UNIVERSITY OF CALIFORNIA

Los Angeles

Landscape of RNA editing across Autism Spectrum Disorders

A dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy in Bioinformatics

by

Stephen Show Tran

2019

ABSTRACT OF THE DISSERTATION

Landscape of RNA editing across Autism Spectrum Disorders

by

Stephen Show Tran
Doctor of Philosophy in Bioinformatics
University of California, Los Angeles, 2019
Professor Xinshu Xiao, Chair

Autism spectrum disorders (ASD) encompass neurodevelopmental diseases that share core deficits in verbal and nonverbal language, reciprocal interactions, and stereotyped and repetitive behaviors. Unfortunately, the molecular etiology of ASD remains incompletely understood. Excitingly, the recent advent of next generation RNA sequencing has now enabled whole-genome characterization of RNA regulation, expression, and modification in ASD. One such RNA modification, that is highly prevalent in mammalian synapses yet not studied in ASD, is RNA editing. Thus, in this dissertation, we perform a comprehensive spatiotemporal and first genome-wide study of RNA editing in ASD across multiple implicated brain regions, genetic etiologies, and developmental time points spanning fetal development to adulthood. Collectively, we uncovered general trends and regulatory mechanisms of RNA editing relevant to neuronal tissue.

We first characterized RNA editing in the largest cohort of ASD postmortem brains to date. Strikingly the ASD patients exhibited convergent trends of global downregulated RNA editing (hypoediting) affecting synaptic development and transmission genes. The global hypoediting was observed across multiple brain regions and multiple syndromic forms of ASD including dup15q11.2-13.1 duplication syndrome patients and Fragile X syndrome patients. Network analyses and experimental work demonstrated that Fragile X proteins, FMRP and FXR1P, regulated many of the dysregulated editing sites in human brain.

Since postmortem brains only provide postnatal time windows for studying ASD, we next used organoid models to characterize the landscape of RNA editing over ASD fetal brain neurodevelopment. We generated the first large-scale dataset of hundreds of organoids modelling cerebral cortex development (cortical spheroids) over multiple time periods and encompassing a myriad of penetrant autism-susceptibility mutations. RNA editing gradually increased over cortical spheroid development both in control spheroids and ASD spheroids. However, at all developmental timepoints, the ASD cohort again displayed global trends of hypoediting. Functional enrichment analyses implicated the hypoedited RNA editing in cellular development and proliferation of radial glia, intermediate progenitors, and newborn neurons.

Throughout these ASD studies, we encountered difficulties running common statistical analyses due to distinctive properties of RNA editing data. Thus, we lastly developed a statistical framework to handle RNA editing data based on a beta-binomial distribution. We developed a method, called REDITs (RNA editing tests), to handle important RNA editing analyses including significance of case-control differences and

regression associations with covariates. REDITs had demonstrably higher sensitivity and specificity on simulated and real RNA editing datasets when compared to the most alternative methods used in the RNA editing field.

Online Supplemental Tables

Chapter 2 Supplemental Tables

Chapter 3 Supplemental Tables

Chapter 4 Supplemental Tables

The dissertation of Stephen Show Tran is approved.

Daniel Geschwind

William Lowry

Gang Li

Xinshu Xiao, Committee Chair

University of California, Los Angeles

2019

To God, my family, and brother, Peter, who has severe Autism

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

First, I would like to thank my thesis advisor, Dr. Xinshu (Grace) Xiao, for her guidance over the past five years. I have never worked with a Professor as open-minded and driven to excel in new and exciting areas of research no matter how difficult or challenging the journey. Grace graciously accepted me as a direct graduate student in her lab even though the Bioinformatics class was already full when I applied. And she instigated me to pursue my interests in studying Autism even though her prior work and training was not in neuroscience. Her acumen as a researcher and principal investigator always amazes me, and I hope that I’ve imbibed the high quality standards which she has always held her lab accountable to. For the many times she has gone above and beyond her responsibility as my advisor, she has my deep gratitude.

I would like to thank former and current members of the Xiao Lab, especially the wetlab members Dr. Jae-Hoon Bahn, Dr. Hyun-Ik Jun, and Adel Azghadi, who dedicated many hours toward discussing and conducting experiments for this work. I also appreciate the many dry lab members of our lab many of whom I enjoyed spending time with both in and out of lab. It has been interesting, humbling, and fun getting to know the interests, opinions, and experiences of our lab members who come from such diverse countries and backgrounds.

Next I would like to thank all my other committee members: Professor Dan Geschwind, Professor William Lowry, and Professor Gang Li for their helpful comments

and advice relating to multiple aspects of my projects. I would like to especially thank Professor Geschwind to provided us with such incredible Autism datasets that have all largely driven my projects. I also appreciate Professor Geschwind for helping me over the entire course of my postdoctoral search.

I'd like to thank members of Professor Daniel Geschwind's lab for code and helpful discussions relating my work in Autism. I also thank members from Professor Gene Yeo's lab for their willingness to generate great datasets for our lab.

Last but not least, I thank my family especially my parents for encouraging me to pursue a PhD and directing, guiding, and molding me intellectually, financially, emotionally, and spiritually my entire life. And to Peter, my younger brother with severe Autism, to whom I dedicate this dissertation. I thank God for my nonverbal brother, Peter, who is a genius writer, an upright man of highest character, and one my most important sources of motivation as a researcher.

This work was partially supported by the Eureka fellowship and the Genomic Analysis Training Program fellowship at UCLA.

Chapter 2 was published in *Nature Neuroscience*, 2019. S.S. Tran, H.I. Jun, J.H. Bahn, A. Azghadi, G. Ramaswami, E.L. Van Nostrand, T.B. Nguyen, Y.H.E. Hsiao, C. Lee, G.A. Pratt, V. Martínez-Cerdeño, R.J. Hagerman, G.W. Yeo, D.H. Geschwind & X. Xiao. "Widespread RNA editing dysregulation in brains from autistic individuals." The dissertation author was the primary author of this paper.

Chapter 4 is submitted to *Oxford Bioinformatics*. S.S. Tran, Q. Zhou, X. Xiao. “Statistical inference of differential RNA editing sites from RNA- sequencing data by hierarchical modeling.” The dissertation author is the primary author of this paper.

