencer, Dr. Jae Hoon Bahn, Dr. Zhiheng Liu, Dr. Ling Zhang and all the past members.

Thanks to my parents Mario Quinones Lozano and Delia Valdez Bustos. They made huge personal sacrifices to let me study abroad since the age of sixteen. Also special thanks to my brother Claudio. His support got me though the hardest time of my PhD while I struggled the most with my mental health. My love and my gratitude to my whole family are eternal.

Finally, a huge thank to Helen Lin. I met Helen in the fourth year of my PhD, and we have been inseparable ever since. Thank you for your love, your friendship, and your company. Having you in my life has made every challenge more bearable and every day more beautiful.

This work was partially supported by the Bioengineering Supplemental and First-Year Research Fellowships.

Chapter 2 was published in *Communications Biology*. 2019. **Quinones-Valdez G,** Tran S., Jun HI, Bahn JH, Yang EW, Zhan L, Brummer A, Wei X, VanNostrand E, Pratt G, Yeo GW, Graveley BR, Xiao X. “Regulation of RNA editing by RNA Binding Proteins in Human Cells”. doi: https://doi.org/10.1038/s42003-018-0271-8. The dissertation author was the primary author of this manuscript.

Chapter 3 is currently in preparation for publication. **Quinones-Valdez G**, Fu T, Chan TW, Xiao X. “scAllele: a versatile tool for the analysis of scRNA-seq”. The dissertation author was the primary author of this manuscript. X.X. and G.Q.V. designed the study and wrote the paper. T.W.C. contributed to the processing of the test data. T.F. performed the experimental validation of the novel variants.

Chapter 4 is currently in preparation for publication. **Quinones-Valdez G**, Amoah K, Xiao X. “Extending scAllele: a method to identify haplotype-specific isoform expression in long-read RNA-seq”. The dissertation author was the primary author of this manuscript. X.X. and G.Q.V. designed the study and wrote the paper. K.A. collaborated in the data analysis.

VITA

Education and Employment

2011 – 2015	B.S. in Agricultural and Biological Engineering, University of Florida, Gainesville, FL
2012 – 2013	Research Assistant, Department of Horticultural Sciences, University of Florida, Gainesville, FL
2014 – 2015	Research Assistant, Department of Molecular Genetics and Microbiology, University of Florida, Gainesville, FL
2015 – 2021	Graduate Student Researcher, Department of Bioengineering, University of California, Los Angeles, Los Angeles, CA
Fall 2017	Teaching Assistant, Department of Microbiology, Immunology and Molecular Genetics, University of California, Los Angeles, Los Angeles, CA
Summer 2020	Course Instructor, Institute for Quantitative and Computational Biosciences: BIG Summer program University of California, Los Angeles, Los Angeles, CA

Honors and Awards

2015 – 2016	Bioengineering Research Fellowship
Spring 2017	Bioengineering Supplemental Fellowship

Publications

1. Hsiao YH, Bahn JH, Yang Y, Lin X, Tran S, Yang EW, **Quinones-Valdez G**, Xiao X. *RNA editing in nascent RNA affects pre-mRNA splicing*. Genome Research. 2018

2. **Quinones-Valdez G,** Tran S., Jun HI, Bahn JH, Yang EW, Zhan L, Brummer A, Wei X, VanNostrand E, Pratt G, Yeo GW, Graveley BR, Xiao X. *Regulation of RNA editing by RNA Binding Proteins in Human Cells*. Comm. Bio. 2019

3. Yang EW, Bahn JH, Hsiao YH, Tan BX, Sun Y, Fu T, Zhou B, Van Nostrand EL, Pratt, Freese P, Wei X, **Quinones-Valdez G,** Urban AE, Graveley BR, Burge CB, Yeo GW, Xiao X. *Allele-specific binding of RNA Binding Proteins Reveal Functional Genetic Variants in the RNA.* Nat. Comm., 2019

4. 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

5. **Quinones-Valdez G**, Fu T, Chan TW, Xiao X. *scAllele: a versatile tool for the analysis of scRNA-seq*. (in preparation)

6. **Quinones-Valdez G**, Amoah K, Xiao X. Extending scAllele: a method to identify haplotype-specific isoform expression in long-read RNA-seq. (in preparation)

7. Liu Z., **Quinones-Valdez G**, Xiao X. *L-GIREMI uncovers RNA editing sites in long-read RNA-seq* (in preparation)

8. ENCODE Consortium. *ENCODE Phase III: Building an Encyclopedia of Regulatory Elements for Human and Mouse.* Nature. 2020

9. ENCODE Consortium. *Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genome.* Nature. 2020

Abstracts

1. Biswal MR, Ildefonso CJ, Rossmiller BP, Mao H, Li H, Han P, Zhu P, Tong Y, **Quinones-Valdez G**, Lewin AS. *Delaying Retinal Degeneration in a Mouse Model of Geographic Atrophy: An Antioxidant Gene Therapy Approach*. Investigative Ophthalmology and Visual Science 2015

2. Biswal MR, Ildefonso CJ, Wang Z, Mao H, Li H, Zhu P, **Quinones-Valdez G**, Lewin AS. *Effectiveness of Antioxidant Gene Therapy in Delaying Retinal Degeneration*. Molecular Therapy 2015

CHAPTER 1 - Background

RNA is an essential molecule of life, expressing the genetic code in the DNA, guiding the production of proteins, and regulating a wide spectrum of cellular processes. Thus, it is very important to fully characterize the forms and function of different types of RNA molecules. In this dissertation, we focus on sequence variations in the RNA, tackling the questions related to their identification, function, and regulation.

1.1 SMALL NUCLEOTIDE VARIANTS IN THE RNA

Broadly speaking, two types of nucleotide variants exist in the RNA: genetic variants and post-transcriptional RNA modifications. In this section, we briefly introduce these two types of RNA variants.

Genetic variants are differences in the DNA sequences between individuals. These differences, in the smaller scale (less than 50 bases), can be broadly classified as base substitutions, insertions and deletions. Small variants occur much more frequently than their large-scale counterparts accounting for 99.9% of the total number of genetic variations and sum up to between 4 and 5 million in a typical genome (Gibbs et al., 2015). Identifying functional genetic variants and characterizing their contributions to phenotypic diversity and human diseases are central tasks of the post-genomic era and essential steps towards precision medicine. To date, the function of most genetic variants remains unknown.

If transcribed, genetic variants show up in the RNA molecules as sequence variations. In addition, another type of sequence variants exists in the RNA, which is resulted from the post-transcriptional process called RNA editing. RNA editing is a base modification catalyzed by the

ADAR and APOBEC enzymes. Such modifications are products of deamination reactions which turn Adenosine to Inosine (A-to-I editing, catalyzed by ADAR) or Cytidine to Uridine (C-to-U editing, catalyzed by APOBEC). To date, over 4.5 million A-to-I editing sites are cataloged in the REDI portal database (Picardi et al., 2017), and more than 1.5 million sites in the RADAR database (Ramaswami & Li, 2014).

Editing sites were also shown to have relevant roles in the fate of the transcript. However, similarly as genetic variants, the function of the majority of them remains unknown.

1.2 IDENTIFICATION OF SMALL NUCLEOTIDE VARIANTS IN THE RNA

1.2.1 Identification of genetic variants in the RNA

High-throughput sequencing of RNA (RNA-seq) is now routinely used for transcriptome profiling in numerous research and clinical applications. RNA-seq is well-suited to study nucleotide variants in the RNA. However, there is still a great demand for improved bioinformatic methods for this purpose.

At present, variant calling from sequencing data is a major topic in bioinformatics. This task is traditionally performed on Whole Genome Sequencing (WGS) data with the purpose of obtaining a complete variant profile of the individual. Nonetheless, with the increasing application of RNA-seq, variant calling from RNA-seq data is often needed. In addition, RNA-seq captures a lot more information than just the nucleotide sequences. Splicing, polyadenylation, gene expression, and allele-specific expression events, to name a few, can also be examined via RNA-seq. Therefore, even though limited to expressed genetic variants, RNA-seq provides the information to study their roles in RNA processing and maturation.

The common practice to call small variants in RNA-seq data is to use variant callers that were originally created for DNA-seq data and tune specific parameters or pre-process the data to remove splice junctions. Commonly used methods that work this way include GATK's Haplotype Caller (McKenna et al., 2010), Platypus (Rimmer et al., 2014), and TransIndel (R. Yang et al., 2018). These methods, however, don't directly consider all the unique features of RNA sequencing data. For example, many variant callers use haplotype-inference to reduce false positive and multi-allelic calls. The problem with this approach in RNA-seq is the presence of RNA editing sites, which disrupts the original haplotype. Furthermore, allele-specific alternative splicing can affect the allelic ratio locally, further obscuring the true haplotype.

Most previous studies examining nucleotide variants in the RNA focused on single nucleotide variants (SNVs), given their relative prevalence in human populations (Gibbs et al., 2015). In contrast, identification of small insertion and deletions (INDELs) remains a significant bioinformatic challenge (Sun et al., 2017). The commonly used pileup-based approach for SNV detection is not suitable for INDELs as they will inevitably require some alignment-correction. The consistency in INDEL calling resulted from different algorithms is low and the sensitivity of existing methods is very limited (Hasan et al., 2015). Nonetheless, INDELs account for up to 20% of the polymorphisms present in the genome, thus supporting the necessity to develop more effective methods for their detection.

1.2.2 Identification of RNA editing sites

The challenge behind identifying RNA editing sites from RNA-seq data is that they are indistinguishable from Single Nucleotide Polymorphism (SNPs). Traditional approaches rely on analyzing matched DNA and RNA samples to discard the genetic variants (Lo Giudice et al.,

2020) or simply removing common SNPs from the data, followed by a series of filters (Lee et al., 2013). Other methods rely on the allelic linkage between SNPs to discriminate editing sites from genetic variants, thereby avoiding the need for matched DNA data (Q. Zhang & Xiao, 2015). In addition, specialized approaches were developed to identify clusters of RNA editing sites, namely, hyperediting sites (Porath et al., 2014). RNA editing sites located in highly repetitive regions or close to splice sites remain challenging to identify. To date, most RNA editing studies carried out post-processing steps to remove putative RNA editing sites from such regions (Lee et al., 2013).

1.3 FUNCTION OF NUCLEOTIDE VARIANTS

1.3.1 Function of genetic variants

In the post-genomic era, the function of the vast number of genetic variants in the human population is a heavily pursued topic. Significant progress has been made in deciphering the function of some variants (Cano-Gamez & Trynka, 2020; Gallagher & Chen-Plotkin, 2018; Shastry, 2009) or discovering biomarkers for diagnosis or treatment design (Vogenberg et al., 2010). On the molecular level, major effort has been dedicated to studying variants located in protein-coding, promoter, and splice site regions due to their apparent impacts on gene expression (Cheung & Spielman, 2009; Cookson et al., 2009; Cooper & Mattox, 1997; Gilad et al., 2008; Griffin & Smith, 2000; M. J. Li et al., 2015; Rockman & Kruglyak, 2006; Stepanova et al., 2006). Yet, a large number of disease-associated variants reside in other non-coding regions, such as introns or 3' UTRs, whose functional roles remain largely unknown. A potentially important mechanism of action of these variants is their impact on post-transcriptional gene regulation, such as pre-mRNA splicing, which only gained attention as a major paradigm influencing transcriptome diversity in the past 15 years or so (Glisovic et al., 2008; Janga & Mittal, 2011; Morris et al., 2010; Thomas & Lieberman, 2013).

To identify functional genetic variants in post-transcriptional processes, RNA-seq has been instrumental since it simultaneously captures the expressed variants and the alternative RNA processing isoforms. Our lab and others have made extensive use of RNA-seq to identify functional genetic variants. One type of approach tackles the allele-specific linkage between genetic variants and splicing, polyadenylation or mRNA expression (G. Li et al., 2012). Additionally, the association reflected in allele-specific splicing can be further analyzed at the population-level to detect the functional variants that drive such a splicing association (Amoah et al., 2021). The pathways, targets, and mechanisms of non-sense mediated decay (NMD) were also studied via RNA-seq (Colombo et al., 2017). More generally, expression and splicing quantitative trait loci (QTL) analyses have uncovered many insights about the genetic basis of RNA expression, although such analyses do not elucidate the functional regulatory variants (Gallagher & Chen-Plotkin, 2018; Gilad et al., 2008; Nica & Dermitzakis, 2013).

1.3.2 Function of RNA editing sites

RNA editing sites play important roles in multiple layers of gene expression and RNA processing. The Inosine introduced by ADAR enzymes is read by the translation machinery as a Guanosine thus potentially altering the final protein sequence. A famous example is the Q/R mutation in the GluR ion channels caused by an editing site. This amino acid substitution affects the calcium permeability of AMPA receptors on neurons, and under-editing can cause severe epilepsy (Brusa et al., 1995). Other studies reported that RNA editing can modulate the splicing outcome of nascent mRNA by altering cis-elements or the stability of dsRNA structures (Y. H.

E. Hsiao et al., 2018). The stabilization of dsRNA structures is also critical to mRNA turnover. A-to-I editing in dsRNAs in the 3’UTR reduces the accessibility of Ago2-miRNA complex (Brümmer et al., 2017).

dsRNA is the preferential target of the ADAR enzymes for editing (Levanon et al., 2004). In the human transcriptome, these double-stranded structures are often formed by inverse repeat elements such as Alus (Levanon et al., 2004). The Inosine-Uridine pairs in this structures act as a molecular tag. Hypoediting of endogenous dsRNAs can activate the innate immune response triggered by the RIG-I and MDA5 proteins. Cytosolic editing levels were also shown to increase during Interferon response (Q. Wang et al., 2017) and I-dsRNA was shown to bind stress-granule protein components and participate in the translation silencing of certain transcripts (Scadden, 2007). Together, these data demonstrate that A-to-I editing is critical to stress response and innate immunity in human cells.

1.4 REGULATORY MECHANISMS OF RNA EDITING

A-to-I editing is primarily carried out by the ADAR proteins. However, it is known that additional mechanisms exist that regulate RNA editing. RNA binding proteins (RBPs) control multiple processes in the maturation and fate of mRNAs. Many of these RBPs physically interact with the ADAR proteins or bind to similar target sites as ADARs. Naturally, the question of their roles in RNA editing needs to be addressed.

A previous study carried out a screen for RNA editing repressors, where 3 RBPs SRSF9, DDX15 and RPS14 were identified. The mechanism by which they alter editing levels was mostly in *cis*, as the altered editing sites were located in regions bound by these proteins (Tariq et al., 2013). Another study identified RNA Helicase A (RHA) as an RNA editing repressor. The repression mechanism involves unwinding of the dsRNA substrate that encompasses both intronic and exonic sequences. The dsRNA unwinding also results in enhanced splicing efficiency of the exon (Bratt & Öhman, 2003). This example beautifully illustrates the coordination of two different RNA processing events.

To date, more than 1500 RBPs were identified in humans (Gerstberger et al., 2014). Most of these RBPs are involved in multiple molecular processes. Whether and how they may regulate RNA editing remains uncharacterized.

1.5 ONGOING QUESTIONS

Although great strides have been made in the field, only a small fraction of the known variants have been characterized in detail. The characterization of variants encompasses their accurate detection, identification of their functional roles in specific cellular contexts, and decoding of the regulatory mechanisms underlying their presence, expression, and function. In this work, we present findings and computational methods that directly address these questions.

In Chapter 2, we studied the role of over 200 RNA Binding Proteins (RBPs) in the regulation of RNA editing at the transcriptome-wide scale. The contribution of these RBPs helps explain the observed editing rate that the expression of ADAR alone could not account for. Also, by combining with other types of data, we reported the molecular mechanisms by which this regulation takes place. We present five RBPs that stood out as strong regulators from our study.

In Chapter 3, we present a novel method called scAllele. This method overcomes the intrinsic limitations of single-cell RNA-seq in identifying small variants. scAllele outperforms other commonly used tools. Furthermore, our method analyzes the allelic linkage between nucleotide variants and splicing isoforms to infer allele-specific alternative splicing. This novel approach can identify direct variant-mediated regulation of splicing solely from RNA-seq data. We applied scAllele to a lung cancer scRNA-seq dataset and observed linkage events that were unique to the tumor-samples. Many of these events were present in biologically relevant genes.

In Chapter 4, we extended the scAllele method and applied it to third-generation sequencing (or long read) RNA-seq data. Long-read data is naturally advantageous in capturing allelic linkage events because the entire RNA transcript is sequenced. We presented technical adaptations made to the original algorithm to accommodate long reads. We employed this new version, called T-allele, to long read data from 8 different cell lines and identified haplotype-specific alternative splicing events.

CHAPTER 2 - Regulation of RNA editing by RNA-binding proteins in Human Cells

2.1 ABSTRACT

Adenosine-to-inosine (A-to-I) editing, mediated by the ADAR enzymes, diversifies the transcriptome by altering RNA sequences. Recent studies reported global changes in RNA editing in disease and development. Such widespread editing variations necessitate an improved understanding of the regulatory mechanisms of RNA editing. Here, we study the roles of >200 RNA-binding proteins (RBPs) in mediating RNA editing in two human cell lines. Using RNA-sequencing and global protein-RNA binding data, we identify a number of RBPs as key regulators of A-to-I editing. These RBPs, such as TDP-43, DROSHA, NF45/90 and Ro60, mediate editing through various mechanisms including regulation of *ADAR1* expression, interaction with ADAR1, and binding to Alu elements. We highlight that editing regulation by Ro60 is consistent with the global up-regulation of RNA editing in systemic lupus erythematosus. Additionally, most key editing regulators act in a cell type-specific manner. Together, our work provides insights for the regulatory mechanisms of RNA editing.

2.2 INTRODUCTION

RNA editing refers to the alteration of RNA sequences through insertion, deletion or substitution of nucleotides (Axel et al., 1999; Shaw et al., 1988). In human cells, the most prevalent type of RNA editing is adenosine to inosine editing (A-to-I editing), catalyzed by the protein family adenosine deaminases acting on RNA (ADARs) (Nishikura, 2016). A-to-I editing can have profound impacts on gene expression through a wide spectrum of mechanisms, including changing protein-coding sequences (Nishikura, 2016), modifying splice sites (Bentley, 2014; Y. H. E. Hsiao et al., 2018), affecting RNA nuclear export (Z. Zhang & Carmichael, 2001) and altering microRNA sequences or their target sites (Nishikura, 2016).

Facilitated by high-throughput RNA sequencing (RNA-seq) methods, millions of A-to-I editing sites have been identified in human cells (Picardi et al., 2017; Ramaswami & Li, 2014). Although the function of most of these sites remains unknown, the involvement of RNA editing in various biological processes is increasingly appreciated (Nishikura, 2010). Recent studies showed that the editing levels of numerous RNA editing sites vary significantly across tissues, developmental stages and disease status (Gallo et al., 2017; Hwang et al., 2016; Tan et al., 2017). These findings prompted many outstanding questions, one of which relates to the regulatory mechanisms that underlie the widespread editing variations. ADAR proteins are the best-known regulators of the human editomes (Nishikura, 2016). However, variations in the expression levels of the ADAR genes alone can only account for some of the observed editing variations (Brümmer et al., 2017; Tan et al., 2017; Washburn & Hundley, 2016). Thus, there is a critical need for identifying and understanding additional regulatory mechanisms of RNA editing.

RNA-binding proteins (RBPs) are important regulators for all steps of RNA maturation. Post-transcriptional RNA processing, such as splicing and polyadenylation, is controlled by the formation of different ribonucleoprotein complexes with RBPs at their core (Gerstberger et al., 2014; Hentze et al., 2018). A number of RBPs have been reported to affect RNA editing (Washburn & Hundley, 2016). For example, some RBPs, such as SRSF9 and RPS14, interact with ADAR2 and affect A-to-I editing of a number of substrates (Tariq et al., 2013). The protein FMRP (encoded by the FMR1 gene) was shown to affect RNA editing by interacting with the

ADAR proteins in multiple organisms (Bhogal et al., 2011; Filippini et al., 2017; Shamay-Ramot et al., 2015; Tran et al., 2019). Another ADAR1-interacting protein, DICER, was reported to inhibit ADAR1 editing activity *in vitro* (Ota et al., 2013). In addition to ADAR-interacting proteins, other RBPs may affect editing by influencing the double-stranded RNA (dsRNA) substrates of ADAR or the interaction between ADAR and dsRNAs. For example, several RNA helicases, including RNA helicase A (DHX9) and DDX15, were reported to repress RNA editing, presumably by disrupting dsRNA structures (Bratt & Öhman, 2003; Tariq et al., 2013). The catalytically inactive ADAR protein, ADAR3, represses RNA editing by competitively binding to dsRNA targets in both human cells and C. *elegans* (Oakes et al., 2017; Washburn et al., 2014). Another dsRNA-binding protein, Staufen, binds to numerous inverted Alu repeats (De Lucas et al., 2014), the most prevalent type of ADAR1 substrates in human cells. As a result, Staufen may also regulate RNA editing, although this topic needs further investigation.

It was estimated that more than 3000 RBPs exist in human cells (Huang et al., 2018). Although the function of the majority of RBPs is unknown, it is now clear that many RBPs play a role in multiple steps of post-transcriptional RNA processing (Hentze et al., 2018). However, compared to other processes such as splicing for which a large number of splicing factors have been cataloged, the number of proteins known to regulate RNA editing remains relatively small. Nevertheless, it is known that ADAR1 potentially interacts with numerous other proteins (Cusick et al., 2009) and many RBPs may recognize Alu elements or dsRNA targets (Saunders & Barber, 2003; Ule, 2013; Washburn & Hundley, 2016). Thus, it is very likely that additional editing regulators remain to be uncovered, which will help to explain the observed editome variability in diseases, between cell types and developmental stages. To this end, we carried out a systematic analysis of the potential involvement of a large panel of RBPs in regulating RNA editing in human cells. We report a number of RBPs as key regulators of RNA editing, most of which function in a cell type-specific manner. Our findings greatly expand the repertoire of known RBPs as RNA editing regulators.

2.3 RESULTS

2.3.1 Global analysis of RNA editing upon knockdown of >200 RBPs

To examine potential regulatory mechanisms of RNA editing, we analyzed a large number of RNA-seq datasets generated upon knockdown of hundreds of RBPs as part of the ENCODE project (Nostrand et al., 2017). Specifically, data from two human cell lines, K562 (chronic myelogenous leukemia) and HepG2 (liver hepatocellular carcinoma) were included. Individual knockdown experiments were carried out for 222 and 225 RBPs in K562 and HepG2 cells, respectively. For each RBP, two biological replicates of knockdown were carried out, followed by polyA-selected RNA-seq, which were accompanied by two replicates of control experiments. An average of 33.5 million pairs of reads (2x100 or 2x101nt) were obtained for each knockdown or control replicate. To identify RNA editing sites, the RNA-seq data were analyzed using our previously developed methods (Ahn & Xiao, 2015; Bahn et al., 2012; Lee et al., 2013; Q. Zhang & Xiao, 2015), followed by batch-normalization (see Methods).

A total of 893,701 and 444,263 distinct editing sites were identified in the K562 and HepG2 samples, respectively. In a single dataset, the number of predicted editing sites ranged from 226 to 16,657, which approximately correlated with RNA-seq read coverage of the samples (Supplementary Fig. 2-1a). An average of 92% of the editing sites in each sample were of the A-to-G type, consistent with A-to-I editing (Supplementary Fig. 2-1a). This high percentage suggests a high accuracy of our RNA editing identification method, as previously shown (Q.

Zhang & Xiao, 2015). As expected, the majority of editing sites were located in Alu regions (Supplementary Fig. 2-1b), and introns or 3’ UTRs (Supplementary Fig. 2-1c). Among all distinct A-to-G editing sites from the two cell lines, 63% and 69% overlapped those in the RADAR (Ramaswami & Li, 2014) and REDIPortal (Picardi et al., 2017) databases, respectively. In this study, we restricted all subsequent analyses to A-to-G sites in order to focus on ADAR-catalyzed editing.

2.3.2 The landscape of differential editing upon RBP knockdown

Next, to assess the impact of RBPs on RNA editing, we identified differentially edited sites upon knockdown of each of these RBPs in each cell line (see Methods). We observed that different RBPs induced variable degrees of editing changes, ranging from being negligible to affecting nearly 50% of all testable sites (see Methods) (Fig. 2-1a, Supplementary Fig. 2-2). As a positive control, ADAR1 knockdown induced the most widespread editing reduction among all RBPs, supporting the effectiveness of our methods.

A comparison of differentially edited (DE) sites associated with different RBPs revealed that ADAR1 shared many sites with other RBPs in both cell lines (Fig. 2-1a, links between RBPs), consistent with the fact that ADAR1 is a main catalytic enzyme of A-to-I editing. It is also apparent that a small number of RBPs were associated with relatively high levels of editing changes in either positive or negative direction (Fig. 2-1a, Supplementary Fig. 2-2c, f). Specifically, 31 and 15 RBPs in K562 and HepG2 cells, respectively, had at least 10% of all testable sites as differentially edited sites (Supplementary Fig. 2-2a, d). Examples of such RBPs include ADAR1, FXR1, DROSHA and TARDBP. Notably, some of these RBPs shared many differentially edited sites (Fig. 2-1a, links between RBPs, and Supplementary Fig. 2-3). For the union of differentially edited sites of a pair of RBPs, we calculated a directional agreement score to evaluate the concordance of the directions of editing changes associated with two RBPs (Methods). Two small clusters of RBPs showed relatively high (same directional changes) or low (opposite directional changes) agreement scores with ADAR1, supporting possible existence of both enhancers and repressors of editing (Fig. 2-1b). Using a linear regression model, we estimated that the top 15 RBPs (including ADAR1), each of which affected ≥10% of editing sites in K562, together accounted for 52% of editing variation in this cell line, and 35% in HepG2 (Fig. 2-1c). This percentage remains high even if the most influential samples (such as ADAR1 and DROSHA knockdown) were excluded (Supplementary Fig. 2-4a). Using ADAR1 expression alone in the regression accounted for 6% and 15% of editing variation in K562 and HepG2, respectively (Supplementary Fig. 2-4b). In contrast, this percentage is 43.8% and 8.7% if 14 RBPs (except ADAR1) were used in K562 and HepG2, respectively (Supplementary Fig. 2-4c). Moreover, the inclusion of an interaction term between ADAR1 and the other RBPs increased this percentage by about 10% in each cell line, although this interaction term was not statistically significant (Supplementary Fig. 2-4d). Together, these results support that RNA editing is regulated by auxiliary proteins besides ADAR1 and the functional impact of such proteins is likely cell type-specific.

To examine the correlative relationship more systematically among RBPs, we calculated the correlation of editing changes of the differentially edited sites associated with a pair of RBPs respectively. For this analysis, we included RBPs whose knockdown induced differential editing among ≥ 2% of all testable sites. We then applied hierarchical clustering on the correlation coefficients (Supplementary Fig. 2-5a, b). In each cell line, we observed one small cluster mainly containing RBPs (including ADAR1) associated with the greatest reduction in editing levels upon their knockdown. A similar pattern was observed when clustering RBPs with WGCNA (weighted-gene co-expression networks) (Supplementary Fig. 2-6a, b), a more robust statistical framework than hierarchical clustering (B. Zhang & Horvath, 2005). Compared to RBPs that induced editing reduction upon knockdown, those associated with editing up-regulation upon their knockdown did not cluster strongly. Importantly, experimental batches did not confound the clusters in either hierarchical clustering or WGCNA (Supplementary Fig. 2-5, Supplementary Fig. 2-6). Together, these results suggest that a relatively small number of RBPs are associated with observable editing changes that are correlated with each other.

Since many RBPs may contribute to multiple RNA processing mechanisms (Hentze et al., 2018), the RNA editing changes observed upon knockdown of an RBP may not reflect a direct involvement of the RBP in regulating RNA editing. To further understand the underlying processes, we next examined whether an RBP may impose editing changes through three types of possible mechanisms: (1) by regulating ADAR1 expression; (2) by interacting with the ADAR proteins; (3) by binding to similar RNA substrates as the ADAR proteins.

2.3.3 TARDBP as a novel regulator of ADAR1 expression

To examine whether any RBPs may regulate ADAR expression, we analyzed ADAR1/2/3 mRNA expression levels in different RBP knockdown samples (Fig. 2-2a, Supplementary Fig. 2-7a, b). ADAR1 is much more abundant than ADAR2 and ADAR3 in both cell lines. Thus, we next focused on potential regulators of ADAR1. We observed that most RBPs did not cause significant changes in ADAR1 mRNA expression. One exception is the gene TARDBP, whose knockdown induced about two-fold reduction in ADAR1 mRNA level in HepG2 cells. In addition, Western blot analysis confirmed that ADAR1 protein level was also reduced upon TARDBP knockdown in HepG2 cells, but not in K562 cells (Fig. 2-2b). Consistent with the observed ADAR1 expression changes, TARDBP knockdown induced a global reduction in RNA editing levels in HepG2 cells (Fig. 2-2c), but not in K562 cells (Fig. 2-1a).

TARDBP encodes for the TDP-43 protein that binds to both DNA and RNA sequences (Lagier-Tourenne et al., 2010). TDP-43 is known to regulate transcription and multiple RNA processing steps (Lagier-Tourenne et al., 2010). To examine the potential regulatory mechanisms of TDP-43 on ADAR1 expression, we asked whether it may regulate ADAR1 transcription. We analyzed existing chromatin immunoprecipitation (ChIP-Seq) data of TDP-43 (HepG2 cells) from the ENCODE consortium (Nostrand et al., 2017) (Fig. 2-2d, Supplementary Fig. 2-7c). A significant ChIP peak (FDR = 0.001, peak 1) was observed overlapping the first exon of a p110 isoform of ADAR1. A second peak (FDR = 0.002, peak 2), although only significant in one replicate, was found upstream of the first exon of another p110 isoform. Note that the latter peak also overlaps the first exon of the p150 isoform of ADAR1. However, ADAR1 p150 expression is undetectable in HepG2 (Fig. 2-2b). Both ChIP peaks are located in open chromatin regions as denoted by the presence of H3K27Ac marks, DNase I hypersensitive sites and transcription factors binding clusters (Pol2 IgR) (Fig. 2-2d).

Based on these data, we hypothesized that the ChIP peak regions are regulatory elements that control transcription of ADAR1. To test this hypothesis, we cloned these regions, respectively, into different luciferase reporters to test whether they may serve as promoter or enhancer elements. These constructs were transiently transfected into HepG2 cells, followed by measurement of luciferase activity. Using the PGL3-Basic vector, we observed a significant increase in luciferase activity with the inclusion of either peak region, compared to empty vectors (Fig. 2-2e). Furthermore, in the PGL3-Enhancer vector that lacks the SV40 promoter, both peak regions induced higher luciferase activity than controls (Fig. 2-2f). In contrast, only the second peak region induced significant luciferase activity in the PGL3-Promoter vector that lacks the SV40 enhancer (Fig. 2-2f). We additionally validated the binding of TDP-43 to the reporter construct using ChIP followed by real-time and semi-quantitative PCR (Supplementary Fig. 2-7d, e and f). The PCR amplification of the TDP-43 Immunoprecipitation products showed significant enrichment of the TDP-43 binding sequences used in the reporter assay (Supplementary Fig. 2-7e and f). Together, these results suggest that TDP-43 binds to multiple regulatory regions of the ADAR1 gene that may serve as promoters or enhancers.

2.3.4 ADAR1-interacting RBPs as RNA editing regulators

ADAR1 is known to interact with a large number of proteins (Cusick et al., 2009) (Supplementary Table 1). A prevailing question in the field is whether ADAR1's interacting partners may confer regulation on RNA editing. To examine this question, we started from known ADAR1-interacting proteins, and asked: (1) whether knockdown of a protein induced a considerable amount of RNA editing changes; and (2) whether this protein binds significantly close to the differential RNA editing sites observed upon its knockdown. The confirmation of the second question serves as a strong indication of the direct involvement of a protein in modulating RNA editing. Such proteins likely affect RNA editing through their known interacting relationships with ADAR1, although the exact mechanisms need to be investigated in the future.

Our study included 18 known ADAR1-interacting proteins whose expression was knockdown in at least one cell line. The majority of known ADAR1-interacting proteins induced differential editing in <10% of the associated testable editing sites upon their knockdown in

K562 or HepG2 cells (Supplementary Fig. 2-8a). This observation suggests that not all ADAR-interacting partners influence RNA editing extensively. Nevertheless, a small number of RBPs were associated with considerable editing changes (≥10% of all testable sites) upon their respective knockdown, including DROSHA, FXR1, XRCC6 and MATR3 in K562 cells, ILF2 and PABPC1 in HepG2 cells (Fig. 2-3a).

Next, we asked whether the above proteins might be direct regulators of RNA editing by examining the RBP binding locations relative to differentially edited sites. Although not a necessary requirement for direct regulators of RNA editing, a significantly close distance between RBP binding and differentially edited sites provides strong support for a direct regulatory relationship. To examine the global protein-RNA binding patterns, we analyzed the enhanced crosslinking immunoprecipitation (eCLIP) data for the above proteins generated by the ENCODE project (Nostrand et al., 2017) (except ILF2 and PABPC1 for which eCLIP data are not available) (Fig. 2-3b). As a positive control, we included our previously published ADAR1 CLIP-Seq data from the U87MG cells (Bahn et al., 2015) in this analysis. Although ADAR1 CLIP was generated using a different cell type, the CLIP peaks were significantly closer than expected by chance to differentially edited sites observed in K562 or HepG2 cells upon ADAR1 knockdown. This data supports the validity of this analysis.

Among all proteins included in this analysis, DROSHA demonstrated the most significant relationship between protein binding and protein knockdown-induced differential editing (Fig. 2-3a, b). DROSHA is best known to bind to double-stranded primary microRNA (miRNA) structures and mediate miRNA biogenesis in the nucleus (Han et al., 2004). In this study, we observed that DROSHA is one of the RBPs that induced the strongest reduction in RNA editing upon its knockdown in K562 cells, affecting ~30% of all testable editing sites (Fig. 2-1a, 2-3a).

To further confirm the involvement of DROSHA in regulating RNA editing, we carried out an immunoprecipitation (IP) experiment for this protein, followed by ADAR1 immunoblotting in K562 cells (Fig. 2-3c, Supplementary Fig. 2-8b). Note that DROSHA's endogenous expression in K562 cells is relatively low, which precluded its detection in the input sample. Nonetheless, the expression of DROSHA is clearly detectable in the IP samples. The presence of ADAR1 in DROSHA IP samples support that these two proteins interact with each other in K562 cells. In addition, RNase A treatment did not affect the observed interaction, suggesting that the interaction between ADAR1 and DROSHA does not depend on single-stranded RNA (the main target of RNase A). This result is consistent with our previous observation in HeLa cells (Bahn et al., 2015). Lastly, we confirmed that DROSHA knockdown did not induce observable change in ADAR1 protein expression in K562 cells (Supplementary Fig. 2-8c). Together, our data support that DROSHA is a strong enhancer of RNA editing, most likely by interacting with ADAR1 in the nucleus.

In addition to DROSHA, ILF2 and ILF3 (also called NF45 and NF90) are likely direct regulators of RNA editing via ADAR1 interactions. Both proteins are well-known RNA-dependent interacting proteins of ADAR1 (Guan et al., 2008; Nie et al., 2005). These proteins are localized in the nucleus (Parrott et al., 2005) and form complexes to regulate multiple aspects of RNA metabolism (Guan et al., 2008; Sakamoto et al., 2009; Wolkowicz & Cook, 2012). We observed that ILF2 knockdown in HepG2 cells caused up-regulation of editing levels in 13% of testable sites (Fig. 2-3a, Supplementary Fig. 2-8a). Its impact on editing in K562 cells is less pronounced than in HepG2 cells, but the same predominant direction of up-regulation was observed, which affected 5% of testable sites (Supplementary Fig. 2-8a). Similarly, ILF3 knockdown induced up-regulation of editing levels in 4% of sites in both cell lines

(Supplementary Fig. 2-8a). ILF3 (but not ILF2) eCLIP-seq data are available in both K562 and HepG2 cells. We observed that ILF3 binds significantly closer to differentially edited sites than expected by chance in both cell lines (Supplementary Fig. 2-8d). Together with previous findings that ILF2 and ILF3 form protein complexes and interact with ADAR1 (Guan et al., 2008; Nie et al., 2005), our data support a model where these proteins repress RNA editing by interacting with ADAR1 and binding close to ADAR1's target sequences. In addition, compared to IFL3, ILF2 may play a more direct role in influencing the editing outcomes of ADAR1.

2.3.5 Alu-binding RBPs as RNA editing regulators

Alu sequences, especially inverted Alu pairs, form dsRNA structures, which are the main ADAR1 substrates for RNA editing in human cells (Nishikura, 2010, 2016). One natural question is whether Alu-binding RBPs in general may regulate RNA editing by facilitating or inhibiting the interaction between ADAR1 and its dsRNA substrates. To address this question, we first analyzed all available ENCODE eCLIP-seq datasets to identify Alu-binding RBPs (Methods). We observed that a small number of RBPs are associated with high levels of Alu-binding, manifested as the high percentage of eCLIP peaks overlapping sense or antisense Alu elements (Fig. 2-4a). As expected, ADAR1 showed the highest level of Alu-binding among all proteins, despite the fact that ADAR1 CLIP was generated using a different cell line (U87MG cells) (Bahn et al., 2015). Notably, some proteins (such as hnRNP C) demonstrated a substantial bias for preference to either sense or antisense Alus, compared to the background sense/antisense Alu composition in expressed genes (~44% sense, ~56% antisense Alus in HepG2 and K562). In contrast, the sense/antisense Alu compositions of ILF3 and ADAR1 peaks are similar to the background, which may indicate that their binding specificity relies on RNA structures more than on the specific sequences.

Next, we asked whether the extent of Alu-binding correlates with the level of impact of each RBP on RNA editing. Surprisingly, there is little correlation between these two variables (Fig. 2-4b). In addition, the direction of editing changes upon RBP knockdown did not show consistent trend for these proteins (Supplementary Fig. 2-9). These observations suggest that Alu-binding alone is not sufficient for an RBP to influence ADAR1 editing. For a subset of these RBPs, eCLIP-seq data are available (Fig. 2-4c). We observed that the binding sites of ILF3 (Supplementary Fig. 2-8d), XRCC6 (Fig. 2-3b), TROVE2, AUH and PUS1 are significantly closer to differentially edited sites than expected by chance (Fig. 2-4c). Note that ILF3 and XRCC6 are also known ADAR1-interacting proteins (Supplementary Table 1), as described in the section above. The third gene, TROVE2, affects ~25% of all testable editing sites, with a bias toward up-regulated editing levels upon TROVE2 knockdown (Fig. 2-4b). We therefore further investigated the possible involvement of this protein in RNA editing regulation, as presented in the next section.

2.3.6 Alu-binding protein Ro60 (TROVE2) in RNA editing regulation

TROVE2 encodes for the protein Ro60, which is present in both the nucleus and cytoplasm of vertebrate cells (F. H. M. Simons et al., 1994). Anti-Ro60 antibodies occur in many patients with systemic lupus erythematosus (SLE), an autoimmune disease characterized by interferon activation, autoantibodies and multi-organ tissue destruction (Rahman & Isenberg, 2008). We analyzed RNA editing patterns using RNA-seq data derived from the blood samples of 99 SLE patients and 18 controls (Hung et al., 2015). Consistent with our findings in K562 cells (Fig. 2-5a), SLE samples, many with loss of Ro60 function, showed a predominant bias of upregulated RNA editing levels (Fig. 2-5b), which was also reported in a recent study (Roth et al., 2018). Moreover, consistent with interferon activation in SLE, we observed that ADAR1, particularly the interferon-inducible p150 isoform, was significantly overexpressed in SLE patients (Fig. 2-5c).

Based on the data above, the up-regulation of RNA editing in SLE may be due to one or both of the following mechanisms: (1) Up-regulated ADAR1 expression as a result of interferon response; (2) loss of Ro60, a repressor of RNA editing via Alu-binding. We observed a correlated pattern between the loss of Ro60 and up-regulation of ADAR1, both correlating with up-regulated RNA editing levels (Fig. 2-5d). To distinguish these two models, we obtained RNA-seq data from K562 cells following TROVE2 overexpression (OE). Compared to control cells, TROVE2-overexpressing cells showed a significant bias toward reduced editing levels (Fig. 2-5e), while no significant change in ADAR expression was observed (Fig. 2-5f). It should be noted that ADAR expression levels did not change upon TROVE2 knockdown in K562 cells either (Supplementary Fig. 2-10). Therefore, our data support a direct role of TROVE2 in repressing RNA editing, most likely by its interaction with Alu elements. Notably, the observed up-regulation of RNA editing in SLE patients likely reflects contribution by both loss of Ro60 function and ADAR1 up-regulation. As a result, the extent of RNA editing changes in SLE is much more pronounced than those observed in TROVE2 knockdown or OE cells (Fig. 2-5a, b, e), consistent with the lack of ADAR1 expression changes in the latter groups.

2.3.7 Cell type differences in RNA editing regulation

While examining RBPs in the above categories, we observed that the two cell lines, K562 and HepG2, were often associated with different RBPs that imposed the largest impact on editing. In total, 199 RBPs have RNA-seq data generated from both cell lines. However, only 3 RBPs, including ADAR1, were found to affect editing in ≥10% of all testable sites in both cell lines. Thus, we next examined whether the impact of RBPs on RNA editing is different between these two cell lines. For this analysis, we focused on the 35 RBPs with available data in both cell lines whose knockdown induced differential editing changes in ≥10% of the testable sites in at least one cell line (Supplementary Fig. 2-11a). It should be noted that, for the majority of these RBPs, their possible mechanisms of action on editing are not clear, including whether the observed editing changes are direct or indirect effects of RBP knockdown.

Since RNA editing is only observable in expressed RNA, one main factor underlying cell type-specificity in editing is the availability of the target transcripts. Thus, we first examined the between-cell-line overlap of differentially edited sites associated with each RBP in groups of genes stratified by their expression levels. As expected, the overlap of differentially edited sites is much higher in genes relatively highly expressed in both cell lines than those that are high in only one cell line (Fig. 2-6a). Thus, cell type-specific gene expression contributes to the observed differences in editing profiles between the two cell lines.

To further compare RBP knockdown-induced editing changes between the two cell lines, we identified the set of testable editing sites (total read coverage ≥ 5 per replicate and editing ratio ≥10% in either knockdown or control) common to both cell lines for each of the 35 RBPs. Indeed, common testable editing sites for most RBPs only constitute <50% of all testable sites in each cell line (Supplementary Fig. 2-11b), again reflecting differences in gene expression levels. Among these common sites, the fraction of differentially edited sites for each RBP is comparable to that among all testable sites in each cell line (Supplementary Fig. 2-11c vs. Supplementary Fig. 2-11a). Next, we evaluated the concordance of editing changes in the common sites between the two cell lines. To this end, we calculated the directional agreement scores as defined in Fig. 2-1b (Methods). A small number of RBPs, such as SRSF5, PABPC1 and PCBP1, had apparent opposite directions (positive or negative) in editing changes upon their knockdown in the two cell lines (Supplementary Fig. 2-11c), which resulted in negative agreement scores (Fig. 2-6b). Additionally, the majority of these RBPs had low agreement scores (e.g., 25 RBP with absolute score < 0.05, meaning that less than 5% of their differentially edited sites agree), including TROVE2 and TARDBP. Together, our analyses of common testable sites suggest that the (direct or indirect) impact of RBPs on RNA editing is different depending on the cell type.