VITA

Education and employment

2010-2014	B.S. in Computational and Systems Biology University of California, Los Angeles Valedictorian Summa Cum Laude College Honors Highest Departmental Honors Phi Beta Kappa Honors Society
2014-2019	Graduate Student Researcher, Interdepartmental Program of Bioinformatics, University of California, Los Angeles
Winter 2018	Teaching Assistant, Department of Integrative Biology and Physiology, University of California, Los Angeles

Honors, Awards, and Presentations

2015-2017	UCLA Genomic Analysis Training Program Grant
2017-2018	Eureka Scholarship, UCLA Department of Integrative Biology and Physiology
Spring 2017	Spotlight presentation for UCLA Big Data Research Symposium
Summer 2017	Spotlight presentation for UCLA Gene Regulation Monthly Intramural Meeting

Patent

Patent title: DISCOVERY OF ENZYMES FROM THE ALPHA-KETO ACID DECARBOXYLASE FAMILY
Patent office: United States
Patent number: WO2017040378A1
Filed: March, 9th 2017

Publications

1. **Stephen Tran***, Qing Zhou, Xinshu Xiao. “Statistical inference of differential RNA editing sites from RNA-sequencing by hierarchical modeling.” Article submitted, 2019

2. **Stephen Tran***, Hyun-Ik Jun, Jae-Hoon Bahn, Adel Azghadi, Gokul Ramaswami, Eric L. Van Nostrand, Thai B. Nguyen, Yun-Hua E. Hsiao, Changhoon Lee, Gabriel A. Pratt, Veronica Martinez-Cerdeno, Randi J. Hagerman, Gene W. Yeo, Daniel Geschwind, Xinshu Xiao. “Widespread RNA editing dysregulation in brains from autistic individuals.” *Nature Neuroscience*, 2019

3. Giovanni Quinones-Valdez, **Stephen S. Tran**, Hyun-Ik Jun, Jae Hoon Bahn, Ei-Wen Yang, Lijun Zhan, Anneke Brümmer, Xintao Wei, Eric L. Van Nostrand, Gabriel A. Pratt, Gene W. Yeo, Brenton R. Graveley, Xinshu Xiao. “Regulation of RNA editing by RNA-binding proteins in human cells.” *Nature Communications Biology*, 2019

4. Minori Ohashi, Elena Korsakova, Denise Allen, Peiyee Lee, Kai Fu, Benni S. Vargas, Jessica Cinkornpumin, Carlos Salas, Jenny C. Park, Igal Germanguz, Justin Langerman, Contantinos Chronis, Edward Kuoy, **Stephen Tran**, Xinshu Xiao, Matteo Pellegrini, Kathrin Plath, William E. Lowry Loss. “Loss of MECP2 Leads to Activation of P53 and Neuronal Senescence.” *Stem Cell Reports*, 2018

5. Yun-Hua Esther Hsiao, Jae Hoon Bahn, Yun Yang, Xianzhi Lin, **Stephen Tran**, Ei-Wen Yang, Giovanni Quinoes-Valdez, Xinshu Xiao. “RNA editing in nascent RNA affects pre-mRNA splicing.” *Genome Research*, 2018

6. Wai Shun Mak* and **Stephen Tran***, Ryan J. Marcheschi, Steve Bertolani, James Thompson, David Baker, James C. Liao, and Justin Siegel. “Integrative genomic mining for enzyme function to enable engineering of a non-natural biosynthetic pathway.” *Nature Communications*, 2015

7. Maskit Maymon, Pilar Martínez-Hidalgo, **Stephen S. Tran**, Tyler Ice, Karena Craemer, Teni Anbarchian, Tiffany Sung, Lin H. Hwang, Minxia Chou, Nancy A. Fujishige, William Villella, Jérôme Ventosa, Johannes Sikorski, Erin R. Sanders, Kym F. Faull, and Ann M. Hirsch. “Mining the phytomicrobiome to understand how bacterial coinoculations enhance plant growth.” *Frontiers in Plant Science*, 2015

8. Ann Hirsch et. al. **Stephen Tran**. “Complete genome sequence of *Micromonospora* L5, a potential plant-growth regulating actinomycete, originally isolated from *Casuarina equisetifolia* root nodules.” *Journal of Bacteriology Genome Announcements Section*, 2013

9. James D. Adams, **Stephen Tran**, Vincent Wong, Pauline Fontaine, and Jirair Petrosyan. “ß-Sitosterol Lithospermate from Salvia Columbariae.” *The Open Natural Products Journal*, 2012

Chapter 1

Background

1.1. General background of RNA editing

RNA editing consists of base conversions of single nucleotides within RNA molecules. The Animalia kingdom has two types of known RNA editing, adenosine converted to inosine (A-to-I), and cytosine converted to uracil (C-to-U), catalyzed by the *ADAR* and *APOBEC* family of proteins respectively[1,2]. The A-to-I type of editing constitutes the predominate majority of RNA editing in human, in part because 10% of the human genome consists of Alu elements[3] which form double stranded secondary structures propitious for *ADAR* binding[4]. The rife expansion of Alu within the human genome 55 million years[3] ago may also explain why RNA editing occurs most prevalently in human than any other studied species including drosophila, mouse, zebrafish, macaque, and chimpanzee[5]. A diverse landscape of RNA editing also presents across human tissues, with overall editing gauging lowest in skeletal muscle and highest in brain tissues[6].

1.2. RNA editing in brain

1.2.1. Function of RNA editing in brain

The above observations suggest particular functional importance for RNA editing in the human brain. Even before the advent of high-throughput sequencing, the vast majority of studied editing sites were found to hold critical roles in neuronal synapses[7]. The first A-to-I editing site in human was serendipitously discovered when there was an observed disparity between RNA and DNA in the coding sequence of *GRIA2*, which encodes a subunit of the GluR2 glutamate channel[8]. This editing site, causing an amino acid change of glutamine into arginine, was found to dramatically alter the calcium permeability of GluR2, and if attenuated, phenotypically led to mouse seizure and death[9]. Many other editing sites in brain were subsequently discovered. Similar calcium permeability modifying editing sites were discovered in GluK1 and GluK2[8]. The *GRIA2*, *GRIA3*, and *GRIA4* genes all have recoding RNA editing sites altering an arginine to glycine which enhances recovery from desensitization after firing currents[10]. Recovery from desensitization is also enhanced by RNA editing of isoleucine to valine in potassium channel *KCNA1*[11]. Some genes such as *CACNA1D* and *HTR2C* contain multiple editing sites in close proximity which combinatorically generate 5 and 24 isoforms respectively[12,13]. All the editing in *CACNA1D* occurs in its IQ domain which reduces calcium dependent CaM binding[12]. The editing in *HTR2C* reduces G-protein coupling efficiency of neurotransmitter serotonin to secondary messengers[13].

RNA editing has also been found to regulate alternative splicing of neuronal transcripts through both trans and cis mechanisms. An editing site substituting glycine to serine in the brain specific splicing factor *NOVA1*[14] increases its half-life. RNA editing in *NOVA1* increases from embryonic to adult stages which could contribute to alternative splicing involved in neurodevelopment[14]. Another interesting example is auto-editing by ADAR2 within one of its introns, which creates an alternative splice-site eventuating in a non-functional isoform[15]. This process serves as an auto-negative regulatory loop controlling overall editing by *ADAR2*. The *GRIA2* gene has an editing site at the end of exon 13 that when edited leads preferentially to inclusion of a downstream mutual exclusive flip exon over the nearer flop exon[16]. The flip exon and G edited site slow receptor desensitization over the flop and A non-edited site[16].

1.2.2. Global landscape of RNA editing in brain

The advent of next generation sequencing (RNA-seq) recently expanded the study of RNA editing from a couple of aforementioned recoding sites, to the identification of hundreds of thousands of sites present in brain[7,17]. Contrasting to earlier work, most editing actually resides in non-coding regions of the human region[17] rather than in exons. These non-coding sites, however, may also serve neural function. The editing levels of most noncoding sites have been found to gradually increase over prenatal to postnatal development[18,19]. Some of these sites reside in the seed regions of multiple microRNAs, likely influencing the microRNA repertoire during brain maturation[20]. The imperative of this temporal regulation was also demonstrated through human neuronal progenitor cells where low editing of *GRIA2* leading to high calcium influx, was necessary for differentiation into neurons; overstimulating editing levels through overexpression of *ADAR2* in progenitor cells abrogated their ability to differentiate[21]. Since this increasing trend was found mostly in samples ranging from infancy to adulthood, the trajectories of RNA editing still require clarification over the various stages of fetal development and in human aging.

Unexpectedly, the developmental increases in RNA editing do not correlate with *ADAR* gene expression or protein level[18,22], indicating existence of other proteins that regulate RNA editing in brain. Some candidates include *KPNA3* and *PIN1* which were found to increasingly localize and stabilize *ADAR2* in the nucleus over neural progenitor cell development[23]. However, the majority of neuronal relevant RNA editing regulators remain undiscovered.

An important question concerns how RNA editing contributes to specific synaptic processes, neural circuits, and cellular specific functions. Higher resolution technologies such as single cell sequencing have begun to delineate RNA editing landscapes across cell types in brain. Overall, neurons have globally higher editing levels than other cell types in brain such as endothelial, oligodendrocytes, astrocytes, and microglia[24]. A recent study examined RNA editing across neuronal subtypes within drosophila brain and uncovered hundreds of differential editing sites that could delineate the different populations of neurons[25]; many sites resided in evolutionarily conserved regions and had likely functions for cell type specific neuronal transmission[25].

1.2.3. RNA editing in brain-related diseases

Given its role in synaptic transmission, RNA editing poses as a candidate contributor towards neurological and brain disorders. In motor neurons of amyotrophic lateral sclerosis patients, editing of the Q/R editing site in *GRIA2* was highly attenuated, likely contributing towards motor neuron death[26]. Editing levels in the serotonin receptor *HTR2C* were found markedly altered in brains of suicide individuals[27]. In Alzheimer's disease, editing in multiple synaptic transmission genes was hypoedited[28]. Some editing levels have also been found altered in brains of Schizophrenic and Bipolar Disorder patients[29] and Autism[30]. Although suggestive of a perpetrating role, these previous studies only investigated a select number of well-characterized recoding sites. A fuller understanding of the role of RNA editing in brain disorders will necessitate studying the global landscape of editing across many samples of diseased brain.

1.3. Autism Spectrum Disorders

1.3.1. Introduction to ASD

Towards this goal, chapters 3 and 4 of this dissertation focuses on comprehensively characterizing the global landscape of RNA editing in ASD (Autism Spectrum Disorders). ASD is a constellation of neurodevelopmental disorders affecting 1 in 68 individuals that share in common core deficits in communication and reciprocal interaction and repetitive and stereotyped behaviors[31]. The cause of Autism is strongly genetic with estimates between 70-80% from twin studies[32]. However, the genetic architecture of ASD is very complex, with strong evidence for an etiological root from both common and hundreds of rare-genetic variants[33].

1.3.2. Transcriptomic landscape of ASD from postmortem brains

Despite a heterogeneous genetic etiology, large scale transcriptomic studies in postmortem brains have found striking convergence across individuals with ASD. Gene co-expression networks repeatedly identify that modules (gene networks) related to immune response are upregulated in ASD while modules related to neuronal and synaptic functions are downregulated[34-38]. Interestingly, microRNAs, a class of small RNA molecules that inhibits gene expression, in postmortem brains also organized into modules; upregulated microRNA modules in ASD targeted the downregulated neuronal gene modules and downregulated microRNA modules targeted the upregulated immune modules[39], suggesting a possible causal relationship. Neuronal enriched alternative splicing has been found dysregulated[34,38], and at least a third of ASD patients displayed widespread dysregulation of microexons (exons <27 nucleotides)[40].

A recent dataset of single-cell sequencing in 15 ASD and matched control postmortem cortex samples clarified how the transcriptomic changes observed in bulk tissue correspond to aberrations amongst brain cell types[41]. Specifically, laminar layer neurons, astrocytes, and microglia cells all had significant differential gene expression burdens from ASD. Interestingly, ASD patients had elevated cellular compositions of astrocyte cells. Furthermore, astrocytes and microglial cells in ASD had gene expression signatures indicative of activated states. Strikingly, even sample-specific differentially expressed genes unanimously had strongest burdens in laminar layer neurons and converged in cellular processes related to synaptic development and transmission[41]. Overall these findings refine transcriptomic convergence of ASD to activated astrocyte and microglia states and prevalent dysregulation of gene expression in laminar neurons and synaptic processes.

1.3.3. Pre-natal transcriptomics of ASD

Autism phenotypically manifests as early as infancy[42], but the molecular etiology likely stems prenatally during fetal development. Unfortunately, the above postmortem studies were restricted to studying ASD at postnatal time points. A couple of studies have circumvented this limitation by testing enrichment of Autism aberrations within fetal developmental gene expression modules derived from control, fetal brain samples[43,44]. Strikingly, rare, de-novo genetic variants from ASD are enriched in fetal modules related to histone modifications and gene regulation[43,44]. The genes within these modules express exclusively during fetal stages before desisting postnatally. In contrast, the genes in the aforementioned downregulated neuronal modules only express highly postnatally[43,44]. These studies prove that gene expression has fetal-specific regulation imperative to ASD progression and not investigable using postmortem brains alone. It is possible that RNA editing also displays multimodal abnormalities in prenatal versus postnatal ASD.