2.4 DISCUSSION

We report a global study to identify RBPs as novel regulators of A-to-I editing in human cells. Using hundreds of RNA-seq datasets derived upon knockdown of individual RBPs, we investigated the influence of each RBP on RNA editing in K562 and HepG2 cells. Complemented by protein-RNA binding analyses using eCLIP-seq data and experimental validations, our study yielded a number of findings that help to fill in the significant gap in our understanding of additional regulators of RNA editing beyond the ADAR proteins.

An important observation of this study is that, among >200 RBPs analyzed in each cell line, only a small number of proteins caused substantial changes in RNA editing upon their knockdown. Since most RBPs contribute to multiple aspects of RNA processing and regulation (Gerstberger et al., 2014; Hentze et al., 2018), it is not surprising that loss of an RBP may cause a myriad of changes in gene expression, including RNA editing, directly or indirectly. Indeed, our data showed that for the vast majority of RBPs, there always existed a small fraction of editing sites with altered editing levels upon RBP knockdown. Such small degrees of changes are most likely consequences of alterations in other aspects of RNA regulation that sporadically and indirectly correlated with an observed RNA editing change. For example, changes in alternative splicing caused by RBP knockdown may affect the observed level of editing for certain sites in the intron. Therefore, we reason that direct regulators of RNA editing, those that affect the expression, function or protein-RNA interactions of ADAR proteins, should cause considerable editing changes that are relatively widespread. It should be noted that the reverse may not always hold – some proteins associated with large editing changes may not be direct regulators of RNA editing.

We focused on three categories of potential direct regulators of RNA editing: (1) proteins that regulate ADAR expression, (2) interact with ADAR1 or (3) bind to Alu elements. One immediate observation is that not all ADAR-interacting or Alu-binding proteins influence RNA editing significantly. Based on previous studies, ADAR1 interacts with many RBPs (Cusick et al., 2009). However, ADAR-interaction studies were carried out in specific cell types. It is possible that these protein-protein interactions are highly cell type-specific, which may explain the lack of RNA editing changes upon knockdown of many known ADAR-interacting proteins in K562 or HepG2. In addition, ADAR proteins were shown to affect post-transcriptional processes other than RNA editing (Bahn et al., 2015). Thus, another explanation for our observation is that some ADAR-interacting proteins may affect other aspects of ADAR1 function. Similarly, Alu-binding proteins may not affect RNA editing if their interactions with Alus are independent of ADAR1 or if they affect other aspects of ADAR1 function.

Within the above categories, we highlighted a few RBPs with significant impact on RNA editing, including DROSHA, ILF2/3, TARDBP and TROVE2. In our previous study (Bahn et al., 2015), we reported that ADAR1 interacts with DROSHA and enhances miRNA production in HeLa cells. Here, we confirmed the interaction between DROSHA and ADAR1 in K562 cells (Fig. 2-3c). This interaction is consistent with the observed significant reduction in RNA editing upon DROSHA knockdown in K562 cells (Fig. 2-3a) and the closer distance than expected by chance between DROSHA binding and the differentially edited sites (Fig. 2-3b). Together, these results suggest that the interaction between DROSHA and ADAR1 enhances the primary functions of these proteins reciprocally. Interestingly, another family of well-known ADAR1-interacting proteins, ILF2 and ILF3, were also reported to affect miRNA biogenesis (Nussbacher & Yeo, 2018; Sakamoto et al., 2009), the knockdown of which caused reduction of RNA editing. Therefore, RNA editing and miRNA biogenesis may be regulated by a common set of RBPs, likely due to the involvement of double-stranded RNA structures in both pathways.

Another protein with a significant role in editing regulation is Ro60 (encoded by the gene TROVE2). We observed that Ro60 binds to Alu elements and TROVE2 knockdown induced an increase in RNA editing for more than 1000 editing sites in K562 cells. This editing change was recapitulated in SLE patients with loss of Ro60 function. The patient samples showed a substantial change in RNA editing, to a higher extent than that observed in K562 cells, possibly due to the combined impact of ADAR1 p150 upregulation and Ro60 loss of function in SLE. Another disease-related protein with a novel involvement in regulating RNA editing is TDP-43 (encoded by the gene TARDBP). TDP-43 is a key player in the pathogenesis of Amyotrophic Lateral Sclerosis (ALS) (Warraich et al., 2010), a neurodegenerative disease caused by the aggregation of TDP-43 in the cytoplasm of neurons (Lagier-Tourenne et al., 2010). We observed that TDP-43 enhances ADAR1 transcription, thus influencing the global levels of RNA editing. For both SLE and ALS, further studies are needed to better understand how aberrant RNA editing profiles may contribute to the disease processes.

In addition to the RBPs highlighted above, there are a number of other proteins that were observed with extensive changes in RNA editing upon their knockdown. For example, FXR1 knockdown led to significant reduction of RNA editing in K562 cells (38% of editing sites were downregulated), but not HepG2 cells (Fig. 2-1). Indeed, our recent study reported that FXR1 reduces RNA editing in the brain and contributes to hypoediting in Autism brains (Tran et al., 2019). Thus, the role of FXR1 in RNA editing depends greatly on the cell type, as similarly observed for other proteins in this study (Fig. 2-6). In addition to direct regulators of RNA editing or ADAR1 expression, knockdown of some RBPs may cause an apparent editing change due to indirect mechanisms. One previously reported mechanism for such indirect effects is editing-dependent stabilization of mRNAs, mediated by the AGO2-miRNA targeting pathway (Brümmer et al., 2017).

Lastly, it should be noted that many other mechanisms may affect RNA editing, which are not studied in this work, including those executed by RNA helicases (Bratt & Öhman, 2003; Tariq et al., 2013), snoRNAs (Doe et al., 2009; Vitali et al., 2005) or proteins that affect ADAR protein modification, degradation or localization (Desterro et al., 2005; Marcucci et al., 2011; Tan et al., 2017).

2.5 METHODS

2.5.1 Datasets

Fastq files of RNA-seq data generated following RBP knockdown or control shRNA transfection were downloaded from the ENCODE data portal (Nostrand et al., 2017) (encodeproject.org). Data released between October 2014 and January 2017 is included in this study. These data were generated in 24 and 26 batches in the K562 and HepG2 cell lines, respectively.

RNA-seq reads were aligned using RASER v0.52 (Ahn & Xiao, 2015) against the human genome (hg19) and Ensembl transcriptome (Hubbard, 2002) (Release 75), with the parameters m = 0.05 and b = 0.03. Only uniquely mapped reads were retained for further analysis. Duplicated reads (those with identical start and end coordinates) were removed from the alignment files.

2.5.2 Identification and analysis of RNA editing

Mismatches in the RNA-seq reads were first examined to ensure the overall quality of the mismatch calls (Bahn et al., 2012). This step removes likely sequencing errors based on base call quality, mismatch nucleotide changes and mismatch position in the reads. We then filtered these mismatches by removing those located in homopolymers, splice sites, simple repeats, and those whose read coverage demonstrated a strand bias (Lee et al., 2013). These sites were further processed using GIREMI (Q. Zhang & Xiao, 2015) to obtain high-confidence editing sites. GIREMI identifies editing sites based on the mutual information between editing sites and/or SNPs. Since the RNA-seq experiments were conducted in multiple batches, we designed a scheme to reduce the potential batch effects. Specifically, within each batch, multiple RBP knockdown experiments and one control shRNA experiment (2 replicates each) were carried out. We assume that only a minority of RBPs, if any, in a batch regulates the editing of a particular site. Based on this assumption, for each editing site identified in any dataset of a batch, we defined the control editing level as the average of its editing level in all RBP knockdown and control experiments in the same batch. This procedure was omitted for a small number of batches where only one RBP was included. Based on clustering results (Supplementary Fig. 2-5, Supplementary 6), this method effectively removed batch effects.

To identify differentially edited sites upon knockdown of an RBP, the editing level of each editing site was compared to the above averaged editing level in the same batch. Since each knockdown experiment had two biological replicates, we estimated the expected variance in the editing level from the two replicates for each editing site using a method similar as in the BEAPR package (E. W. Yang et al., 2019). Significant differentially edited sites were identified using a normal distribution parameterized by the mean editing level between two replicates and the expected variance calculated above. The FDR was calculated using the Benjamini-Hochberg method (Benjamini & Hochberg, 1995). Differentially edited sites were called by requiring FDR ≤10% and the absolute change in the editing level between knockdown and control ≥ 5%. The code for the identification of differentially edited sites is available at https://github.com/gxiaolab/RNA_editing/tree/master/RBP_regulation, together with all differentially edited sites identified in this study.

2.5.3 Overlap scores of differential editing of two RBPs

We calculated overlap scores to represent the degrees of overlap among differentially edited sites associated with a pair of RBPs. For each pair of RBPs, two overlap scores were defined, represented by two links in the CIRCOS plots (Fig. 2-1a). The scores correspond to the thickness and color of the links in the plots. To calculate these scores, we first obtained the number of shared differentially edited sites of two RBPs. The numbers of differentially edited sites with the same (n_1) or opposite (n_2) directions in their changes of editing levels upon knockdown were obtained. Pairs of RBPs with less than 20 total shared differentially edited sites were not considered (thus with no links in the plot). Then, we obtained the number (t) of shared testable sites in the datasets of the two RBPs. The ratios n_1/t and n_2/t were then calculated, where n_2/t was reported as $-n_2/t$ to represent the opposite directions in editing changes. The final overlap scores are defined as the Z-score of these ratios across all RBP pairs for each cell line.

2.5.4 Global direction of editing regulation

We tested whether there exists a significant bias in the direction of editing changes (higher or lower relative to controls) caused by the knockdown of an RBP using a bootstrap sampling approach (Figs. 2-2, 2-3 and 2-5). For each RBP-knockdown sample, we obtained the total number of differentially edited sites (n) and the fraction of these differentially edited sites with increased editing level upon knockdown (r). We then randomly sampled n sites from all testable sites of the same RBP-knockdown dataset and calculated a similar fraction (r_i^*). We repeated this random sampling process 100,000 times to obtain an empirical distribution of the ratios: $r^* = r_1^*, r_2^*, \dots, r_{100,000}^*$. The z-score of r was therefore defined as $z = \frac{r - \widehat{r^*}}{\sigma_{r*}}$ where $\widehat{r^*}$ and σ_{r*} were the mean and standard deviation of r^*, respectively. Finally, the empirical p value of r was calculated by comparing to r^*.

For TARDBP (Fig. 2-2c), we additionally tested the significance of change in the global editing levels for all testable sites. We performed a similar test as described above, but by randomly sampling sites from all testable sites of all RBP knockdown datasets in the same batch as TARDBP.

2.5.5 WGCNA clustering

To examine whether subsets of RBPs function similarly in regulating RNA editing, we carried out a clustering analysis of RBPs using the Weighted Gene Co-expression Network Analysis (Langfelder & Horvath, 2008). This method finds networks (modules) of nodes based on their topological overlap. For each cell line, the nodes of the WGNCA network consisted of all the RBPs with knockdown data. The edge scores between the nodes (i.e. RBPs) were calculated using pairwise correlation (bicorrelation as recommended by WGCNA) between their differential editing levels between knockdown and controls. We employed WGCNA to create signed networks, which required a soft threshold of 12 to satisfy scale-free topology (B. Zhang & Horvath, 2005). Modules in the resulting dendrograms were then examined manually (Supplementary Fig. 2-6).

2.5.6 eCLIP-seq analysis

eCLIP-seq data of 126 and 109 RBPs in K562 and HepG2 cells, respectively, were adapter-trimmed and de-multiplexed (Van Nostrand et al., 2016). For each RBP, we obtained eCLIP-seq data from two biological replicates and one size-matched (SM) Input control (Van Nostrand et al., 2016).

To accommodate potential Alu-binding proteins whose eCLIP reads may not align uniquely, the eCLIP data were analyzed using a step-wise mapping procedure (Bahn et al., 2015). Specifically, the reads were aligned to rRNA sequences first. This step helps to control for spurious artifacts possibly caused by reads derived from rRNA. Those that did not align to rRNAs were retained and aligned to the Alu sequences located in RefSeq genes. This step allows up to 100 multiple alignments per read, maximizing the number of reads that map to Alu elements. Subsequently, reads that did not map to Alu sequences were aligned to the human genome (hg19), where only uniquely mapped reads were retained. All the alignments were performed by the STAR aligner(Dobin et al., 2013) with ENCODE standard parameters (as specified in the STAR manual). All alignments were required to be end-to-end without soft-clipping. eCLIP peaks were called using a Poisson model (Bahn et al., 2015) by requiring a Bonferroni-corrected p value cutoff of 0.01.

Next, we examined whether the distance between differentially edited sites upon an RBP knockdown and the eCLIP peaks of the RBP is significantly closer than expected by chance. For each differentially edited site, we calculated its distance to the closest eCLIP peak within the same gene. differentially edited sites in genes that do not have an eCLIP peak were discarded. As control sites, we used known editing sites from the REDIportal database (Picardi et al., 2017) that satisfy the following: (1) ≥15 combined total read coverage from the two replicates; (2) located in the same gene as the differentially edited site; (3) not identified with edited reads in any dataset of our study. For ADAR1, we used non-REDIportal A's as random controls, instead of known editing sites.

For each RBP, we randomly selected the same number of control sites as that of differentially edited sites and calculated the distance between a control site and the closet eCLIP peak within the same gene. We repeated this process 200 times to generate 200 sets of controls. The distances of each set of controls to eCLIP peaks were visualized via empirical cumulated distribution function (eCDF), similarly for the distances of the actual differentially edited sites to eCLIP peaks. Next, we calculated the area under the curve (AUC) of each distance eCDF and compared the AUC corresponding to differentially edited sites and those resulted from the 200 sets of control sites. It is expected that smaller distances lead to larger AUCs. Thus, to determine whether the differentially edited sites were significantly closer to eCLIP peaks than expected by chance, we calculated the two-sided p-value by fitting the AUC values of the controls with a normal distribution. In addition, a fold-change (FC) was calculated as the ratio between the AUC associated with differentially edited sites and the mean AUC of the control eCDFs.

2.5.7 Directional agreement score

For each pair of RBPs tested in the same cell line, we took the union of their associated differentially edited sites and further retained only those sites that are testable in both RBP knockdown datasets. Testable sites were defined as those with ≥5 total reads per replicate, and with ≥10% editing level in either knockdown or control. Using these editing sites, we asked whether the directions of editing changes upon knockdown of the two RBPs are concordant by calculating the directional agreement score. Specifically, for each of the above differentially edited site, we labeled it as + or - if its change in editing level upon RBP knockdown is a positive or a negative value, respectively. For sites with the same label for both RBPs, a +1 agreement score was assigned. Otherwise, a score of -1 was given. If the editing site is differentially edited in only one of the two RBP knockdowns, a score 0 was given. The final directional agreement score of a pair of RBPs is defined as the average value of the score of each included editing site in this analysis.

The directional agreement score of the same RBP between K562 and HepG2 was calculated similarly.

2.5.8 Co-immunoprecipitation

Cells were maintained with DMEM supplemented with 10% FBS and 100 U ml^{-1} penicillin/streptomycin at 37 °C and 5% CO_2. Ten million cells were collected and lysed in 1 ml non-denaturing lysis buffer at pH 8.0, containing 20 mM Tris-HCl, 125 mM NaCl, 1% NP-40, and 2 mM EDTA supplemented with complete protease inhibitor cocktail. Extracted proteins were incubated overnight with DROSHA antibody (Bethyl Laboratories, A301-886A) at 4 °C; precipitation of the immune complexes was performed with Dynabeads Protein G (Thermo Fisher Scientific, 1003D) for 4 hours at 4 °C, according to the manufacturer's instructions. After immunoprecipitation, the beads were washed three times with the lysis buffer at 4 °C and eluted from the Dynabeads using elute buffer (0.2 M glycine, at pH 2.8). Twenty microliters were loaded onto the gel and the samples were processed by SDS-polyacrylamide gel electrophoresis (SDS-PAGE) and analyzed by Western blot. The following antibodies were used for the Western blots: ADAR1 antibody (Santa Cruz, sc-73408) and DROSHA antibody (Bethyl Laboratories, A301-886A). The HRP-linked secondary antibodies were used and the blots were visualized with the ECL kit (GE, RPN2232).

2.5.9 Constructs, transfection, luciferase reporter assay

TDP-43 ChIP peak regions were cloned into a firefly luciferase reporter pGL3 vectors (Promega). The pSV40-Renilla vector (Promega) encoding the Renilla luciferase reporter gene Rluc (*Renilla reniformis*) was used for transfection efficiency. Transfections were performed with the use of Lipofectamine 3000 (Invitrogen). HepG2 cells were seeded into 12 well plates at a density of 2.0 x 10^5 cells per well the day before transfection. For each well of cells 1.0 μg of the pGL3 constructs were co-transfected with 0.1 μg of the pSV40-Renilla vectors. The transfected cells were collected after 48 h. Luciferase activities were measured with the Dual-

Luciferase Reporter Assay System (Promega, E1910). To normalize for transfection efficiency, the reporter activity was expressed as the ratio of firefly activity to renilla activity. For each construct, three independent experiments were performed in triplicate.

2.6 CODE AVAILABILITY

Scripts for differential editing analysis (and related results) are available at https://github.com/gxiaolab/RNA_editing/tree/master/RBP_regulation.

2.7 DATA AVAILABILITY

All data sets used in this study can be obtained from the ENCODE project website at http://www.encodeproject.org. We used shRNA RNA-seq and eCLIP-Seq datasets in HepG2 and K562 cells with release dates between October 2014 and January 2017. The data underlying the main figures are available in Supplementary Data 1.

2.8 ACKNOWLEDGEMENTS

We thank members of the Xiao, Graveley and Yeo laboratories for helpful discussions and comments on this work. This work was funded by the National Human Genome Research Institute ENCODE Project, grants U01HG009417 to X.X. and U54HG007005 to B.R.G. and G.W.Y. X.X. was partially supported by grants from the NIH (R01HG006264, R01AG056476). S.S.T. was supported by the NIH Training Grant in Genomic Analysis and Interpretation T32HG002536 and the Eureka fellowship of UCLA. E.L.V.N. is a Merck Fellow of the Damon Runyon Cancer Research Foundation (DRG-2172-13) and is supported by a K99 grant from the NIH (HG009530). G.A.P. is supported by the NSF graduate research fellowship.

2.9 AUTHOR CONTRIBUTIONS

G.Q.V., S.S.T., E.W.Y, A.B, X.W., E.L.V.N, and G.A.P conducted the bioinformatic analyses. H.I.J, J.H.B., and L.Z. conducted the molecular biology experiments. X.X. designed and supervised the study in collaboration with G.W.Y. and B.R.G. All authors contributed to the writing of the paper.

2.10 COMPETING INTERESTS

G.W.Y, is a co-founder of Locana and Eclipse Bioinnovations and member of the scientific advisory boards of Locana, Eclipse Bioinnovations and Aquinnah Pharmaceuticals. E.V.N, is a co-founder and member of the scientific advisory board of Eclipse BioInnovations. The terms of these arrangements have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies. The rest of the authors declare no competing interests.

2.11 FIGURES

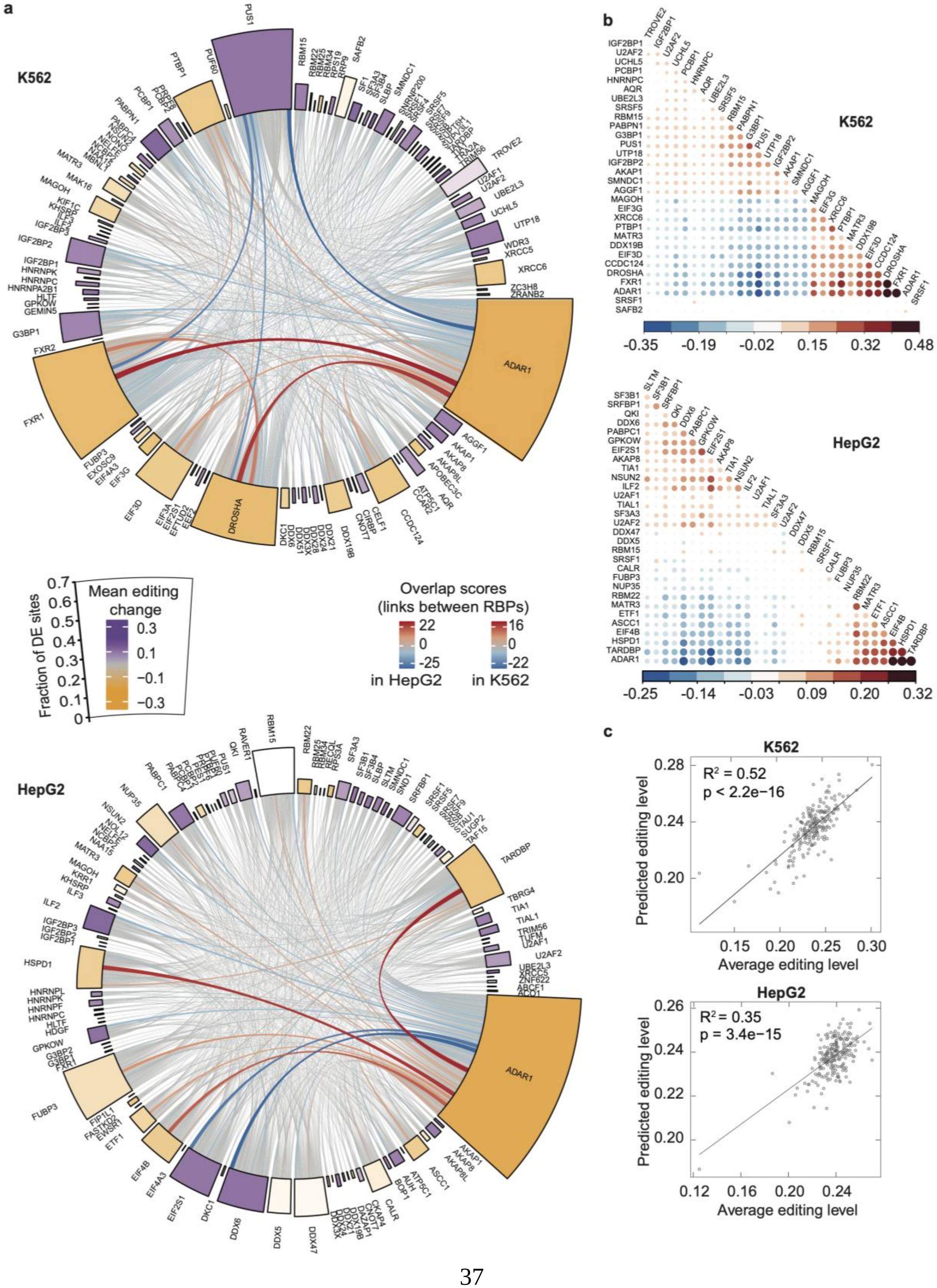

a
K562
PUS1
PUF60
PTBP1
RBM15
RBM22
RBM25
RBM34
RPS19
RRP9
SAFB2
SF1
SF3A3
SF3B4
SLBP
SMNDC1
SNRNP200
SRSF5
TROVE2
U2AF1
U2AF2
UBE2L3
UCHL5
UTP18
WDR3
XRCC5
XRCC6
ZC3H8
ZRANB2
ADAR1
AGGF1
AKAP1
AKAP8
AKAP8L
APOBEC3C
AQR
ATP5C1
CCAR2
CCDC124
CELF1
CIRBP
CNOT7
DDX19B
DDX21
DKC1
DDX6
DROSHA
EEF2
EFTUD2
EIF2S1
EIF3A
EIF3D
EIF3G
EIF4A3
EXOSC9
FUBP3
FXR1
FXR2
G3BP1
GEMIN5
GPKOW
HLTF
HNRNPA2B1
HNRNPC
HNRNPK
IGF2BP1
IGF2BP2
IGF2BP3
ILF2
ILF3
KHSRP
KIF1C
MAGOH
MAK16
MATR3
PABPN1
PCBP1
PCBP2
PRPF8
Fraction of DE sites
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Mean editing change
0.3
0.1
−0.1
−0.3
Overlap scores (links between RBPs)
22
0
-25
in HepG2
16
0
-22
in K562
HepG2
RBM15
RBM22
RBM25
RBM34
RECQL
RPS3A
SF3A3
SF3B1
SF3B4
SLBP
SLTM
SMNDC1
SND1
SRFBP1
SRSF1
SRSF5
SRSF7
SRSF9
STAU1
SUGP2
TAF15
TARDBP
TBRG4
TIA1
TIAL1
TRIM56
TUFM
U2AF1
U2AF2
UBE2L3
XRCC5
ZNF622
ABCF1
ACO1
ADAR1
AKAP1
AKAP8
AKAP8L
ASCC1
ATP5C1
AUH
BOP1
CALR
CKAP4
CNOT7
DAZAP1
DDX19B
DDX21
DDX3X
DDX47
DDX5
DDX6
DKC1
EIF2S1
EIF4A3
EIF4B
ETF1
EWSR1
FASTKD2
FIP1L1
FUBP3
FXR1
G3BP1
G3BP2
GPKOW
HDGF
HLTF
HNRNPC
HNRNPF
HNRNPK
HNRNPL
HSPD1
IGF2BP1
IGF2BP2
IGF2BP3
ILF2
ILF3
KHSRP
KRR1
MAGOH
MATR3
NAA15
NCBP2
NOL12
NSUN2
NUP35
PABPC1
PCBP1
PCBP2
PUM1
PUS1
QKI
RAVER1
b
K562
TROVE2
IGF2BP1
U2AF2
UCHL5
PCBP1
HNRNPC
AQR
UBE2L3
SRSF5
RBM15
PABPN1
G3BP1
PUS1
UTP18
IGF2BP2
AKAP1
SMNDC1
AGGF1
MAGOH
EIF3G
XRCC6
PTBP1
MATR3
DDX19B
EIF3D
CCDC124
DROSHA
FXR1
ADAR1
SRSF1
SAFB2
-0.35
-0.19
-0.02
0.15
0.32
0.48
HepG2
SLTM
SF3B1
SRFBP1
QKI
DDX6
PABPC1
GPKOW
EIF2S1
AKAP8
TIA1
NSUN2
ILF2
U2AF1
TIAL1
SF3A3
U2AF2
DDX47
DDX5
RBM15
SRSF1
CALR
FUBP3
NUP35
RBM22
MATR3
ETF1
ASCC1
EIF4B
HSPD1
TARDBP
ADAR1
-0.25
-0.14
-0.03
0.09
0.20
0.32
c
K562
$R^2 = 0.52$
$p < 2.2e{-16}$
Predicted editing level
0.28
0.24
0.20
Average editing level
0.15
0.20
0.25
0.30
HepG2
$R^2 = 0.35$
$p = 3.4e{-15}$
Predicted editing level
0.26
0.24
0.22
0.20
Average editing level
0.12
0.16
0.20
0.24

Figure 2-1. Global overview of RNA editing regulation by RBPs. **(a)** CIRCOS plot illustrating differential editing (DE) patterns upon knockdown (KD) of each RBP in each cell line. For each cell line, 100 RBPs are shown as those with the highest percentage of differentially edited sites among all testable sites (represented by the height of the outer box). The color of the box denotes the average changes in editing levels of differentially edited sites relative to controls upon RBP knockdown. For the links between RBPs, the thickness and color of the line both reflect an overlap score, that is, the fraction of shared differentially edited sites among shared testable sites between two RBPs (Methods). Positive overlap scores reflect concordant direction in the editing changes induced by the pair of RBPs, while negative values represent the opposite. The width of the box is set automatically to accommodate all the links associated with each box. ADAR1 has the greatest impact on global editing in both cell lines, which serves as a positive control. **(b)** Hierarchical clustering of RBPs using pair-wise directional agreement scores (Methods). The top 31 RBPs with the highest percentage of differentially edited sites among all testable sites per cell line are shown. The size of the dot and its color both reflect the directional agreement score. These RBPs cluster in two main groups composed of those associated with positive or negative editing changes upon their knockdown. **(c)** Correlation between actual average editing levels per sample and predicted RNA editing levels calculated via linear regression of gene expression levels of the top 15 RBPs in each cell line (including ADAR1). Each dot represents one sample and all RBP knockdown samples are included. R^2 and p values were calculated by Pearson correlation. The expression of these RBPs explained about 35% and 52% of the total variance in editing in HepG2 and K562 cells, respectively.

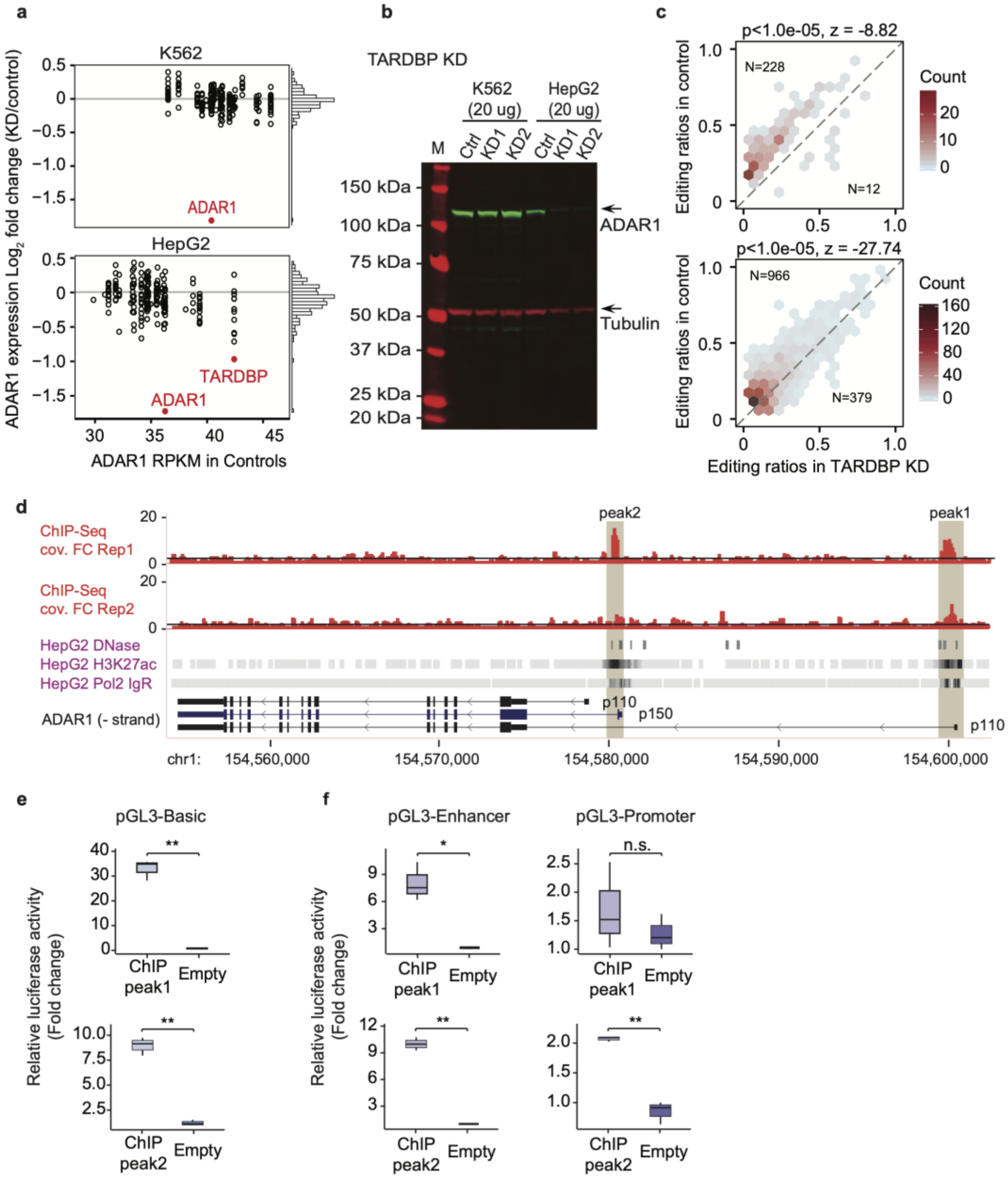


Figure 2-2. Regulation of ADAR1 expression by TARDBP. (a) Differential ADAR1 mRNA expression upon RBP knockdown in K562 and HepG2 cells compared to controls. Histograms on the right show the distributions of log-fold-change (LFC) values. TARDBP is the only RBP

whose knockdown caused differential expression of ADAR1 (absolute value of LFC knockdown/Control ≥1 and DESeq q-value < 10^{-9}). **(b)** Western blot of shRNA-mediated TARDBP knockdown (KD) and control (Ctrl) cells. Blots were probed with antibodies detecting ADAR1 and Tubulin (as a control). TARDBP knockdown significantly reduced ADAR1 expression in HepG2 cells only. **(c)** Editing ratios in TARDBP knockdown samples in HepG2 compared to their respective control values. The numbers (N) of editing sites with decreased and increased editing upon TARDBP knockdown are shown. The top panel includes only differentially edited sites and the bottom panel shows all testable editing sites. P-values and z scores were calculated using a bootstrap sampling strategy to evaluate the bias in the numbers of up- vs. down-regulated editing sites (Methods). Knockdown of TARDBP caused a global downregulation in editing levels. **(d)** TDP-43 (encoded by the TARDBP gene) ChIP-seq peaks in HepG2 cells overlapping ADAR1 transcripts. Fold change of read coverage relative to input control is shown for two replicated experiments (Rep1 and 2). Black line denotes fold change = 2. Three representative ADAR1 transcripts (RefSeq annotation) are shown, coding for the p110 and p150 forms of the ADAR1 protein. In addition, H3K27Ac, DNase hypersensitivity data in HepG2 cells and RNA polymerase 2 binding data generated by the ENCODE consortium are shown. **(e and f)** Luciferase assays of a series of pGL3 constructs containing the TDP-43 ChIP peak regions. The ChIP sequence for peak1 (top) and peak2 (bottom) were built into the construct. Basic pGL3 construct **(e)** and pGL3-Enhancer (lacking the SV40 promoter) and pGL3-Promoter (lacking the SV40 enhancer) constructs **(f)** are shown. The ratio of firefly luciferase to renilla luciferase was calculated for each experiment. The mean value (3 replicates) for each test construct was normalized to the activity of the empty vector. (Unpaired, two-tailed Student's t-test, *P < 0.05, **P<0.01, n.s.: not significant).

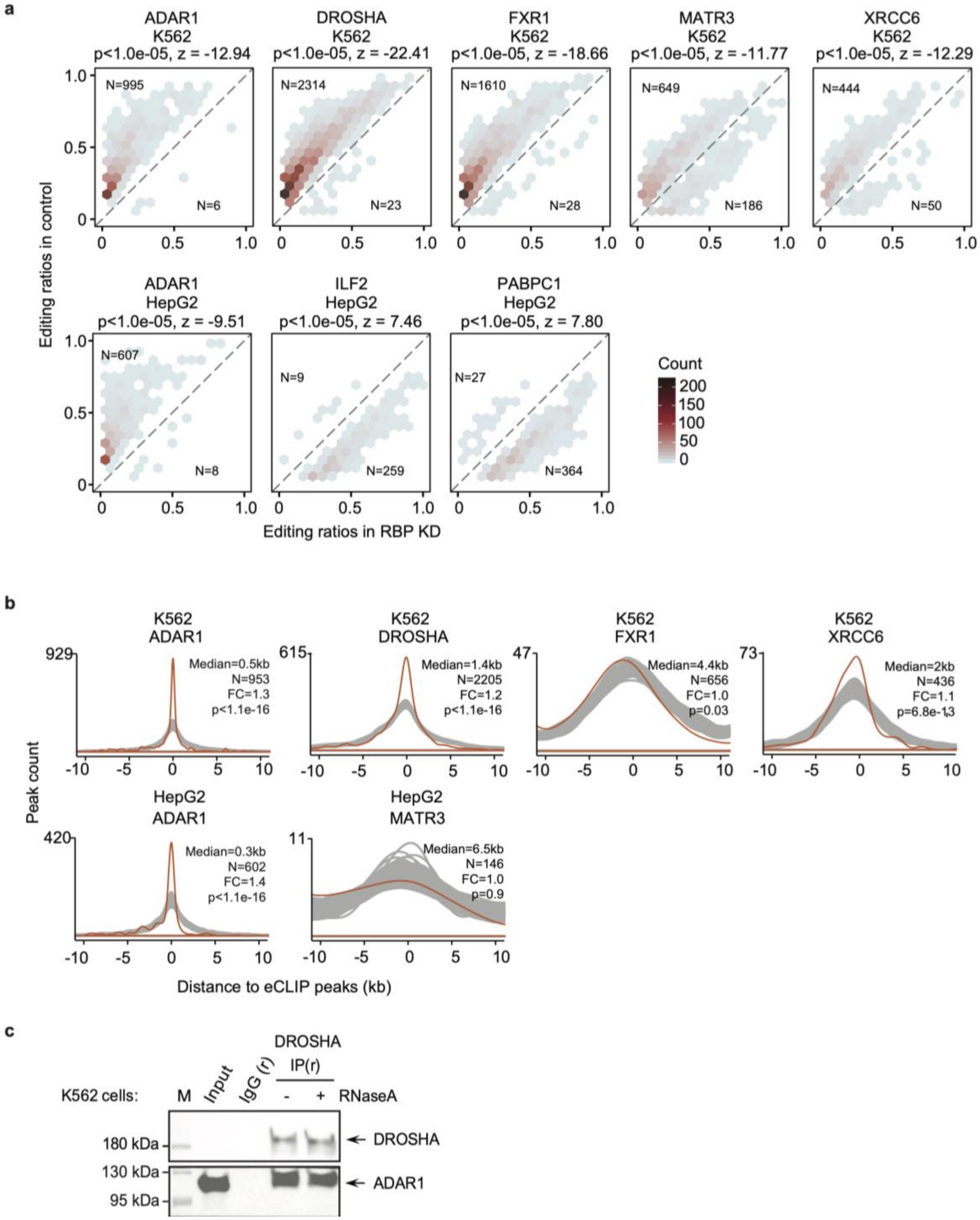


Figure 2-3. Regulation of RNA editing by ADAR1-interacting RBPs. (a) Editing ratios of differentially edited sites in RBP knockdown samples and their corresponding controls for ADAR1 and ADAR1-interacting proteins, similar as **Fig. 2-2c**. **(b)** Distribution of distances

between eCLIP peaks and their closest differentially edited sites (orange) or control sites (gray) (Methods). The median distance is shown. N represents the number of differentially edited sites used in the calculation. Calculations of FC (fold change) and p values are described in Methods. ADAR1, DROSHA and XRCC6 bind significantly closer to differentially edited sites than to control sites. **(c)** Co-IP experiment with and without RNase A treatment in K562 cells demonstrates interaction between ADAR1 and DROSHA. IP was performed using DROSHA antibody or corresponding rabbit (r) isotype IgG. Blots for IP samples were probed with antibodies detecting ADAR1 and DROSHA.

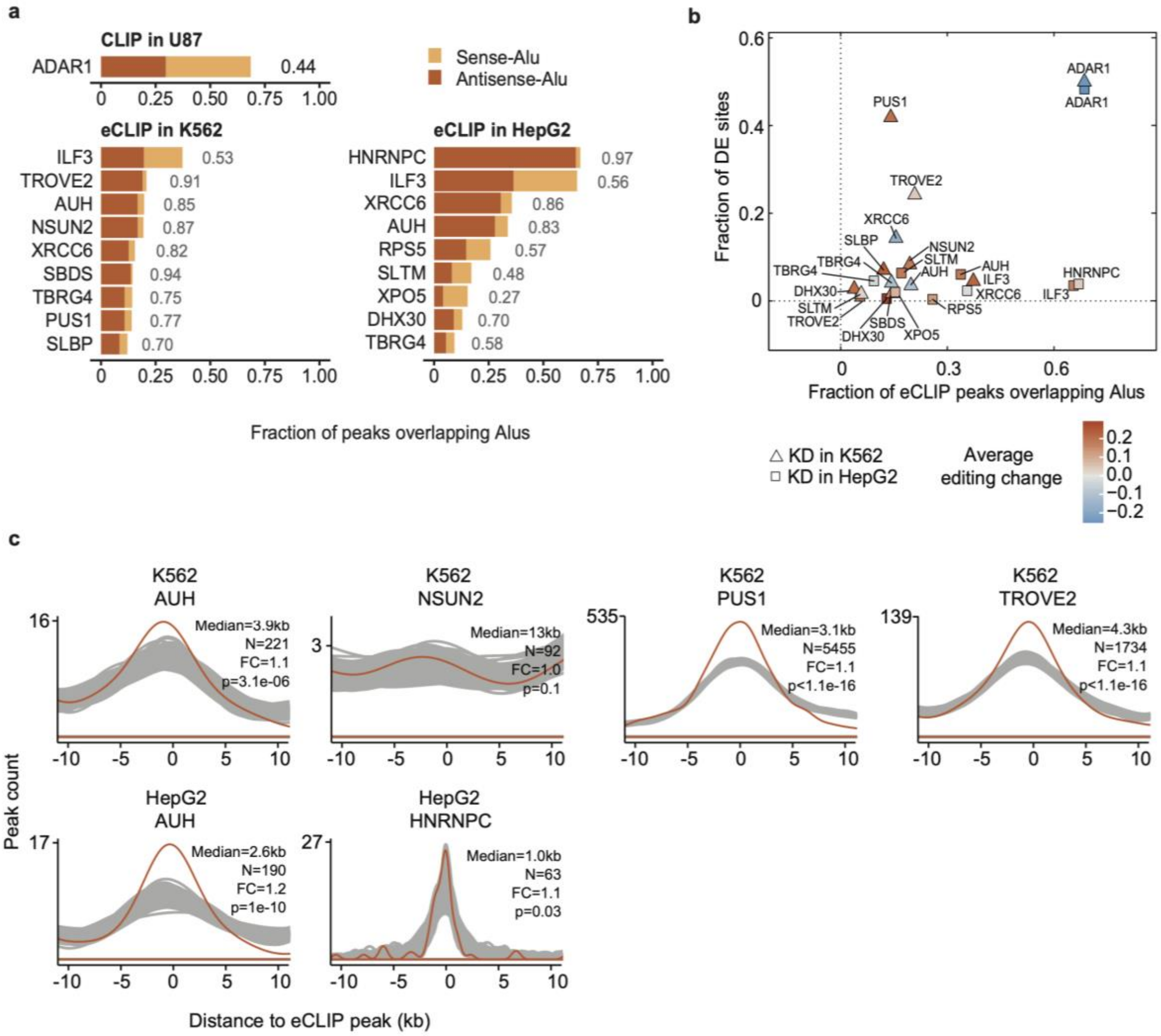


Figure 2-4. Regulation of RNA editing by Alu-binding RBPs. (a) Top RBPs ranked by their fractions of eCLIP peaks that overlap Alu elements. RBPs were required to have both RNA-seq and eCLIP-seq data in the same cell line. ADAR CLIP-seq data was obtained in the U87MG cells. Sense and antisense Alus denote Alu elements with consensus sequence on the same or opposite strand as the CLIP or eCLIP peaks, respectively. The fraction of antisense Alus among all Alu-overlapping peaks is shown for each protein (next to the bar). **(b)** For the RBPs in **(a)**, comparison between the fraction of differentially edited sites among all testable sites and the fraction of eCLIP peaks that overlap Alu elements. The dots are colored according to the average

editing changes of differentially edited sites (knockdown - control). There is no direct correlation between Alu-binding frequency and editing regulation for the tested RBPs. **(c)** Distribution of distances between eCLIP peaks and their closest differentially edited sites (orange) or control sites (gray) (Methods), for RBPs in **(a)** with eCLIP data, calculated similarly as in **Fig. 2-3b**. AUH, PUS1 and TROVE2 bind significantly closer to differentially edited sites than to control sites.