One promising avenue for modeling and studying prenatal ASD etiology is organoids. Organoids capture diverse cell types, molecular cell to cell interactions, and three dimensional structures which more realistically model brain development than iPSC cell models. One study produced telecephalic organoids from iPSC cells from 4 families with idiopathic ASD probands[45]. The organoids transcriptionally recapitulated development of 8-9 post conception weeks of neocortex and hippocampus. Compared to control fathers, the ASD probands organoids had upregulated genes involved in transcriptional regulation and synaptic assembly and overproduction of GABAergic neurons. The study attempted to augment homogeneity of their disease cohort by only using probands that had enlarged head circumferences[45]. This, however, raises an issue of distinguishing which observed aberrations are specific to their cohort and which are reflective of core ASD etiology. To ensure generalizability to core ASD deficits, future organoid studies will need to include larger sample sizes harboring both idiopathic ASD and mutations in known autism-susceptibility genes. Furthermore, understanding the fetal progression of ASD will require propagating organoids across multiple differentiation time points.

1.3.4. Necessity of studying RNA editing in Autism

Despite massive progress made in detailing the transcriptomic landscape of ASD, RNA editing has been largely neglected. A single publication on a small sample size of postmortem cortex (11 ASD samples) measured editing at a dozen recoding RNA editing sites using targeted sequencing[30]. ASD patients tended to have outlier levels of editing relative to controls. However, a more complete understanding of the contribution of editing to ASD will require global analysis of RNA editing across many samples. Important questions include how RNA editing presents in adult ASD, which molecule processes regulate or are affected by editing, and how RNA editing projects within ASD fetal development.

1.4. Limitations of data and technologies for studying RNA editing

Recent efforts have generated a prodigious amount of postmortem and single cell data for studying ASD[34,41]. However, the datasets are generally designed for gene expression studies which can complicate repurposing for RNA editing. Single cell studies in brain often use single-nucleus RNA sequencing because of the harsh conditions needed to dissociate brain tissue and for scalability[41,46]. However, the current technologies for single-nucleus sequencing employ STAMPs (single-cell transcriptomes attached to microparticles) which only capture the 3’ ends of mRNA transcripts[46,47]. While STAMPs are suitable for calculating gene expression, they lack detection of RNA editing sites in the middle and 5’ ends of transcripts.

Bulk tissue RNA-seq, though not utilizing STAMPs, suffers from generally nonuniform or sparse transcript coverage. Across multiple studies, the number of detected RNA editing sites linearly correlated with the number of RNA-sequencing reads[19,48], indicating that editing level quantification did not reach saturation under typical sequencing coverages. This complicates comparing editing sites between multiple samples from RNA-sequencing since the coverage per editing site can dramatically vary per sample. Fundamentally it is of interest whether RNA editing levels differ between ASD versus control samples. However, as RNA editing level is calculated as the number of reads harboring a “G” nucleotide divided by the total number of reads covering the editing site, studies will inevitably have to cope with samples either lacking coverage or having less accurate editing level quantification due to small coverage.

Unfortunately, commonly utilized methods for gene expression studies, such as the t-test[18] or linear regression[49] of gene expression values, do not handle this issue. Thus more development of advanced statistical frameworks for comparing RNA editing data is needed and is developed in chapter 4 of this dissertation.

1.5 References

1 Blanc, V. & Davidson, N. O. APOBEC-1-mediated RNA editing. *Wiley interdisciplinary reviews. Systems biology and medicine* **2**, 594-602, doi:10.1002/wsbm.82 (2010).

2 Nishikura, K. A-to-I editing of coding and non-coding RNAs by ADARs. *Nature reviews. Molecular cell biology* **17**, 83-96, doi:10.1038/nrm.2015.4 (2016).

3 Hasler, J. & Strub, K. Alu elements as regulators of gene expression. *Nucleic acids research* **34**, 5491-5497, doi:10.1093/nar/gkl706 (2006).

4 Daniel, C., Silberberg, G., Behm, M. & Ohman, M. Alu elements shape the primate transcriptome by cis-regulation of RNA editing. *Genome biology* **15**, R28, doi:10.1186/gb-2014-15-2-r28 (2014).

5 Hung, L. Y. *et al.* An Evolutionary Landscape of A-to-I RNA Editome across Metazoan Species. *Genome biology and evolution* **10**, 521-537, doi:10.1093/gbe/evx277 (2018).

6 Tan, M. H. *et al.* Dynamic landscape and regulation of RNA editing in mammals. *Nature* **550**, 249-254, doi:10.1038/nature24041 (2017).

7 Ramaswami, G. & Li, J. B. Identification of human RNA editing sites: A historical perspective. *Methods (San Diego, Calif.)* **107**, 42-47, doi:10.1016/j.ymeth.2016.05.011 (2016).

8 Sommer, B., Kohler, M., Sprengel, R. & Seeburg, P. H. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell* **67**, 11-19, doi:10.1016/0092-8674(91)90568-j (1991).

9 Higuchi, M. *et al.* Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature* **406**, 78-81, doi:10.1038/35017558 (2000).

10 Krampfl, K. *et al.* Control of kinetic properties of GluR2 flop AMPA-type channels: impact of R/G nuclear editing. *The European journal of neuroscience* **15**, 51-62, doi:10.1046/j.0953-816x.2001.01841.x (2002).

11 Bhalla, T., Rosenthal, J. J., Holmgren, M. & Reenan, R. Control of human potassium channel inactivation by editing of a small mRNA hairpin. *Nature structural & molecular biology* **11**, 950-956, doi:10.1038/nsmb825 (2004).

12 Huang, H. *et al.* RNA editing of the IQ domain in Ca(v)1.3 channels modulates their Ca(2)(+)-dependent inactivation. *Neuron* **73**, 304-316, doi:10.1016/j.neuron.2011.11.022 (2012).

13 Niswender, C. M., Copeland, S. C., Herrick-Davis, K., Emeson, R. B. & Sanders-Bush, E. RNA editing of the human serotonin 5-hydroxytryptamine 2C receptor silences constitutive activity. *The Journal of biological chemistry* **274**, 9472-9478, doi:10.1074/jbc.274.14.9472 (1999).

14 Irimia, M. *et al.* Evolutionarily conserved A-to-I editing increases protein stability of the alternative splicing factor Nova1. *RNA biology* **9**, 12-21, doi:10.4161/rna.9.1.18387 (2012).

15 Rueter, S. M., Dawson, T. R. & Emeson, R. B. Regulation of alternative splicing by RNA editing. *Nature* **399**, 75-80, doi:10.1038/19992 (1999).

16 Grosskreutz, J. *et al.* Kinetic properties of human AMPA-type glutamate receptors expressed in HEK293 cells. *The European journal of neuroscience* **17**, 1173-1178, doi:10.1046/j.1460-9568.2003.02531.x (2003).

17 Picardi, E., D'Erchia, A. M., Lo Giudice, C. & Pesole, G. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic acids research* **45**, D750-d757, doi:10.1093/nar/gkw767 (2017).