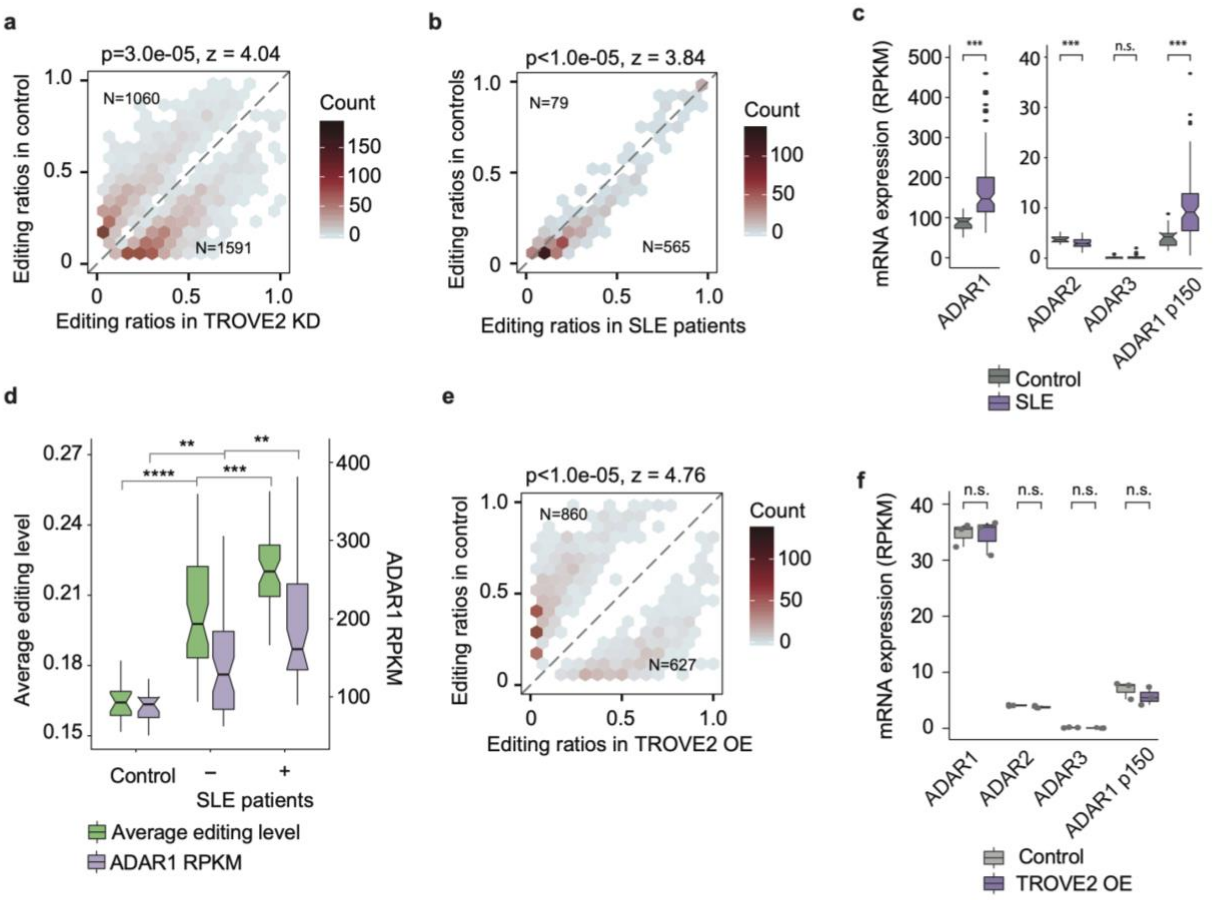


Figure 2-5. Regulation of RNA editing by TROVE2. (a) Editing ratios of differentially edited sites in TROVE2 knockdown samples and the corresponding controls in K562 cells, similar as **Fig. 2-2c**. **(b)** Similar as **(a)**, editing ratios of differentially edited sites in blood samples of SLE patients and control subjects. The data demonstrate reduced editing levels in SLE patients. **(c)** mRNA expression of ADAR1, ADAR2, ADAR3 and the ADAR1 p150 isoform in blood samples of SLE patients and control subjects (18 Controls, 99 SLE Patients, P values were calculated using two-tailed Wilcoxon ranksum test, ***P<0.001,****P<10^{-5}). **(d)** Average editing levels and ADAR1 mRNA expression (RPKM) of control subjects, SLE patients with medium to high Ro60 antibody levels (+) and SLE patients without detectable Ro60 antibody levels (-). Increasing levels of Ro60 antibody correlated with higher ADAR1 expression and

editing levels (P-values were calculated using a two-tailed Wilcoxon rank sum test and corrected for multiple testing using Bonferroni correction, **P<0.01, ***P<0.001, ****P<10^{-8}). **(e)** Similar as **(a)**, editing levels of differentially edited sites in TROVE2 overexpression (OE) samples and the corresponding controls in K562 cells. OE of TROVE2 led to a bias toward lower editing levels. **(f)** mRNA expression of ADAR genes in TROVE2 OE samples in K562 cells (N = 3 biological replicates, P values were calculated using a two-tailed Wilcoxon ranksum test). No significant change was observed in ADAR expression in TROVE2 OE samples compared to controls.

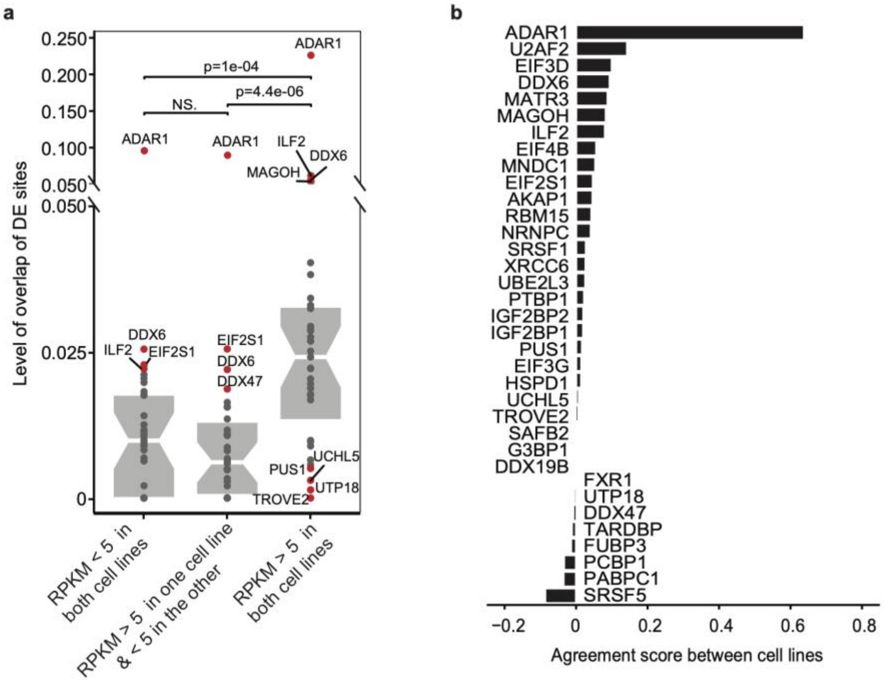
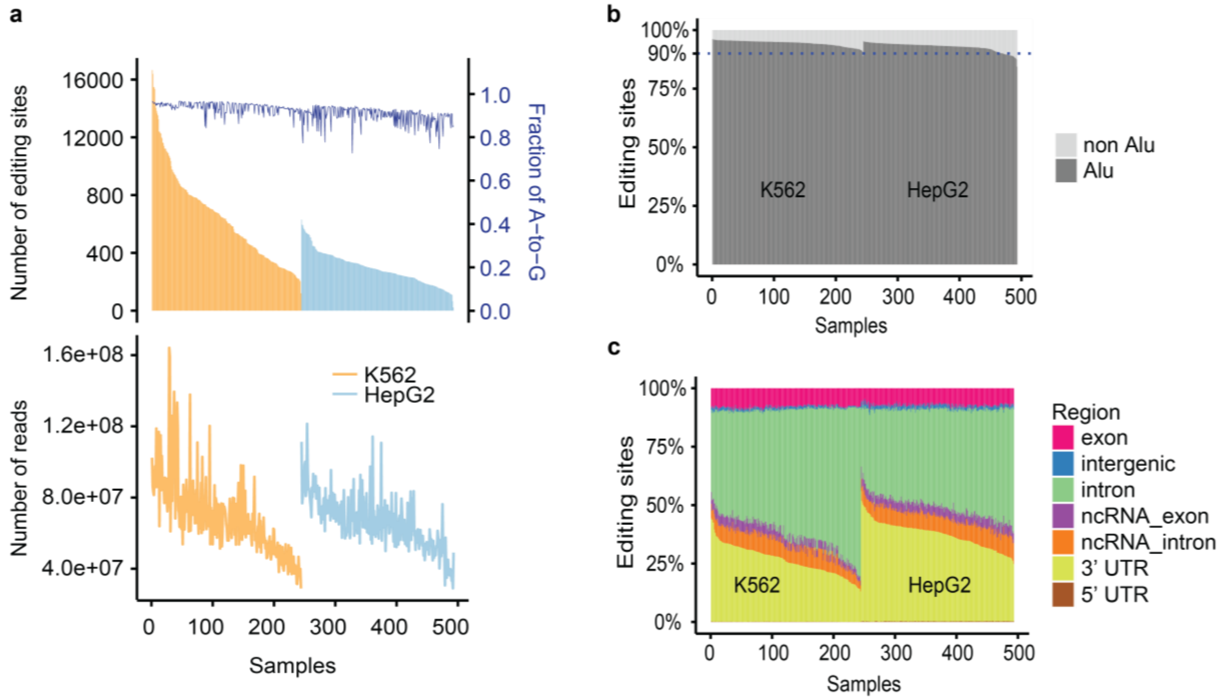
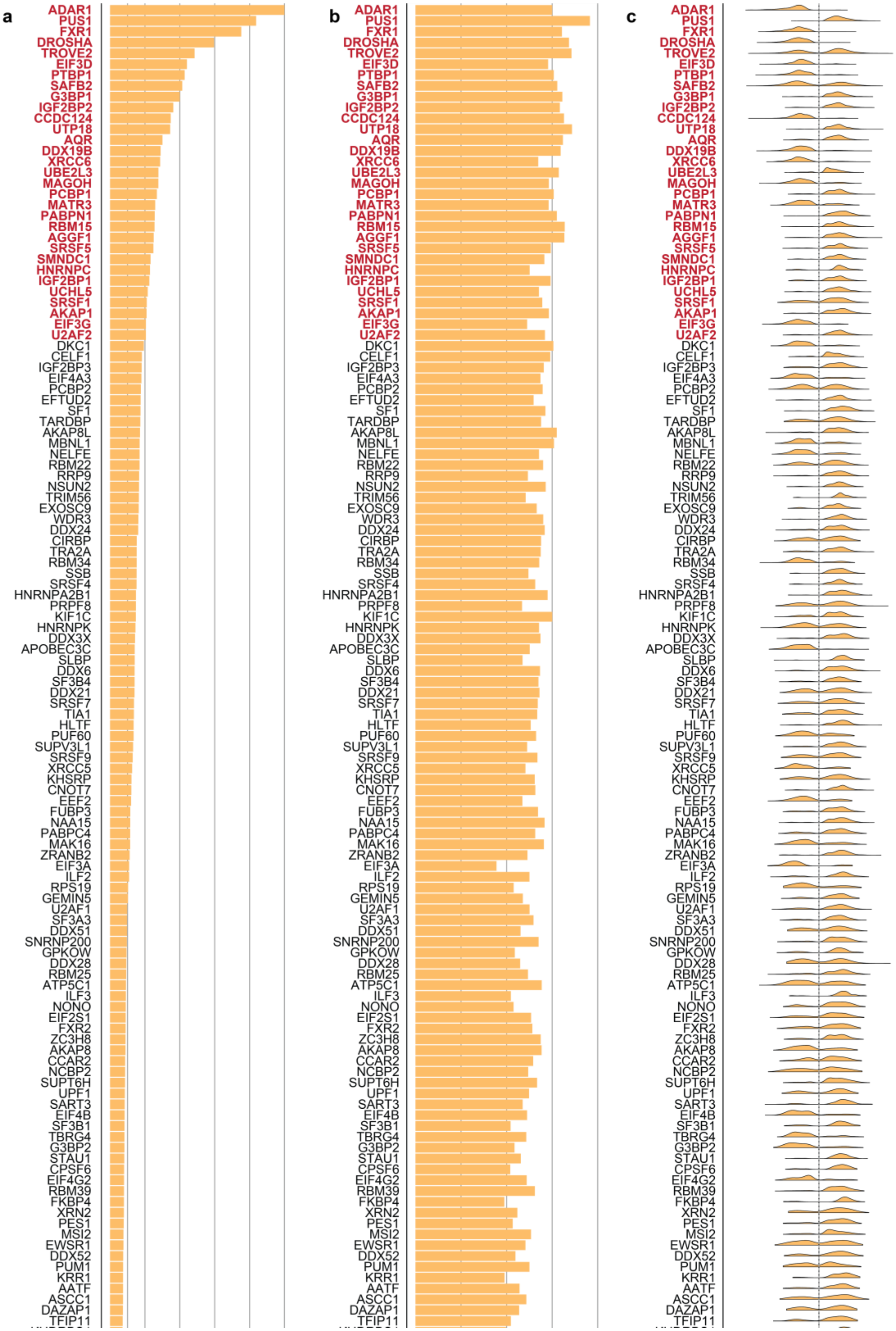


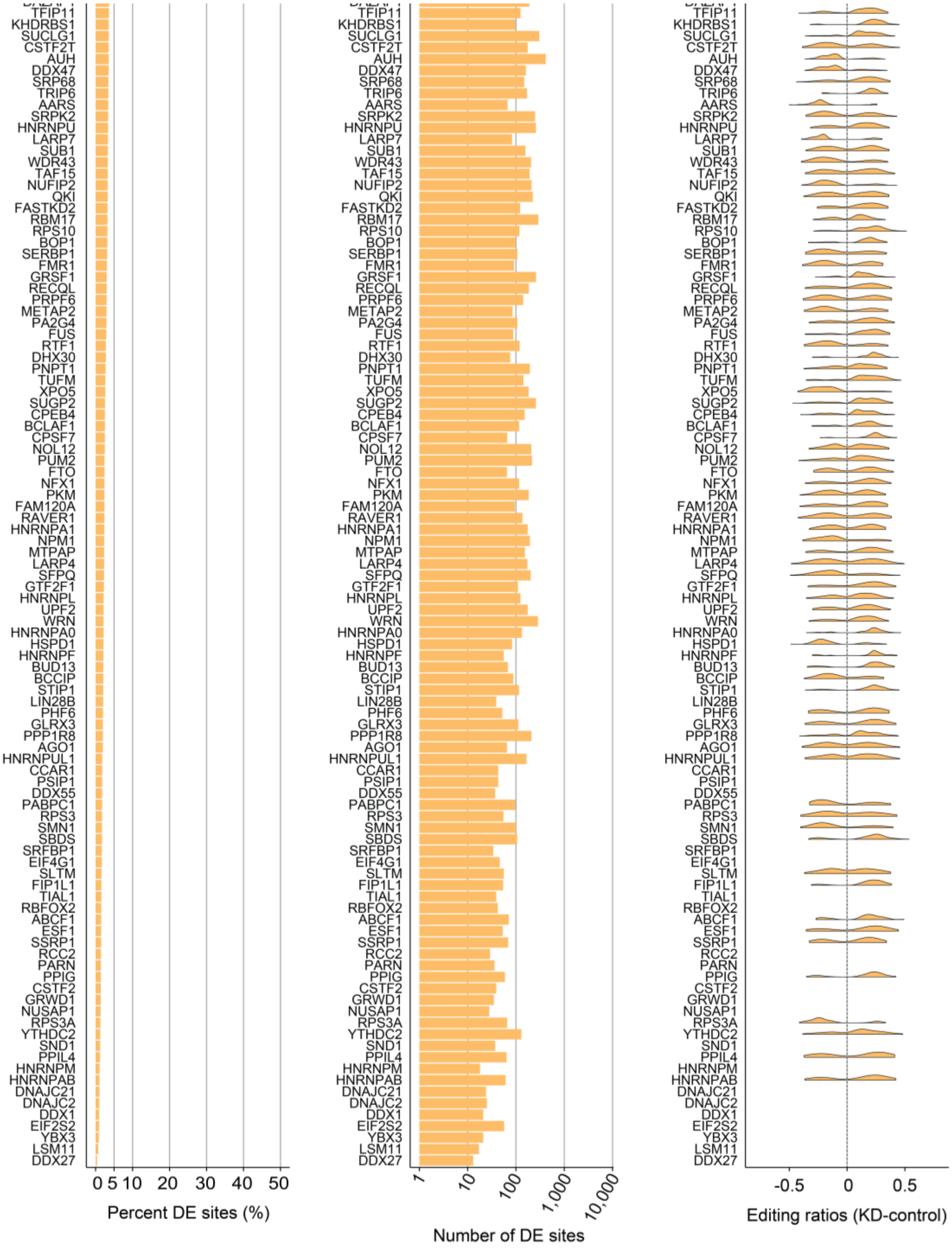
Figure 2-6. Cell type-specific impact of RBPs on RNA editing. Overlap of differentially edited sites between K562 and HepG2 cells for RBPs with at least 10% testable sites being differentially edited sites in at least one cell line. The differentially edited sites are separated into 3 groups based on the expression levels (RPKM) of the corresponding genes in the two cell lines. In each group, a ratio (level of overlap) was calculated for each RBP between the number of shared differentially edited sites and the number of the union of differentially edited sites between the two cell lines. P values were calculated using Wilcoxon rank sum test. Editing sites in highly expressed genes common to the two cell lines had the highest overlap between the two cell lines among the 3 groups of genes. **(b)** For RBPs in **(a)**, directional agreement scores (Methods) of their differentially edited sites between K562 and HepG2 cells.

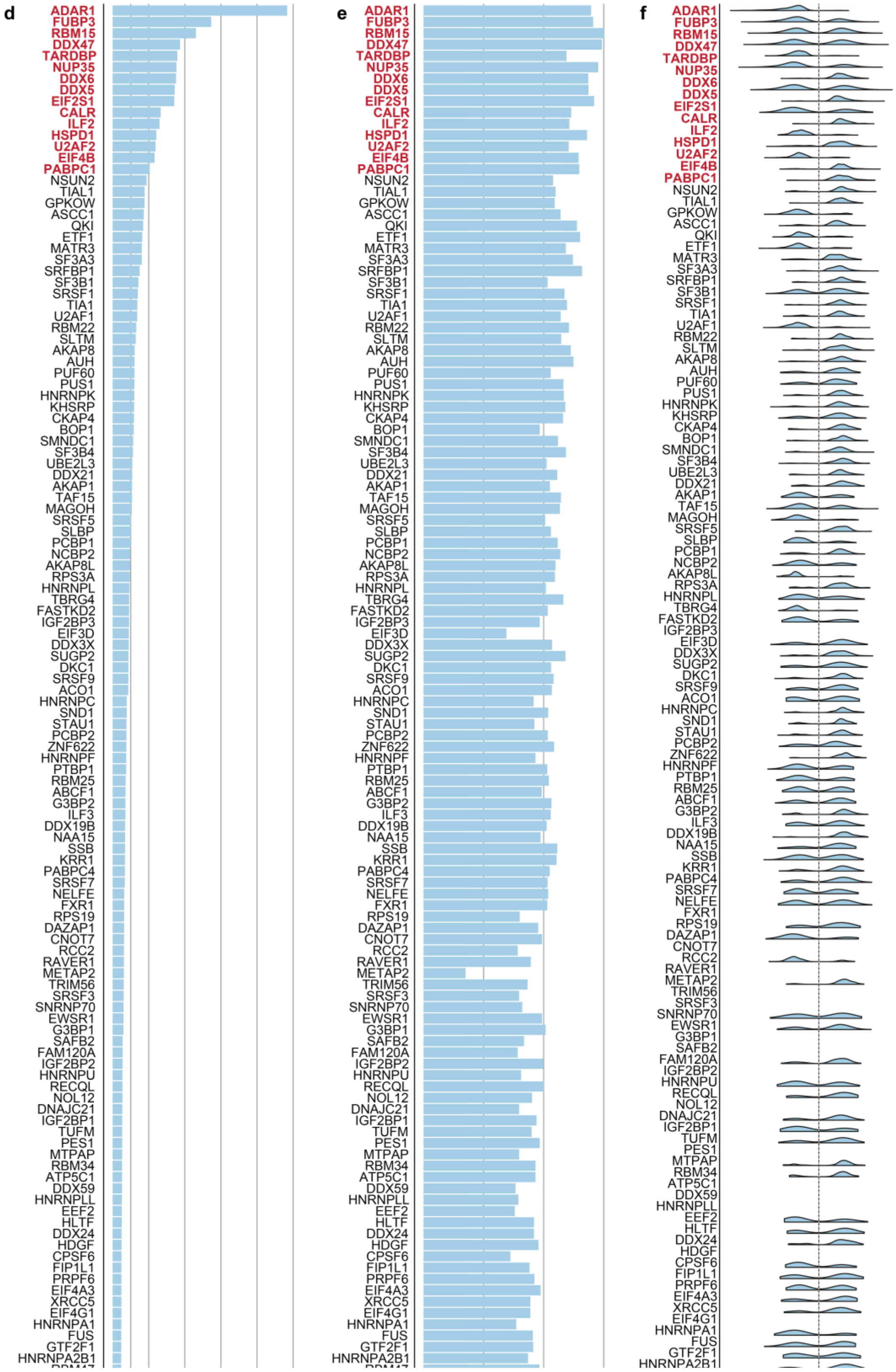
2.12 SUPPLEMENTAL FIGURES

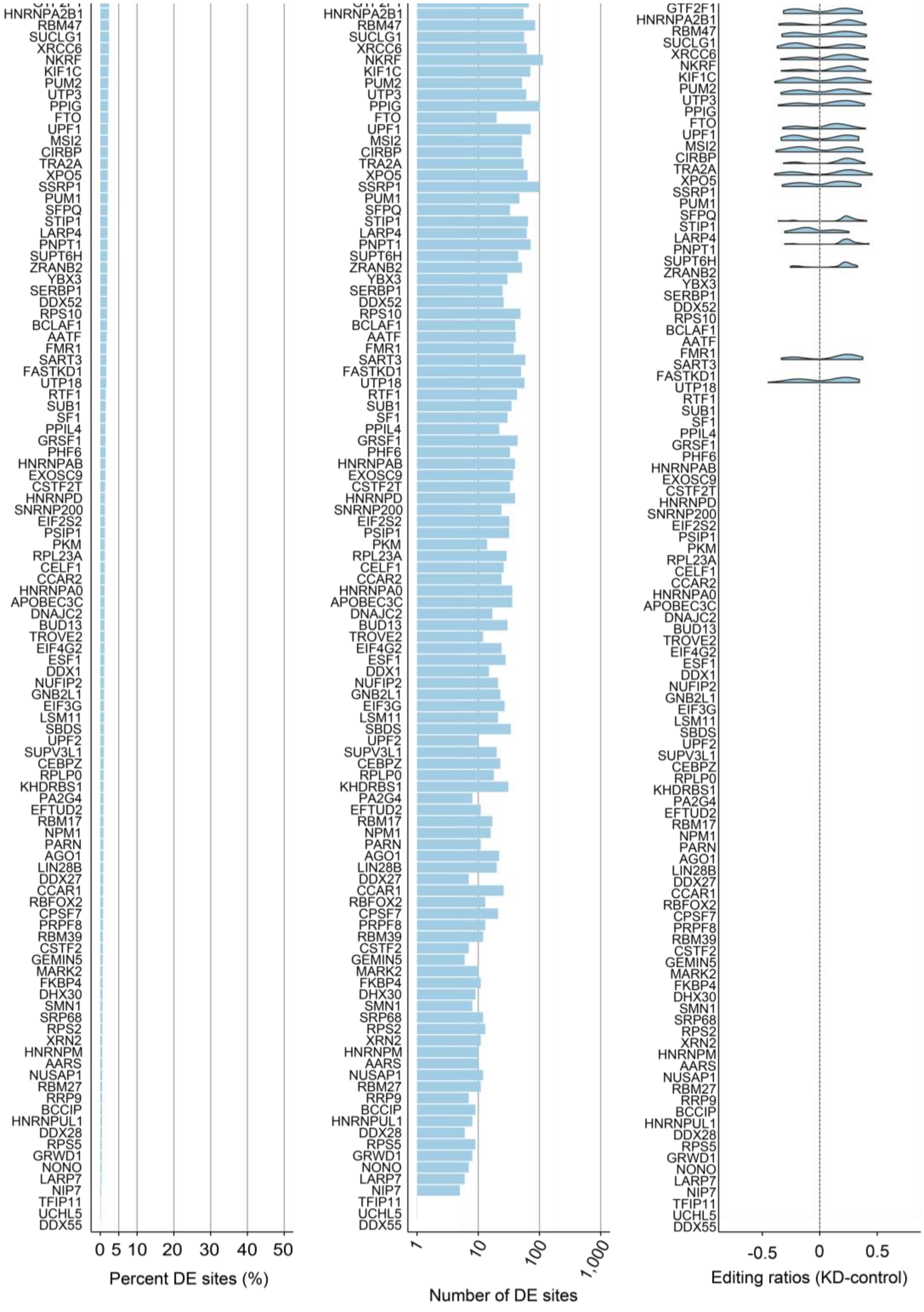


Supplemental Figure 2-1. Summary of editing sites. (a) Number of testable editing sites (top) and total RNA sequencing depth per samples in K562 and HepG2 cells. Fraction of A-to-G editing sites also shown for each sample (top, blue line). **(b)** Percentage of editing sites in Alu regions for each sample. **(c)** Genomic distribution of editing sites. ncRNA refers to non-coding transcripts.


a
ADAR1
PUS1
FXR1
DROSHA
TROVE2
EIF3D
PTBP1
SAFB2
G3BP1
IGF2BP2
CCDC124
UTP18
AQR
DDX19B
XRCC6
UBE2L3
MAGOH
PCBP1
MATR3
PABPN1
RBM15
AGGF1
SRSF5
SMNDC1
HNRNPC
IGF2BP1
UCHL5
SRSF1
AKAP1
EIF3G
U2AF2
DKC1
CELF1
IGF2BP3
EIF4A3
PCBP2
EFTUD2
SF1
TARDBP
AKAP8L
MBNL1
NELFE
RBM22
RRP9
NSUN2
TRIM56
EXOSC9
WDR3
DDX24
CIRBP
TRA2A
RBM34
SSB
SRSF4
HNRNPA2B1
PRPF8
KIF1C
HNRNPK
DDX3X
APOBEC3C
SLBP
DDX6
SF3B4
DDX21
SRSF7
TIA1
HLTF
PUF60
SUPV3L1
SRSF9
XRCC5
KHSRP
CNOT7
EEF2
FUBP3
NAA15
PABPC4
MAK16
ZRANB2
EIF3A
ILF2
RPS19
GEMIN5
U2AF1
SF3A3
DDX51
SNRNP200
GPKOW
DDX28
RBM25
ATP5C1
ILF3
NONO
EIF2S1
FXR2
ZC3H8
AKAP8
CCAR2
NCBP2
SUPT6H
UPF1
SART3
EIF4B
SF3B1
TBRG4
G3BP2
STAU1
CPSF6
EIF4G2
RBM39
FKBP4
XRN2
PES1
MSI2
EWSR1
DDX52
PUM1
KRR1
AATF
ASCC1
DAZAP1
TFIP11
b
ADAR1
PUS1
FXR1
DROSHA
TROVE2
EIF3D
PTBP1
SAFB2
G3BP1
IGF2BP2
CCDC124
UTP18
AQR
DDX19B
XRCC6
UBE2L3
MAGOH
PCBP1
MATR3
PABPN1
RBM15
AGGF1
SRSF5
SMNDC1
HNRNPC
IGF2BP1
UCHL5
SRSF1
AKAP1
EIF3G
U2AF2
DKC1
CELF1
IGF2BP3
EIF4A3
PCBP2
EFTUD2
SF1
TARDBP
AKAP8L
MBNL1
NELFE
RBM22
RRP9
NSUN2
TRIM56
EXOSC9
WDR3
DDX24
CIRBP
TRA2A
RBM34
SSB
SRSF4
HNRNPA2B1
PRPF8
KIF1C
HNRNPK
DDX3X
APOBEC3C
SLBP
DDX6
SF3B4
DDX21
SRSF7
TIA1
HLTF
PUF60
SUPV3L1
SRSF9
XRCC5
KHSRP
CNOT7
EEF2
FUBP3
NAA15
PABPC4
MAK16
ZRANB2
EIF3A
ILF2
RPS19
GEMIN5
U2AF1
SF3A3
DDX51
SNRNP200
GPKOW
DDX28
RBM25
ATP5C1
ILF3
NONO
EIF2S1
FXR2
ZC3H8
AKAP8
CCAR2
NCBP2
SUPT6H
UPF1
SART3
EIF4B
SF3B1
TBRG4
G3BP2
STAU1
CPSF6
EIF4G2
RBM39
FKBP4
XRN2
PES1
MSI2
EWSR1
DDX52
PUM1
KRR1
AATF
ASCC1
DAZAP1
TFIP11
c
ADAR1
PUS1
FXR1
DROSHA
TROVE2
EIF3D
PTBP1
SAFB2
G3BP1
IGF2BP2
CCDC124
UTP18
AQR
DDX19B
XRCC6
UBE2L3
MAGOH
PCBP1
MATR3
PABPN1
RBM15
AGGF1
SRSF5
SMNDC1
HNRNPC
IGF2BP1
UCHL5
SRSF1
AKAP1
EIF3G
U2AF2
DKC1
CELF1
IGF2BP3
EIF4A3
PCBP2
EFTUD2
SF1
TARDBP
AKAP8L
MBNL1
NELFE
RBM22
RRP9
NSUN2
TRIM56
EXOSC9
WDR3
DDX24
CIRBP
TRA2A
RBM34
SSB
SRSF4
HNRNPA2B1
PRPF8
KIF1C
HNRNPK
DDX3X
APOBEC3C
SLBP
DDX6
SF3B4
DDX21
SRSF7
TIA1
HLTF
PUF60
SUPV3L1
SRSF9
XRCC5
KHSRP
CNOT7
EEF2
FUBP3
NAA15
PABPC4
MAK16
ZRANB2
EIF3A
ILF2
RPS19
GEMIN5
U2AF1
SF3A3
DDX51
SNRNP200
GPKOW
DDX28
RBM25
ATP5C1
ILF3
NONO
EIF2S1
FXR2
ZC3H8
AKAP8
CCAR2
NCBP2
SUPT6H
UPF1
SART3
EIF4B
SF3B1
TBRG4
G3BP2
STAU1
CPSF6
EIF4G2
RBM39
FKBP4
XRN2
PES1
MSI2
EWSR1
DDX52
PUM1
KRR1
AATF
ASCC1
DAZAP1
TFIP11

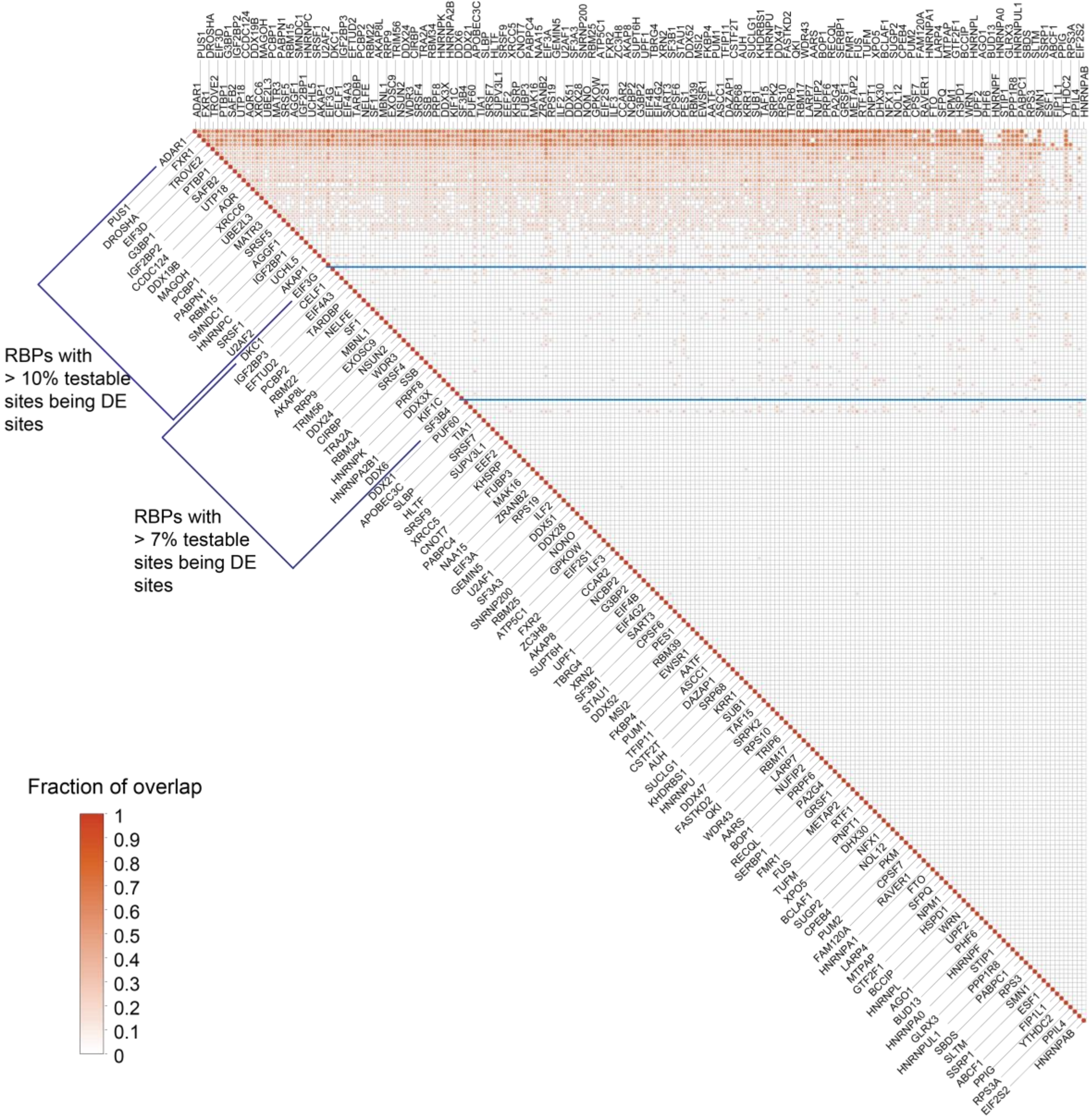

TFIP11
KHDRBS1
SUCLG1
CSTF2T
AUH
DDX47
SRP68
TRIP6
AARS
SRPK2
HNRNPU
LARP7
SUB1
WDR43
TAF15
NUFIP2
QKI
FASTKD2
RBM17
RPS10
BOP1
SERBP1
FMR1
GRSF1
RECQL
PRPF6
METAP2
PA2G4
FUS
RTF1
DHX30
PNPT1
TUFM
XPO5
SUGP2
CPEB4
BCLAF1
CPSF7
NOL12
PUM2
FTO
NFX1
PKM
FAM120A
RAVER1
HNRNPA1
NPM1
MTPAP
LARP4
SFPQ
GTF2F1
HNRNPL
UPF2
WRN
HNRNPA0
HSPD1
HNRNPF
BUD13
BCCIP
STIP1
LIN28B
PHF6
GLRX3
PPP1R8
AGO1
HNRNPUL1
CCAR1
PSIP1
DDX55
PABPC1
RPS3
SMN1
SBDS
SRFBP1
EIF4G1
SLTM
FIP1L1
TIAL1
RBFOX2
ABCF1
ESF1
SSRP1
RCC2
PARN
PPIG
CSTF2
GRWD1
NUSAP1
RPS3A
YTHDC2
SND1
PPIL4
HNRNPM
HNRNPAB
DNAJC21
DNAJC2
DDX1
EIF2S2
YBX3
LSM11
DDX27
0 5 10 20 30 40 50
Percent DE sites (%)
1
10
100
1,000
10,000
Number of DE sites
-0.5
0
0.5
Editing ratios (KD-control)

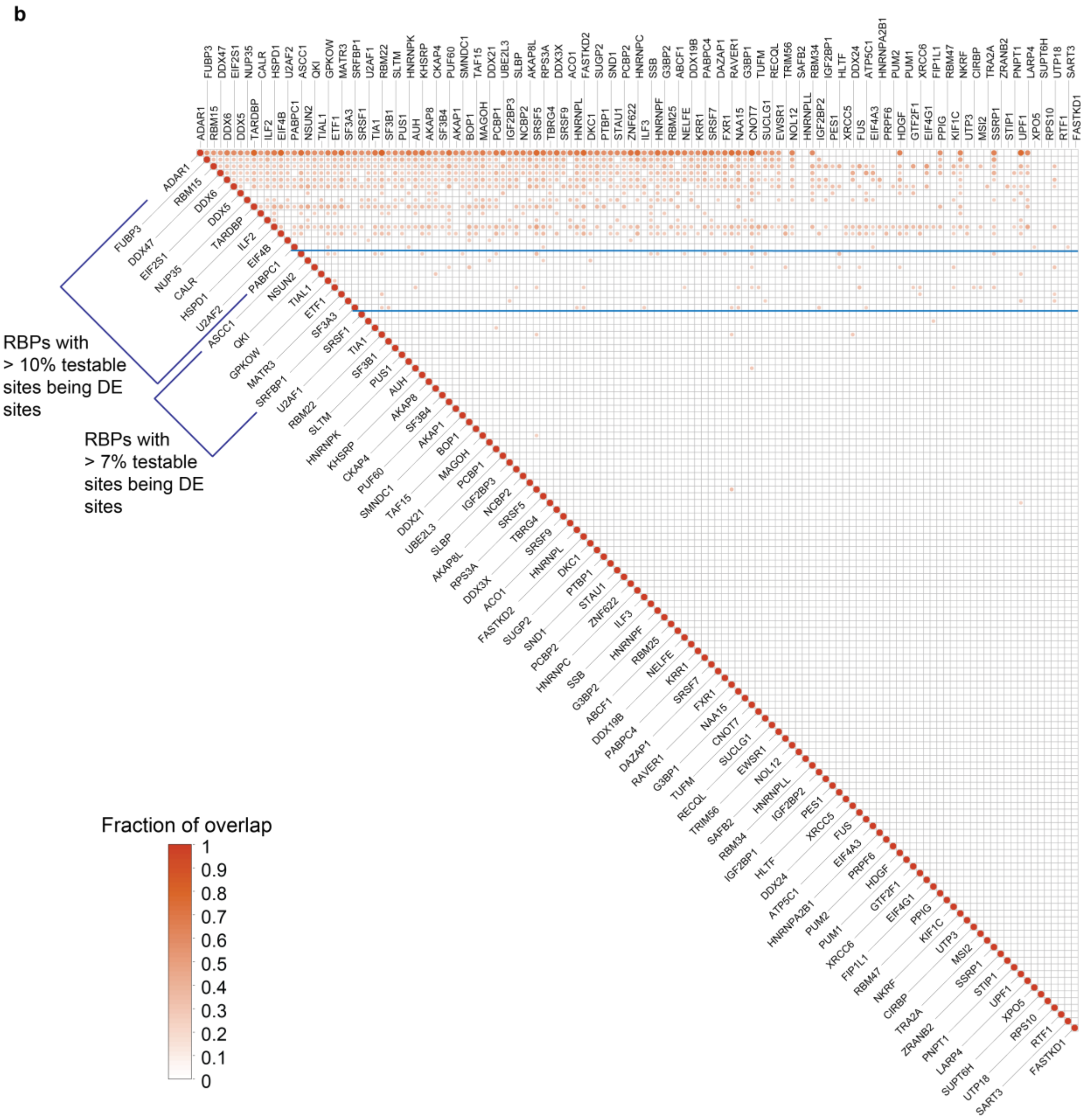

d
ADAR1
FUBP3
RBM15
DDX47
TARDBP
NUP35
DDX6
DDX5
EIF2S1
CALR
ILF2
HSPD1
U2AF2
EIF4B
PABPC1
NSUN2
TIAL1
GPKOW
ASCC1
QKI
ETF1
MATR3
SF3A3
SRFBP1
SF3B1
SRSF1
TIA1
U2AF1
RBM22
SLTM
AKAP8
AUH
PUF60
PUS1
HNRNPK
KHSRP
CKAP4
BOP1
SMNDC1
SF3B4
UBE2L3
DDX21
AKAP1
TAF15
MAGOH
SRSF5
SLBP
PCBP1
NCBP2
AKAP8L
RPS3A
HNRNPL
TBRG4
FASTKD2
IGF2BP3
EIF3D
DDX3X
SUGP2
DKC1
SRSF9
ACO1
HNRNPC
SND1
STAU1
PCBP2
ZNF622
HNRNPF
PTBP1
RBM25
ABCF1
G3BP2
ILF3
DDX19B
NAA15
SSB
KRR1
PABPC4
SRSF7
NELFE
FXR1
RPS19
DAZAP1
CNOT7
RCC2
RAVER1
METAP2
TRIM56
SRSF3
SNRNP70
EWSR1
G3BP1
SAFB2
FAM120A
IGF2BP2
HNRNPU
RECQL
NOL12
DNAJC21
IGF2BP1
TUFM
PES1
MTPAP
RBM34
ATP5C1
DDX59
HNRNPLL
EEF2
HLTF
DDX24
HDGF
CPSF6
FIP1L1
PRPF6
EIF4A3
XRCC5
EIF4G1
HNRNPA1
FUS
GTF2F1
HNRNPA2B1
e
ADAR1
FUBP3
RBM15
DDX47
TARDBP
NUP35
DDX6
DDX5
EIF2S1
CALR
ILF2
HSPD1
U2AF2
EIF4B
PABPC1
NSUN2
TIAL1
GPKOW
ASCC1
QKI
ETF1
MATR3
SF3A3
SRFBP1
SF3B1
SRSF1
TIA1
U2AF1
RBM22
SLTM
AKAP8
AUH
PUF60
PUS1
HNRNPK
KHSRP
CKAP4
BOP1
SMNDC1
SF3B4
UBE2L3
DDX21
AKAP1
TAF15
MAGOH
SRSF5
SLBP
PCBP1
NCBP2
AKAP8L
RPS3A
HNRNPL
TBRG4
FASTKD2
IGF2BP3
EIF3D
DDX3X
SUGP2
DKC1
SRSF9
ACO1
HNRNPC
SND1
STAU1
PCBP2
ZNF622
HNRNPF
PTBP1
RBM25
ABCF1
G3BP2
ILF3
DDX19B
NAA15
SSB
KRR1
PABPC4
SRSF7
NELFE
FXR1
RPS19
DAZAP1
CNOT7
RCC2
RAVER1
METAP2
TRIM56
SRSF3
SNRNP70
EWSR1
G3BP1
SAFB2
FAM120A
IGF2BP2
HNRNPU
RECQL
NOL12
DNAJC21
IGF2BP1
TUFM
PES1
MTPAP
RBM34
ATP5C1
DDX59
HNRNPLL
EEF2
HLTF
DDX24
HDGF
CPSF6
FIP1L1
PRPF6
EIF4A3
XRCC5
EIF4G1
HNRNPA1
FUS
GTF2F1
HNRNPA2B1
f
ADAR1
FUBP3
RBM15
DDX47
TARDBP
NUP35
DDX6
DDX5
EIF2S1
CALR
ILF2
HSPD1
U2AF2
EIF4B
PABPC1
NSUN2
TIAL1
GPKOW
ASCC1
QKI
ETF1
MATR3
SF3A3
SRFBP1
SF3B1
SRSF1
TIA1
U2AF1
RBM22
SLTM
AKAP8
AUH
PUF60
PUS1
HNRNPK
KHSRP
CKAP4
BOP1
SMNDC1
SF3B4
UBE2L3
DDX21
AKAP1
TAF15
MAGOH
SRSF5
SLBP
PCBP1
NCBP2
AKAP8L
RPS3A
HNRNPL
TBRG4
FASTKD2
IGF2BP3
EIF3D
DDX3X
SUGP2
DKC1
SRSF9
ACO1
HNRNPC
SND1
STAU1
PCBP2
ZNF622
HNRNPF
PTBP1
RBM25
ABCF1
G3BP2
ILF3
DDX19B
NAA15
SSB
KRR1
PABPC4
SRSF7
NELFE
FXR1
RPS19
DAZAP1
CNOT7
RCC2
RAVER1
METAP2
TRIM56
SRSF3
SNRNP70
EWSR1
G3BP1
SAFB2
FAM120A
IGF2BP2
HNRNPU
RECQL
NOL12
DNAJC21
IGF2BP1
TUFM
PES1
MTPAP
RBM34
ATP5C1
DDX59
HNRNPLL
EEF2
HLTF
DDX24
HDGF
CPSF6
FIP1L1
PRPF6
EIF4A3
XRCC5
EIF4G1
HNRNPA1
FUS
GTF2F1
HNRNPA2B1