18 Hwang, T. *et al.* Dynamic regulation of RNA editing in human brain development and disease. *Nat Neurosci* **19**, 1093-1099, doi:10.1038/nn.4337 (2016).

19 Tran, S. S. *et al.* Widespread RNA editing dysregulation in brains from autistic individuals. *Nat Neurosci* **22**, 25-36, doi:10.1038/s41593-018-0287-x (2019).

20 Ekdahl, Y., Farahani, H. S., Behm, M., Lagergren, J. & Ohman, M. A-to-I editing of microRNAs in the mammalian brain increases during development. *Genome research* **22**, 1477-1487, doi:10.1101/gr.131912.111 (2012).

21 Whitney, N. P. *et al.* Calcium-permeable AMPA receptors containing Q/R-unedited GluR2 direct human neural progenitor cell differentiation to neurons. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **22**, 2888-2900, doi:10.1096/fj.07-104661 (2008).

22 Wahlstedt, H., Daniel, C., Enstero, M. & Ohman, M. Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. *Genome research* **19**, 978-986, doi:10.1101/gr.089409.108 (2009).

23 Behm, M., Wahlstedt, H., Widmark, A., Eriksson, M. & Ohman, M. Accumulation of nuclear ADAR2 regulates adenosine-to-inosine RNA editing during neuronal development. *Journal of cell science* **130**, 745-753, doi:10.1242/jcs.200055 (2017).

24 Picardi, E., Horner, D. S. & Pesole, G. Single-cell transcriptomics reveals specific RNA editing signatures in the human brain. *RNA (New York, N.Y.)* **23**, 860-865, doi:10.1261/rna.058271.116 (2017).

25 Sapiro, A. L. *et al.* Illuminating spatial A-to-I RNA editing signatures within the Drosophila brain. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 2318-2327, doi:10.1073/pnas.1811768116 (2019).

26 Hideyama, T. *et al.* Profound downregulation of the RNA editing enzyme ADAR2 in ALS spinal motor neurons. *Neurobiology of disease* **45**, 1121-1128, doi:10.1016/j.nbd.2011.12.033 (2012).

27 Weissmann, D. *et al.* Region-specific alterations of A-to-I RNA editing of serotonin 2c receptor in the cortex of suicides with major depression. *Transl Psychiatry* **6**, e878, doi:10.1038/tp.2016.121 (2016).

28 Khermesh, K. *et al.* Reduced levels of protein recoding by A-to-I RNA editing in Alzheimer's disease. *RNA (New York, N.Y.)* **22**, 290-302, doi:10.1261/rna.054627.115 (2016).

29 Silberberg, G., Lundin, D., Navon, R. & Ohman, M. Deregulation of the A-to-I RNA editing mechanism in psychiatric disorders. *Human molecular genetics* **21**, 311-321, doi:10.1093/hmg/ddr461 (2012).

30 Eran, A. *et al.* Comparative RNA editing in autistic and neurotypical cerebella. *Mol Psychiatry* **18**, 1041-1048, doi:10.1038/mp.2012.118 (2013).

31 Association, A. P. *Diagnostic and statistical manual of mental disorders (4th ed., text rev.)*. 4th edn, (2000).

32 Geschwind, D. H. Genetics of autism spectrum disorders. *Trends in cognitive sciences* **15**, 409-416, doi:10.1016/j.tics.2011.07.003 (2011).

33 Ramaswami, G. & Geschwind, D. H. Genetics of autism spectrum disorder. *Handbook of clinical neurology* **147**, 321-329, doi:10.1016/b978-0-444-63233-3.00021-x (2018).

34 Parikshak, N. N. *et al.* Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* **540**, 423-427, doi:10.1038/nature20612 (2016).

35 Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380-384, doi:10.1038/nature10110 (2011).

36 Gupta, S. *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat Commun* **5**, 5748, doi:10.1038/ncomms6748 (2014).

37 Gandal, M. J. *et al.* Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science (New York, N.Y.)* **359**, 693-697, doi:10.1126/science.aad6469 (2018).

38 Gandal, M. J. *et al.* Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science (New York, N.Y.)* **362**, doi:10.1126/science.aat8127 (2018).

39 Wu, Y. E., Parikshak, N. N., Belgard, T. G. & Geschwind, D. H. Genome-wide, integrative analysis implicates microRNA dysregulation in autism spectrum disorder. *Nat Neurosci* **19**, 1463-1476, doi:10.1038/nn.4373 (2016).

40 Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**, 1511-1523, doi:10.1016/j.cell.2014.11.035 (2014).

41 Velmeshev, D. *et al.* Single-cell genomics identifies cell type-specific molecular changes in autism. *Science (New York, N.Y.)* **364**, 685-689, doi:10.1126/science.aav8130 (2019).

42 Jones, W. & Klin, A. Attention to eyes is present but in decline in 2-6-month-old infants later diagnosed with autism. *Nature* **504**, 427-431, doi:10.1038/nature12715 (2013).

43 Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997-1007, doi:10.1016/j.cell.2013.10.020 (2013).

44 Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008-1021, doi:10.1016/j.cell.2013.10.031 (2013).

45 Mariani, J. *et al.* FOXG1-Dependent Dysregulation of GABA/Glutamate Neuron Differentiation in Autism Spectrum Disorders. *Cell* **162**, 375-390, doi:10.1016/j.cell.2015.06.034 (2015).

46 Habib, N. *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature methods* **14**, 955-958, doi:10.1038/nmeth.4407 (2017).

47 Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214, doi:10.1016/j.cell.2015.05.002 (2015).

48 Quinones-Valdez, G. *et al.* Regulation of RNA editing by RNA-binding proteins in human cells. *Communications biology* **2**, 19, doi:10.1038/s42003-018-0271-8 (2019).

49 Brummer, A., Yang, Y., Chan, T. W. & Xiao, X. Structure-mediated modulation of mRNA abundance by A-to-I editing. *Nat Commun* **8**, 1255, doi:10.1038/s41467-017-01459-7 (2017).

Chapter 2

Widespread RNA editing dysregulation in Autism Spectrum Disorder Patient Brains

2.1 Introduction

Autism spectrum disorder (ASD) is characterized by a developmental deficit in social communication accompanied by repetitive and restrictive interests[1], with a strong neuropathology implicating glutamatergic[2] and serotonergic[3] circuits, aberrant structural development in multiple brain regions[4], excitatory and inhibitory imbalance[5], and abnormal synaptogenesis[6]. The genetic etiology of ASD remains incompletely understood and shows substantial heterogeneity[7]. Nevertheless, recent studies, leveraging the increasing availability of postmortem samples, have revealed shared patterns of transcriptome dysregulation affecting neuronal and glial coding and non-coding gene expression[8,9], neuronal splicing including microexons[10], and microRNA targeting[11] across approximately 2/3 of ASD patients. These studies highlight down-regulation of activity-dependent genes in neurons and up-regulation of astrocyte and microglial genes as key points of convergence in ASD pathology.