GTF2F1
HNRNPA2B1
RBM47
SUCLG1
XRCC6
NKRF
KIF1C
PUM2
UTP3
PPIG
FTO
UPF1
MSI2
CIRBP
TRA2A
XPO5
SSRP1
PUM1
SFPQ
STIP1
LARP4
PNPT1
SUPT6H
ZRANB2
YBX3
SERBP1
DDX52
RPS10
BCLAF1
AATF
FMR1
SART3
FASTKD1
UTP18
RTF1
SUB1
SF1
PPIL4
GRSF1
PHF6
HNRNPAB
EXOSC9
CSTF2T
HNRNPD
SNRNP200
EIF2S2
PSIP1
PKM
RPL23A
CELF1
CCAR2
HNRNPA0
APOBEC3C
DNAJC2
BUD13
TROVE2
EIF4G2
ESF1
DDX1
NUFIP2
GNB2L1
EIF3G
LSM11
SBDS
UPF2
SUPV3L1
CEBPZ
RPLP0
KHDRBS1
PA2G4
EFTUD2
RBM17
NPM1
PARN
AGO1
LIN28B
DDX27
CCAR1
RBFOX2
CPSF7
PRPF8
RBM39
CSTF2
GEMIN5
MARK2
FKBP4
DHX30
SMN1
SRP68
RPS2
XRN2
HNRNPM
AARS
NUSAP1
RBM27
RRP9
BCCIP
HNRNPUL1
DDX28
RPS5
GRWD1
NONO
LARP7
NIP7
TFIP11
UCHL5
DDX55
0 5 10 20 30 40 50
Percent DE sites (%)
1
10
100
1,000
Number of DE sites
-0.5
0
0.5
Editing ratios (KD-control)

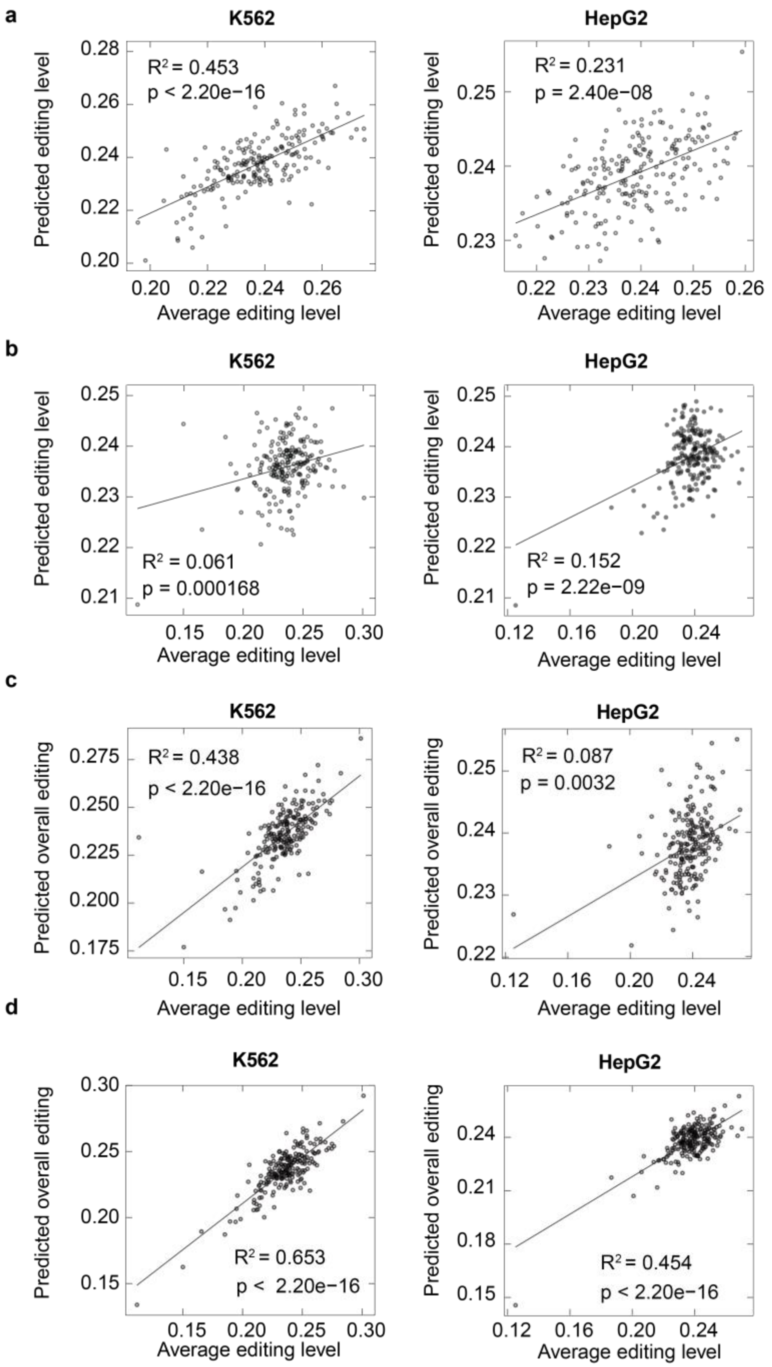
Supplemental Figure 2-2. Summary of DE sites in each RBP. (a - c) Differential editing associated with RBPs in K562 cells. **(d - f)** Differential editing associated with RBPs in HepG2 cells. **(a and d)** Percentage of differentially edited (DE) sites among all testable sites associated with each RBP. **(b and e)** Number of DE sites associated with each RBP. **(c and f)** Distribution of editing changes of all DE sites associated with each RBP. Only RBPs with more than 50 DE sites are shown. RBPs whose DE sites comprise at least 10% of all testable editing sites are highlighted in red.

a


PUS1
DROSHA
EIF3D
G3BP1
IGF2BP2
CCDC124
DDX19B
MAGOH
PCBP1
PABPN1
RBM15
SMNDC1
HNRNPC
SRSF1
U2AF2
DKC1
IGF2BP3
EFTUD2
PCBP2
RBM22
AKAP8L
RRP9
TRIM56
DDX24
CIRBP
TRA2A
RBM34
HNRNPK
HNRNPA2B1
DDX6
DDX21
APOBEC3C
SLBP
HLTF
SRSF9
XRCC5
CNOT7
PABPC4
NAA15
EIF3A
GEMIN5
U2AF1
SF3A3
SNRNP200
RBM25
ATP5C1
FXR2
ZC3H8
AKAP8
SUPT6H
UPF1
TBRG4
XRN2
SF3B1
STAU1
DDX52
MSI2
FKBP4
PUM1
TFIP11
CSTF2T
AUH
SUCLG1
KHDRBS1
HNRNPU
DDX47
FASTKD2
QKI
WDR43
AARS
BOP1
RECQL
SERBP1
FMR1
FUS
TUFM
XPO5
BCLAF1
SUGP2
CPEB4
PUM2
FAM120A
HNRNPA1
LARP4
MTPAP
GTF2F1
BCCIP
HNRNPL
AGO1
BUD13
HNRNPA0
GLRX3
HNRNPUL1
SBDS
SLTM
SSRP1
ABCF1
PPIG
RPS3A
EIF2S2
ADAR1
FXR1
TROVE2
PTBP1
SAFB2
UTP18
AQR
XRCC6
UBE2L3
MATR3
SRSF5
AGGF1
IGF2BP1
UCHL5
AKAP1
EIF3G
CELF1
EIF4A3
TARDBP
NELFE
SF1
MBNL1
EXOSC9
NSUN2
WDR3
SRSF4
SSB
PRPF8
DDX3X
KIF1C
SF3B4
PUF60
TIA1
SRSF7
SUPV3L1
EEF2
KHSRP
FUBP3
MAK16
ZRANB2
RPS19
ILF2
DDX51
DDX28
NONO
GPKOW
EIF2S1
ILF3
CCAR2
NCBP2
G3BP2
EIF4B
EIF4G2
SART3
CPSF6
PES1
RBM39
EWSR1
AATF
ASCC1
DAZAP1
SRP68
KRR1
SUB1
TAF15
SRPK2
RPS10
TRIP6
RBM17
LARP7
NUFIP2
PRPF6
PA2G4
GRSF1
METAP2
RTF1
PNPT1
DHX30
NFX1
NOL12
PKM
CPSF7
RAVER1
FTO
SFPQ
NPM1
HSPD1
WRN
UPF2
PHF6
HNRNPF
STIP1
PPP1R8
PABPC1
RPS3
SMN1
ESF1
FIP1L1
YTHDC2
PPIL4
HNRNPAB
RBPs with > 10% testable sites being DE sites
RBPs with > 7% testable sites being DE sites
Fraction of overlap
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

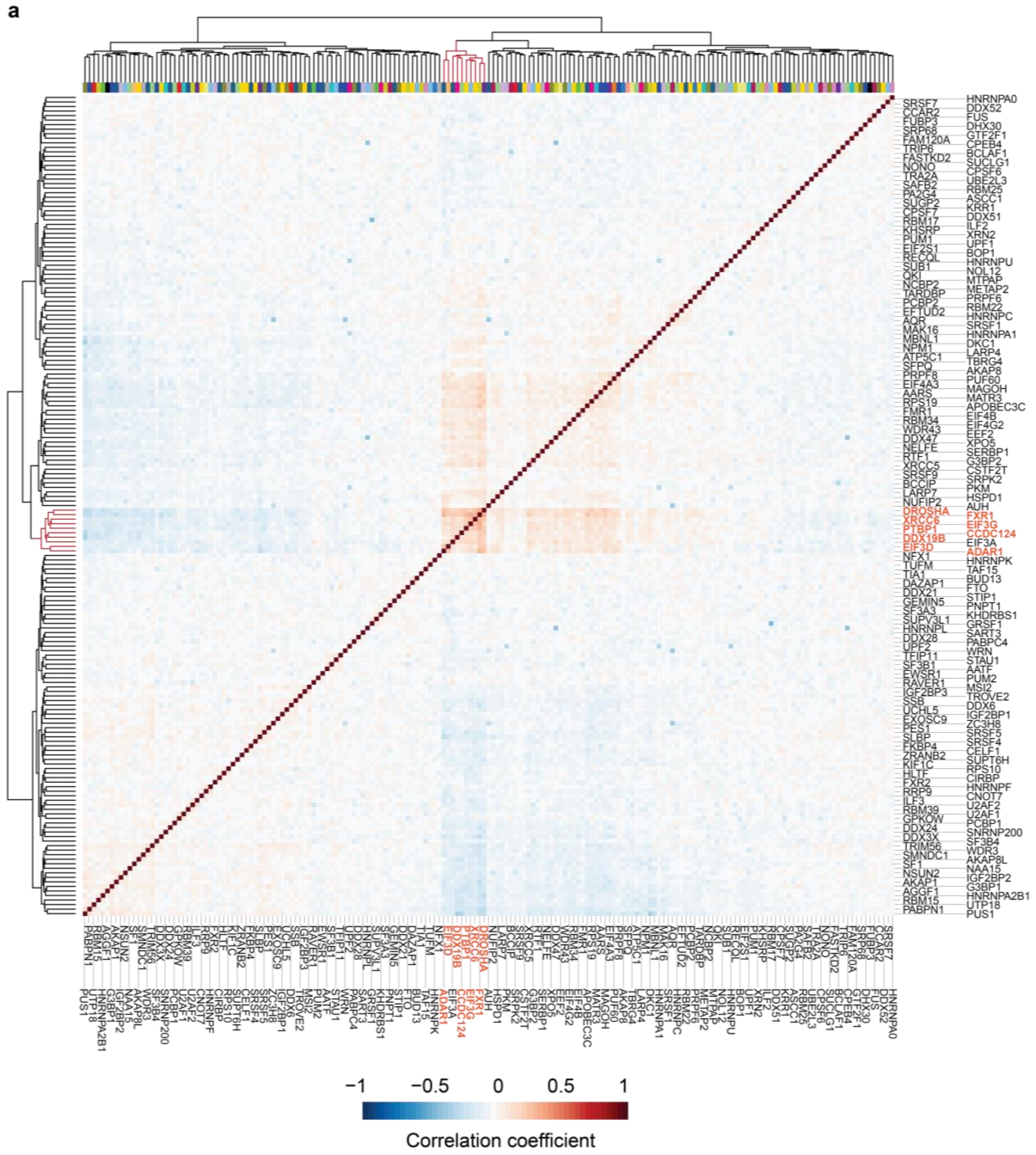


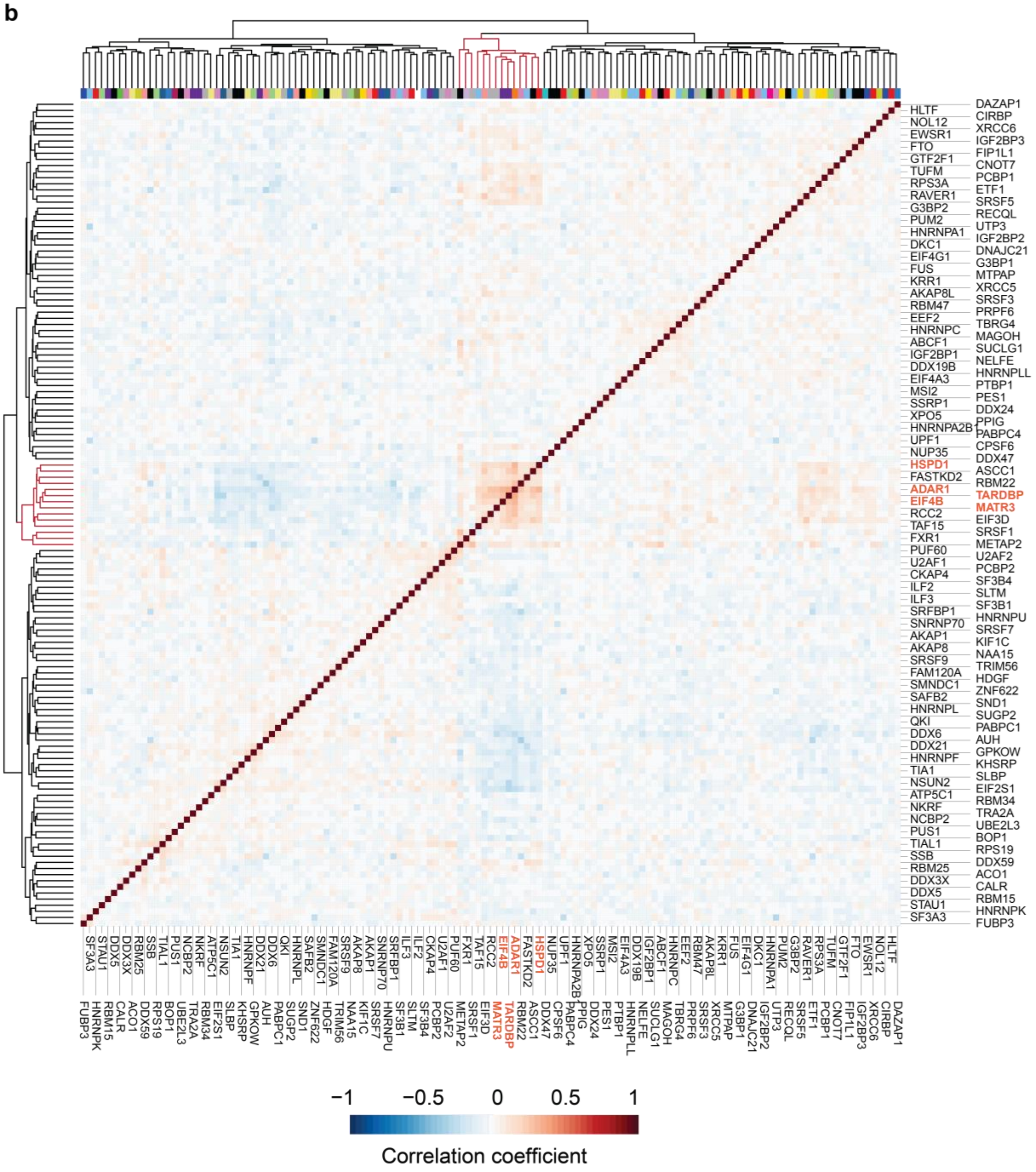
Supplemental Figure 2-3. Overlap of DE sites between RBP knock down samples. RBPs in K562 cells **(a)** and HepG2 cells **(b)**. The fraction of overlap was calculated as the number of common differentially edited sites over the minimum number of differentially edited sites between the two samples. Values < 0.25 were not shown for visualization purpose. RBPs are sorted from top to bottom based on their average fraction of overlap with all other RBPs (RBPs at the top have the highest values). RBPs whose differentially edited sites comprise >10% or 7% of all testable sites are highlighted.



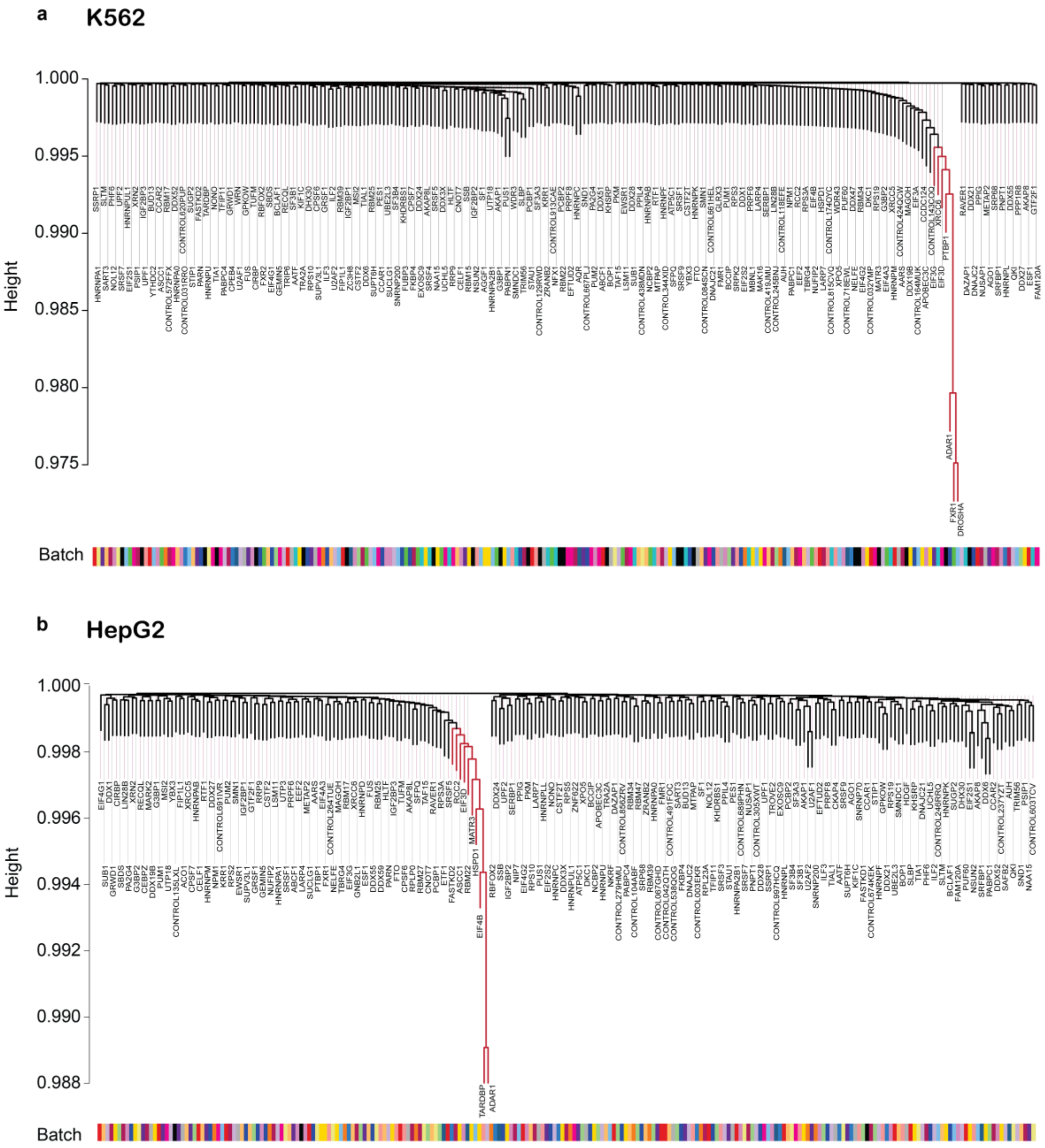
Supplemental Figure 2-4. Regression of predicted vs observed editing ratios. Regression analysis between the observed average editing levels in each RNA-seq dataset and the predicted editing level (Similar to Fig. 2-1c). The predicted editing level is based on: **(a)** the mRNA expression of the top 15 RBPs (including ADAR1) with greatest impact on editing regulation.

The top 10 highest leverage points were removed to test their influence on the regression. **(b)** the mRNA expression of ADAR1 alone. **(c)** the expression of the top 14 RBPs (excluding ADAR1). **(d)** the expression of the top 15 RBPs (including ADAR1) and ADAR1-RBP interaction terms.


a
HNRNPA0
SRSF7
DDX52
CCAR2
FUS
FUBP3
DHX30
SRP68
GTF2F1
FAM120A
CPEB4
TRIP6
BCLAF1
FASTKD2
SUCLG1
NONO
CPSF6
TRA2A
UBE2L3
SAFB2
RBM25
PA2G4
ASCC1
SUGP2
KRR1
CPSF7
DDX51
RBM17
ILF2
KHSRP
XRN2
PUM1
UPF1
EIF2S1
BOP1
RECQL
HNRNPU
SUB1
NOL12
QKI
MTPAP
NCBP2
METAP2
TARDBP
PRPF6
PCBP2
RBM22
EFTUD2
HNRNPC
AQR
SRSF1
MAK16
HNRNPA1
MBNL1
DKC1
NPM1
LARP4
ATP5C1
TBRG4
SFPQ
AKAP8
PRPF8
PUF60
EIF4A3
MAGOH
AARS
MATR3
RPS19
APOBEC3C
FMR1
EIF4B
RBM34
EIF4G2
WDR43
EEF2
DDX47
XPO5
NELFE
SERBP1
RTF1
G3BP2
XRCC5
CSTF2T
SRSF9
SRPK2
BCCIP
PKM
LARP7
HSPD1
NUFIP2
AUH
DROSHA
FXR1
XRCC6
EIF3G
PTBP1
CCDC124
DDX19B
EIF3A
EIF3D
ADAR1
NFX1
HNRNPK
TUFM
TAF15
TIA1
BUD13
DAZAP1
FTO
DDX21
STIP1
GEMIN5
PNPT1
SF3A3
KHDRBS1
SUPV3L1
GRSF1
HNRNPL
SART3
DDX28
PABPC4
UPF2
WRN
TFIP11
STAU1
SF3B1
AATF
EWSR1
PUM2
RAVER1
MSI2
IGF2BP3
TROVE2
SSB
DDX6
UCHL5
IGF2BP1
EXOSC9
ZC3H8
PES1
SRSF5
SLBP
SRSF4
FKBP4
CELF1
ZRANB2
SUPT6H
KIF1C
RPS10
HLTF
CIRBP
FXR2
HNRNPF
RRP9
CNOT7
ILF3
U2AF2
RBM39
U2AF1
GPKOW
PCBP1
DDX24
SNRNP200
DDX3X
SF3B4
TRIM56
WDR3
SMNDC1
AKAP8L
SF1
NAA15
NSUN2
IGF2BP2
AKAP1
G3BP1
AGGF1
HNRNPA2B1
RBM15
UTP18
PABPN1
PUS1
PUS1
PABPN1
UTP18
RBM15
HNRNPA2B1
AGGF1
G3BP1
AKAP1
IGF2BP2
NSUN2
NAA15
SF1
AKAP8L
SMNDC1
WDR3
TRIM56
SF3B4
DDX3X
SNRNP200
DDX24
PCBP1
GPKOW
U2AF1
RBM39
U2AF2
ILF3
CNOT7
RRP9
HNRNPF
FXR2
CIRBP
HLTF
RPS10
KIF1C
SUPT6H
ZRANB2
CELF1
FKBP4
SRSF4
SLBP
SRSF5
PES1
ZC3H8
EXOSC9
IGF2BP1
UCHL5
DDX6
SSB
TROVE2
IGF2BP3
MSI2
RAVER1
PUM2
EWSR1
AATF
SF3B1
STAU1
TFIP11
WRN
UPF2
PABPC4
DDX28
SART3
HNRNPL
GRSF1
SUPV3L1
KHDRBS1
SF3A3
PNPT1
GEMIN5
STIP1
DDX21
FTO
DAZAP1
BUD13
TIA1
TAF15
TUFM
HNRNPK
NFX1
ADAR1
EIF3D
EIF3A
DDX19B
CCDC124
PTBP1
EIF3G
XRCC6
FXR1
DROSHA
AUH
NUFIP2
HSPD1
LARP7
PKM
BCCIP
SRPK2
SRSF9
CSTF2T
XRCC5
G3BP2
RTF1
SERBP1
NELFE
XPO5
DDX47
EEF2
WDR43
EIF4G2
RBM34
EIF4B
FMR1
APOBEC3C
RPS19
MATR3
AARS
MAGOH
EIF4A3
PUF60
PRPF8
AKAP8
SFPQ
TBRG4
ATP5C1
LARP4
NPM1
DKC1
MBNL1
HNRNPA1
MAK16
SRSF1
AQR
HNRNPC
EFTUD2
RBM22
PCBP2
PRPF6
TARDBP
METAP2
NCBP2
MTPAP
QKI
NOL12
SUB1
HNRNPU
RECQL
BOP1
EIF2S1
UPF1
PUM1
XRN2
KHSRP
ILF2
RBM17
DDX51
CPSF7
KRR1
SUGP2
ASCC1
PA2G4
RBM25
SAFB2
UBE2L3
TRA2A
CPSF6
NONO
SUCLG1
FASTKD2
BCLAF1
TRIP6
CPEB4
FAM120A
GTF2F1
SRP68
DHX30
FUBP3
FUS
CCAR2
DDX52
SRSF7
HNRNPA0
−1
−0.5
0
0.5
1
Correlation coefficient

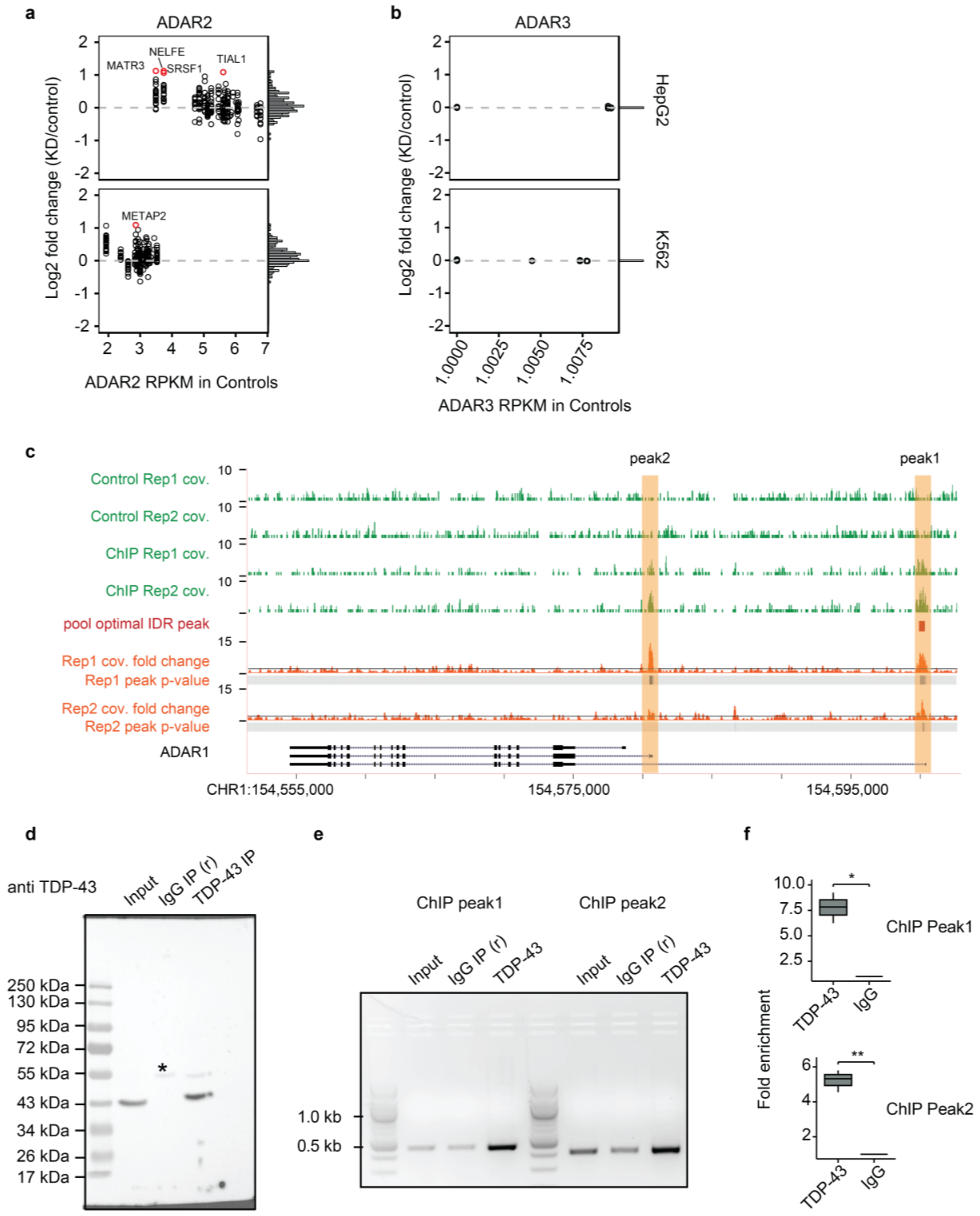


Supplemental Figure 2-5. Clustering of RBPs based on RNA editing alteration upon KD. Hierarchical clustering of pair-wise Spearman correlation of editing changes upon RBP knockdown in K562 cells **(a)** and HepG2 cells **(b)**. The union of all differentially edited sites identified in the RBP knockdown samples is used. For each pair of RBPs, only differentially edited sites that are testable in both datasets are included. The small cluster shown in red is associated with the highest correlation coefficients. This cluster contains RBPs associated with most significant reduction in editing (based on percentage of differentially edited sites among all testable sites) upon their knockdown. RBPs are labeled in orange in this cluster if they are associated with >10% differentially edited among all testable sites. The color labels on top denote experimental batches of each RBP.

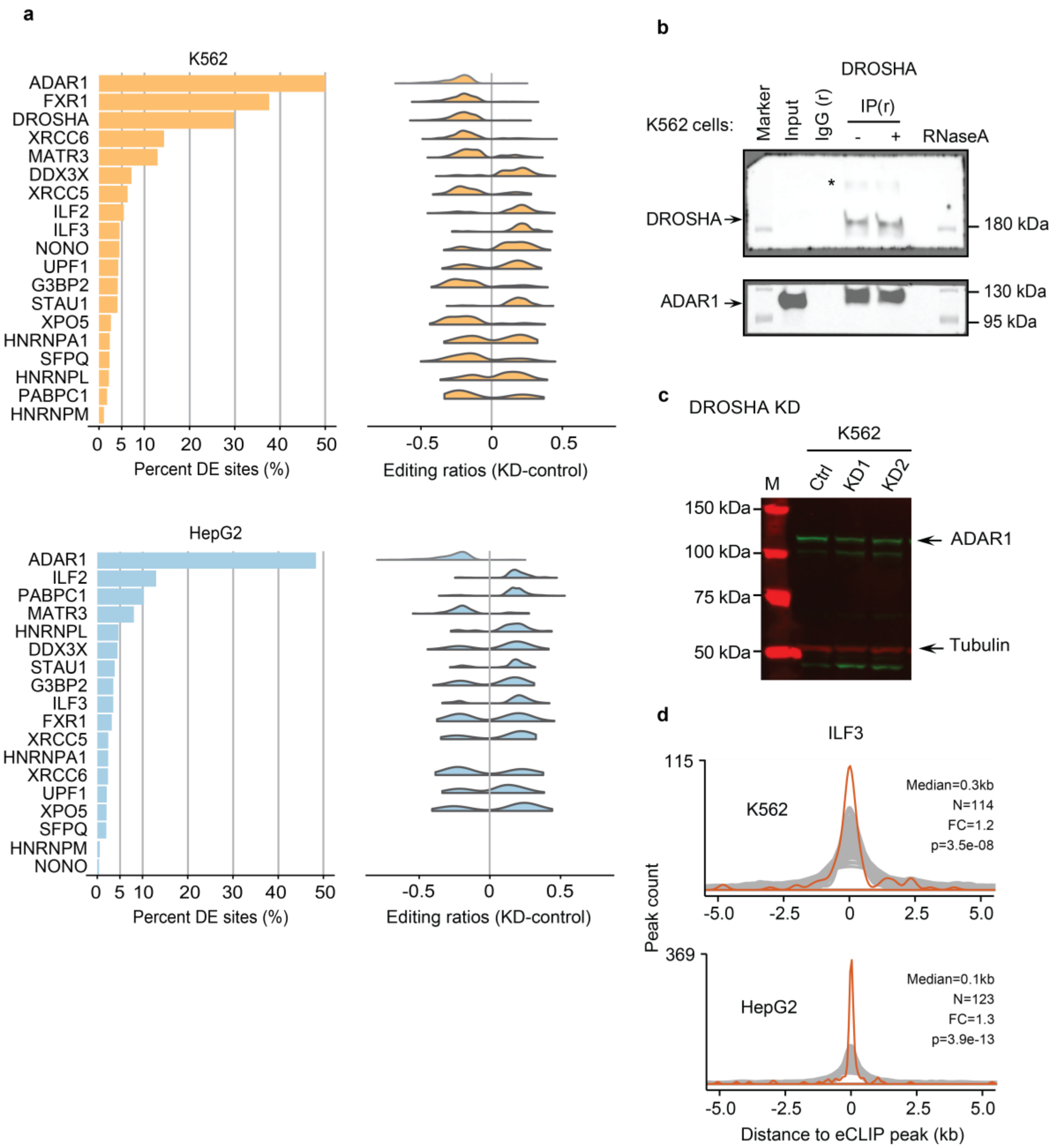


Supplemental Figure 2-6. WGCNA clustering of RBP knockdown data and controls.

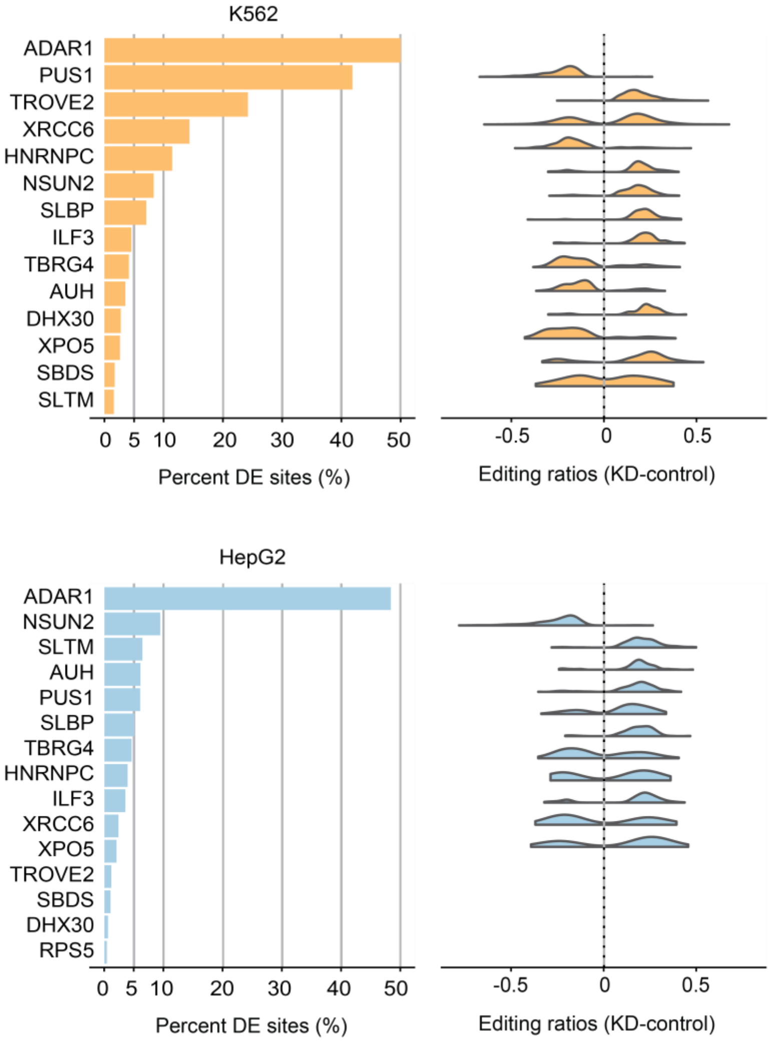
Correlation between RBPs were calculated via biweight mid-correlation using the change in editing levels upon RBP knockdown (see Methods). The clusters in red contain RBPs with high percentage of differentially edited sites among all testable sites. These RBPs are associated with reduction in editing upon their knockdown. RBPs in K562 **(a)** and HepG2 **(b)** cells shown. The color labels at the bottom denote experimental batches of each RBP.



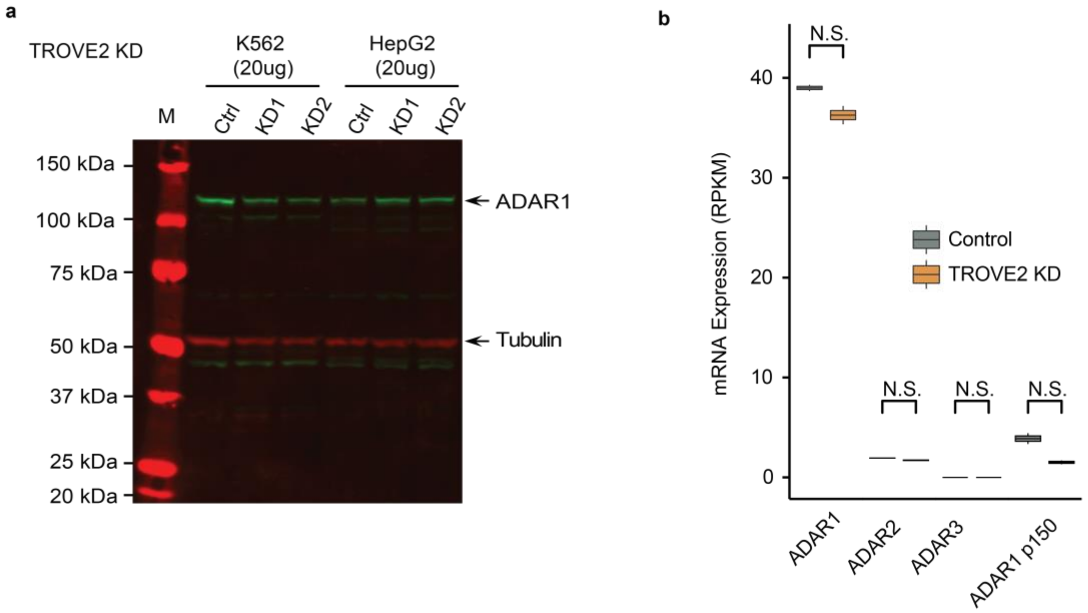
Supplemental Figure 2-7. Regulation of ADAR proteins expression. mRNA expression (RPKM) of **(a)** ADAR2. **(b)** ADAR3. Log2 fold changes (LFCs) of expression levels (knockdown/control) are shown. RBPs whose knockdown induced LFC > 1 and DESeq p value < 1e-9 are labeled in red. Histograms of LFC values are shown on the side of the plots. RPKM values have a pseudo count of 1. **(c)** Genome Browser view of TDP-43 ChIP-Seq datasets. In green, the read coverage of the ChIP samples and controls. In orange, the coverage fold change (ChIP vs control) of the two replicates. A line is drawn at fold change = 2. The peak coordinates and p-values were obtained using spp peak caller by the ENCODE consortium. P-values smaller than 0.01 are shown in dark gray tones. The peak (peak1) that passed the optimal irreproducible discovery rate (IDR) criteria of ENCODE is illustrated in the track named "pool optimal IDR peak" in red. **(d)** Western Blot of TDP-43 Immuno precipitation in HepG2 cells. The asterisk mark cross reaction with the antibody heavy chain. **(e)** Semi-quantitative PCR products from TDP-43 ChIP, input control and IgG using primers targeting peak1 and peak2 respectively. **(f)** Real-time qPCR of products from TDP-43 ChIP and IgG (BioRad qPCR Analysis Software, CFX Maestro Software). The values correspond to the fold enrichment of the band intensity (IP/input). Error bars represent standard deviation calculated from 3 biological replicates (Student's t-test, *P<0.05, **P<0.01)



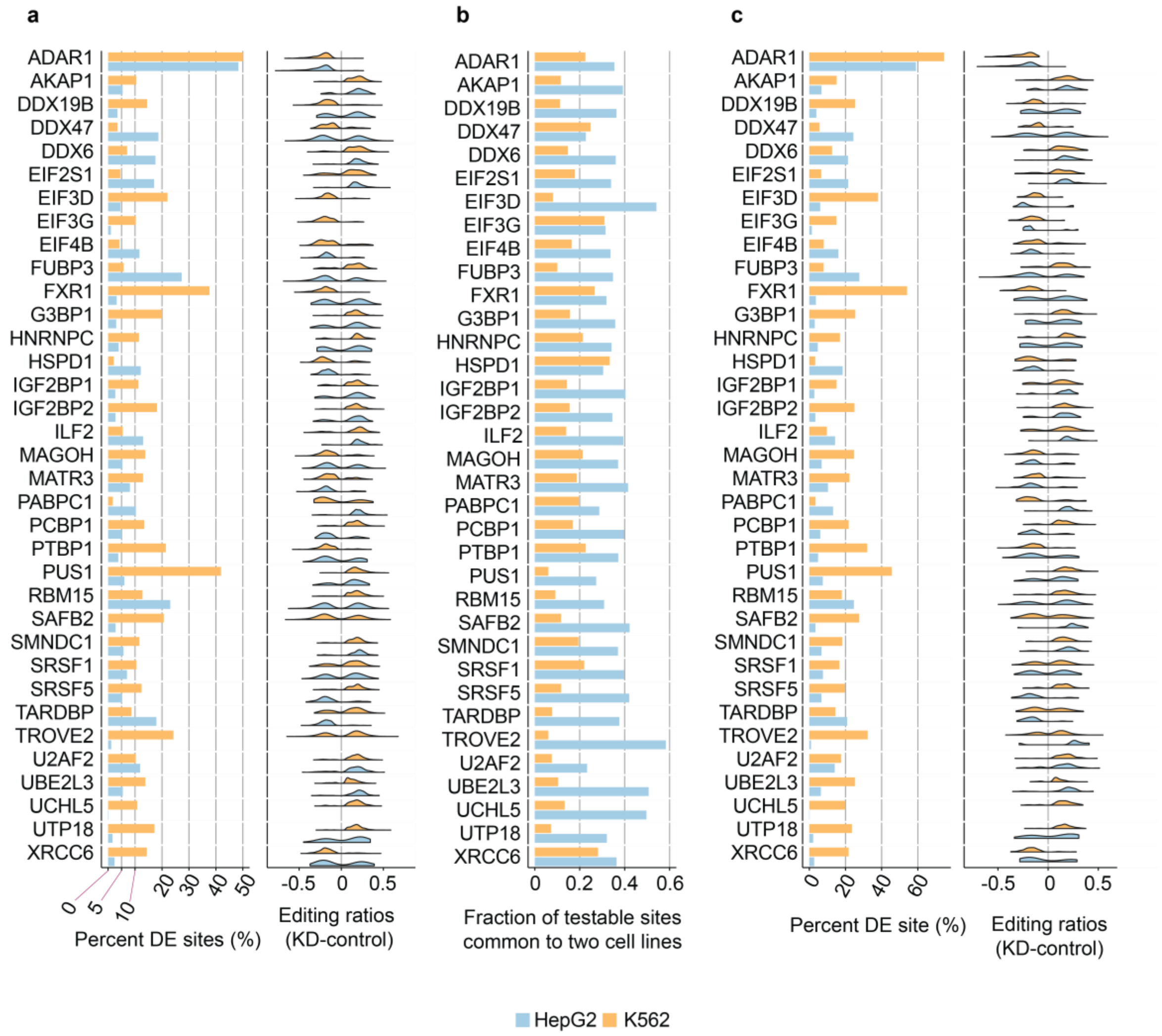
Supplemental Figure 2-8. RNA editing regulation by ADAR1-interacting RBPs. (a) RNA editing regulation by ADAR1-interacting RBPs. RBPs with shRNA RNA-seq data from ENCODE known to interact with ADAR1 in the literature were selected. Left: percentages of testable editing sites that are differentially edited (DE) upon RBP knockdown (KD) are shown. Right: distributions of editing changes of the differentially edited sites associated with each RBP are shown (except for RBPs with less than 50 differentially edited sites). **(b)** Uncropped blot image of the DROSHA and ADAR1 Co-Immunoprecipitation from Fig. 2-3c. The asterisk indicates non-specific bands. The gel was cut to use for multiple antibodies. **(c)** Western blot of shRNA-mediated DROSHA knockdown. Both K562 and HepG2 cells were transduced with lentiviruses expressing pLKO.1-DROSHA shRNA (knockdown) and pLKO.1-control shRNA (control), followed by selection with puromycin to establish stable cell lines. Blots were probed with antibodies detecting ADAR1 and Tubulin control. **(d)** Distance between ILF3 eCLIP peaks and their closest differentially edited sites (orange). Gray curves represent such distances relative to control sites (see Methods).



Supplemental Figure 2-9. RNA editing regulation by Alu-binding RBPs. The top 10 RBPs with highest fraction of eCLIP peaks overlapping alu regions for each cell line were selected. Left: percentages of testable editing sites that are differentially edited (DE) upon RBP knockdown (KD) are shown. Right: distributions of editing changes of the differentially edited sites associated with each RBP are shown (except for RBPs with less than 50 differentially edited sites).



Supplemental Figure 2-10. ADAR proteins expression upon TROVE2 KD. (a) Western blot of shRNA-mediated TROVE2 knockdown (KD). Both K562 and HepG2 cells were transduced with lentiviruses expressing pLKO.1-TROVE2 shRNA (knockdown) and pLKO.1-control shRNA (control), followed by selection with puromycin to establish stable cell lines. Blots were probed with antibodies detecting ADAR1 and Tubulin (as a control). **(b)** Expression of ADAR transcripts in TROVE2 knockdown K562 cells (N = 2 biological replicates, P-values were obtained using DESeq).



Supplemental Figure 2-11. RBP regulation of RNA editing across cell lines. RBPs with at least 10% of testable sites being differentially edited sites (DE) upon knockdown (KD) in at least one cell line. **(a)** Left: percentages of testable editing sites that are differentially edited upon RBP knockdown are shown. Right: distributions of editing changes of the differentially edited sites associated with each RBP are shown (except for RBPs with less than 50 differentially edited sites). **(b)** Fraction of testable editing sites that are common to both cell lines. **(c)** Percentage of the common testable editing sites that are differentially edited upon RBP knockdown (left) and editing changes of these sites (right).

CHAPTER 3 - scAllele: a versatile tool for the detection and analysis of variants in scRNA-seq

3.1 ABSTRACT

Single-cell RNA sequencing (scRNA-seq) data contain rich information at the gene, transcript, and nucleotide levels. Most analyses of scRNA-seq have focused on gene expression profiles, and it remains challenging to extract nucleotide variants and isoform-specific information. Here, we present scAllele, an integrative approach that detects single nucleotide variants, insertions, deletions, and their allelic linkage with splicing patterns in scRNA-seq. We demonstrate that scAllele achieves better performance in identifying nucleotide variants than other commonly used tools. The read-specific variant calls by scAllele enables allele-specific splicing analysis. Applied to a lung cancer scRNA-seq data set, scAllele identified variants with strong allelic linkage to alternative splicing, some of which being cancer-specific. scAllele represents a versatile tool to uncover multi-layer information and novel biological insights from scRNA-seq data.

3.2 INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) affords a unique glimpse into the transcriptome at the single-cell resolution, revealing great cellular heterogeneity (Papalexi & Satija, 2018). Although this data harbors rich information of a cell's transcriptome, most studies focus exclusively on gene expression. Fewer studies have tackled other important attributes of the data such as single-nucleotide variants (SNVs) (Zafar et al., 2016) and allele-specific expression (K. Choi et al., 2019; Jiang et al., 2017; Kim et al., 2015). In addition, most analyses of genetic variants in scRNA-seq data used methods originally designed for bulk RNA-seq or DNA-seq (Liu et al., 2019; Schnepp et al., 2019), due to the lack of tools specifically designed for variants calls in scRNA-seq.

Variant calling in RNA poses significant computational challenges. Often, variant callers rely on resolving the haplotypes from the NGS reads (Garrison & Marth, 2012; McKenna et al., 2010). This practice however is not applicable at the RNA level where transcripts can be edited by ADAR or other RNA editing enzymes (Nishikura, 2016), introducing nucleotide modifications, indistinguishable from SNPs. Additionally, allele-specific expression and alternative splicing will affect the minor-allele frequency of the variants further obscuring the true haplotype proportion in the data. Furthermore, in scRNA-seq, most expressed genes have shallow coverage (Yamawaki et al., 2021; M. J. Zhang et al., 2020), highlighting the importance of accurate detection with fewer reads available. To date, no method exists that explicitly focuses on both SNVs and insertions/deletions (INDELs) in scRNA-seq.

Following the identification of nucleotide variants, the next major challenge is to link them to their potential molecular function. To this end, RNA-seq data possess unique advantages given the afforded multi-level information: gene expression, transcript isoforms and sequence variations. Using bulk RNA-seq, numerous studies leveraged this strength to uncover allelic bias of genetic variants in gene expression or splicing (Fan et al., 2020; G. Li et al., 2012) which, for example, can lead to discovery of functional cis-acting variants that alter splicing (Amoah et al., 2021; Y.-H. E. Hsiao et al., 2016; Y. H. E. Hsiao et al., 2018; G. Li et al., 2012). Despite their typical low coverage, scRNA-seq data provide similar multi-level information. However, no method exists to exploit such features of scRNA-seq to uncover functionally relevant variants.

Here, we introduce scAllele, a versatile tool that performs both variant calling and functional analysis of the variants in alternative splicing using scRNA-seq. As a variant caller, scAllele reliably identifies SNVs and microindels (less than 20 bases) with low coverage. It implements RNA-friendly haplotype filtering by accounting for potential RNA editing sites and allele-specific splicing. Following variant calling, scAllele identifies significant associations between variant alleles and alternative splicing, which provides direct evidence of allele-specific splicing.

Using scRNA-seq data associated with well-characterized genotypes, we show that scAllele significantly outperforms other commonly used methods, even for difficult-to-call regions and lowly-covered transcripts. We apply scAllele to scRNA-seq data derived from lung cancer samples. Our analysis identifies variants that have significant allelic linkage to splicing isoforms, some of which are enriched in cancer cells. Thus, scAllele is an integrative analysis tool that uncovers multi-level information in scRNA-seq.

3.3 RESULTS

3.3.1 Algorithm Overview

scAllele calls nucleotide variants via local reassembly (Fig. 3-1a). To scan variants in the entire transcriptome, we split the mapped reads into read clusters (RC), which we defined as genomic intervals containing overlapping reads. The reads from each RC are subsequently decomposed into overlapping k-mers and reassembled into a directed de-Bruijn graph. The reference genomic sequence is included in the reassembly to serve as the reference haplotype in the RC. The nodes of the graph represent k-mers derived from the read sequence. Two nodes with k-mers overlapping by k-1 bases are connected with a directed edge. The ‘bubbles’ in the graph represent differences among the read sequences which include the genome reference sequence.

To identify nucleotide variants, we first traverse the graph with a depth-first search to identify nodes marking the beginning and the end of each bubble (source and sink nodes) and their respective pairing (Fig. 3-1a, Local reassembly). Hereafter, we perform a per-read analysis of the graph. We obtain the walk in the graph that best matches the read sequence. Based on the previously identified source-sink pairs, we can detect the variants present in each read. The presence of repeats or low-complexity regions significantly complicates the detection of variants since the de-Bruijn graph can be traversed in multiple ways. scAllele overcomes this challenge by performing a Dijkstra-based traversal of the graph with the assumption that the walk with the smallest editing distance best represents the set of variants present in the read. Finally, we collect the variants from the RC reads and score them using a generalized linear model (GLM) (Fig. 3-1b) where the following features are included: read position, base quality, number of neighboring tandem repeats, allelic ratio, sequencing error rate, and haplotype fitting (see Methods: *Variant scoring*).

An important feature of scAllele is the detection of variants at the read level. This feature enables a direct analysis of allelic linkage between the variants and other attributes of the reads. Here, we focus on identifying allelic linkage with alternative splicing via mutual information, similarly as in our previous work for RNA editing identification (Q. Zhang & Xiao, 2015). We consider overlapping introns as “alleles” of the same intronic part (Fig. 3-1c) and calculate the read coverages of each “haplotype” (between introns and nucleotide variants). In this way, we can incorporate splicing isoforms in the mutual information calculation to identify allele-specific splicing (see Methods: *Linkage Analysis*).

It should be noted that scAllele is a stand-alone tool and only requires a bam file to conduct variant calling and linkage detection. However, pre-processing of the bam file is recommended to achieve optimal results (Supplemental Fig. 3-1).

3.3.2 Evaluation of variant calls in GM12878 and iPSC cells

We evaluated the variant-calling function of scAllele using scRNA-seq (Smart-seq2) of GM12878 cells and iPSC cells from 3 individuals (Tung et al., 2017). These individuals were carefully genotyped by the GIAB (Zook et al., 2019) and 1000 Genomes projects (Clarke et al., 2017), thus providing a "ground truth" for method evaluation. We compared the performance of scAllele to those of three other popular variant callers: Freebayes (v.1.3.4), GATK (v.4.2.0.0) and Platypus (v.1.0) (Garrison & Marth, 2012; McKenna et al., 2010; Rimmer et al., 2014). The performance evaluation followed previously published guidelines (Krusche et al., 2019) with some modifications to accommodate RNA variants (see Methods). For GM12878, we used three benchmark datasets: GIAB's list of all genetic variants, GIAB's list of high-confidence genetic variants and the variant calls based on long-read DNA sequencing (Oxford Nanopore) (Karst et al., 2021).

For each data set and each method, we calculated the true positive counts at specific false positive cutoffs for both microindels and SNPs (Fig. 3-2a, Supplemental Fig. 3-2a). Overall, scAllele achieves the best performance for microindels among all methods, and its performance on SNPs is on par with the others. The strength of scAllele in microindel identification is notable as these variants are known for their challenging detection (Sun et al., 2017). Furthermore, scAllele also demonstrated superior performance in capturing microindels in "difficult regions" (Fig. 3-2b, Supplemental Fig. 3-2b). These difficult regions were defined by GIAB (Zook et al.,

2019) as the union of regions with low mappability, high GC-content, low complexity or presence of repeats, and segment duplication among others .

It should be noted that although the above cells have been analyzed by the GIAB and 1000 Genomes projects, their genotype calls may still miss some true positives. As examples, we experimentally confirmed 4 microindels categorized as false positives according to the "ground truth" (Fig. 3-2c, Supplemental Fig. 3-3). The 4 microindels were identified by scAllele and Platypus (3 by GATK, 3 by Freebayes, See Methods). Thus, the above performance of scAllele (and the other methods) may be a conservative estimation.

One of the hallmarks of scRNA-seq is the limited read coverage per gene. Thus, it is highly desirable to develop variant callers with superior performance at low read coverage. scAllele meets this demand and demonstrates a performance gain relative to the other methods in lowly covered variants (Fig. 3-2d, Supplemental Fig. 3-2c). Indeed, about 95% of the ground truth variants were covered by less than 5 reads in each data set. Thus, scAllele affords a unique advantage for scRNA-seq data.

Unique to RNA-seq, the allelic read counts of genetic variants reflect their allelic expression levels. Thus, in addition to variant calling, it is necessary to accurately estimate the allelic quantification of each variant. To test the performance of scAllele in this regard, we segregated the ground truth variants into heterozygous and homozygous groups. The heterozygous variants are expected to exhibit an approximately normal distribution in their ALT allelic ratios (variant allele read number/total read number) centered around 0.5 (G. Li et al., 2012). For both microindels and SNPs, this expected distribution was observed, with scAllele and Platypus showing best performance across all data sets (Fig 3-2e, Supplemental Fig. 3-2d). For homozygous variants, the allelic ratios should be 1. In most data sets, all methods show a predominant value of 1 for homozygous SNPs. In contrast, for microindels, scAllele clearly demonstrates consistent desirable performance, which was not observed for the other methods.

Overall, the above evaluations support the superior performance of scAllele for scRNA-seq variant analysis, especially in handling microindels, an aspect that is much more challenging compared to the most-often tackled SNP identification.

3.3.3 Linkage calculation between variants

In addition to variant calling, scAllele enables read-level allelic linkage analysis. Such analysis is not possible with other variant callers as read-level information is not extracted. In scAllele, the degree of allelic linkage is quantified as the mutual information (MI) between two types of variants: nucleotide variants and alternatively spliced junctions representing introns (Fig. 3-1). This metric is expected to require a relatively high number of reads harboring both types of variants. To achieve an understanding of the read coverage requirements, we calculated the MI between pairs of known genetic variants in the GM12878 and iPSC data used in the last section. As expected, the MI of these variant pairs in the RNA is generally high, regardless of read coverage, reflecting the associated DNA haplotypes (Fig. 3-3a).

As a comparison, we also calculated the MI between pairs of nucleotide variants where either was known genetic variant (Fig. 3b). Since GM12878 has been well-genotyped, unknown variants observed in the RNA-seq reads are likely RNA editing sites or sequencing errors. The MI of these variant pairs is expected to be low in general (Q. Zhang & Xiao, 2015), unless rare allele-specific RNA editing exists. This expectation of low MI was met at relatively high read coverage (>10). However, at lower read coverage, the MI is inflated to high values due to low number of transcripts captured in the library. Thus, it is necessary to impose a minimum read coverage requirement for MI calculation. In this study, we set this cutoff to be 10 to avoid false positive linkage calls. Additionally, we require a minimum MI of 0.52 to call significant linkage events, as 95% of the known genetic variant pairs (with ≥10 reads) had an MI of 0.52 or greater, and 90% of the unknown variant pairs (with ≥10 reads) failed this MI cutoff (Fig. 3-3c). The read coverage and MI cutoffs can be altered by the user in scAllele.

3.3.4 scAllele unveils nucleotide variants and allele-specific splicing events in lung cancer cells

Next, we applied scAllele to scRNA-seq data of lung cancer (Smart-seq2) (Maynard et al., 2020). We focused on cancer cells and their normal counterparts, epithelial cells, in tumor and matched normal samples of two patients (TH179 and TH238; n = 574 cells). We first carried out variant calling in each cell. An SNV or microindel was retained if it was detected in at least 3 cells. Furthermore, we compared the presence of the variants in normal epithelial or cancer cells. A variant was defined as cancer-enriched if it was not detected in normal cells or its presence is significantly more frequent in cancer compared to normal cells (BH-corrected $p < 0.1$, Methods). Otherwise, the variant was labeled as a cancer-normal common variant. As a sanity check, we note that no variant was found to be enriched in normal cells relative to cancer cells.

As shown in Fig. 3-4a, >15,000 variants were identified in each patient, with the majority being SNVs. Most SNVs are annotated SNPs in the dbSNP (b151) or Cosmic (human cancer mutations) database. Importantly, Cosmic variants constitute a larger fraction among cancer-enriched variants compared to the common category ($p < 1e-14$ in both patients, Chi-Squared test). Among the unannotated (i.e., novel) SNVs, some may be novel genetic variants and others may reflect RNA editing events. Indeed, a large fraction (74% in TH179, 76% in TH238) of the novel SNVs corresponded to A-to-G or C-to-T RNA editing types. A relatively large fraction of microindels was not annotated in either database, probably reflecting the incomplete knowledge of this type of variants in current databases. Interestingly, cancer-enriched SNVs were more often located in coding exons, less often in introns, compared with cancer-normal common SNVs (Fig. 3-4b). In contrast, the fraction of microindels in coding exons is approximately the same for the two categories of variants. Nonetheless, one patient (TH179) showed enrichment of cancer-enriched indels in 3' UTRs (depletion in introns), relative to the common microindels.

Following variant calling, we identified the allele-specific splicing events in each cell. As examples, Fig. 3-4c shows two significant linkage events. In these cases, the SNPs demonstrated strong allelic linkage with alternative splicing patterns (exon skipping and alternative 5' splice site, respectively). Across cancer and normal cells, the number of allele-specific splicing events varied greatly, ranging from 0 to 49 events (Fig. 3-4d), most of which involved SNVs. This number correlates approximately with the number of spliced junction reads present in each cell (Fig. 3-4d insets). Based on down-sampling of a few deeply sequenced cells, we observed that 1M total reads can enable identification of up to 11 events per cell (Fig. 3-4e). In some cells, the number of events plateaued at around 5M reads. Thus, to afford power for splicing analysis, a relatively large number of scRNA-seq reads is needed.

3.3.5 Cancer and normal cells exhibit unique and differential linkage events

Next, we asked whether cancer and normal cells harbor different allele-specific splicing events. Among all events identified in this study, 27 were observed in both cancer and normal cells, whereas more events were exclusive to one of the two classes of cells (Fig. 3-5a). In general, most events were observed in a small number of cells (< 5).

To identify differential linkage events between cancer and normal cells, we focused on two scenarios. The first scenario includes variants present and testable for splicing linkage in both cancer and normal cells, but significant linkage was detected with higher prevalence (p-value < 0.05 Fisher's exact test) in one cell class than the other. For this scenario, we observed 2 events, both of which occurred in higher proportion in the normal than cancer cells (Normal-differential, Fig. 3-5b, Table 3-1). These events may reflect a loss of function of the variants in the cancer cells. Figure 3-5c (left) shows an example of such an event in the CTSE gene, where the C allele of the variant is linked to skipping of the middle exon, whereas the T allele is associated with exon inclusion. This linkage was only observed in normal cells, but not cancer cells (despite the presence of the variant and adequate read coverage in cancer cells). Notably, the gene CTSE encodes for Cathepsin E, an aspartic protease with a vital role in protein degradation, bioactive protein generation, antigen processing and presentation (Yamamoto et al., 2012).

In the second scenario, the variants were not present/testable in the normal cells but had significant linkage in the cancer cells, or vice versa (labeled as "cancer-specific" or "normal-specific", Fig. 3-5b, Table S1). For this scenario, we observed 67 events that were cancer-specific and a smaller number of events (29) that were normal-specific. Notably, 11 cancer-specific events were observed in at least 3 cancer cells, whereas no normal-specific events exceeded this level of prevalence (Fig. 3-5b). Figure 3-5c (right) shows an example in the gene IFI44L (interferon-induced protein 44-like), a type I interferon stimulated gene with a role in host antiviral response (DeDiego et al., 2019). The allele-specific splicing event and the associated variant were only observed in cancer cells.

3.4 DISCUSSION

scRNA-seq affords unprecedented views of single cell transcriptomes. Similar to bulk RNA-seq, scRNA-seq provides information on the single-nucleotide level. However, identification of nucleotide variants in scRNA-seq is challenging due to the limited read coverage per cell. Here, we present scAllele, a versatile tool that not only enables variant calling in scRNA-seq but also uncovers variants involved in allele-specific RNA processing.

We showed that scAllele outperforms other popular methods in variant calling, especially for microindels, the class of variants that are less well characterized than SNPs. Built upon local reassembly, scAllele refines read alignments and corrects possible misalignments in each read, thus enhancing variant detection accuracy per read. Additionally, scAllele uses a GLM model to detect high confidence variants. These features together confer an advantage in performance given low sequencing depth.

The read-level variant calling by scAllele enables another advantage, that is, facilitating a detailed view of the allelic bias linked to alternative RNA isoforms. In this work, we focused on allele-specific splicing patterns. A similar approach can be extended to examine other aspects of RNA expression, such as alternative polyadenylation. We note that this type of analysis requires a relatively high read coverage per event, as it simultaneously quantifies alternative alleles of nucleotide variants, alternative RNA isoforms and their combined linkage patterns. We showed that the number of such events increased with higher scRNA-seq sequencing depth, indicating that RNA representation in scRNA-seq was not saturated at lower depth, such as 1M reads. With the continued drop in sequencing cost, we expect to see wide applications of allele-specific and alternative RNA isoform analyses, such as those enabled by scAllele.

We applied scAllele to a lung cancer scRNA-seq dataset (with matched controls). Our analysis identified a large number of nucleotide variants, many of which have enriched presence in cancer cells. Compared to variants common to both normal and cancer cells, cancer-enriched variants were more often cataloged in Cosmic, supporting the validity of the scAllele variant calls. Additionally, cancer-enriched variants were more often located in coding regions, which suggests a potential role in altering protein sequences or producing neoantigens. Although microindels are not as abundant as SNVs, they may also have critical roles in human diseases (Chen & Guo, 2020), an area remains under-explored. In general, compared to SNVs, we observed a relative enrichment of microindels in 3' UTRs. Given the existence of numerous regulatory elements in the 3' UTRs (Mayr, 2017), microindels in these regions may alter many processes, such as mRNA stability, translation, or mRNA localization, which should be investigated in the future.

In the cancer and normal epithelial cells, we identified more than 150 allele-specific splicing or linkage events. We further categorized these events based on their relative prevalence in cancer or normal cells. Even though most events were observed in a small number of cells, likely due to low read coverage in single cells, such categorization provides an approximate overview of their relative enrichments. Among these events, many have important relevance to cancer, such as those in the CTSE and IFI44L genes in Fig. 3-5c. Our results suggest that scRNA-seq data possess useful information to uncover important alternative splicing events, linking genotypes to this molecular phenotype.

In summary, scAllele offers a unique approach to maximize the information extracted from scRNA-seq data sets. With the emergence of scRNA-seq data from a large spectrum of samples, scAllele will lead to a granular view of the genetic landscape of each cell and the potential genetic drivers of gene expression complexity.

3.5 METHODS

3.5.1 scAllele: detailed outline

To scan variants in the entire transcriptome, we grouped the sequencing reads into Read Clusters (RC). An RC is made up of a group of overlapping reads. It should be noted that the entire sequence of each read was included, not only the segment that overlaps the RC. Multi-mapped, chimeric, and low mapping-quality reads were removed as well as reads with ≥ 5 soft-clipped bases or trailing homopolymers (n ≥ 15). An additional "reference read" was included as part of the RC. This read contains the genomic reference sequence of the entire range spanned by the RC.

In each RC, the reads were decomposed into overlapping k-mers (k-1 overlap) which are the nodes of the de Bruijn Graph (dBG). The edges represent consecutive nodes (i.e., two k-mers overlapping by k-1) in the reads. Every edge was labeled with the name of the reads that contained this consecutive pair of k-mers and the position in the read where the k-mers were located.

The graph was then processed by compacting and removing certain nodes. Walks on the graph that contain consecutive nodes of *in-degree = 1* and *out-degree = 1* can be merged into a single node that contains a sequence length of $k + n - 1$ bases where n is the number of nodes being merged. In addition, subsequences in the "reference read" that did not overlap with other reads (which are usually intronic segments) were also compressed. This step greatly simplifies the graph because the intronic regions are generally several thousand bases long, much longer than the average RC. Other nodes were removed from the graph if they did not provide useful information. For example, we defined the actual start and end of the RC as the first and last nodes that originate from the "reference read". By definition, these nodes have *in-degree = 0* and *out-degree = 0* respectively. Additional nodes that complied with the degree requirement but did not originate from the “reference read” were labeled as alternative starts and ends. These alternative starts and ends also represent differences among read sequences. However, since they do not form a bubble, it’s not possible to infer the variant causing such difference.

Subsequently, scAllele inferred the walk on the graph that matched the original sequence of each read. These walks were named “read walks”. Because some nodes were removed or merged in the previous step, this walk is not necessarily the same sequence of nodes obtained from the initial read decomposition. As a result, many of the original reads will be matched by the same “read walk”, reducing the number of distinct reads to process.

In the compacted/cleaned dBG, we identified the bubble structures by locating the source nodes, the sink nodes and the walks connecting them via depth-first search (DFS) of the graph. These structures represent variants, and with the DFS we can identify which specific source node, sink node, and connecting walk correspond to each allele. This information was then used to identify the variants and their alleles present on each “read walk”. In the case of highly interconnected/cyclic graphs (due to existence of repeats or low complexity regions), this assignment was aided by a Dijkstra-like algorithm which identifies the most likely set of variants on a “read walk” by minimizing the editing distance between the read walk sequence and the reference sequence. More specifically, first, all the end-to-end “read walks” were identified. Then, by calculating the cumulative edit distance at every node and traversing the graph through different walks, we can select the best walk. Note that introns were also considered a type of variants in these intermediate steps and are processed in the same way as the nucleotide variants. However, we did not assign an edit distance to them.

The variants were further processed by normalizing, left aligning and atomizing them. Different features were collected for each variant including read counts for each allele, base qualities, read positions, and count of tandem repeats flanking the variant. An additional feature, namely haplotype-fitting was calculated using the entire set of reads and variants from each RC. These features were then used to score the quality of the variant (see Variant Scoring). At this point, the variant-calling step was complete. Since scAllele identifies variants at the read level, this information was stored in memory for subsequent mutual information analysis, based on which the linkage between nucleotide variants and splicing isoforms was calculated (see linkage analysis).

3.5.2 Variant scoring

We trained a generalized linear model (GLM) using "ground truth" genetic variants and various features obtained from the main algorithm of scAllele. The features were the variant's ALT allelic ratio, the number of tandem repeats neighboring the variant, the sequencing error rate (SER), the median base quality in the variant's proximity, the read position, and the haplotype-fitting.

Low ALT allelic ratio is often indicative of a false variant. If a variant results from a sequencing error, then it is likely to have low frequency in the reads. In fact, we can calculate the probability of observing an allelic ratio (AB) if it is resulted from a sequencing error using the binomial distribution.

$$\mathrm{p(AB \mid seq.error)} = \binom{DP}{AC} f^{AC}(1-f)^{DP-AC}$$

Where *AC* is the ALT allele counts, and *DP* is the total read count. The value of *f* is the probability of error which in most cases corresponds to the SER. The value used for SER was

0.01 which is the maximum error rate for the Illumina sequencing platforms(Pfeiffer et al., 2018). On tandem repeats, however, the probability of error is expected to increase due to the propensity of PCR slippage. We then defined *f* as follows:

$$f = \begin{cases} SER + 0.075 \; if \; TandemRep \geq 5 \; and \; varLength = 1 \\ SER + 0.035 \; if \; TandemRep \geq 5 \; and \; varLength \; \geq 2 \\ SER \; otherwise \end{cases}$$

These values are approximations of the empirical estimation of stutter noise made in lobSTR (Gymrek et al., 2012). The stutter noise was found to be a function of the variant length (*varLength*) and the number of tandem repeats (*TandemRep*). This p-value was used as an additional feature in the GLM and served as an interaction term between *AB* and *TandemRep.*

The base quality was also used to identify sequencing errors. In the case of SNVs, we simply used the base quality at the mismatch position. In the case of INDELs, we used the median base quality in the neighboring region of the variant (+/- 7 bases) since the original position of the INDEL is, in many cases, ambiguous. The median read position was also used as a metric, not only because the 3' ends of the reads tends to have lower base quality, but also because the bubble structures in the dBG are less reliable if they only use the ends of the "read walks". Lastly, scAllele calculated another metric called "haplotype-fitting". This refers to the ability of clustering the variants alleles into two potential haplotypes based on their colocalization in the reads. We clarify that we do not aim to infer the actual haplotype since RNA-seq data is not ideal for this task. This step simply checks for multi-allelic variants and allele combinations that result in more than two haplotypes. For this step, we discarded potential RNA editing sites and we perform the clustering at the RC level which, most of the time, matches with exonic coordinates. In this way, the haplotype is not disturbed by non-genetic variants or allele-specific splicing.

The regression of the GLM was performed using the scikit-learn package (Pedregosa et al., 2011) from Python. The training data consist of genetic variants identified in scRNA-seq data originated from the GM12878 cell line. We used the ground truth from GIAB (Zook et al., 2019) and trained the model on a subset of the dataset used in the "*Evaluation of variant calls in GM12878 and iPSC cells"* section. It's important to note that the training data was not used to derive the performance results in that section. We trained three separate models for SNPs, insertions, and deletions respectively. The data was shuffled and split 25 times for cross-validation. In each round, the model was fitted and tested using AUROC and F1 as performance metrics. Overall, the performance scores were very consistent across iterations with a mean F1 of 0.78, 0.87, and 0.97 and AUROC of 0.85, 0.88, 0.68 for deletions, insertions, and SNPs respectively. The standard deviation of these values over different iterations was less than 0.02. The score from the regression is the following log-likelihood:

$$log\left(\frac{p(Variant = True)}{p(Variant = False)}\right) = GLM(feature1, feature2, \dots)$$

The specifications for a standard VCF file (specified by VCFtools, v.4.2) require the variant's quality score (QUAL) to be Phred-scaled in the form:

$$QUAL = -10 \times log_{10}(p(Variant = False))$$

Thus, a regression score of zero (QUAL = 3.01) represents equal probability of a variant being true or false and is the theoretical cutoff. However, from the benchmark evaluation, we learned that F1 was usually maximized at regression scores between 1 and 2 (10.4 ≤ QUAL ≤ 20.0). This cutoff can be user-defined according to the stringency demand of their analyses and the score format they prefer. The default score format is QUAL and the cutoff is 10.

3.5.3 Experimental validation of novel INDELs

Variants from the GM12878 scRNA-seq data which had high QUAL score but were classified as false positives were selected for experimental validation. The variants loci were amplified from gDNA of GM12878 cells and cloned into vectors (pCR2.1-TOPO or pCag-EGFP). Single colonies were picked for plasmid isolation and Sanger sequencing. (Supplemental Fig. 3-3).

3.5.4 Linkage analysis

scAllele detects variants at the read level allowing for allelic linkage detection between them. For every RC, all the reads that overlap a variant position were collected with their corresponding allele (REF or ALT). The reads from different RC's were pooled together after scanning an entire chromosome. In paired-end data, a read cluster may not contain both mates of the reads. Thus, by merging reads from different RC's we can increase the number of potential linkages.

For every pair of variants that were less than 100 kb apart, scAllele retrieved the reads that overlap both variants and their corresponding alleles. From these reads, one array per variant was constructed with the allele information. We used Python's scikit-learn package(Pedregosa et al., 2011) to calculate the mutual information between these two arrays.

For the calculation of linkage between nucleotide variants and splicing isoforms, scAllele first grouped all the introns found based on their overlap. For example, if there was a genomic interval that overlapped all intron 1, intron 2, and intron 3, then such interval was labeled as "intronic part" and the overlapping introns were considered "alleles" of that "intronic part". With this conversion, introns can be used to calculate mutual information in the same way as a nucleotide variant.

Based on our analysis in the section "*Linkage calculation between variants"* we determined that at least 10 common reads were necessary for a pair of variants to be testable. Additionally, a minimum mutual information of 0.52 was required for significant linkage.

3.5.5 scRNA-seq processing and mapping

Raw scRNA-seq fastq files from GM12878 cell line was retrieved from ENCODE (Accession: ENCSR000AIZ). These replicates were deeply sequenced (about 30 million reads); thus, we down sampled each replicate into 30 alignment files with roughly 1 million reads each. The goal was to resemble a shallowly sequenced sample to test our method on low coverage data.

The scRNA-seq data from iPSC cells, corresponding to individuals NA19098, NA19101, and NA19239, was obtained from NCBI (Accession number: GSE77288) (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE77288)

The lung cancer (Maynard et al., 2020) dataset is available through NCBI (BioProject PRJNA591860) (https://www.ncbi.nlm.nih.gov/bioproject/28889).

Raw reads from all samples were pre-processed using fastqc (v.0.11.7) (A. Simons, 2010) to check for adapter content and over-represented sequences. If present, these sequences were removed using cutadapt (v.1.9) (Martin, 2011). 3'end of reads with low base quality were also trimmed using sickle (v.1.33) (Joshi & Fass, 2011). The reads were mapped to hg38 (GM12878 and IPSC samples) and to hg19 (Lung Cancer samples) using two-pass STAR alignment (v.2.7.0c) (Dobin et al., 2013). Finally, we marked PCR duplicates using the tool MarkDuplicates from Picard Tools (v.2.25.2) (Broad Institute, 2019) (Supplemental Fig. 3-1).

3.5.6 Benchmarking

Variant calling performed by Platypus, GATK-HaplotypeCaller and Freebayes were first filtered to remove variants with alternative ALT count < 2 and A-to-G, C-to-T mismatches as they may represent RNA editing sites. Variants with labels indicating low quality were also removed.

We evaluated the performance of the variant callers using the vcfeval function from rtg tools (v.3.12) (Cleary et al., 2015) as suggested by the benchmarking standards determined previously (Krusche et al., 2019), with the following parameters:

```
RTG vcfeval -T 1 -b $TRUTH_VCF -c $QUERY_VCF -o $OUTPUT -t $REF_SDF -f 'QUAL' --bed-regions=$RC_BED --all-records --decompose --ref-overlap --sample ALT --output-mode='annotate'
```

The variable RC_BED is a bed file containing all the genomic regions covered by at least one read. We employ it to reduce the running time of the software. The option "--sample ALT" was used to skip the genotype matching and is more appropriate for RNA data.

The performance metric used in this study was true positive count at a fixed false-positive thresholds. Given that the ground truth genetic variants were obtained from WGS, the sensitivity of variant calling in RNA-seq is in principle restricted to the number of genetic variants that are transcribed and sequenced in RNA. Hence, sensitivity, true-positive rate and F1 scores won't accurately reflect the performance of variant callers in RNA-seq.

For the evaluation of variant calling in difficult regions, we used the union of difficult region bed files from GIAB which merges regions of low-mappability, high GC-content, Segment duplication, low complexity, functional regions, and other difficult regions.

3.5.7 Detection of cancer-enriched variants and annotation

We used scAllele on the lung cancer dataset from Maynard et al (Maynard et al., 2020) and selected the cells corresponding to two individuals (TH238, TH179) where both cancer and normal tissue biopsies were obtained. Then, we focused on variants present in at least three cells of an individual. We calculated the prevalence of each variant across cells. Using the hypergeometric test, we evaluated the enrichment of each variant in cancer cells compared to normal. The p-value obtained was then corrected for multiple testing using the BH method. We defined a given variant as cancer-enriched if the BH corrected p-value is ≤ 0.1 or if the variant is not present in any normal cell.

We further overlapped the variants with the COSMIC (cancer.sanger.ac.uk) (Tate et al., 2019) and, subsequently, dbSNP (b151) (Sherry et al., 2001) databases to annotate them. From the COSMIC annotation, we only selected variants that were confirmed to be somatic and were found in lung tissue. A variant was labeled "novel" if it is not present in either database.

3.5.8 Detection of differential linkage events

To detect differential linkage, we selected common linkage events (same nucleotide variant and same introns) between the cells of the same individual. We then classified them into four categories explained by the two proposed scenarios. For scenario 1, we selected linkage events that were present in cancer cells, but not in normal cells, or vice versa. These events were grouped into the categories “cancer-specific” and “normal-specific”, respectively. For scenario 2, we selected linkage events that were present in both types of cells, but significantly more prevalent in one compared to the other. These events were grouped into the categories “cancer-differential” and “normal-differential”. To detect differential prevalence, we used the Fisher’s Exact test with the number of cells with the linkage event in each biopsy and the number of cells that were testable for linkage.

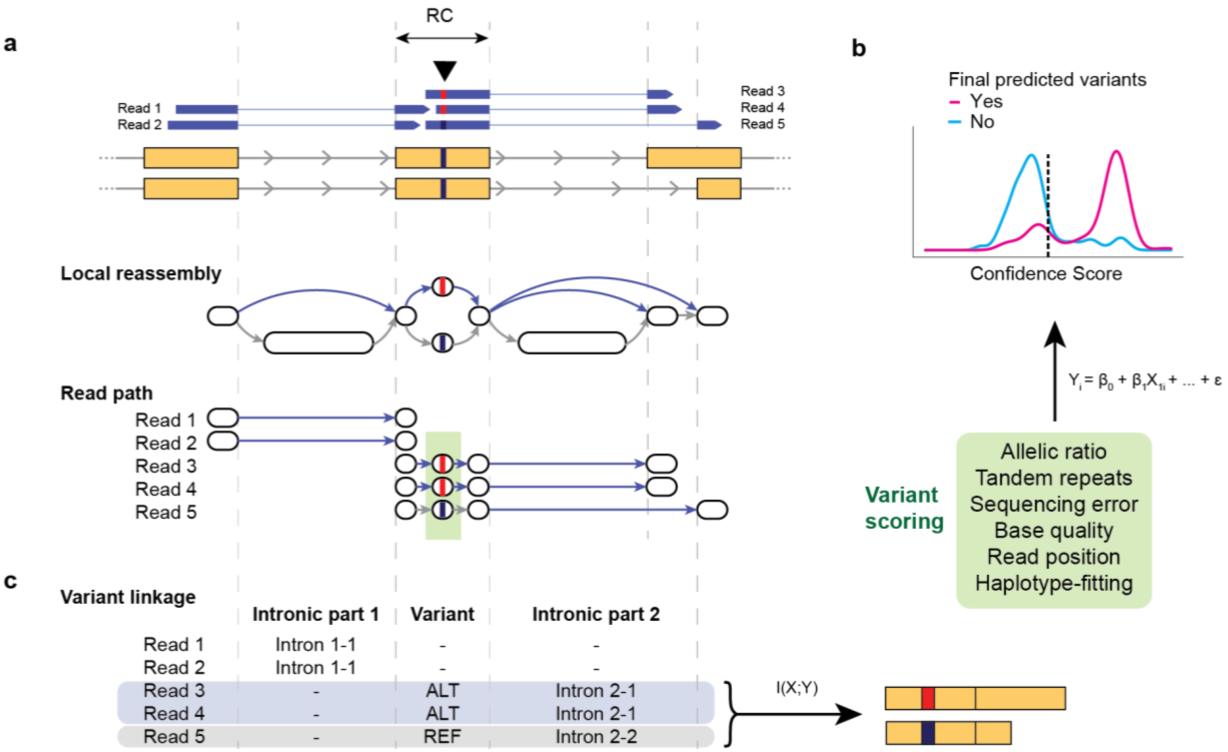
3.6 CODE AVAILABILITY

The scAllele software is available at PyPI (https://pypi.org/user/giovanniquinones/scAllele). The scripts used for the analyses in this work are available in our github repository: https://github.com/gxiaolab/scAllele/Manuscript

3.7 DATA AVAILABILITY

Variant calls and linkage events from the GM12878, IPSC cells for individuals NA19098, NA19101 and NA19239, and the lung cancer cells are available in our github repository: https://github.com/gxiaolab/scAllele/data

3.8 FIGURES



	Intronic part 1	Variant	Intronic part 2
Read 1	Intron 1-1	-	-
Read 2	Intron 1-1	-	-
Read 3	-	ALT	Intron 2-1
Read 4	-	ALT	Intron 2-1
Read 5	-	REF	Intron 2-2

Figure 3-1. scAllele algorithm outline. (a) Illustration of the main algorithm of scAllele for variant calling. The reads and the reference genomic sequence overlapping a read cluster (RC) are decomposed into k-mers and are reasembled into a de Bruijn graph (Local reassembly). The graph shown here is a compacted version. The ‘bubbles’ in the graph indicate a sequence mismatch i.e. a variant. For each read, scAllele obtains a path for the original read sequence and infers the allele of each variant (including introns) (Read Path). **(b)** Variants (green box in a) identified from the graph are then scored using a generalized linear model (GLM). The GLM was trained with different features (green box) to assign a confidence score to the variants. (See Methods for details). **(c)** To identify allele-specific splicing (i.e., variant linkage), scAllele performs a mutual information calculation between nucleotide variants (SNVs, microindels) and

intronic parts (where the ‘alleles’ are the different overlapping introns), to calculate allelic linkage of splicing isoforms.