Another major RNA processing mechanism is RNA editing, which refers to the alteration of RNA sequences through insertion, deletion or substitution of nucleotides. Catalyzed by the ADAR family of enzymes, adenosine-to-inosine (A-to-I) editing is the most prevalent type of RNA editing in humans, affecting the majority of human genes[12]. As inosines in RNA are recognized as guanosines by cellular machinery, A-to-I editing can alter gene expression in different ways, for example, through amino acid substitutions, modulation of RNA stability, alteration of alternative splicing, and modifications of regulatory RNAs or *cis*-regulatory motifs[12,13].

RNA editing plays important roles in neurodevelopment and maintenance of normal neuronal function[13]. Indeed, a number of A-to-I editing sites alone are imperative in modulating excitatory responses and permeability of ionic channels and other neuronal signaling functions[13]. Not surprisingly, aberrant RNA editing has been reported in several neurological disorders, such as schizophrenia, bipolar disorder, amyotrophic lateral sclerosis,[14] and Alzheimer's disease.[15] In ASD, a previous study analyzed a few known RNA editing sites in synaptic genes and reported altered editing patterns in a small cohort of ASD cerebella[16]. Yet, it remains unaddressed if global patterns of RNA editing may contribute to the neuropathology of ASD, a question that requires larger patient cohorts and multiple implicated brain regions. In addition, the regulatory mechanisms of aberrant editing in neurological disorders including ASD remain largely unknown.

Here we report global patterns of dysregulated RNA editing across the largest cohort of ASD brain samples to date, spanning multiple brain regions. We identified a core set of down-regulated RNA editing sites, enriched in genes of glutamatergic and synaptic pathways and ASD susceptibility genes. Multiple lines of evidence associate a distinct set of these hypoedited sites with Fragile X proteins: FMRP and FXR1P. Through transcriptome-wide protein-RNA binding analyses and detailed molecular assays, we show that FMRP and FXR1P interact with ADAR and modulate A-to-I editing. It is known that mutations in FMRP lead to the Fragile X syndrome, a disease with high comorbidity with ASD[7]. Indeed, we observed convergent dysregulated patterns of RNA editing in Fragile X and ASD patients, which is consistent with the findings that genes harboring ASD risk mutations are enriched in FMRP targets[17,18]. Overall, we provide global insights regarding RNA editing in ASD pathogenesis and elucidate a regulatory function of Fragile X proteins in RNA editing that additionally serves as a molecular link between ASD and Fragile X Syndrome.

2.2 Results

2.2.1 RNA editing analysis of ASD postmortem brain samples

From 69 unique post-mortem subjects, we obtained rRNA-depleted total RNA-Seq (50 base paired-end, non-strand-specific) from three brain regions implicated in ASD-susceptibility: frontal cortex, temporal cortex, and cerebellum (Supplementary Table 1). In total, there were 29, 30, and 31 ASD samples, and 33, 27, and 29 control samples from frontal cortex, temporal cortex, and cerebellum, respectively, with 45 subjects in common across 3 brain regions and 20 subjects in common across 2 regions (Supplementary Fig. 1). These datasets were generated as part of our transcriptomic study of ASD brain[9]. Overall, the ASD and control groups did not have significant differences in variables that might confound RNA editing analysis (e.g. age, gender, etc.) (Supplementary Fig. 2). Each brain sample was sequenced to an average of 70 million raw read pairs and averaged 55 million uniquely mapped pairs (Supplementary Fig. 3)[9].

We applied our previously developed methods to identify RNA editing sites using the RNA-Seq data[19], and implemented additional steps to capture editing sites located in “hyperedited” regions, which were likely missed by regular methods[20] (Methods). Combining these approaches, we identified a total of 98,477, 97,994, and 134,085 predicted RNA editing sites from frontal cortex, temporal cortex, and cerebellum, respectively. As expected, the number of predicted RNA editing sites per sample correlated with read coverage approximately (Supplementary Fig. 4).

On average, >95% predicted RNA editing sites were A-to-G and T-to-C editing types per sample, and the remaining 5% mainly consisted of C-to-T and G-to-A types, consistent with canonical A-to-I and C-to-U editing reflected in non-strand-specific RNA-Seq data (Fig. 1a, Supplementary Fig. 5). Notably, most (84%) of the A-to-I editing sites are listed in the REDIportal database[21] (Supplementary Fig. 5). The majority of RNA editing sites were located in Alu sequences[21] (Supplementary Fig. 5) and in intronic regions[21] (Supplementary Fig. 5), and the sequence context of A-to-G sites was consistent with the typical sequence signature known for ADAR substrates[22] (Supplementary Fig. 5). Examination of correlation between ADAR expression levels and various partitions of editing sites (Alu, non-Alu repetitive, non-repetitive regions) showed overall positive correlation with ADAR1 and ADAR2 across the editome and weakly negative or no correlation with ADAR3 (Supplementary Fig. 6). These findings are consistent with known properties of RNA editing established in the literature[23], and altogether, strongly support the validity of our predicted A-to-I editing sites.

The frontal and temporal cortex shared more than 70% of their sites, while the two cortical regions and cerebellum shared 50-55% (Supplementary Fig. 5). Furthermore, the editing levels of common editing sites between two brain regions were highly consistent (correlation coefficient 0.96 between cortices, and 0.89 to 0.90 between cortex and cerebellum, Supplementary Fig. 5). Thus, the three brain regions demonstrated similarities and differences in RNA editomes, with the cortices having more similarities in RNA editomes than with cerebellum, likely reflecting the substantial differences in cellular composition and physiology between these two regions[24].

2.2.2 Reduction of RNA editing in ASD frontal cortex

Given the observed difference in RNA editing between brain regions, we first focused on analysis of RNA editing dysregulation in frontal cortex, a region with strong transcriptomic alterations in ASD[8,9]. We identified a total of 3,314 differential editing sites in ASD ($p < 0.05$, and editing level difference ≥ 5% or editing prevalence difference ≥ 5%, see Methods and Supplementary Table 2), which were robust to the choice of statistical models and parameters (Methods, Figs. S7-9). For each individual, 2.6-10.5% of all editing sites were identified as differential (Fig. 1b). Strikingly, the differentially edited sites showed a bias of hypoediting in ASD samples (Fig. 1c); the number of down-regulated RNA editing sites in ASD far outnumbered those that were upregulated ($p = 1.3e-59$, Chi-squared test, Fig. 1c).