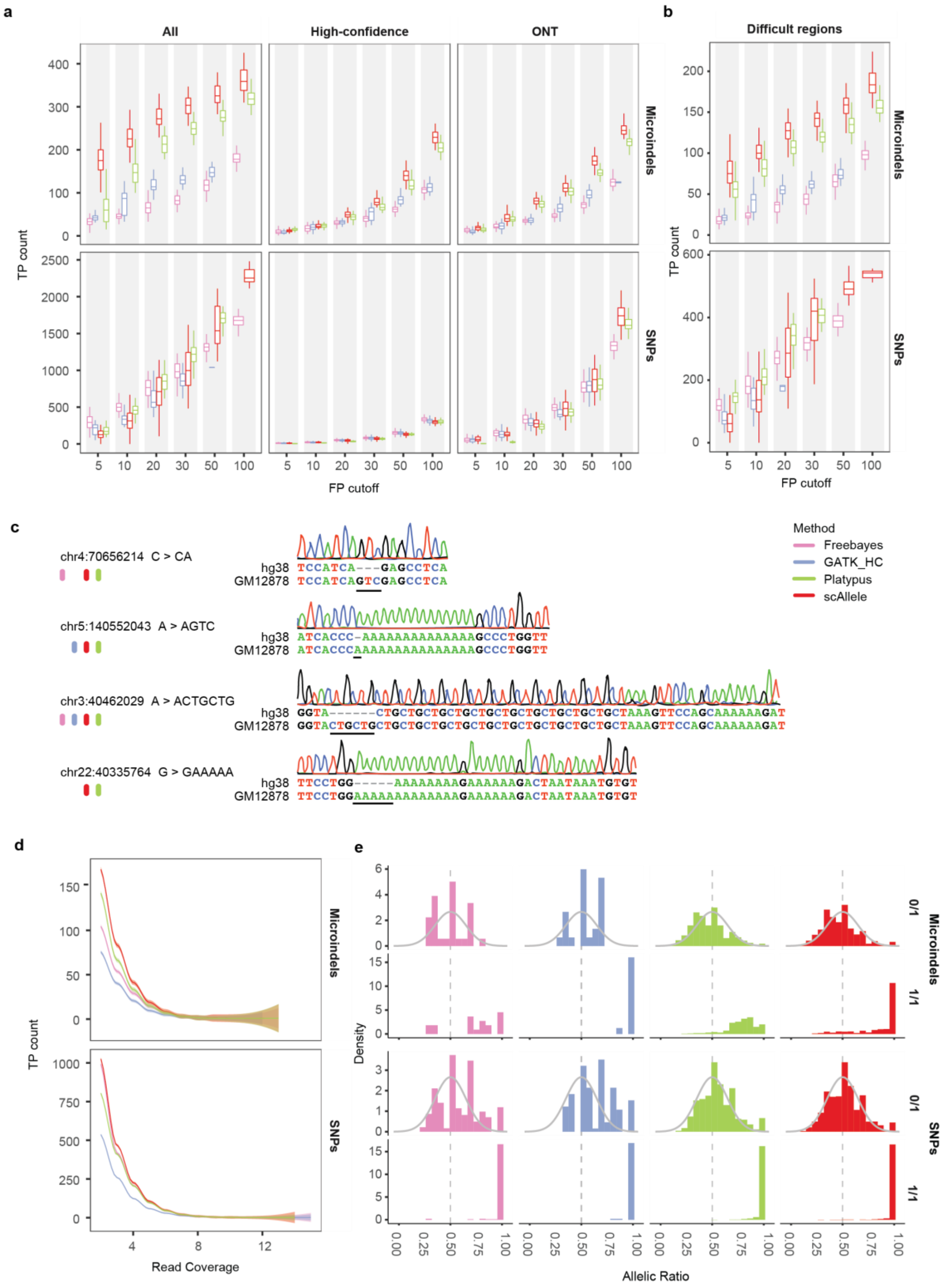

a
All
High-confidence
ONT
Microindels
SNPs
TP count
FP cutoff
5 10 20 30 50 100
0 100 200 300 400
0 500 1000 1500 2000 2500
b
Difficult regions
Microindels
SNPs
TP count
FP cutoff
0 50 100 150 200
0 200 400 600
c
chr4:70656214 C > CA
hg38 TCCATCA---GAGCCTCA
GM12878 TCCATCAGTCGAGCCTCA
chr5:140552043 A > AGTC
hg38 ATCACCC-AAAAAAAAAAAAAAGCCCTGGTT
GM12878 ATCACCCAAAAAAAAAAAAAAAGCCCTGGTT
chr3:40462029 A > ACTGCTG
hg38 GGTA------CTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTAAAGTTCCAGCAAAAAAGAT
GM12878 GGTACTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTAAAGTTCCAGCAAAAAAGAT
chr22:40335764 G > GAAAAA
hg38 TTCCTGG-----AAAAAAAAGAAAAAAGACTAATAAATGTGT
GM12878 TTCCTGGAAAAAAAAAAAAAGAAAAAAGACTAATAAATGTGT
Method
Freebayes
GATK_HC
Platypus
scAllele
d
Microindels
SNPs
TP count
0 50 100 150
0 250 500 750 1000
4 8 12
Read Coverage
e
Density
0/1
1/1
Microindels
SNPs
0 2 4 6
0 5 10 15
0 1 2 3
0.00 0.25 0.50 0.75 1.00
Allelic Ratio

Figure 3-2. Performance of scAllele in variant calling. **(a)** scRNA-seq data of GM12878 cells are analyzed. True positive (TP) count of four variant callers evaluated at different false positive (FP) thresholds. Performance shown for microindels (top) and SNPs (bottom). Three different ground truth sets of genetic variants were used for this evaluation: all GIAB-reported small variants (all), GIAB's high-confidence variants (high-confidence) and variants called via Nanopore sequencing by Karst et al. (ONT). **(b)** Performance in 'difficult regions' defined by GIAB. **(c)** Experimental validation of novel INDELS via Sanger sequencing. The dots under the variant's information are the methods that detected them. **(d)** True positive count of four variant callers (at maximum F1 and specificity > 0.9) across different read coverages. The line indicates the average of all cells. **(e)** Allelic ratios in the RNA of true positive variants (at maximum F1 and specificity > 0.9) segregated by their true genotype. 0/1: heterozygous; 1/1: homozygous variant allele. A theoretical normal distribution of mean = 0.5 and std. dev = 0.15 is included as a reference distribution of heterozygous (0/1) variants.

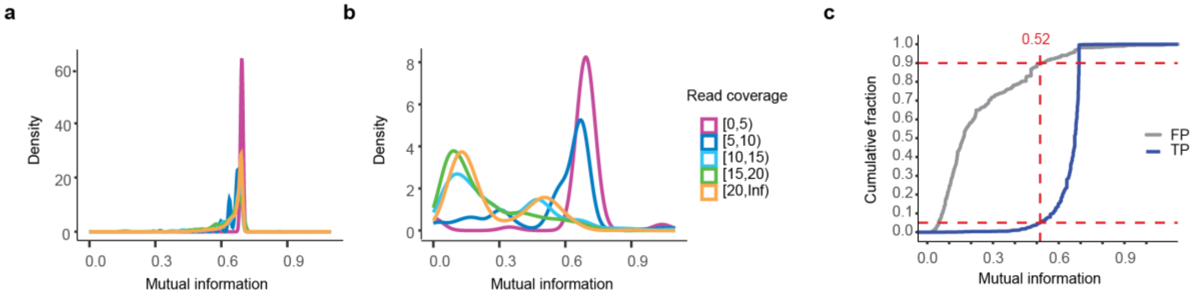


Figure 3-3. MI Linkage calculation between variant pairs. (a) Mutual information (MI) distribution (natural log-based) of pairs of true genetic variants (namely, true positives (TPs)) from the GM12878 scRNA-seq segregated by the number of reads covering the pair. Most of these pairs have values close to the theoretical maximum for two alleles ln(2) = 0.693 regardless of coverage. **(b)** MI distribution of pairs of variants where at least one is not a known genetic variant (namely, false positives (FPs)). **(c)** Cumulative distribution of MI of TP and FP pairs with a minimum read coverage of 10. The MI cutoff of 0.52 was selected as the minimum value for significant linkage between variants. This cutoff rejects 90% FP pairs and 5% TP pairs (dashed lines).

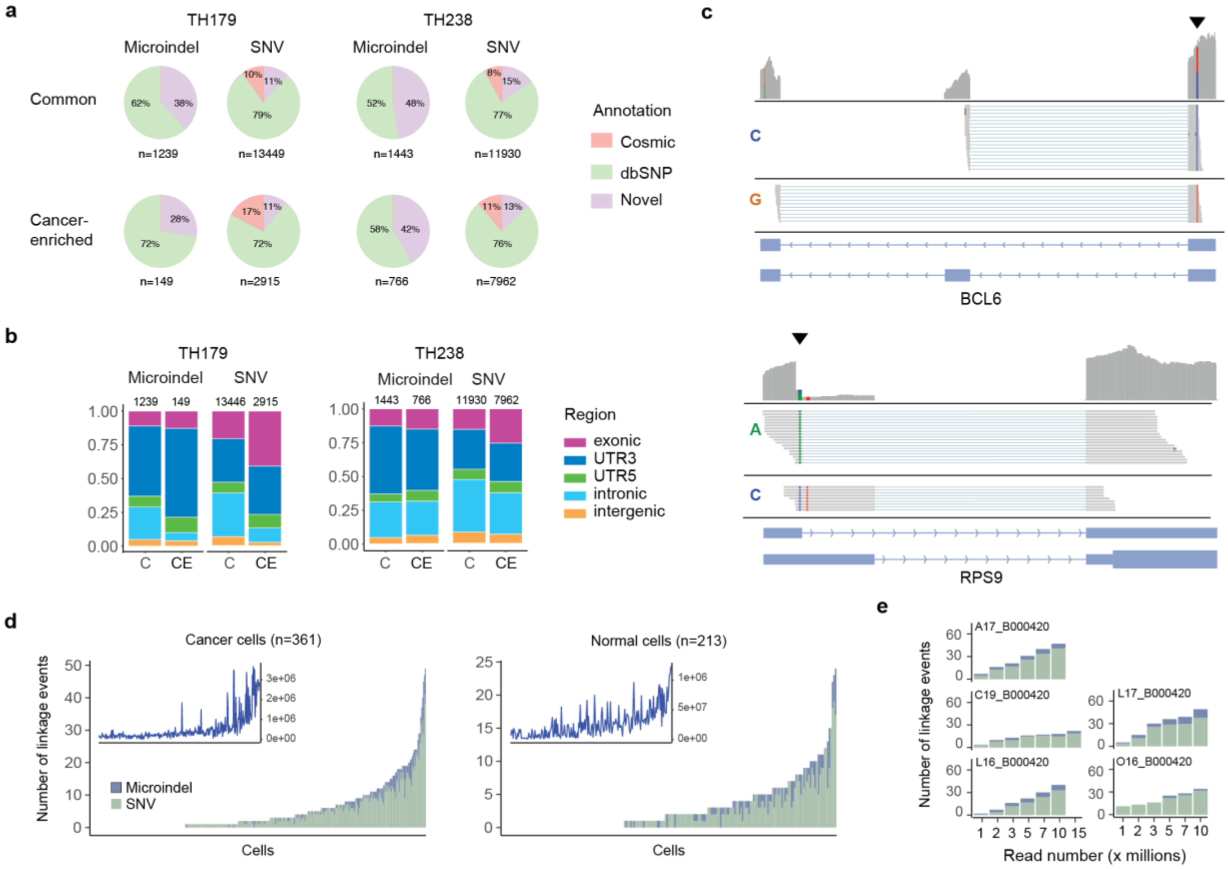


Figure 3-4. Summary of variants and linkage events detected by scAllele in the Lung Cancer dataset. (a) Variants identified in each individual (TH179 and TH238) (≥3 cells). Common: variants common to cancer and normal cells. Cancer-enriched: variants enriched in cancer cells (see Text). Novel: variants not annotated in dbSNP or Cosmic. Variants labelled as "Cosmic" may also present in dbSNP. **(b)** Distribution of variants in a in different types of genomic regions. C: common. CE: cancer-enriched. **(c)** IGV view of two example allele-specific splicing events. The location of the variant is denoted by the black arrowhead. The two alleles of each variant are colored differently. **(d)** Number of linkage events identified in cancer or normal cells (union of cells from the two individuals). The cells are ranked by their total number of events. Events where the nucleotide variant is microindels and SNVs are shown in different

colors. The insets show the total number of junction reads in the scRNA-seq data of the cells sorted in the same order as the main panel. **(e)** Number of linkage events identified in 5 deeply sequenced cells at different down sampled total read coverage.

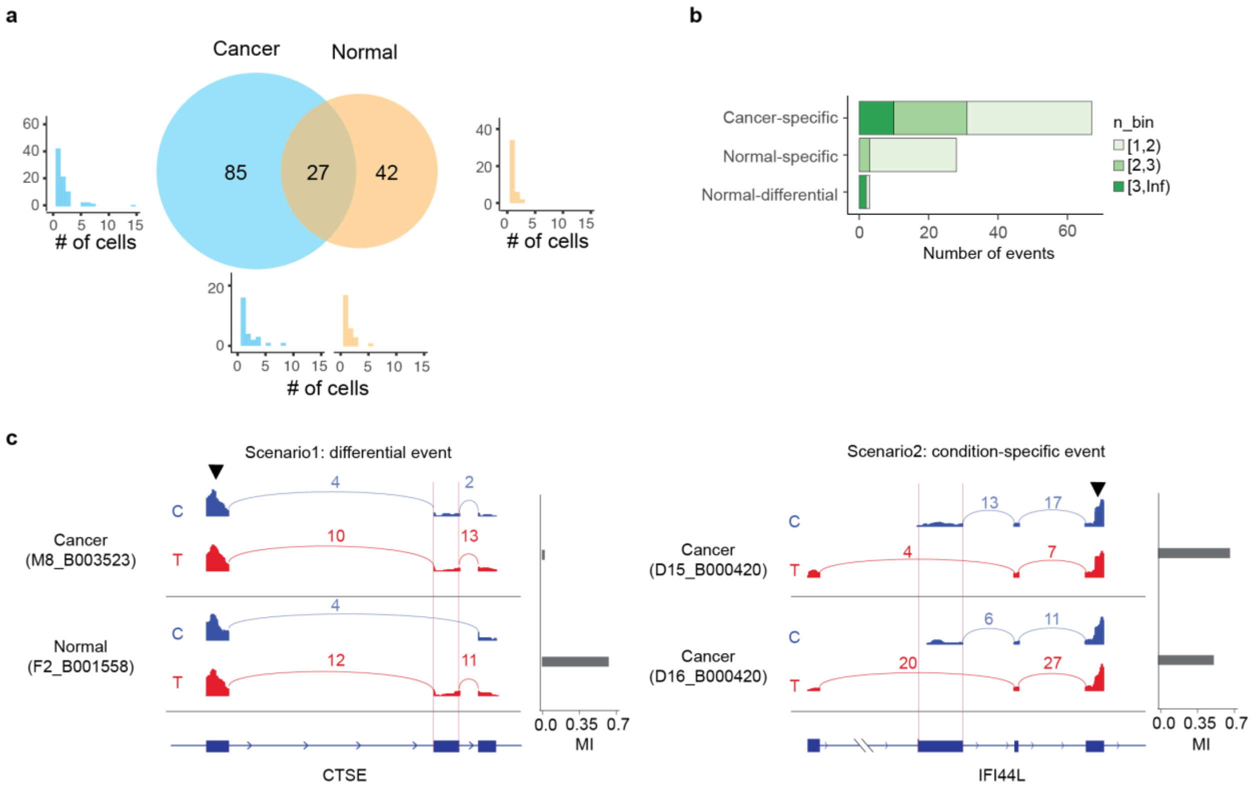
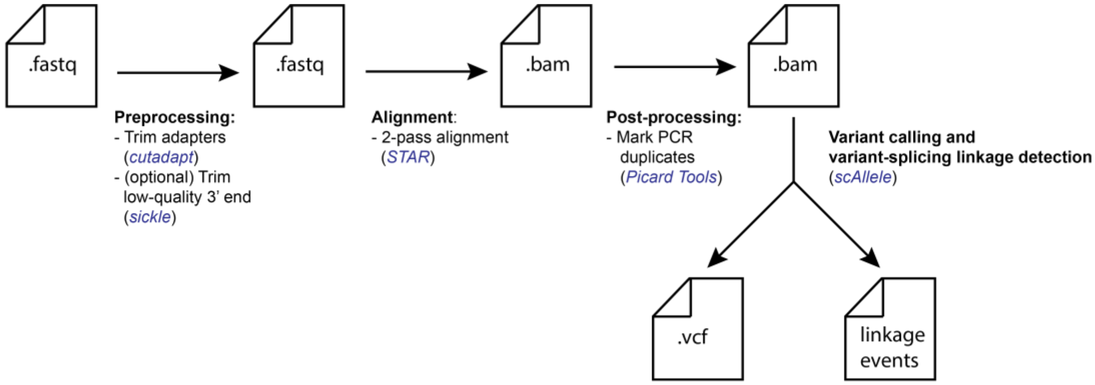
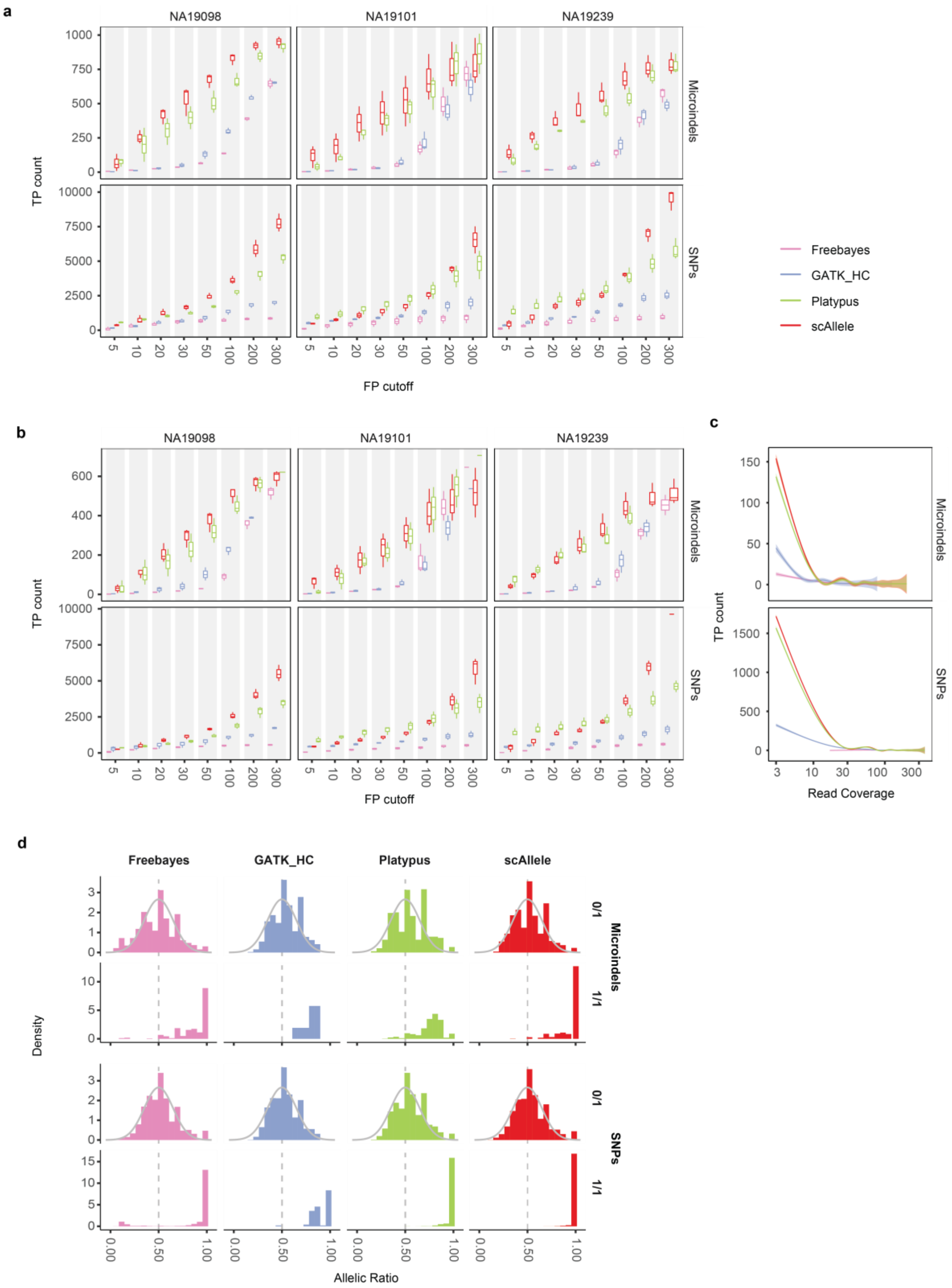


Figure 3-5. Comparison of allele-specific splicing events in cancer and normal cells. (a) Number of events shared by cancer and normal cells, or exclusive to one of the two classes. Histograms of the number of cells harboring each type of events are also shown. **(b)** Number of events categorized as cancer-specific, normal-specific, or differential (see Text for details). c. Example allele-specific splicing events.

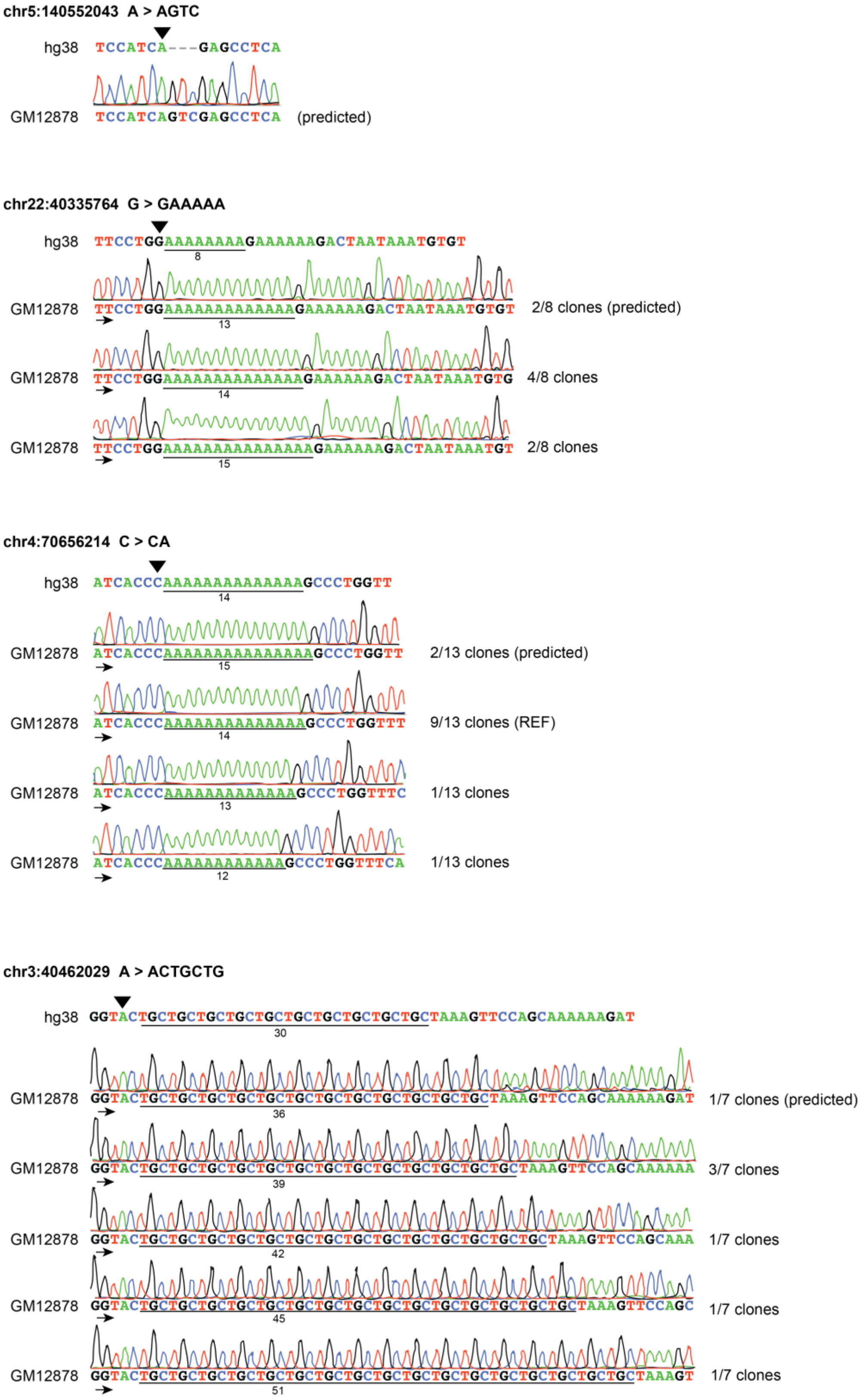
3.9 SUPPLEMENTAL FIGURES



Supplemental Figure 3-1. Recommended step-by-step pipeline for variant-calling and variant-splicing linkage detection using scAllele. Reads from raw fastq files were processed to trim adapters and low quality 3’ ends (names of the tools shown in blue). These reads were then aligned by STAR using the 2-pass alignment option. This option allows for the accurate mapping of splice junctions which is important for accurate identification of variant-splicing linkage detection. After alignment, PCR duplicates were marked using Picard Tools. Duplicated reads amplify the frequency of random errors which can lead to false-positive calls. Finally, scAllele calls variants and identifies variant-splicing linkage.


a
NA19098
NA19101
NA19239
Microindels
SNPs
TP count
FP cutoff
0
250
500
750
1000
2500
5000
7500
10000
5
10
20
30
50
100
200
300
Freebayes
GATK_HC
Platypus
scAllele
b
NA19098
NA19101
NA19239
Microindels
SNPs
TP count
FP cutoff
200
400
600
c
TP count
Read Coverage
Microindels
SNPs
3
30
100
150
1500
d
Freebayes
GATK_HC
Platypus
scAllele
0/1
1/1
Microindels
SNPs
Density
Allelic Ratio
0.00
0.50
1.00
15

Supplemental Figure 3-2. Variant call performance evaluation on iPSC scRNA-seq of three individuals. (a) True positive (TP) count of four variant callers evaluated at different false positive (FP) limits. Performance shown for microindels (top) and SNPs (bottom). **(b)** Performance in ‘difficult regions’ defined by GIAB. **(c)** True positive count of four variant callers (at maximum F1 and specificity > 0.9) across different read coverages. The line indicates the average of all cells. **(d)** Allelic ratios in the RNA of true positive variants (at maximum F1 and specificity > 0.9) segregated by their true genotype. 0/1: heterozygous; 1/1: homozygous variant allele. A theoretical normal distribution of mean = 0.5 and std. dev = 0.15 is included as a reference distribution of heterozygous (0/1) variants.


chr5:140552043 A > AGTC
hg38 TCCATCA---GAGCCTCA
GM12878 TCCATCAGTCGAGCCTCA (predicted)
chr22:40335764 G > GAAAAA
hg38 TTCCTGGAAAAAAAAGAAAAAAGACTAATAAATGTGT
8
GM12878 TTCCTGGAAAAAAAAAAAAAGAAAAAAGACTAATAAATGTGT 2/8 clones (predicted)
13
GM12878 TTCCTGGAAAAAAAAAAAAAAGAAAAAAGACTAATAAATGTG 4/8 clones
14
GM12878 TTCCTGGAAAAAAAAAAAAAAAGAAAAAAGACTAATAAATGT 2/8 clones
15
chr4:70656214 C > CA
hg38 ATCACCCAAAAAAAAAAAAAAGCCCTGGTT
14
GM12878 ATCACCCAAAAAAAAAAAAAAAGCCCTGGTT 2/13 clones (predicted)
15
GM12878 ATCACCCAAAAAAAAAAAAAAGCCCTGGTTT 9/13 clones (REF)
14
GM12878 ATCACCCAAAAAAAAAAAAAGCCCTGGTTTC 1/13 clones
13
GM12878 ATCACCCAAAAAAAAAAAAGCCCTGGTTTCA 1/13 clones
12
chr3:40462029 A > ACTGCTG
hg38 GGTACTGCTGCTGCTGCTGCTGCTGCTGCTGCTAAAGTTCCAGCAAAAAAGAT
30
GM12878 GGTACTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTAAAGTTCCAGCAAAAAAGAT 1/7 clones (predicted)
36
GM12878 GGTACTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTAAAGTTCCAGCAAAAAA 3/7 clones
39
GM12878 GGTACTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTAAAGTTCCAGCAAA 1/7 clones
42
GM12878 GGTACTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTAAAGTTCCAGC 1/7 clones
45
GM12878 GGTACTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTAAAGT 1/7 clones
51

Supplemental Figure 3-3. Experimental validation of novel INDELs from GM12878 scRNA-seq. The predicted variant information is shown (chromosome, position, reference allele, and alternative allele). The triangle indicates the position of the variant in the hg38 reference sequence. The different traces indicate the INDELs present in GM12878 cells (PCR amplified and SANGER sequenced, see Methods for details). The number of clones containing each INDEL is also shown. 3 of the 4 variants are adjacent to homopolymer or tandem repeat sequences. The black line underneath indicates the length of the repeat. The arrow indicates the direction of sequencing.

Table 3-1. List of linkage events identified in the Lung Cancer dataset. Nucleotide variant alleles and introns "alleles" (SP) are shown for each linkage event. The differential classification for each event, as calculated based on occurrence (See Methods).

INDIVI DUAL	VARIANT	GENE	INTRONS (SP = ALLELE)	DIFFERENTIAL CLASSIFICATION
TH179	chr1:206327510:T>C	CTSE	SP1:206327597-206328961 SP2:206328860-206328962	NORMAL_diff
TH179	chr1:206327510:T>C	CTSE	SP1:206327597-206328961 SP2:206327597-206328718	TUMOR_specific
TH179	chr11:308363:G>C	IFITM2	SP1:308439-314921 SP2:308438-309011	NORMAL_specific
TH179	chr14:94849420:G>A	SERPINA1	SP1:94849579-94854896 SP4:94849579-94856793	NORMAL_specific
TH179	chr15:40330518:T>G	SRP14	SP1:40329213-40330475 SP2:40329220-40330482	NORMAL_specific
TH179	chr19:10395683:A>G	ICAM1	SP1:10395684-10395788 SP2:10395705-10395790	TUMOR_specific
TH179	chr19:19011244:T>C	COPE	SP1:19011258-19015578 SP2:19011259-19014076	TUMOR_specific
TH179	chr20:57478807:C>T	GNAS	SP1:57470740-57478582 SP2:57474040-57478584	NORMAL_specific
TH179	chr9:130630749:C>G	AK1	SP1:130630792-130634101 SP2:130630755-130634101	NORMAL_specific
TH179	chr7:95041016:G>C	PON2	SP1:95041056-95041623 SP2:95041092-95041623	NORMAL_specific
TH179	chr11:33757941:A>G	CD59	SP1:33753015-33757927 SP2:33755157-33757927 SP3:33744008-33757925 SP4:33753008-33757927	NORMAL_specific
TH179	chr11:102192574:A>C	BIRC3	SP1:102190737-102192566 SP2:102188300-102192564	TUMOR_specific
TH179	chr10:7849688:T>C	ATP5C1	SP1:7848974-7849621 SP2:7844818-7849621	NORMAL_specific
TH179	chr11:64009879:G>A	FKBP2	SP1:64008711-64009855 SP2:64008604-64009854	NORMAL_specific

TH179	chr11:64089216:C>G	PRDX5	SP1:64088263-64088334 SP2:64085858-64088335	NORMAL_specific
TH179	chr12:120933946:T>G	DYNLL1	SP1:120934019-120934203 SP2:120934019-120934217	NORMAL_specific
TH179	chr16:54953071:G>A	CRNDE	SP1:54954322-54957495 SP2:54954239-54957495	NORMAL_specific
TH179	chr17:16052816:C>T	NCOR1	SP1:16052832-16055259 SP2:16052832-16054903	NORMAL_specific
TH179	chr22:50315382:C>G	CRELD2	SP1:50315407-50315938 SP2:50315409-50316258	TUMOR_specific
TH179	chr1:43737863:G>A	TMEM125	SP2:43736924-43737853 SP3:43736465-43737854	NORMAL_specific
TH179	chr8:57219269:T>TA	SDR16C5	SP1:57218282-57219234 SP2:57214133-57219234	TUMOR_specific
TH179	chr8:37728017:A>G	RAB11FIP1	SP1:37728047-37732032 SP2:37728047-37728795	NORMAL_specific
TH179	chr8:37728019:T>G	RAB11FIP1	SP1:37728047-37732032 SP2:37728047-37728795	NORMAL_specific
TH179	chr2:238449007:T>C	MLPH	SP1:238434448-238448989 SP2:238436159-238448989 SP3:238443290-238448989	NORMAL_specific
TH179	chr2:238449023:T>C	MLPH	SP1:238434448-238448989 SP2:238436159-238448989 SP3:238443290-238448989	NORMAL_specific
TH179	chr2:238449023:T>C	MLPH	SP1:238443290-238448989 SP2:238436160-238448990	NORMAL_specific
TH179	chr2:238449107:A>G	MLPH	SP1:238449177-238451209 SP2:238449176-238449443	NORMAL_specific
TH179	chr1:47280859:G>A	CYP4B1	SP1:47280936-47334500 SP2:47282853-47283635	TUMOR_specific
TH238	chr10:81317836:G>A	SFTPA2	SP1:81317862-81371607 SP2:81317878-81371661 SP5:81319754-81320118	TUMOR_specific
TH238	chr10:81317836:G>A	SFTPA2	SP1:81319753-81320117 SP2:81317877-81372130 SP6:81317884-81372137 SP8:81317862-81371607	TUMOR_specific

TH238	chr10:81318721:C>T	SFTPA2	SP6:81319206-81372960 SP8:81319753-81320117 SP11:81318747-81372067	TUMOR_specific
TH238	chr10:81318736:T>C	SFTPA2	SP4:81319109-81372117 SP6:81319206-81372960 SP8:81319753-81320117 SP11:81318747-81372067	TUMOR_specific
TH238	chr10:81318736:T>C	SFTPA2	SP5:81318747-81372067 SP8:81319206-81372960 SP19:81319116-81372124 SP20:81319149-81372941 SP21:81319753-81320117	TUMOR_specific
TH238	chr10:81318737:G>A	SFTPA2	SP4:81319109-81372117 SP5:81319754-81320118 SP6:81319206-81372960 SP8:81319149-81372941	TUMOR_specific
TH238	chr10:81318748:C>T	SFTPA2	SP4:81319109-81372117 SP5:81319754-81320118 SP6:81319206-81372960	TUMOR_specific
TH238	chr10:81319184:G>A	SFTPA2	SP1:81319206-81372960 SP8:81319263-81320119	TUMOR_specific
TH238	chr10:81319184:G>A	SFTPA2	SP3:81319263-81320119 SP5:81319263-81319724	TUMOR_specific
TH238	chr10:81319184:G>A	SFTPA2	SP4:81319245-81372073 SP6:81319753-81320117 SP8:81319240-81370724 SP10:81319262-81320118	TUMOR_specific
TH238	chr10:81319184:G>A	SFTPA2	SP2:81319206-81372960 SP4:81319753-81320117 SP13:81319262-81320118	TUMOR_specific
TH238	chr11:72465877:T>C	STARD10	SP1:72466798-72468810 SP2:72466799-72469574	TUMOR_specific
TH238	chr17:76167047:C>T	SYNGR2	SP1:76167135-76167818 SP2:76167730-76167818	NORMAL_specific
TH238	chr10:32141460:T>C	ARHGAP12	SP1:32141526-32142993 SP2:32141526-32150322	TUMOR_specific

TH238	chr10:43891993:G>A	HNRNPF	SP1:43892039-43904578 SP2:43892039-43903164	TUMOR_specific
TH238	chr11:17298365:G>C	NUCB2	SP1:17298376-17304335 SP2:17298396-17304335	TUMOR_specific
TH238	chr14:94582130:T>TGGCCATGGC	IFI27	SP1:94577143-94581195 SP2:94578120-94581196	NORMAL_diff
TH238	chr17:6917703:C>T	RNASEK	SP1:6916077-6916984 SP2:6916835-6916983	TUMOR_specific
TH238	chr20:43530234:A>C	YWHAB	SP1:43514528-43530171 SP2:43516384-43530171	NORMAL_diff
TH238	chr6:2959513:C>T	SERPINB6	SP1:2959575-2971764 SP2:2959577-2971724 SP3:2959577-2962127	TUMOR_specific
TH238	chr6:2959513:C>T	SERPINB6	SP1:2959575-2971764 SP2:2959577-2971350	TUMOR_specific
TH238	chr7:12727826:G>C	ARL4A	SP1:12727353-12727789 SP2:12726669-12727790	TUMOR_specific
TH238	chr18:77724726:A>C	HSBP1L1	SP1:77724787-77726611 SP2:77724786-77730438	TUMOR_specific
TH238	chr14:24615435:T>C	PSME2	SP1:24615105-24615415 SP2:24614981-24615415	TUMOR_specific
TH238	chr2:10537018:A>G	HPCAL1	SP1:10537046-10548686 SP2:10537046-10559858	TUMOR_specific
TH238	chr16:88875928:T>C	APRT	SP1:88876248-88876476 SP2:88876115-88876477	TUMOR_specific
TH238	chr1:75190452:C>T	CRYZ	SP2:75190519-75196065 SP3:75190519-75198639	TUMOR_specific
TH238	chr2:87820915:C>T	LINC00152	SP1:87755164-87820724 SP2:87779942-87820724	TUMOR_specific
TH238	chr10:88730374:C>T	ADIRF-AS1;ADIRF	SP1:88728363-88730232 SP2:88728362-88729955	TUMOR_specific
TH238	chr20:32664864:C>CCAG	RALY	SP1:32661441-32663678 SP2:32661669-32663675	NORMAL_specific
TH238	chr1:8021778:T>C	PARK7	SP1:8021796-8022822 SP2:8021853-8022821	TUMOR_specific
TH238	chr16:4524155:C>G	NMRAL1	SP1:4524168-4524410 SP2:4524168-4524554	TUMOR_specific

TH238	chr2:56150864:C>T	EFEMP1	SP1:56150075-56150845 SP2:56149583-56150845	TUMOR_specific
TH238	chr22:42018038:C>T	XRCC6	SP1:42017348-42017991 SP2:42017594-42017991 SP3:42017414-42017992	TUMOR_specific
TH238	chr11:65664346:T>C	FOSL1	SP1:65661593-65664279 SP2:65660767-65664278	TUMOR_specific
TH238	chr17:42402119:T>C	SLC25A39	SP1:42400959-42402078 SP2:42400946-42402078	TUMOR_specific
TH238	chr19:1036444:G>A	CNN2	SP1:1036246-1036414 SP2:1032696-1036414	TUMOR_specific
TH238	chr19:1032689:G>T	CNN2	SP1:1032695-1036413 SP2:1032695-1036127	TUMOR_specific
TH238	chr8:67834876:G>GT	SNHG6	SP1:67834349-67834847 SP2:67834628-67834848	NORMAL_specific
TH238	chr6:49522865:G>T	C6orf141	SP1:49522870-49528245 SP2:49522871-49529416	TUMOR_specific
TH238	chr14:24686214:T>A	NEDD8	SP1:24687638-24701455 SP2:24687638-24700861	TUMOR_specific
TH238	chr14:104381513:A>G	C14orf2	SP1:104381527-104387276 SP2:104381527-104387806	TUMOR_specific
TH238	chr15:41146880:T>C	SPINT1	SP1:41146721-41146836 SP2:41146284-41146835	TUMOR_specific
TH238	chr7:43810764:G>A	BLVRA	SP1:43798333-43810735 SP2:43799709-43810735	TUMOR_specific
TH238	chr7:95045559:T>TA	PON2	SP1:95041764-95045553 SP2:95041789-95045552 SP3:95044585-95045553	NORMAL_specific
TH238	chr2:238449107:A>G	MLPH	SP1:238449176-238449443 SP2:238449175-238451207	NORMAL_specific
TH238	chr2:238449107:A>G	MLPH	SP1:238449599-238451207 SP2:238449176-238451208	TUMOR_specific
TH238	chr2:238449108:G>T	MLPH	SP1:238449176-238449443 SP2:238449175-238451207	NORMAL_specific
TH238	chr2:238449108:G>T	MLPH	SP1:238449599-238451207 SP2:238449176-238451208	TUMOR_specific

TH238	chr1:46774783:A>G	UQCRH	SP1:46774800-46775826 SP2:46774799-46775567	TUMOR_specific
TH238	chr13:43639845:A>C	DNAJC15	SP1:43597870-43639820 SP2:43636431-43639820	TUMOR_specific
TH238	chr17:79089590:A>G	BAIAP2	SP1:79082308-79089567 SP2:79084756-79089569	TUMOR_specific
TH238	chr3:187446211:C>T	BCL6	SP1:187444687-187446147 SP2:187443417-187446146	NORMAL_specific
TH238	chr2:228197238:G>A	MFF	SP1:228190143-228195340 SP2:228193506-228195341	TUMOR_specific
TH238	chr19:54704760:A>C	RPS9	SP1:54704830-54705027 SP2:54704757-54705027	TUMOR_specific
TH238	chr1:79095581:T>C	IFI44L	SP1:79094078-79094634 SP2:79086256-79094634	TUMOR_specific
TH238	chr20:2637071:T>C	NOP56	SP1:2636680-2637045 SP2:2636860-2637045	TUMOR_specific
TH238	chr20:390646:G>A	RBCK1	SP1:390668-391053 SP2:390669-398168	TUMOR_specific
TH238	chr14:93172903:C>T	LGMN	SP1:93170706-93172826 SP2:93171053-93172827	TUMOR_specific
TH238	chr3:107937408:C>T	IFT57	SP1:107934236-107937381 SP2:107932868-107937380	TUMOR_specific
TH238	chr5:131826322:G>A	IRF1	SP1:131825176-131826236 SP2:131822823-131826236	TUMOR_specific
TH238	chr19:20828522:A>G	ZNF626	SP1:20828255-20828489 SP3:20808457-20828489	TUMOR_specific
TH238	chr2:170460175:G>GT	PPIG	SP1:170441409-170460150 SP2:170441000-170460150	TUMOR_specific
TH238	chr5:131826413:G>C	IRF1	SP1:131825176-131826236 SP2:131822823-131826236	TUMOR_specific
TH238	chr6:41757339:C>CA	PRICKLE4; TOMM6	SP1:41757073-41757303 SP2:41757074-41757267	TUMOR_specific
TH238	chr17:16052816:C>T	NCOR1	SP1:16052832-16054903 SP2:16052832-16055259	TUMOR_specific
TH238	chr4:184367260:C>CA	CDKN2AIP	SP1:184366787-184367239 SP3:184366818-184367239	TUMOR_specific

TH238	chr11:32611129:G>GA	EIF3M	SP1:32605475-32611091 SP2:32610682-32611092	TUMOR_specific
TH238	chr8:86129663:G>GT	C8orf59	SP1:86127267-86129614 SP2:86126861-86129614	NORMAL_specific
TH238	chr14:70795929:G>GT	COX16	SP1:70795954-70809374 SP2:70795954-70826235	TUMOR_specific

CHAPTER 4 - Extending scAllele: a method to identify haplotype-specific isoform expression in long-read RNA-seq

4.1 INTRODUCTION

Third-Generation Sequencing (TGS) allows for the identification and discovery of full-length isoforms. Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) are the dominant technologies to perform long-read sequencing. Since their first release in 2010 and 2014 respectively, both technologies have been continuously improved, and a large number of computational methods have been synergistically developed for the analyses of this data. However, the sequencing error rate of these platforms remains an important issue. While short read sequencing has a reported error rate of 0.01% to 0.05% (Ardui et al., 2018; Fox et al., 2014), long-read sequencing error rate ranges from 1% to 15% (Ardui et al., 2018; Jain et al., 2018; Wenger et al., 2019). Such high error rate significantly complicates variant calling.

To address the issue of high error rate, a ‘read polishing’ or ‘error correction’ step is usually performed. Current approaches include k-mer-based reassembly of the circular consensus sequences (CCS)(Warren et al., 2019), complementing the CCSs with matched short reads (so called hybrid methods) (Koren et al., 2017; Vaser et al., 2017), and protecting known variants from exclusion by the polishing (Wyman & Mortazavi, 2019). To date, there is no clear consensus on the best practices to alleviate the high error rate in long-read RNA-seq (Amarasinghe et al., 2020). Many current methods have significant limitations and, if not applied carefully, they can significantly reduce the sensitivity in detecting sequence variants (Schmidt et al., 2017).

Variant calling in TGS DNA data has been thoroughly studied. In particular, DeepVariant is an effective method that enabled highly sensitive and precise variant calls for PacBio's long HiFi DNA-seq reads (Olsen et al., 2020; Wenger et al., 2019). This method represents a major milestone in the field. However, compared to DNA-seq, much less progress has been made towards TGS RNA-seq data.

We previously introduced the method scAllele, which is a tool that performs variant-calling and allele-specific alternative splicing (ASAS) identification in scRNA-seq data. The versatility and efficiency of this method motivated us to extend it to TGS RNA-seq data. Here, we introduce T-Allele, a TGS version of scAllele that calls variants with high accuracy and identifies haplotype-specific alternative splicing (HSAS). We explore the technical challenges intrinsic to this type of data and the approach used by T-Allele to overcome such challenges. We also study the necessity and impact of read polishing on the variant calls. We applied T-Allele to PacBio RNA-seq data generated from 8 different cell types and identified up to 44 genes with HSAS per sample. Additionally, we observed that most of these events were cell type specific.

4.2 RESULTS

4.2.1 The T-Allele Algorithm

T-Allele makes use of the main scAllele algorithm with a few modifications to accommodate for TGS reads. The main challenge in processing long reads is that they form very complex de Bruijn Graphs (dBG). With increasing read length, the chances of multiple occurrences of the same k-mer increases, thus creating cyclic structures in the graph. Although the original scAllele algorithm can handle this type of structures, it comes at the cost of long computational time which would be significantly increased with longer reads.

To overcome this issue, T-Allele splits a long RNA-seq read into overlapping partial reads made up of 3 consecutive mapped segments (Fig 4-1a). Partial reads overlap each other by 2 mapped segments. For every RC, the partial read whose middle segment overlaps the RC is retained. Variant calling takes place at the RC level. Thus, in principle, only the segment that overlaps the RC is needed. However, the flanking segments are also included to identify the flanking introns (See Methods). The full-length haplotype and isoform information are retained. The collection of nucleotide variants and introns identified in each read is updated after scanning every RC in the chromosome (Fig. 4-1a Pool variants from RC).

Some of the features used in the GLM classifier in scAllele do not apply to TGS reads. For example, PacBio's CCS reads do not provide base quality scores, which was one of the features used in the GLM of scAllele. In addition, another scAllele feature, the position of the variants in the read, is not relevant here since only partial reads are considered. On the other hand, mutual information (MI) between nucleotide variants is a very relevant feature in the long reads (Supplemental Fig. 4-1a). Since each read in theory reflects a full-length mRNA, all *bona fide* genetic variants or RNA editing sites in the same mRNA can be interrogated for their mutual information. Thus, compared to short read RNA-seq, long TGS reads help to capture many more linkage relationships (Supplemental Fig. 4-1b). As a result, MI in long reads serves as an effective means to segregate true genetic variants (with high mutual information) from sequencing errors (with low mutual information). Given the above reasons, the features used in the GLM of T-Allele include allelic ratio, number of neighboring tandem repeats, sequencing error rate, haplotype-fitting, and mutual information (Fig. 4-1b).

We trained 3 different classifiers for SNPs, deletions, and insertions, respectively. The GLM for SNPs affords the best precision (F1 = 0.95) compared to those for insertions (F1 =

0.87) or deletions (F1 = 0.85) (Fig. 4-1b). True positive SNPs are overwhelmingly identified compared to false positives. INDELs are known to be challenging to capture in RNA-seq reads (Sun et al., 2017). Since the most frequent sequencing errors in PacBio data are small INDELs, the frequency of false positive INDELs remains high even after read polishing (Fig. 4-1b).

4.2.2 Impact of read polishing on variant calling

A critical step in the processing of TGS data is read polishing (Amarasinghe et al., 2020; Rhoads & Au, 2015). This step removes spurious variants that persist after obtaining the consensus sequences (CCS). And although these tools were proved to be highly efficient (Amarasinghe et al., 2020), we evaluate the effect of this process on the calling of lowly covered variants, or variants with low allelic ratio.

We compared T-Allele's variant calling performance on TGS data before ('raw') and after ('filtered') polishing. To this end, we used PacBio RNA-seq data of the GM12878 cell line retrieved from the ENCODE portal (Accession number: ENCSR634AKC). The software TranscriptClean was used for the polishing (Wyman & Mortazavi, 2019) step and genotypes of GM12878 were obtained from the Genome in a Bottle (GIAB) database as the ground truth (Zook et al., 2019). Overall, the variant calls on the 'filtered' reads had more true positives at fixed precision (false positive count) compared to the 'raw' reads (Fig. 4-2a). The sensitivity of SNPs and INDELs detection is, however, affected differently. The total number of true positive SNPs is similar for the 'raw' and 'filtered' methods (Supplemental Fig 4-2a), but the 'filtered' variants reach this count with higher precision. INDELs, on the other hand, have very closely overlapping ROCs for the 'raw' and 'filtered' variants (Supplemental Fig 4-2a), although the 'raw' ROC shows a larger number of false positives. This observation suggests that similar performance can be achieved for both sets of variants if different QUAL score (from the GLM) cutoffs were selected. Specifically, a higher QUAL score cutoff for the ‘raw’ variants could yield an ROC identical to the ‘filtered’ curve (note ROCs shown in Fig. 4-2a and Supplemental Fig. 4-2a used a QUAL cutoff of 0). This discrepancy in the QUAL score is due to the fact the classifier was trained on ‘filtered’ data, which had a different proportion of false positives.

We also tested whether read polishing may affect the estimated allelic ratio (AB) of the identified genetic variants. We extracted the true positive variants identified at fixed precision (false positive count = 100) by both the ‘raw’ and ‘filtered’ methods or only by one method. The allelic ratio distributions of the 3 sets of variants (common, ‘raw’ reads only, and ‘filtered’ only) were bimodal with means at around 0.5 and 1.0 corresponding to heterozygous and ALT homozygous variants respectively. The distributions of common variants are similar for the two methods. However, heterozygous variants had a high peak at AB = 1.0 (Supplemental Fig. 4-2b), especially for the ‘raw’ reads. Most of heterozygous variants with AB = 1.0 had low read coverage (data not shown). Together, the data suggest that lowly-covered variants were not detected as frequently in the ‘filtered’ data as in the ‘raw’ data, likely due to the loss of reads during polishing.

4.2.3 Calculation of allelic linkage and allele-specific splicing

To carry out the analysis of linkage between pairs of variants (including intronic parts), we first determined the best parameters to use, similarly as in scAllele. We calculated the mutual information between pairs of variants and grouped them based on the number of common reads covering the pair. Pairs of false positives variants are expected to have low MI since they are likely sequencing errors that occur randomly. High MI between pairs of false positives is due to low transcript count which, in turn, results in a low number of common reads (Fig. 4-3a). Based on this analysis, we observed that a minimum number of 6 common reads was required for reliable linkage calculation. We then selected all pairs of variants with at least 6 common reads and calculated their MI. At a minimum MI score of 0.48, we can confidently classify true and false linkage events with a sensitivity of 95% and specificity of 91% (Fig. 4-3b). We thus used this MI score cutoff (and a min read coverage of 6) to select significant linkage events between nucleotide variant and intronic parts (See Chapter 3 Methods: *Linkage analysis*).

Allele-specific alternative splicing is the result of one or more functional genetic variants that regulate the splicing of an exon. Due to linkage disequilibrium, multiple genetic variants (including the functional one) in the same haplotype are linked to the same allele-specific splicing event. Similarly, as scAllele, T-Allele identifies linkages between a nucleotide variant and an intronic part. T-Allele then examines whether multiple variants are linked to the same splicing event, as expected for haplotype-specific splicing. Nonetheless, the nucleotide-level linkages are adequate evidence for the existence of haplotype-specific splicing. In addition, it offers the flexibility to accommodate noise in the data where certain variants may not be significantly linked to a splicing event, despite *bona fide* haplotype-specific splicing. This arrangement also accommodates local linkage, where the alternative splicing of some exons in the isoform is not linked to the haplotype. Thus, the term HSAS captures both alternative splicing at the local level and at the full isoform level.

4.2.4 Allele-specific splicing in different cell lines

We applied T-Allele to PacBio RNA-seq data generated from 8 cell lines (HepG2, K562, GM12878, PANC1, PC3, HCT116, MCF7, and IMR90). The number of genes identified with

HSAS events ranges from 14 (MCF-7) to 44 (HepG2) (Fig. 4-4a). This number is not strongly correlated to the total number of reads containing a splice junction (Fig. 4-4a). The overlap of HSAS-containing genes between different cell lines is very low as indicated by the Jaccard index (Fig. 4-4b). To rule out the possibility that distinct gene expression profiles dictate this low overlap, we obtained the Jaccard index of genes that are testable for linkage analysis (Supplemental Fig. 4-3a). A testable gene is required to 1) have a heterozygous variant detected in the reads and 2) have at least 6 reads covering the variant and a spliced junction. This overlap is also fairly low (Jaccard index ~ 0.2), suggesting that HSAS is generally cell type-specific.

As an example, an HSAS event in the gene RIPK2 identified in K562 cells is shown in Fig. 4-4c. The splicing of exon number 4 (red box, gene on the - strand) is strongly linked to a variant in exon number 6. That variant is a SNP at position chr8: 89772750 with alleles T and C. Reads with the reference allele (T, 'hap2') skipped exon number 4 while those with the alternative allele (C, 'hap1') included the exon. Interestingly, exon number 10 is also alternatively spliced, but it is not linked to the variant. This observation highlights the fact that not all alternative splicing events are regulated by genetic variants.

Another example HSAS event in the CROCCP2 gene is shown in Supplemental Fig. 4-3b. This gene was identified in 5 of the 8 cell lines included in our study. In this gene, 8 variants were identified, and the IGV alignment track shows that the reads can clearly be segregated by haplotypes. The haplotype denoted as "hap1" includes an unannotated exon (relative to RefSeq, red box) that is not included in hap2. Note that the splicing of this region includes two types of alternative patterns: exon skipping and intron retention, as shown by the IGV read alignments.

4.3 DISCUSSION

The advent of Third Generation Sequencing has allowed researchers to examine RNA transcripts at full-length, elucidating novel splicing isoforms and facilitating the study of splicing regulation. However, variant calling in TGS RNA-seq remains an under-explored area thus limiting our ability to uncover genetic regulation of alternative transcript expression. Here we present a new method called T-Allele to address this challenge.

T-Allele achieves high precision in the variant calls, approximately 99.3% for SNPs and 95.2% for INDELs (at FP = 100, “all” truth set, ‘filtered’ reads), which is comparable to the performance of a comprehensive RNA-seq workflow combining short and long-read data (Sahraeian et al., 2017). We also studied the effect of read polishing on variant calling performance. Overall, read polishing does not significantly affect the sensitivity of the variant calls and it greatly improves the specificity, especially that of SNP calling. The precision of INDEL calling is also affected, mostly by shifting the QUAL score of the variants. This suggests that T-Allele reliably identifies INDELs with and without polishing, with the minor downside of assigning them a score higher in the ‘raw’ data than in the ‘filtered’ one. This observation reflects the fact that scAllele has superior performance on INDELs (Chapter 4). Thus, even if given reads with high error rate, this algorithm can identify INDELs with high precision. The issue with the QUAL score shifting is nonetheless relevant, and it’s something worth addressing in future releases.

The allelic ratio of genetic variants in the RNA before and after polishing was not significantly affected. The only issue we observed was the loss of lowly-covered true positives. This loss, although reducing sensitivity of variant calls, does not impair the linkage analysis, which requires a relatively high read coverage.

The main limitation of scAllele was that ASAS identification is limited to a variant and intronic part that reside within the same short read. Although the functional variant regulating splicing is likely located close to the splice site, longer range regulation does exist. Using TGS long reads, T-Allele overcomes this limitation by pooling variants from different RCs and matching the name of the reads containing each variant in each RC.

Using true and false positive pairs of variants from the GM12878 data, we evaluated the optimal parameters for linkage detection. We find that a minimum MI score of 0.48 and at least 6 common reads are required. The application of T-Allele to PacBio data of 8 cell lines uncovered 14 to 44 genes with HSAS events per cell line, with a potentially high level of cell type-specificity. This cell type-specificity is likely rooted from the diverse genetic background of the cell lines, because previous studies showed that ASAS events tend to share across cell lines if the same causal genetic variant is present in all cell lines (Y.-H. E. Hsiao et al., 2016).

Although, there exist other methods that perform allele-specific splicing analysis on TGS data (B. Wang et al., 2020), they do not perform *de novo* variant calling and are limited to substitution alleles. Additionally, they do not account for the complex splicing regulation (mixture of allele-specific and other factors) the way T-Allele does.

In summary, T-Allele is an effective method for both variant calling and identification of HSAS events in TGS data. It leverages the advantage of long read RNA-seq in revealing both full-length transcripts and genetic variants in the RNA. With the increasing application of long read RNA-seq technologies, we expect T-Allele will lead to discovery of a large number of functional variants in splicing regulation, thus improving our knowledge of the genetic basis of molecular phenotypes.

4.4 METHODS

4.4.1 Detailed algorithm

T-Allele is built upon the scAllele algorithm presented in Chapter 4. A main modification here is the pre-processing of long reads RNA-seq. A typical TGS RNA-seq read is composed of multiple mapped segments split by introns. The read length in this type of data can easily surpass 10 thousand bases (Pollard et al., 2018). Hence, decomposing and assembling the entire read will inevitably lead to highly interconnected and cyclic dBGs. With increasing read length, the probability of multiple occurrences of the same k-mer increases, thus increasing the number of cycles in the graph.

To address this challenge, we leverage the feature of scAllele where variant calling is performed locally (at the RC level). As a result, only the segment of the read that overlaps the RC is needed. T-Allele splits the long read into overlapping partial reads made up of 3 consecutive mapped segments (Fig. 4-1a). For each RC, only the partial read whose middle segment overlaps the RC is included. The segments of the partial read that flank the RC are used to 1) identify the flanking introns, and 2) avoid ‘alternative ends’ in the dBG for that RC in case there exist variants close to the beginning and end of the interval.

The sequencing error rate in TGS is higher than second-generation sequencing, resulting in high frequency of mismatches and INDELs in the reads. To further facilitate dBG processing, low frequency edges were removed from the graph, defined as those with read count being ≤5% of the read count of the edges in the predecessor source node. This cutoff was determined based on the sequencing error rate of 0.05. Although real variants exist with low allelic ratios, it is challenging to distinguish them from sequencing errors.

4.4.2 Model training

We used HepG2 genotype data obtained from ENCODE (Accession number: ENCSR319QHO) to train the GLMs of T-Allele. The training was performed on ‘filtered’ reads. Based on the result comparison between ‘filtered’ and ‘raw’ reads, we reasoned that it is more convenient to used polished reads for downstream analyses. In addition, the high prevalence of mismatches and INDELs in ‘raw’ reads made the training less reliable since the proportion of false positives is much higher than that of true positives. Such a proportion would skew the model to prioritize specificity over sensitivity. With the rapidly improving sequencing technologies and the many options to polish reads, it is reasonable to train the classifiers with the ‘filtered’ reads.

4.4.3 Benchmarking

For the variant calls, we use a minimum read coverage of 5 reads and minimum ALT allele read count of 2. We also removed potential editing sites and INDELs longer than 20 bases. The RTG command and the definition of performance metrics are the same as used for scAllele (Chapter 4).

4.5 CODE AVAILABILITY

The T-Allele software is available at PyPI (https://pypi.org/user/giovanniquinones/TAllele). The scripts used for the analyses in this work are available in our github repository: https://github.com/gxiaolab/TAllele/Manuscript

4.6 DATA AVAILABILITY

Variant calls and linkage events from the different cell lines are available in our github repository: https://github.com/gxiaolab/TAllele/data

4.7 FIGURES

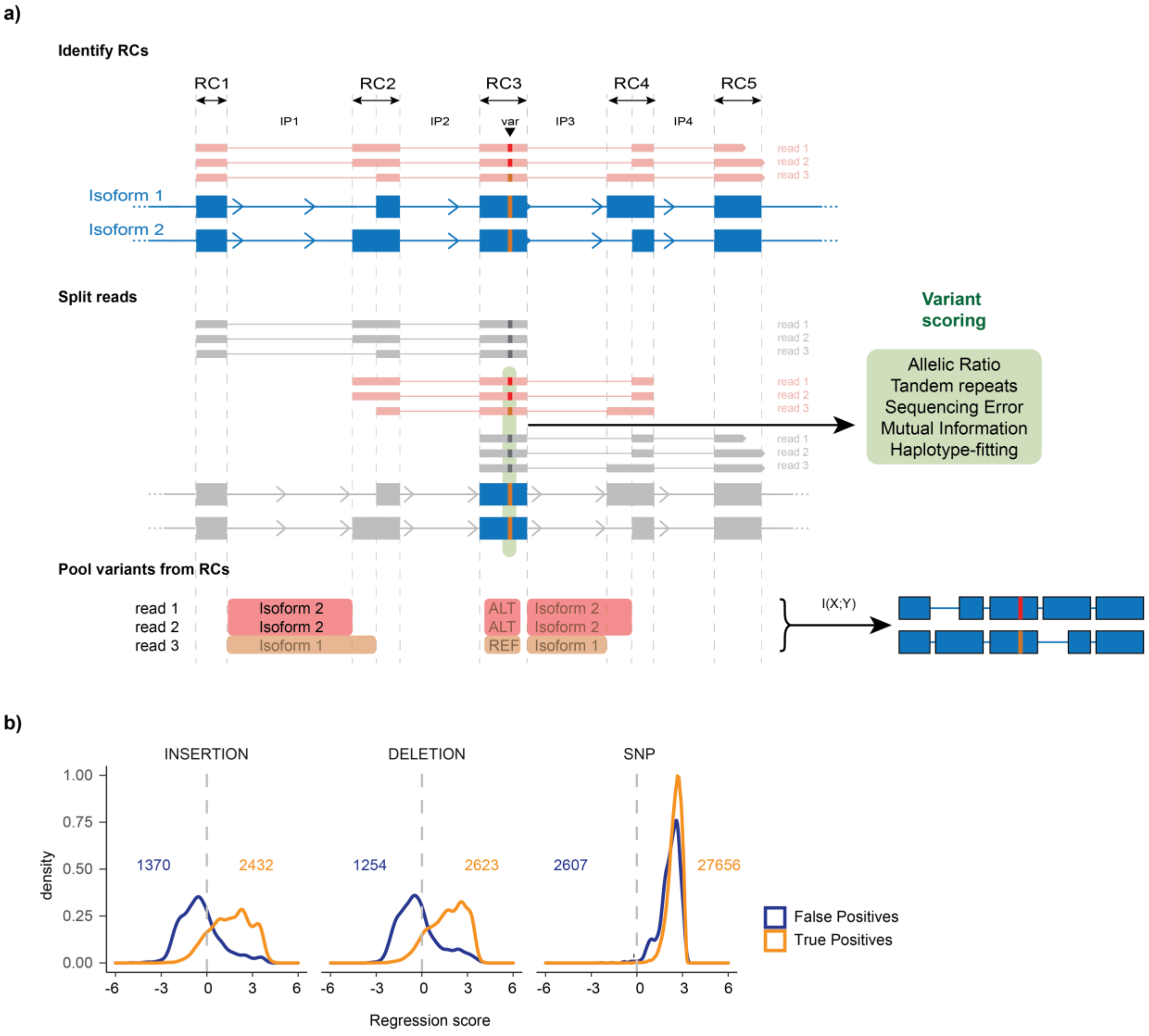


Figure 4-1. T-Allele algorithm outline. **(a)** Read Clusters (RCs) are defined as overlapping mapped segments from the TGS reads. The long reads are split into overlapping partial reads of 3 consecutive segments (Split reads). For each RC, only the partial reads whose middle segment overlaps is considered. Variant calling is performed similarly as scAllele, but with different features (Variant scoring). The variants and introns identified from each RC are pooled together and the linkage analysis via mutual information is performed. **(b)** Training of the GLM classifier

using HepG2 PacBio RNA-seq and truth variant set from ENCODE. 3 models were trained for the 3 types of genetic variants (insertions, deletions, and SNPs) respectively. The middle line represents the theoretical cutoff between true and false variants (regression score = 0). The number of true and false positive variants used in each model is shown.

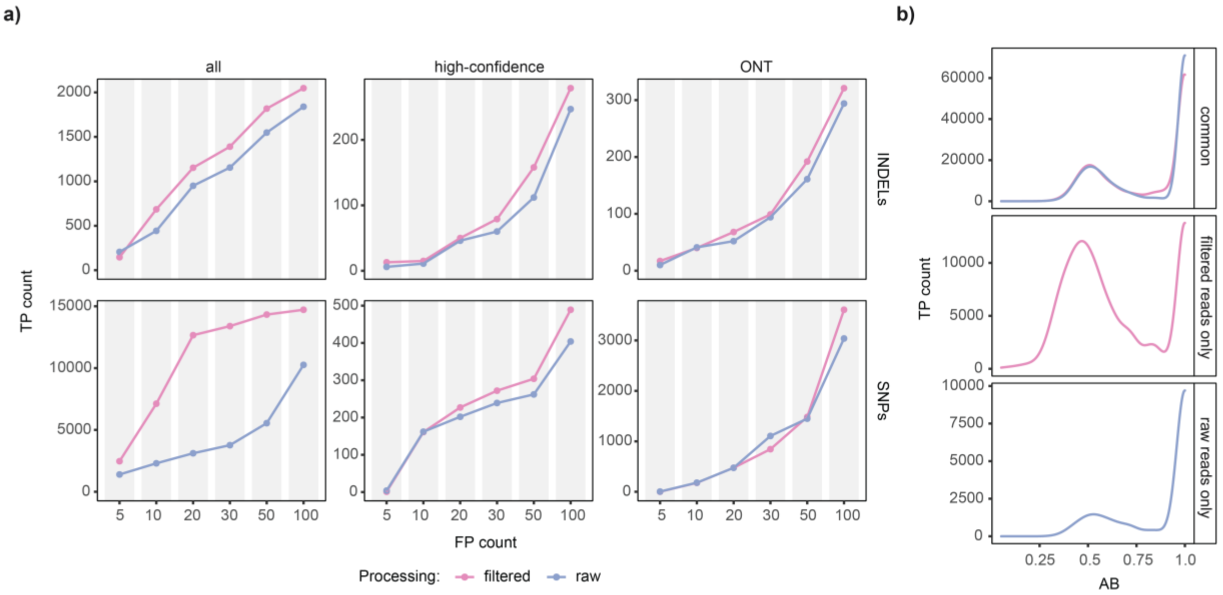


Figure 4-2. Variant calling evaluation on GM12878 PacBio's RNA-seq. (a) True positive (TP) count of T-Allele at different false positive counts. We compare the performance in the variant calls before (raw) and after (filtered) read polishing with TranscriptClean. Additionally, we use 3 different benchmark truth sets: GIAB set of all variants (all), GIAB's set of high-confidence variants (high-confidence) and variant calls made by Karst et al on Nanopore Sequencing (ONT). Overall, the variant calls on the filtered data is more precise, but the magnitude depends on the truth set. **(b)** True positive count (at FP count = 100) at various AB (allelic ratio) values split by the dataset where the variants are present. Allelic ratio (AB) is bimodally distributed with means at 0.5 and 1 corresponding to heterozygous and ALT homozygous variants.

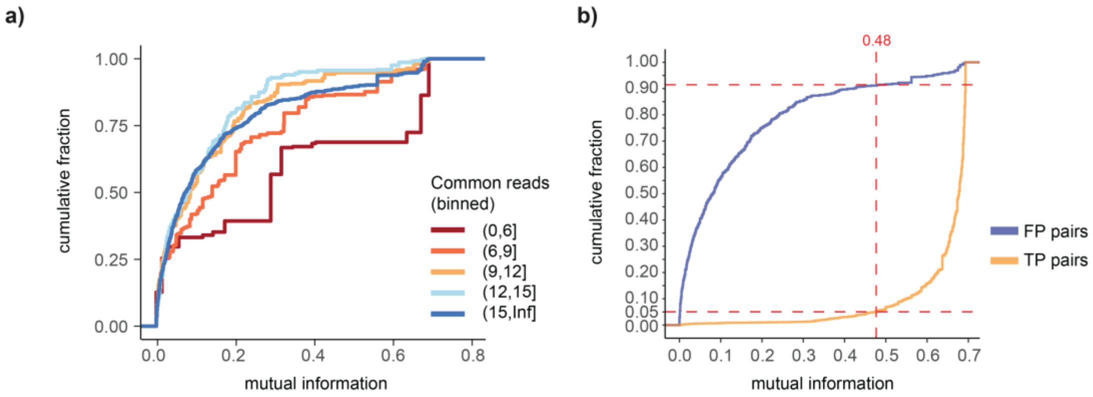


Figure 4-3. Parameter estimation for mutual information calculation. (a) ECDF distribution of mutual information between pairs of false positive variants binned by the number of common reads. The curves converge to mostly low MI values with larger number of common reads. **(b)** ECDF distribution of mutual information between pairs of true positive (TP) and false positive (FP) variants. Variant pairs with at least 6 common reads were included. The dashed vertical line at MI = 0.48 is the define cutoff. This value rejects around 91% of the FP pairs and maintains 95% of the TP pairs.

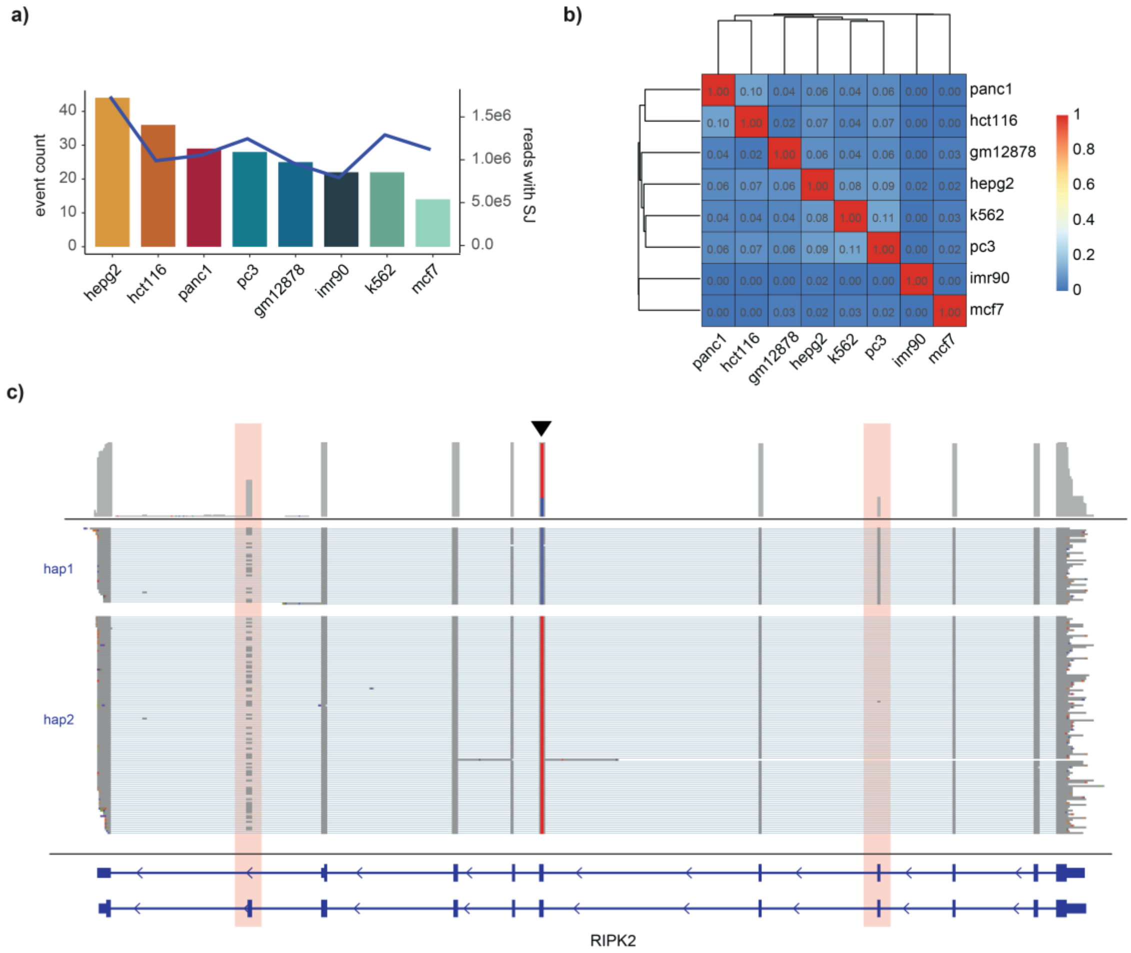
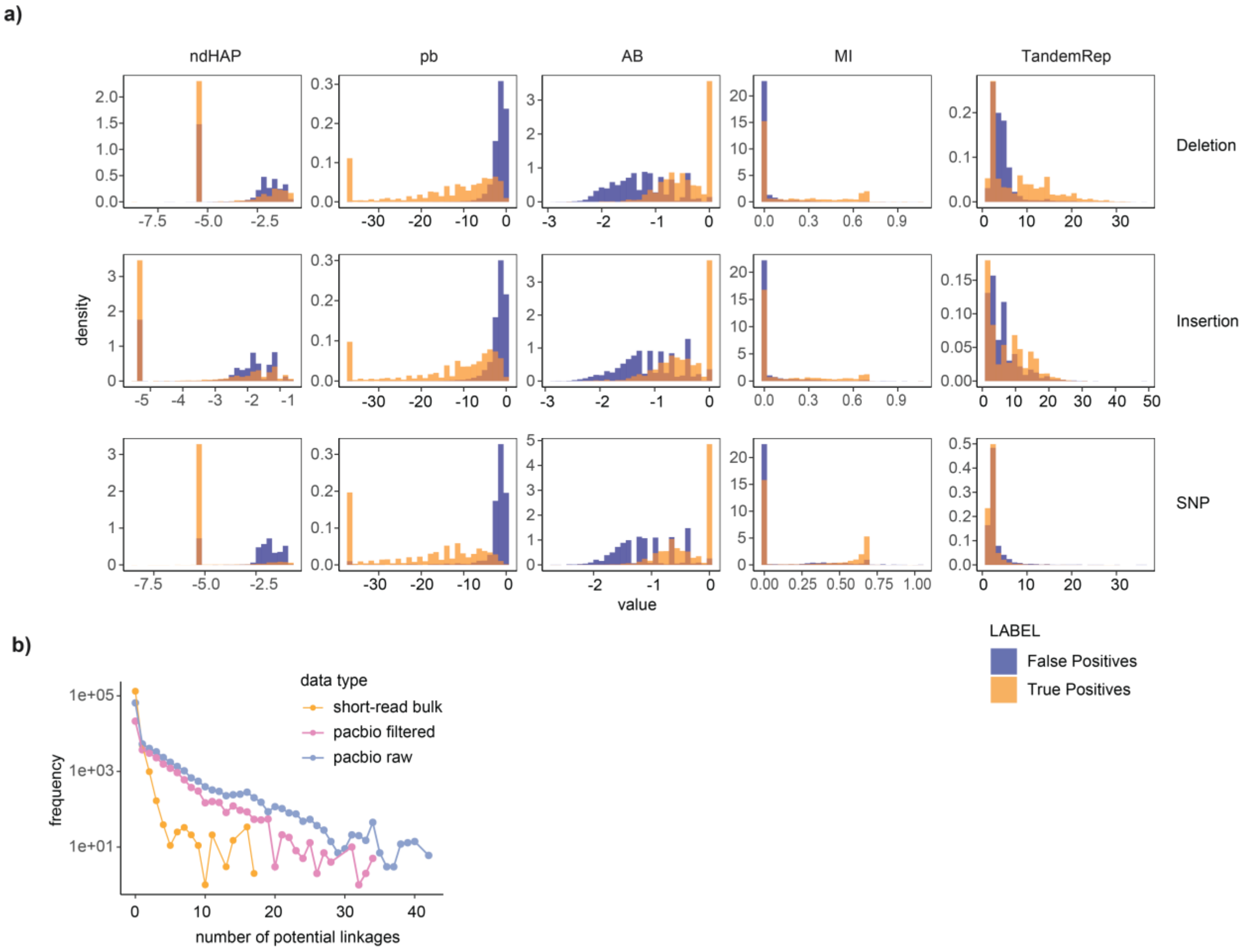
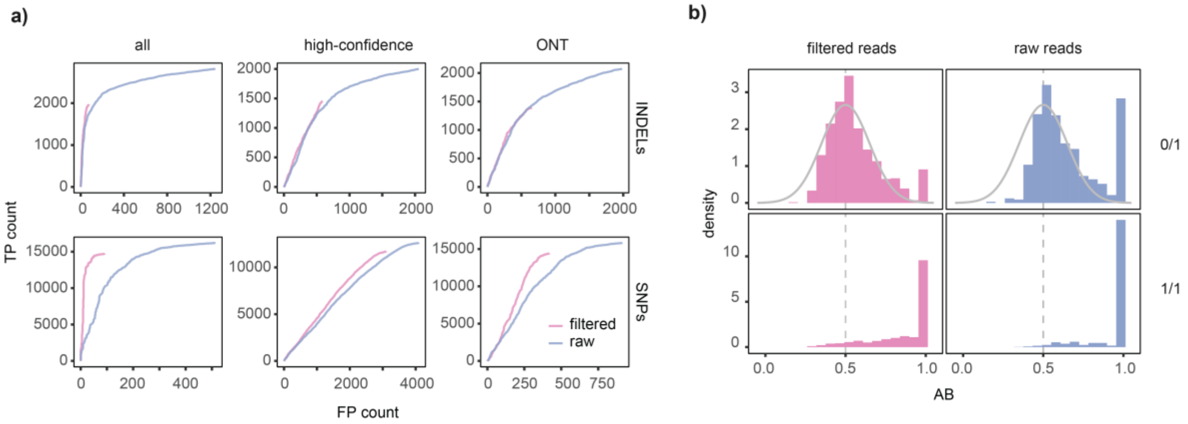


Figure 4-4. Detection of linkage effects. (a) Number of Allele-Specific Alternative Splicing (HSAS) events identified in PacBio's RNA-seq from 8 cell lines. The secondary axis (blue line) shows the number of reads splice junction (SJ) present in the data. **(b)** Jaccard index (JI) of HSAS events among cell lines. Overall, the events found are mostly cell-line specific. K562 and PC-3 cells have the highest overlap (JI = 0.11). **(c)** An example of HSAS identified in the K562 RNA-seq in the RIPK2 gene. Exon number 4 and exon number 10 (from the right, in the highlighted area) are alternatively spliced. The splicing of exon number 4 is allele specific. The heterozygous nucleotide variant is located on exon number 6 (shown with the black arrow). The reads are split based on the alleles of this variant into haplotype 1 (hap1) and haplotype 2 (hap2).

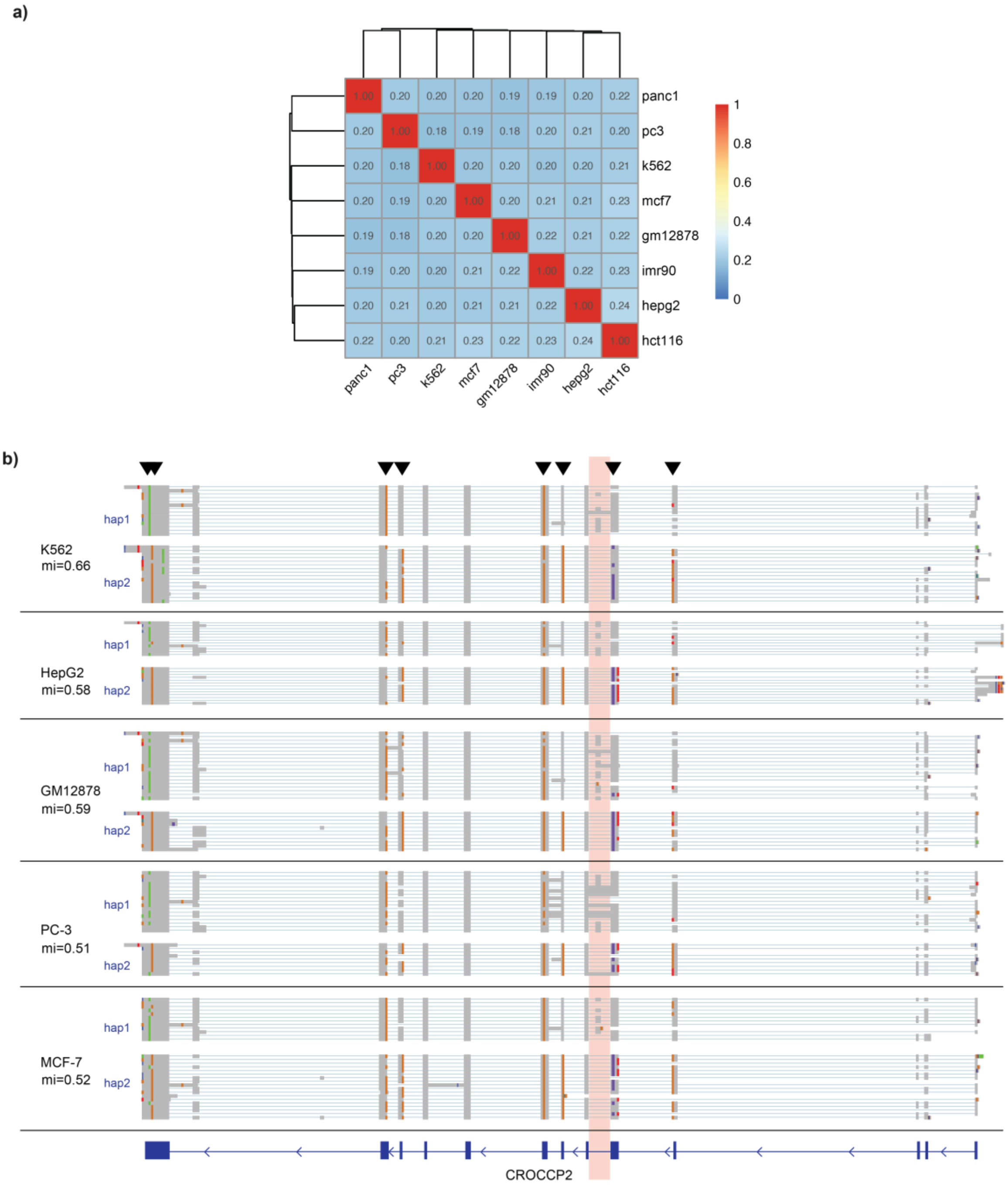
4.8 SUPPLEMENTAL FIGURES



Supplemental Figure 4-1. Features of T-Allele's classifier. (a) Features used to train the classifiers for the 3 variant types (deletion, insertions, and SNPs): "ndHAP" = log10(haplotype fitting); "pb" = log10(binomial allelic ratio test); "AB" = log10(allelic ratio); "MI" = mutual information, "TandemRep" = number of neighboring tandem repeats. The true and false positives variants were obtained from the variant call benchmark using ENCODE's PacBio HepG2 long-read RNA-seq and ENCODE WGS HepG2 truth variant set. **(b)** Number of variants colocalized in at least 4 reads (potential linkages) for GM12878 RNA-seq data from different platforms (Illumina short read and PacBio long read). Additionally, PacBio's data was evaluated before (raw) and after (filtered) read polishing with TranscriptClean. As expected, the long-read data allows for more potential linkages for every variant.



Supplemental Figure 4-2. Performance evaluation on polished and unpolished reads. ROC curves for the INDEL (top) and SNP (bottom) calls in the GM12878 long-read RNA-seq. Variants with QUAL score > 0 were selected. We use 3 different benchmark truth sets: GIAB set of all variants (all), GIAB's set of high-confidence variants (high-confidence) and variant calls made by Karst et al on Nanopore Sequencing (ONT). **(b)** True positive variants (at fixed FP count = 100) distributed based on allelic ratio (AB) and split by their genotype. The "filtered" reads have a more "normal" distribution of heterozygous (0/1) variants compared to the "raw", but the "raw" variants have a better distribution for the ALT homozygous (1/1).



Supplemental Figure 4-3. HSAS identification in cell lines. (a) Jaccard Index of genes testable for linkage among the cell lines. A gene is defined as testable if it contains enough coverage (6 reads) and contains a heterozygous variant. **(b)** IGV view of the alignment track (down sampled) of an HSAS event in the CROCCP2 gene. This event is found in 5 of the 8 cell lines examined.

The exon between in the highlighted strip (unannotated) is alternatively spliced based on the haplotypes identified (hap1 and hap2). These haplotypes correspond to the alleles of the 8 variants identified on this gene (indicated with black triangles). Additionally, K562, GM12878, and PC-3 cell lines display some degree of retention of the introns in the highlighted strip. This intron retention event is also linked to the haplotypes because “hap2” reads do not show any retention.

CHAPTER 5 - Concluding remarks

5.1 SUMMARY

The study of nucleotide variants is key to the understanding of phenotypic diversity and the underlying biological mechanisms. RNA molecules with nucleotide variation exhibit differences in RNA processing and maturation. In the present work, we focus on tackling the questions regarding identification, function, and regulation of nucleotide variants in RNA.