Across potentially confounding biological and technical variables, diagnosis (i.e., ASD or control) was the only variable with significant association (Supplementary Fig. 10), allowing differential editing sites to substantially separate the two groups of subjects (Fig. 1d). Also, genes harboring the differential editing sites had minimal gene expression differences between ASD and control groups (Supplementary Fig. 10), evidencing that differential editing was unlikely secondary to differential gene expression.

We utilized Sanger sequencing to confirm the observed editing differences of 8 sites (Supplementary Table 3), covering an expansive range of editing level differences (Fig. 1e). Each editing site was tested in eight postmortem frontal cortex samples (4 ASD, 4 controls), selected based on sample availability (Supplementary Table 1b). The editing differences calculated from RNA-Seq strongly correlated with those from Sanger sequencing (Fig. 1e, $R^2 = 0.75$), validating the accuracy of our editing level quantification.

The set of genes harboring at least one differential editing site in frontal cortex (total of 1,189) exhibited significant gene ontology (GO) enrichment for categories including ionotropic glutamate receptor activity, glutamate gated ion channel activity, and synaptic transmission (Fig. 1f). Consistently, genes (e.g., *KCNIP4*, *PCDH9*, *RBFOX1*, and *CNTNAP2*) with the largest number of differential editing sites (both before or after correction for gene length, Supplementary Fig. 11) were involved in the above functional categories, and a number of genes with differential editing were also known ASD susceptibility genes[25] (Supplementary Fig. 11). For a relatively small number of genes (Supplementary Table 4), such as *KCND2* and *GRIK2*, that harbored differential editing sites associated with their gene expression, we observed strong enrichment in synaptic functions, including presynaptic and postsynaptic membrane, synaptic transmission, cell junction, dendrites, and similar categories (Supplementary

Fig. 11). Lastly, we observed that differential editing sites were significantly enriched in clusters of editing sites that abruptly increase between fetal and infant stages of cortical development[26] (Supplementary Fig. 12). Together, these results indicate that RNA editing could contribute towards aberrant synaptic formation in ASD.

2.2.3 Replication of reduced RNA editing in an independent cohort of ASD frontal cortex

For replication, we analyzed an independent cohort of ASD patients[27]. After balancing technical covariates, we analyzed RNA-Seq data from frontal cortex of 22 ASD and 23 controls (Supplementary Fig. 13, Supplementary Table 1c). This data set had single-end reads from polyA primed libraries and low sequencing depth (< 12 million total reads per sample, which led to slight 5' to 3' bias), constricting sufficient coverage to only 4952 editing sites. We, nevertheless, identified differential editing in 185 sites, with 65% exhibiting reduced editing in ASD (Supplementary Fig. 14, Chi-squared test p = 0.0085), thus reproducing the hypoediting pattern of our main dataset. Differential editing sites in the replication dataset were likewise enriched in genes involved in synapse and cell junction (Supplementary Fig. 14), and the levels of differential editing significantly correlated with those in our study (Supplementary Fig. 14). Replication of the editing landscape using data from a different cohort collected by a different lab strongly supports the validity of our observed ASD editing profiles.

2.2.4 Global analysis of potential regulators of hypoediting in ASD

To elucidate the regulatory mechanisms of hypoediting in ASD brains, we examined the mRNA and protein expression levels of the *ADAR* genes but did not observe significant differences of *ADAR1* and *ADARB1 (ADAR2)* expression in frontal cortex (Fig. 2a-c). Although *ADARB2* (*ADAR3*) protein was undetectable in the brain samples (Supplementary Fig. 15), its mRNA was slightly downregulated in ASD (Fig. 2a), which, as an RNA editing inhibitor[23], cannot explain the observed hypoediting in ASD. The *ADAR* genes did not exhibit differential splicing in these samples, as determined previously[9], and have no reported rare or common variants associated with ASD.

Given the absence of explanatory variation by *ADARs*, we hypothesized other trans-regulators must causally contribute. Given the large-scale editome profiles in this study, if a prevailing mechanism exists for hypoediting in ASD, then a significant number of editing sites should demonstrate correlated variation across the subjects. We applied weighted gene co-expression network analysis (WGCNA)[28] to search for highly correlated clusters of editing sites (i.e., modules) (Methods).

Remarkably, we identified a module enriched in editing sites that had significant association with diagnosis (Fig. 2d, Supplementary Table 5) and enrichment with differential editing sites between ASD and controls in frontal cortex of this study (Fig. 2e), and those from the replication cohort (Supplementary Fig. 14). Correlation between the module “eigengene” (i.e., eigen-editing site) and expression of potential *trans*-regulators (Supplementary Fig. 15) identified strong association between the turquoise module and Fragile X-relevant genes (*FMR1* and *FXR1*) (Fig. 2d). *FMR1* demonstrated positive (i.e. enhancing) correlation with editing changes, while *FXR1* displayed negative (i.e. inhibitory) correlation. This module is significantly enriched with genes related to synaptic ontology (Fig. 2f), consistent with a primary known function of FMRP in localization and maintenance of synapses[29], and previous reports showing enrichment of FMRP binding targets in ASD risk genes[17,18].

2.2.5 Interaction between Fragile X proteins and ADARs

To experimentally inspect the involvement of Fragile X proteins in RNA editing regulation, we first conducted subcellular fractionation experiment followed by Western blot and reciprocal co-immunoprecipitation (co-IP) experiments in HeLa cells to determine the localization and protein interactions of Fragile X proteins and ADARs. Consistent with previous literature, the ADAR proteins were enriched in the nuclear fraction[12], while FMRP and FXR1P were detected substantially in the cytoplasmic fraction[30] (Fig. 3a). Interestingly, FMRP and FXR1P were also highly detectable in nucleus, which was corroborated using immunofluorescence experiments (Supplementary Fig. 16). Subcellular distribution of ADAR proteins remained unchanged upon FMRP or FXR1P knockdown (Fig. 3a). Reciprocal co-IP experiments showed that FMRP interacts with both ADAR1 and ADAR2 in an RNA-independent manner (Fig. 3b), while FXR1P interacted with ADAR1 but not with ADAR2. Additionally, we observed interaction between FMRP and FXR1P (consistent with previous literature)[31], but not between ADAR1 and ADAR2.

2.2.6 FMRP and FXR1P binding relative to dysregulated editing sites

Next, we captured the transcriptome-wide binding patterns of FMRP and FXR1P to RNA transcripts using enhanced UV crosslinking and immunoprecipitation (eCLIP)[32]. Data from two eCLIP experiments and an input control experiment were obtained for each protein using postmortem frontal cortex from control subjects (Methods) (Supplementary Fig. 17).