In Chapter two, we studied the role of over 200 RNA binding proteins (RBPs) in A-to-I editing regulation in HepG2 and K562 cells. Making use of knockdown RNA-seq data of each RBP, we studied the transcriptome-wide A-to-I editing variation. Most of the RBPs caused non-significant changes in the global editing ratio, but we identified a subset that strongly repress or enhance global editing, including DROSHA, ILF2/3, TARDBP and TROVE2. DROSHA and ILF2/3 RBPs were previously reported to interact with the ADAR1 protein. Based on co-immunoprecipitation assays and analysis of their RNA-binding patterns, we confirmed that the editing regulation by these proteins occur through direct protein-protein interaction. Interestingly, these RBPs are involved in miRNA processing, which suggests that these pathways may be closely related. We also identified TROVE2 as an important regulator of RNA editing. This finding and the direction of editing change upon TROVE2 knockdown is consistent with studies of SLE patients which reported a loss of Ro60 (encoded by TROVE2) function and reduced editing levels. Finally, we reported TDP-43 (encoded by TARDBP) as a regulator of editing by enhancing ADAR1 transcription. Together, these findings offer new insights towards a better understanding of RNA editing regulation and a more complete view of the interplay between RBPs and RNA processing.

In Chapter three, we presented scAllele, a versatile tool that enables variant calling in scRNA-seq and uncovers variants involved in allele-specific alternative splicing. Motivated by the lack of specialized tools, we developed a variant caller that is highly sensitive given shallow sequencing and outperforms commonly used methods, especially in the detection of micro INDELS. scAllele uses de Bruin Graph-based reassembly of reads, leading to the identification of nucleotide variants and intronic junction at the read level. This unique feature facilitates a detailed view of the allelic linkage between variants and splicing isoforms via mutual information. We then studied the ideal parameters (read coverage and mutual information threshold) and applied scAllele to a lung cancer scRNA-seq dataset (with matched controls). We identify a large number of variants and more than 150 allele-specific splicing events that were highly enriched in one condition (cancer or normal) or that were differentially prevalent in one of them. Among these events, many have important relevance to cancer, such as those in the CTSE and IFI44L genes. We thus demonstrated that scAllele offers a unique approach to utilize multi-level information inherent to scRNA-seq data. This tool will aid the discovery of novel variants and those that modulate splicing, unveiling meaningful biological insights from the data.

In Chapter four, we introduce an extension of scAllele, namely T-Allele, adapted to tackle Third Generation Sequencing (TGS) RNA-seq data. Given that TGS reads can span several thousands of bases in length, and the sequencing error rate is higher than that in short-read sequencing, it is necessary to incorporate new algorithms and processing steps. We evaluated the variant-calling performance of T-Allele with and without read polishing. We observed that, even though the precision of SNP calls is greatly improved by read polishing, INDELs can be reliably identified either way. This result supports that T-Allele can overcome the high sequencing error rate of TGS data and reinforces the conclusion that scAllele is advantageous in the identification of INDELs. The identification of allele-specific splicing events is greatly facilitated by the long reads. Specifically, long reads allow discovery of not only variant-intron pairs, but also multi-intron or even full isoform association with haplotypes. We applied T-Allele to PacBio RNA-seq data of 8 cell lines and uncovered 14 to 44 genes with allele-specific splicing events in each one, with a potentially high level of cell type-specificity. We demonstrate that events identified in long reads exhibited haplotype-level alternative splicing and, at the same time, discriminates between alternative splicing modulated by nucleotide variants from those that may involve additional factors. Together, we showed high precision in variant calls by T-Allele and its capability of identifying allele-specific isoform expression.

5.2 FUTURE DIRECTIONS

Characterization of all known variants remains an ambitious goal; however, the rapid advance in sequencing technologies, decreasing sequencing costs, and development of new molecular techniques are making this goal more and more attainable.

Transcriptome sequencing at the single-cell level was first introduced in the year 2013. Thanks to this technology, vast cellular heterogeneity within tissues has been discovered (Y. H. Choi & Kim, 2019; Papalexi & Satija, 2018; Zeisel et al., 2015). Metrics such as gene expression was shown to be different even within cells of similar type (Haque et al., 2017; Zeisel et al., 2015). Expression signatures of each cell, never considered previously, are starting to play a crucial role in transcriptomic research. For example, the spatial position of cells with respect to each other was found to be critical for intracellular communication in the study of diseases such as Alzheimer’s disease. This motivated the development of spatial transcriptomics techniques (Longo et al., 2021). Another relevant feature is the timing of transcription. Novel techniques using nucleotide-labeling such as ‘Bru-Seq’ and ‘BruChase-Seq’ capture nascent RNAs and can select cells based on their ‘age’. This approach has greatly increased our understanding of splicing kinetics, and RNA turnover (Paulsen et al., 2014).

Moreover, in order to maximize the information captured from a cell, researchers are now using multi-omics approaches which combine data from different molecular assays on the same cell. In 2019, Argelaguet et al. combined RNA-seq with methylation and chromatin accessibility data and reported that, in the absence of external stimuli, embryonic stem cells become ectoderm (Argelaguet et al., 2019). This is a highly significant contribution to the field of developmental biology and was only possible with the combined analysis of these three datasets.

The ongoing development of sequencing technologies and molecular techniques offer numerous new avenues for the functional characterization of nucleotide variants. These advances will motivate development of new bioinformatic methods. As a result, nucleotide variant analysis in the RNA will become instrumental to developing new strategies towards effective diagnosis, prevention, and treatment of multiple diseases.

REFERENCES

Ahn, J., & Xiao, X. (2015). RASER: Reads aligner for SNPs and editing sites of RNA. *Bioinformatics*, *31*(24), 3906–3913. https://doi.org/10.1093/bioinformatics/btv505

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology 2020 21:1*, *21*(1), 1–16. https://doi.org/10.1186/S13059-020-1935-5

Amoah, K., Hsiao, Y.-H. E., Bahn, J. H., Sun, Y., Burghard, C., Tan, B. X., Yang, E.-W., & Xiao, X. (2021). Allele-specific alternative splicing and its functional genetic variants in human tissues. *Genome Research*, *31*(3), gr.265637.120. https://doi.org/10.1101/GR.265637.120

Ardui, S., Ameur, A., Vermeesch, J. R., & Hestand, M. S. (2018). Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Research*, *46*(5), 2159–2168. https://doi.org/10.1093/NAR/GKY066

Argelaguet, R., Clark, S. J., Mohammed, H., Stapel, L. C., Krueger, C., Kapourani, C. A., Imaz-Rosshandler, I., Lohoff, T., Xiang, Y., Hanna, C. W., Smallwood, S., Ibarra-Soria, X., Buettner, F., Sanguinetti, G., Xie, W., Krueger, F., Göttgens, B., Rugg-Gunn, P. J., Kelsey, G., … Reik, W. (2019). Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature*, *576*(7787), 487–491. https://doi.org/10.1038/s41586-019-1825-8

Axel, B., Anita, M., & Stefan, B. (1999). RNA editing. *FEMS Microbiology Reviews*, *23*(3), 297–316. https://doi.org/10.1111/j.1574-6976.1999.tb00401.x

Bahn, J. H., Ahn, J., Lin, X., Zhang, Q., Lee, J. H., Civelek, M., & Xiao, X. (2015). Genomic analysis of ADAR1 binding and its involvement in multiple RNA processing pathways. *Nature Communications*, *6*(1), 6355. https://doi.org/10.1038/ncomms7355

Bahn, J. H., Lee, J. H., Li, G., Greer, C., Peng, G., & Xiao, X. (2012). Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Research*, *22*(1), 142–150. https://doi.org/10.1101/gr.124107.111

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300. https://doi.org/10.2307/2346101

Bentley, D. L. (2014). Coupling mRNA processing with transcription in time and space. In *Nature Reviews Genetics* (Vol. 15, Issue 3, pp. 163–175). Nature Publishing Group. https://doi.org/10.1038/nrg3662

Bhogal, B., Jepson, J. E., Savva, Y. A., Pepper, A. S., Reenan, R. A., & Jongens, T. A. (2011). Modulation of dADAR-dependent RNA editing by the Drosophila fragile X mental retardation protein. *Nature Neuroscience*, *14*(12), 1517–1524. https://doi.org/10.1038/nn.2950

Bratt, E., & Öhman, M. (2003). Coordination of editing and splicing of glutamate receptor pre-mRNA. *RNA*, *9*(3), 309–318. https://doi.org/10.1261/rna.2750803

Broad Institute. (2019). *Picard toolkit*.

Brümmer, A., Yang, Y., Chan, T. W., & Xiao, X. (2017). Structure-mediated modulation of mRNA abundance by A-to-I editing. *Nature Communications*, *8*(1), 1255. https://doi.org/10.1038/s41467-017-01459-7

Brusa, R., Zimmermann, F., Koh, D. S., Feldmeyer, D., Gass, P., Seeburg, P. H., Sprengel, R., & Sakmann, B. (1995). Early-onset epilepsy and postnatal lethality associated with an editing-deficient GluR-B allele in mice. *Science*, *270*(5242), 1677–1680. https://doi.org/10.1126/science.270.5242.1677

Cano-Gamez, E., & Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics*, *0*, 424. https://doi.org/10.3389/FGENE.2020.00424

Chen, J., & Guo, J. (2020). Comparative assessments of indel annotations in healthy and cancer genomes with next-generation sequencing data. *BMC Medical Genomics 2020 13:1*, *13*(1), 1–11. https://doi.org/10.1186/S12920-020-00818-6

Cheung, V. G., & Spielman, R. S. (2009). Genetics of human gene expression: Mapping DNA variants that influence gene expression. In *Nature Reviews Genetics* (Vol. 10, Issue 9, pp. 595–604). Nat Rev Genet. https://doi.org/10.1038/nrg2630

Choi, K., Raghupathy, N., & Churchill, G. A. (2019). A Bayesian mixture model for the analysis of allelic expression in single cells. *Nature Communications 2019 10:1*, *10*(1), 1–11. https://doi.org/10.1038/s41467-019-13099-0

Choi, Y. H., & Kim, J. K. (2019). Dissecting Cellular Heterogeneity Using Single-Cell RNA Sequencing. *Molecules and Cells*, *42*(3), 189. https://doi.org/10.14348/MOLCELLS.2019.2446

Clarke, L., Fairley, S., Zheng-Bradley, X., Streeter, I., Perry, E., Lowy, E., Tassé, A.-M., & Flicek, P. (2017). The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. *Nucleic Acids Research*, *45*(D1), D854–D859. https://doi.org/10.1093/NAR/GKW829

Cleary, J. G., Braithwaite, R., Gaastra, K., Hilbush, B. S., Inglis, S., Irvine, S. A., Jackson, A., Littin, R., Rathod, M., Ware, D., Zook, J. M., Trigg, L., & Vega, F. M. D. La. (2015). Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. *BioRxiv*, 023754. https://doi.org/10.1101/023754

Colombo, M., Karousis, E. D., Bourquin, J., Bruggmann, R., & Mühlemann, O. (2017). Transcriptome-wide identification of NMD-targeted human mRNAs reveals extensive redundancy between SMG6- and SMG7-mediated degradation pathways. *RNA, 23*(2), 189. https://doi.org/10.1261/RNA.059055.116

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., & Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nature Reviews Genetics 2009 10:3*, *10*(3), 184–194. https://doi.org/10.1038/nrg2537

Cooper, T. A., & Mattox, W. (1997). The regulation of splice-site selection, and its role in human disease. *American Journal of Human Genetics*, *61*(2), 259–266. https://doi.org/10.1086/514856

Cusick, M. E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.-R., Simonis, N., Rual, J.-F., Borick, H., Braun, P., Dreze, M., Vandenhaute, J., Galli, M., Yazaki, J., Hill, D. E., Ecker, J. R., Roth, F. P., & Vidal, M. (2009). Literature-curated protein interaction datasets. *Nature Methods, 6*(1), 39–46. https://doi.org/10.1038/nmeth.1284

De Lucas, S., Oliveros, J. C., Chagoyen, M., & Ortín, J. (2014). Functional signature for the recognition of specific target mRNAs by human Staufen1 protein. *Nucleic Acids Research, 42*(7), 4516–4526. https://doi.org/10.1093/nar/gku073

DeDiego, M. L., Martinez-Sobrido, L., & Topham, D. J. (2019). Novel Functions of IFI44L as a Feedback Regulator of Host Antiviral Responses. *Journal of Virology*, *93*(21). https://doi.org/10.1128/JVI.01159-19

Desterro, J. M. P., Keegan, L. P., Jaffray, E., Hay, R. T., O'Connell, M. A., & Carmo-Fonseca, M. (2005). SUMO-1 modification alters ADAR1 editing activity. *Molecular Biology of the Cell, 16*(11), 5115–5126. https://doi.org/10.1091/mbc.e05-06-0536

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Doe, C. M., Relkovic, D., Garfield, A. S., Dalley, J. W., Theobald, D. E. H., Humby, T., Wilkinson, L. S., & Isles, A. R. (2009). Loss of the imprinted snoRNA mbii-52 leads to increased 5htr2c pre-RNA editing and altered 5HT2CR-mediated behaviour. *Human Molecular Genetics*, *18*(12), 2140–2148. https://doi.org/10.1093/hmg/ddp137

Fan, J., Hu, J., Xue, C., Zhang, H., Susztak, K., Reilly, M. P., Xiao, R., & Li, M. (2020). ASEP: Gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. *PLOS Genetics*, *16*(5), e1008786. https://doi.org/10.1371/JOURNAL.PGEN.1008786

Filippini, A., Bonini, D., Lacoux, C., Pacini, L., Zingariello, M., Sancillo, L., Bosisio, D., Salvi, V., Mingardi, J., La Via, L., Zalfa, F., Bagni, C., & Barbon, A. (2017). Absence of the Fragile X Mental Retardation Protein results in defects of RNA editing of neuronal mRNAs in mouse. *RNA Biology*, *14*(11), 1580–1591. https://doi.org/10.1080/15476286.2017.1338232

Fox, E. J., Reid-Bayliss, K. S., Emond, M. J., & Loeb, L. A. (2014). Accuracy of Next Generation Sequencing Platforms. *Next Generation, Sequencing & Applications*, *1*(01). https://doi.org/10.4172/JNGSA.1000106

Gallagher, M. D., & Chen-Plotkin, A. S. (2018). The Post-GWAS Era: From Association to Function. *American Journal of Human Genetics*, *102*(5), 717. https://doi.org/10.1016/J.AJHG.2018.04.002

Gallo, A., Vukic, D., Michalík, D., O'Connell, M. A., & Keegan, L. P. (2017). ADAR RNA editing in human disease; more to it than meets the I. In *Human Genetics* (Vol. 136, Issue 9, pp. 1265–1278). Springer Berlin Heidelberg. https://doi.org/10.1007/s00439-017-1837-0

Garrison, E., & Marth, G. (2012). *Haplotype-based variant detection from short-read sequencing*. https://arxiv.org/abs/1207.3907v2

Gerstberger, S., Hafner, M., & Tuschl, T. (2014). A census of human RNA-binding proteins. *Nature Reviews Genetics*, *15*(12), 829–845. https://doi.org/10.1038/nrg3813

Gibbs, R. A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J. G., Zhu, Y., Wang, J., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., … Rasheed, A. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Gilad, Y., Rifkin, S. A., & Pritchard, J. K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. In *Trends in Genetics* (Vol. 24, Issue 8, pp. 408–415). Trends Genet. https://doi.org/10.1016/j.tig.2008.06.001

Glisovic, T., Bachorik, J. L., Yong, J., & Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. In *FEBS Letters* (Vol. 582, Issue 14, pp. 1977–1986). https://doi.org/10.1016/j.febslet.2008.03.004

Griffin, T. J., & Smith, L. M. (2000). Genetic identification by mass spectrometric analysis of single-nucleotide polymorphisms: Ternary encoding of genotypes. *Analytical Chemistry*, *72*(14), 3298–3302. https://doi.org/10.1021/ac991390e

Guan, D., Altan-Bonnet, N., Parrott, A. M., Arrigo, C. J., Li, Q., Khaleduzzaman, M., Li, H., Lee, C.-G., Pe'ery, T., & Mathews, M. B. (2008). Nuclear Factor 45 (NF45) Is a Regulatory Subunit of Complexes with NF90/110 Involved in Mitotic Control. *Molecular and Cellular Biology*, *28*(14), 4629–4641. https://doi.org/10.1128/MCB.00120-08

Gymrek, M., Golan, D., Rosset, S., & Erlich, Y. (2012). lobSTR: A short tandem repeat profiler for personal genomes. *Genome Research*, *22*(6), 1154–1162. https://doi.org/10.1101/gr.135780.111

Han, J., Lee, Y., Yeom, K. H., Kim, Y. K., Jin, H., & Kim, V. N. (2004). The Drosha-DGCR8 complex in primary microRNA processing. *Genes and Development*, *18*(24), 3016–3027. https://doi.org/10.1101/gad.1262504

Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine 2017 9:1*, *9*(1), 1–12. https://doi.org/10.1186/S13073-017-0467-4

Hasan, M. S., Wu, X., & Zhang, L. (2015). Performance evaluation of indel calling tools using real short-read data. *Human Genomics*, *9*(1), 20. https://doi.org/10.1186/s40246-015-0042-2

Hentze, M. W., Castello, A., Schwarzl, T., & Preiss, T. (2018). A brave new world of RNA-binding proteins. *Nature Reviews Molecular Cell Biology*, *19*(5), 327–341. https://doi.org/10.1038/nrm.2017.130

Hsiao, Y.-H. E., Bahn, J. H., Lin, X., Chan, T.-M., Wang, R., & Xiao, X. (2016). Alternative splicing modulated by genetic variants demonstrates accelerated evolution regulated by highly conserved proteins. *Genome Research*, *26*(4), 440–450. https://doi.org/10.1101/GR.193359.115

Hsiao, Y. H. E., Bahn, J. H., Yang, Y., Lin, X., Tran, S., Yang, E. W., Quinones-Valdez, G., & Xiao, X. (2018). RNA editing in nascent RNA affects pre-mRNA splicing. *Genome Research*, *28*(6), 812–823. https://doi.org/10.1101/gr.231209.117

Huang, R., Han, M., Meng, L., & Chen, X. (2018). Transcriptome-wide discovery of coding and noncoding RNA-binding proteins. *Proceedings of the National Academy of Sciences*, 201718406. https://doi.org/10.1073/pnas.1718406115

Hubbard, T. (2002). The Ensembl genome database project. *Nucleic Acids Research*, *30*(1), 38–41. https://doi.org/10.1093/nar/30.1.38

Hung, T., Pratt, G. A., Sundararaman, B., Towsend, M. J., Chaivorapol, C., Bhangale, T., Graham, R. R., Ortmann, W., Criswell, L. A., Yeo, G. W., & Behrens, T. W. (2015). The Ro60 autoantigen binds endogenous retroelements and regulates inflammatory gene expression. *Science*, *350*(6259), 455–459. https://doi.org/10.1126/science.aac7442

Hwang, T., Park, C.-K., Leung, A. K. L., Gao, Y., Hyde, T. M., Kleinman, J. E., Rajpurohit, A., Tao, R., Shin, J. H., & Weinberger, D. R. (2016). Dynamic regulation of RNA editing in human brain development and disease. *Nature Neuroscience*, *19*(8), 1093–1099. https://doi.org/10.1038/nn.4337

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., … Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology 2018 36:4*, *36*(4), 338–345. https://doi.org/10.1038/nbt.4060

Janga, S. C., & Mittal, N. (2011). Construction, structure and dynamics of post-transcriptional regulatory network directed by RNA-binding proteins. *Advances in Experimental Medicine and Biology*, *722*, 103–117. https://doi.org/10.1007/978-1-4614-0332-6_7

Jiang, Y., Zhang, N. R., & Li, M. (2017). SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biology 2017 18:1*, *18*(1), 1–15. https://doi.org/10.1186/S13059-017-1200-8

Joshi, N., & Fass, J. (2011). *Sickle: A sliding-window, adaptive, quality-based trimming tool for*

FastQ files. https://github.com/najoshi/sickle

Karst, S. M., Ziels, R. M., Kirkegaard, R. H., Sørensen, E. A., McDonald, D., Zhu, Q., Knight, R., & Albertsen, M. (2021). High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nature Methods*, 1–5. https://doi.org/10.1038/s41592-020-01041-y

Kim, J. K., Kolodziejczyk, A. A., Illicic, T., Teichmann, S. A., & Marioni, J. C. (2015). Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nature Communications*, *6*(1), 1–9. https://doi.org/10.1038/ncomms9687

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/GR.215087.116

Krusche, P., Trigg, L., Boutros, P. C., Mason, C. E., De La Vega, F. M., Moore, B. L., Gonzalez-Porta, M., Eberle, M. A., Tezak, Z., Lababidi, S., Truty, R., Asimenos, G., Funke, B., Fleharty, M., Chapman, B. A., Salit, M., & Zook, J. M. (2019). Best practices for benchmarking germline small-variant calls in human genomes. *Nature Biotechnology*, *37*(5), 555–560. https://doi.org/10.1038/s41587-019-0054-x

Lagier-Tourenne, C., Polymenidou, M., & Cleveland, D. W. (2010). TDP-43 and FUS/TLS: Emerging roles in RNA processing and neurodegeneration. *Human Molecular Genetics*, *19*(R1), R46–R64. https://doi.org/10.1093/hmg/ddq137

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 559. https://doi.org/10.1186/1471-2105-9-559

Lee, J.-H., Ang, J. K., & Xiao, X. (2013). Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants. *RNA*, *19*(6), 725–732. https://doi.org/10.1261/rna.037903.112

Levanon, E. Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z. Y., Shoshan, A., Pollock, S. R., Sztybel, D., Olshansky, M., Rechavi, G., & Jantsch, M. F. (2004). Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nature Biotechnology 2004 22:8*, *22*(8), 1001–1005. https://doi.org/10.1038/nbt996

Li, G., Bahn, J. H., Lee, J. H., Peng, G., Chen, Z., Nelson, S. F., & Xiao, X. (2012). Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Research*, *40*(13). https://doi.org/10.1093/nar/gks280

Li, M. J., Yan, B., Sham, P. C., & Wang, J. (2015). Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression. *Briefings in Bioinformatics*, *16*(3), 393–412. https://doi.org/10.1093/BIB/BBU018

Liu, F., Zhang, Y., Zhang, L., Li, Z., Fang, Q., Gao, R., & Zhang, Z. (2019). Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data. *Genome Biology*, *20*(1), 242. https://doi.org/10.1186/s13059-019-1863-4

Lo Giudice, C., Tangaro, M. A., Pesole, G., & Picardi, E. (2020). Investigating RNA editing in deep transcriptome datasets with REDItools and REDIportal. *Nature Protocols 2020 15:3*, *15*(3), 1098–1131. https://doi.org/10.1038/s41596-019-0279-7

Longo, S. K., Guo, M. G., Ji, A. L., & Khavari, P. A. (2021). Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nature Reviews Genetics 2021*, 1–18. https://doi.org/10.1038/s41576-021-00370-8

Marcucci, R., Brindle, J., Paro, S., Casadio, A., Hempel, S., Morrice, N., Bisso, A., Keegan, L. P., Del Sal, G., & O'Connell, M. A. (2011). Pin1 and WWP2 regulate GluR2 Q/R site RNA editing by ADAR2 with opposing effects. *EMBO Journal*, *30*(20), 4211–4222. https://doi.org/10.1038/emboj.2011.303

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12. https://doi.org/10.14806/EJ.17.1.200

Maynard, A., McCoach, C. E., Rotow, J. K., Harris, L., Haderk, F., Kerr, D. L., Yu, E. A., Schenk, E. L., Tan, W., Zee, A., Tan, M., Gui, P., Lea, T., Wu, W., Urisman, A., Jones, K., Sit, R., Kolli, P. K., Seeley, E., … Bivona, T. G. (2020). Therapy-Induced Evolution of Human Lung Cancer Revealed by Single-Cell RNA Sequencing. *Cell*, *182*(5), 1232-1251.e22. https://doi.org/10.1016/J.CELL.2020.07.017

Mayr, C. (2017). Regulation by 3′-Untranslated Regions. *Https://Doi.Org/10.1146/Annurev-Genet-120116-024704*, *51*, 171–194. https://doi.org/10.1146/ANNUREV-GENET-120116-024704

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Morris, A. R., Mukherjee, N., & Keene, J. D. (2010). Systematic analysis of posttranscriptional gene expression. In *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* (Vol. 2, Issue 2, pp. 162–180). Wiley Interdiscip Rev Syst Biol Med. https://doi.org/10.1002/wsbm.54

Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future.

Philosophical Transactions of the Royal Society B: Biological Sciences, *368*(1620). https://doi.org/10.1098/RSTB.2012.0362

Nie, Y., Ding, L., Kao, P. N., Braun, R., & Yang, J.-H. (2005). ADAR1 Interacts with NF90 through Double-Stranded RNA and Regulates NF90-Mediated Gene Expression Independently of RNA Editing. *Molecular and Cellular Biology*, *25*(16), 6956–6963. https://doi.org/10.1128/MCB.25.16.6956-6963.2005

Nishikura, K. (2010). Functions and Regulation of RNA Editing by ADAR Deaminases. *Annual Review of Biochemistry*, *79*(1), 321–349. https://doi.org/10.1146/annurev-biochem-060208-105251

Nishikura, K. (2016). A-to-I editing of coding and non-coding RNAs by ADARs. In *Nature Reviews Molecular Cell Biology* (Vol. 17, Issue 2, pp. 83–96). Nature Publishing Group. https://doi.org/10.1038/nrm.2015.4

Nostrand, E. L. Van, Freese, P., Pratt, G. A., Wang, X., Wei, X., Blue, S. M., Dominguez, D., Cody, N. A. L., Olson, S., Sundararaman, B., Xiao, R., Zhan, L., Bazile, C., Bouvrette, L. P. B., Chen, J., Duff, M. O., Garcia, K., Gelboin-Burkhart, C., Hochman, A., … Yeo, G. W. (2017). A Large-Scale Binding and Functional Map of Human RNA Binding Proteins. *BioRxiv*, 179648. https://doi.org/10.1101/179648

Nussbacher, J. K., & Yeo, G. W. (2018). Systematic Discovery of RNA Binding Proteins that Regulate MicroRNA Levels. *Molecular Cell*, *69*(6), 1005-1016.e7. https://doi.org/10.1016/j.molcel.2018.02.012

Oakes, E., Anderson, A., Cohen-Gadol, A., & Hundley, H. A. (2017). Adenosine deaminase that acts on RNA 3 (adar3) binding to glutamate receptor subunit B Pre-mRNA Inhibits RNA editing in glioblastoma. *Journal of Biological Chemistry*, *292*(10), 4326–4335.

https://doi.org/10.1074/jbc.M117.779868

Olsen, N. D., Wagner, J., McDaniel, J., Stephens, S. H., Westreich, S. T., Prasanna, A., Johanson, E., Boja, E., Maier, E. J., Serang, O., Jáspez, D., Lorenzo-Salazar, J. M., Muñoz-Barrera, A., Rubio-Rodríguez, L. A., Flores, C., Kyriakidis, K., Malousi, A., Shafin, K., Pesout, T., … Zook, J. M. (2020). precisionFDA Truth Challenge V2: Calling variants from short- and long-reads in difficult-to-map Regions. *BioRxiv*, *5*, 2020.11.13.380741. https://doi.org/10.1101/2020.11.13.380741

Ota, H., Sakurai, M., Gupta, R., Valente, L., Wulff, B.-E., Ariyoshi, K., Iizasa, H., Davuluri, R. V, & Nishikura, K. (2013). ADAR1 forms a complex with Dicer to promote microRNA processing and RNA-induced gene silencing. *Cell*, *153*(3), 575–589. https://doi.org/10.1016/j.cell.2013.03.024

Papalexi, E., & Satija, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology*, *18*(1), 35–45. https://doi.org/10.1038/nri.2017.76

Parrott, A. M., Walsh, M. R., Reichman, T. W., & Mathews, M. B. (2005). RNA binding and phosphorylation determine the intracellular distribution of nuclear factors 90 and 110. *Journal of Molecular Biology*, *348*(2), 281–293. https://doi.org/10.1016/j.jmb.2005.02.047

Paulsen, M. T., Veloso, A., Prasad, J., Bedi, K., Ljungman, E. A., Magnuson, B., Wilson, T. E., & Ljungman, M. (2014). Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA. *Methods*, *67*(1), 45–54. https://doi.org/10.1016/j.ymeth.2013.08.015

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J. L., & Mayer, G. (2018). Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports 2018 8:1*, *8*(1), 1–14. https://doi.org/10.1038/s41598-018-29325-6

Picardi, E., D'Erchia, A. M., Giudice, C. Lo, & Pesole, G. (2017). REDIportal: A comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Research*, *45*(D1), D750–D757. https://doi.org/10.1093/nar/gkw767

Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T., & Sandhu, M. S. (2018). Long reads: their purpose and place. *Human Molecular Genetics*, *27*(R2), R234–R241. https://doi.org/10.1093/HMG/DDY177

Porath, H. T., Carmi, S., & Levanon, E. Y. (2014). A genome-wide map of hyper-edited RNA reveals numerous new sites. *Nature Communications 2014 5:1*, *5*(1), 1–10. https://doi.org/10.1038/ncomms5726

Rahman, A., & Isenberg, D. A. (2008). Systemic Lupus Erythematosus. *New England Journal of Medicine*, *358*(9), 929–939. https://doi.org/10.1056/NEJMra071297

Ramaswami, G., & Li, J. B. (2014). RADAR: A rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Research*, *42*(D1), D109–D113. https://doi.org/10.1093/nar/gkt996

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, *13*(5), 278–289. https://doi.org/10.1016/J.GPB.2015.08.002

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., Wilkie, A. O. M., McVean, G.,

& Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics 2014 46:8*, *46*(8), 912–918. https://doi.org/10.1038/ng.3036

Rockman, M. V., & Kruglyak, L. (2006). Genetics of global gene expression. In *Nature Reviews Genetics* (Vol. 7, Issue 11, pp. 862–872). Nature Publishing Group. https://doi.org/10.1038/nrg1964

Roth, S. H., Danan-Gotthold, M., Ben-Izhak, M., Rechavi, G., Cohen, C. J., Louzoun, Y., & Levanon, E. Y. (2018). Increased RNA Editing May Provide a Source for Autoantigens in Systemic Lupus Erythematosus. *Cell Reports*, *23*(1), 50–57. https://doi.org/10.1016/j.celrep.2018.03.036

Sahraeian, S. M. E., Mohiyuddin, M., Sebra, R., Tilgner, H., Afshar, P. T., Au, K. F., Bani Asadi, N., Gerstein, M. B., Wong, W. H., Snyder, M. P., Schadt, E., & Lam, H. Y. K. (2017). Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nature Communications*, *8*(1), 1–15. https://doi.org/10.1038/s41467-017-00050-4

Sakamoto, S., Aoki, K., Higuchi, T., Todaka, H., Morisawa, K., Tamaki, N., Hatano, E., Fukushima, A., Taniguchi, T., & Agata, Y. (2009). The NF90-NF45 Complex Functions as a Negative Regulator in the MicroRNA Processing Pathway. *Molecular and Cellular Biology*, *29*(13), 3754–3769. https://doi.org/10.1128/MCB.01836-08

Saunders, L. R., & Barber, G. N. (2003). The dsRNA binding protein family: critical roles, diverse cellular functions. *The FASEB Journal*, *17*(9), 961–983. https://doi.org/10.1096/fj.02-0958rev

Scadden, A. D. J. (2007). Inosine-Containing dsRNA Binds a Stress-Granule-like Complex and

Downregulates Gene Expression In trans. *Molecular Cell*, *28*(3), 491–500. https://doi.org/10.1016/j.molcel.2007.09.005

Schmidt, M. H.-W., Vogel, A., Denton, A. K., Istace, B., Wormit, A., van de Geest, H., Bolger, M. E., Alseekh, S., Maß, J., Pfaff, C., Schurr, U., Chetelat, R., Maumus, F., Aury, J.-M., Koren, S., Fernie, A. R., Zamir, D., Bolger, A. M., & Usadel, B. (2017). De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing. *The Plant Cell*, *29*(10), 2336–2348. https://doi.org/10.1105/TPC.17.00521

Schnepp, P. M., Chen, M., Keller, E. T., & Zhou, X. (2019). SNV identification from single-cell RNA sequencing data. *Human Molecular Genetics*, *28*(21), 3569–3583. https://doi.org/10.1093/hmg/ddz207

Shamay-Ramot, A., Khermesh, K., Porath, H. T., Barak, M., Pinto, Y., Wachtel, C., Zilberberg, A., Lerer-Goldshtein, T., Efroni, S., Levanon, E. Y., & Appelbaum, L. (2015). Fmrp Interacts with Adar and Regulates RNA Editing, Synaptic Density and Locomotor Activity in Zebrafish. *PLoS Genetics*, *11*(12), e1005702. https://doi.org/10.1371/journal.pgen.1005702

Shastry, B. S. (2009). SNPs: impact on gene function and phenotype. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 578, pp. 3–22). Methods Mol Biol. https://doi.org/10.1007/978-1-60327-411-1_1

Shaw, J. M., Feagin, J. E., Stuart, K., & Simpson, L. (1988). Editing of kinetoplastid mitochondrial mRNAs by uridine addition and deletion generates conserved amino acid sequences and AUG initiation codons. *Cell*, *53*(3), 401–411. https://doi.org/10.1016/0092-8674(88)90160-2

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K.

(2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308. https://doi.org/10.1093/NAR/29.1.308

Simons, A. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*.

Simons, F. H. M., Pruijn, G. J. M., & Van Venrooij, W. J. (1994). Analysis of the intracellular localization and assembly of Ro ribonucleoprotein particles by microinjection into Xenopus laevis oocytes. *Journal of Cell Biology*, *125*(5), 981–988. https://doi.org/10.1083/jcb.125.5.981

Stepanova, M., Tiazhelova, T., Skoblov, M., & Baranova, A. (2006). Potential regulatory SNPs in promoters of human genes: A systematic approach. *Molecular and Cellular Probes*, *20*(6), 348–358. https://doi.org/10.1016/j.mcp.2006.03.007

Sun, Z., Bhagwate, A., Prodduturi, N., Yang, P., & Kocher, J. P. A. (2017). Indel detection from RNA-seq data: Tool evaluation and strategies for accurate detection of actionable mutations. *Briefings in Bioinformatics*, *18*(6), 973–983. https://doi.org/10.1093/bib/bbw069

Tan, M. H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A. N., Liu, K. I., Zhang, R., Ramaswami, G., Ariyoshi, K., Gupte, A., Keegan, L. P., George, C. X., Ramu, A., Huang, N., Pollina, E. A., Leeman, D. S., Rustighi, A., Goh, Y. P. S., … Li, J. B. (2017). Dynamic landscape and regulation of RNA editing in mammals. *Nature*, *550*(7675), 249–254. https://doi.org/10.1038/nature24041

Tariq, A., Garncarz, W., Handl, C., Balik, A., Pusch, O., & Jantsch, M. F. (2013). RNA-interacting proteins act as site-specific repressors of ADAR2-mediated RNA editing and fluctuate upon neuronal stimulation. *Nucleic Acids Research*, *41*(4), 2581–2593. https://doi.org/10.1093/nar/gks1353

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole,

C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., … Forbes, S. A. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, *47*(D1), D941–D947. https://doi.org/10.1093/NAR/GKY1015

Thomas, M. P., & Lieberman, J. (2013). Live or let die: Posttranscriptional gene regulation in cell stress and cell death. In *Immunological Reviews* (Vol. 253, Issue 1, pp. 237–252). Immunol Rev. https://doi.org/10.1111/imr.12052

Tran, S. S., Jun, H. I., Bahn, J. H., Azghadi, A., Ramaswami, G., Van Nostrand, E. L., Nguyen, T. B., Hsiao, Y. H. E., Lee, C., Pratt, G. A., Martínez-Cerdeño, V., Hagerman, R. J., Yeo, G. W., Geschwind, D. H., & Xiao, X. (2019). Widespread RNA editing dysregulation in brains from autistic individuals. *Nature Neuroscience*, *22*(1), 25–36. https://doi.org/10.1038/s41593-018-0287-x

Tung, P.-Y., Blischak, J. D., Hsiao, C. J., Knowles, D. A., Burnett, J. E., Pritchard, J. K., & Gilad, Y. (2017). Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports 2017 7:1*, *7*(1), 1–15. https://doi.org/10.1038/srep39921

Ule, J. (2013). Alu elements: at the crossroads between disease and evolution. *Biochemical Society Transactions*, *41*(6), 1532–1535. https://doi.org/10.1042/BST20130157

Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., Blue, S. M., Nguyen, T. B., Surka, C., Elkins, K., Stanton, R., Rigo, F., Guttman, M., & Yeo, G. W. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, *13*(6), 508–514. https://doi.org/10.1038/nmeth.3810

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*(5), 737–746. https://doi.org/10.1101/GR.214270.116

Vitali, P., Basyuk, E., Le Meur, E., Bertrand, E., Muscatelli, F., Cavaillé, J., & Huttenhofer, A. (2005). ADAR2-mediated editing of RNA substrates in the nucleolus is inhibited by C/D small nucleolar RNAs. *The Journal of Cell Biology*, *169*(5), 745–753. https://doi.org/10.1083/jcb.200411129

Vogenberg, F. R., Barash, C. I., & Pursel, M. (2010). Personalized Medicine: Part 1: Evolution and Development into Theranostics. *Pharmacy and Therapeutics*, *35*(10), 560. /pmc/articles/PMC2957753/

Wang, B., Tseng, E., Baybayan, P., Eng, K., Regulski, M., Jiao, Y., Wang, L., Olson, A., Chougule, K., Buren, P. Van, & Ware, D. (2020). Variant phasing and haplotypic expression from long-read sequencing in maize. *Communications Biology 2020 3:1*, *3*(1), 1–11. https://doi.org/10.1038/s42003-020-0805-8

Wang, Q., Li, X., Qi, R., & Billiar, T. (2017). RNA Editing, ADAR1, and the Innate Immune Response. *Genes*, *8*(1). https://doi.org/10.3390/GENES8010041

Warraich, S. T., Yang, S., Nicholson, G. A., & Blair, I. P. (2010). TDP-43: A DNA and RNA binding protein with roles in neurodegenerative diseases. In *International Journal of Biochemistry and Cell Biology* (Vol. 42, Issue 10, pp. 1606–1609). Pergamon. https://doi.org/10.1016/j.biocel.2010.06.016

Warren, R. L., Coombe, L., Mohamadi, H., Zhang, J., Jaquish, B., Isabel, N., Jones, S. J. M., Bousquet, J., Bohlmann, J., Birol, I., & Berger, B. (2019). NtEdit: Scalable genome sequence polishing. *Bioinformatics*, *35*(21), 4430–4432. https://doi.org/10.1093/bioinformatics/btz400

Washburn, M. C., & Hundley, H. A. (2016). Controlling the Editor: The Many Roles of RNA-Binding Proteins in Regulating A-to-I RNA Editing. In G. W. Yeo (Ed.), *RNA Processing* (pp. 189–214). Springer Nature.

Washburn, M. C., Kakaradov, B., Sundararaman, B., Wheeler, E., Hoon, S., Yeo, G. W., & Hundley, H. A. (2014). The dsRBP and Inactive Editor ADR-1Utilizes dsRNA Binding to Regulate A-to-I RNA Editing across the C.elegans Transcriptome. *Cell Reports*, *6*(4), 599–607. https://doi.org/10.1016/j.celrep.2014.01.011

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C. S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Wolkowicz, U. M., & Cook, A. G. (2012). NF45 dimerizes with NF90, Zfr and SPNR via a conserved domain that has a nucleotidyltransferase fold. *Nucleic Acids Research*, *40*(18), 9356–9368. https://doi.org/10.1093/nar/gks696

Wyman, D., & Mortazavi, A. (2019). TranscriptClean: variant-aware correction of indels, mismatches and splice junctions in long-read transcripts. *Bioinformatics*, *35*(2), 340–342. https://doi.org/10.1093/BIOINFORMATICS/BTY483

Yamamoto, K., Kawakubo, T., Yasukochi, A., & Tsukuba, T. (2012). Emerging roles of cathepsin E in host defense mechanisms. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1824*(1), 105–112. https://doi.org/10.1016/J.BBAPAP.2011.05.022

Yamawaki, T. M., Lu, D. R., Ellwanger, D. C., Bhatt, D., Manzanillo, P., Arias, V., Zhou, H., Yoon, O. K., Homann, O., Wang, S., & Li, C. M. (2021). Systematic comparison of high-throughput single-cell RNA-seq methods for immune cell profiling. *BMC Genomics*, *22*(1). https://doi.org/10.1186/s12864-020-07358-4

Yang, E. W., Bahn, J. H., Yun-Hua Hsiao, E., Tan, B. X., Sun, Y., Fu, T., Zhou, B., Van Nostrand, E. L., Pratt, G. A., Freese, P., Wei, X., Quinones-Valdez, G., Urban, A. E., Graveley, B. R., Burge, C. B., Yeo, G. W., & Xiao, X. (2019). Allele-specific binding of RNA-binding proteins reveals functional genetic variants in the RNA. *Nature Communications*, *10*(1), 396275. https://doi.org/10.1101/396275

Yang, R., Van Etten, J. L., & Dehm, S. M. (2018). Indel detection from DNA and RNA sequencing data with transIndel. *BMC Genomics*, *19*(1), 270. https://doi.org/10.1186/s12864-018-4671-4

Zafar, H., Wang, Y., Nakhleh, L., Navin, N., & Chen, K. (2016). Monovar: Single-nucleotide variant detection in single cells. *Nature Methods*, *13*(6), 505–507. https://doi.org/10.1038/nmeth.3835

Zeisel, A., M͡oz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., Manno, G. La, Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., Rolny, C., Castelo-Branco, G., Hjerling-Leffler, J., & Linnarsson, S. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, *347*(6226), 1138–1142. https://doi.org/10.1126/SCIENCE.AAA1934

Zhang, B., & Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*(1). https://doi.org/10.2202/1544-6115.1128

Zhang, M. J., Ntranos, V., & Tse, D. (2020). Determining sequencing depth in a single-cell

RNA-seq experiment. *Nature Communications 2020 11:1*, *11*(1), 1–11. https://doi.org/10.1038/s41467-020-14482-y

Zhang, Q., & Xiao, X. (2015). Genome sequence-independent identification of RNA editing sites. *Nature Methods*, *12*(4), 347–350. https://doi.org/10.1038/nmeth.3314

Zhang, Z., & Carmichael, G. G. (2001). The fate of dsRNA in the Nucleus: A p54nrb-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. *Cell*, *106*(4), 465–475. https://doi.org/10.1016/S0092-8674(01)00466-4

Zook, J. M., McDaniel, J., Olson, N. D., Wagner, J., Parikh, H., Heaton, H., Irvine, S. A., Trigg, L., Truty, R., McLean, C. Y., De La Vega, F. M., Xiao, C., Sherry, S., & Salit, M. (2019). An open resource for accurately benchmarking small variant and reference calls. *Nature Biotechnology 2019 37:5*, *37*(5), 561–566. https://doi.org/10.1038/s41587-019-0074-6