We first confirmed the quality of our eCLIP experiments. eCLIP peaks identified in each replicate (Methods, Supplementary Table 6) demonstrated highly correlated read abundance (Supplementary Fig. 17), prompting us to combine peaks from the replicate experiments to maximize the sensitivity of peak detection. The binding sites of both proteins were predominantly distributed in genic 3' UTRs, introns and exons (Supplementary Fig. 17), consistent with previous literature[30,33]. Sequence motif analyses identified ACUG as the most enriched motif among the FMRP eCLIP peaks (Supplementary Fig. 17), which matches a FMRP binding motif previously reported[33], and CAUGC in FXR1P (Supplementary Fig. 17), which is consistent with a previous report that FXR1P tends to associate with AU-rich elements[34].

Next, we examined the proximity of FMRP and FXR1P binding peaks relative to dysregulated editing sites in ASD frontal cortex. Remarkably, the FMRP and FXR1P eCLIP peaks were significantly enriched around editing sites in the turquoise module (Fig. 3c, Methods), a finding that replicated in the FMRP eCLIP data generated from K562 cells by ENCODE[35] (Supplementary Fig. 18), but, importantly, not for proteins lacking evidence for RNA editing regulation (Supplementary Fig. 18). Additionally, FMRP and FXR1P eCLIP target genes significantly overlapped with genes harboring differential editing sites or sites in the turquoise module (Supplementary Fig. 19). These results suggest that FMRP and FXR1P proteins may regulate RNA editing directly in ASD.

2.2.7 FMRP directly modulates RNA editing

To investigate whether FMRP directly affects RNA editing, we conducted a series of minigene reporter assays (Supplementary Fig. 21, Methods) on two example editing sites in HeLa cells (Supplementary Table 3). These editing sites, located in the 3' UTRs of the *TEAD1* and *EEF2K* genes, were chosen due to close proximity with putative FMRP binding motifs (Supplementary Fig. 20). The *TEAD1* and *EEF2K* editing sites are likely site-specific editing sites, since no other sites were observed in their immediate neighborhood.

Knockdown of *FMR1* and *ADAR2* caused significant reduction of editing at the *TEAD1* editing site (Fig. 3d). Similarly, knockdown of *FMR1* caused significant reduction of *EEF2K* editing level (Fig. 3e) and a trend of reduction upon *ADAR1* knockdown (p = 0.06). *EEF2K* is also endogenously edited in HeLa cells, and responded to *FMR1* and *ADAR1* knockdown significantly, concordant with the minigene assays (Supplementary Fig. 21). These data are consistent with our observation that FMRP multifariously interacts with both ADAR1 and ADAR2 proteins and corroborates the positive association of turquoise eigen-editing site with *FMR1* expression levels (Fig. 2d).

Next, we introduced mutations to the FMRP binding motifs in the minigenes in order to weaken the protein-RNA interaction (Supplementary Fig. 20). Loss of these

FMRP binding sites caused significant reduction in RNA editing (Fig. 3f, g), importantly, without changing the predicted double-stranded RNA (dsRNA) structures (Supplementary Fig. 20). Our results suggest that FMRP directly regulates editing of these two site-specific sites through mediated interaction between ADAR and the RNA.

2.2.8 FXR1P regulates hyperedited sites

In contrast to site-specific editing, another class of editing sites consists of hyperedited sites that tend to cluster together[20]. We conducted minigene experiments on three genes (*CNTNAP4*, *NLGN1*, and *TENM2*) that all had manifold editing sites within long double-stranded intronic regions (Supplementary Fig. 22, Supplementary Table 3), two of which (*CNTNAP4* and *NLGN1*) are ASD risk genes. Consistent with its known role in hyperedited RNA editing[20], *ADAR1* knockdown caused reduction in all the detectable editing sites (Fig. 3h), though interestingly *ADAR2* knockdown did too to a lesser degree. Remarkably, the hyperedited sites in these genes showed increased editing levels in *FXR1* (but not in *FMR1*) knockdown cells, which was again consistent with the WGCNA results that showed negative correlation between *FXR1* expression and RNA editing (Fig. 2d). RNA immunoprecipitation experiments supported that FXR1P binds to the regions harboring the editing sites in these target genes (Supplementary Fig. 23). Additionally, mutations in predicted FXR1 binding motifs induced higher editing levels at a majority of sites in two of the three minigenes (Fig. 3i, Supplementary Fig. 23). These results potentially indicate direct inhibitory regulation of hyperediting sites by FXR1P through mediated interaction between ADAR1 and RNA.

2.2.9 Concomitant regulation of RNA editing by FMRP and FXR1P

To further substantiate the above findings, we validated 6 more differential editing sites in two neuroblastoma cell lines (Supplementary Table 3). These candidate sites were chosen based on their propinquity to FMRP or FXR1P eCLIP sites and their nominal correlation with the turquoise module, *FMR1*, or *FXR1* gene expression. As expected, *ADAR1* and *ADAR2* shRNA knockdown reduced editing at all editing sites (Supplementary Fig. 24). Strikingly, *FMR1* shRNA knockdown caused significant reduction of editing of all sites, while *FXR1* knockdown caused significant augmentation of editing in 10 of the 12 sites (Fig. 3j, S24). These results were reproducible between the two cell lines, further substantiating the inhibitory role of FXR1P and enhancing role of FMRP in editing regulation, and demonstrating concomitant regulation of RNA editing by these proteins at some editing sites. Together, our experimental results clearly validate that FMRP and FXR1P are important regulators of RNA editing.

2.2.10 Convergent RNA editing alterations between ASD and Fragile X patients

Loss of FMRP manifests in Fragile X syndrome, the most prevalent monogenic cause of ASD (1-2% of all ASD)[7,36] in which approximately 50% of patients have co-diagnoses or features of ASD[37]. To investigate a possible role for RNA editing contributing to shared molecular pathologies, we generated RNA-Seq data from the frontal cortex of four patients with Fragile X syndrome and four Fragile X carriers or controls (Supplementary Fig. 25). The samples were obtained and separately analyzed from two brain banks.

Western blot confirmed that FMRP expression was absent or reduced in the Fragile X samples relative to carriers or controls, and the expression levels of *ADAR1* and *ADAR2* were similar between the two groups (Supplementary Fig. 25).

Strikingly, differential editing sites identified in the Fragile X dataset (Supplementary Fig. 25, Methods) showed the same trends as those from ASD: they demonstrated a predominant trend of hypoediting in Fragile X patients and strong enrichment in genes related to synaptic transmission, cellular junctions, and ionic transmission (Fig. 4a,b, Supplementary Table 7), and were also significantly enriched around FMRP and FXR1P eCLIP peaks (Fig. 4c). Moreover, a statistically significant overlap was observed between the differentially edited genes in Fragile X patients and those in the turquoise module identified from ASD frontal cortex (Fig. 4d), the module that is correlated with *FMR1* expression (Fig. 2d). In addition, a significant overlap exists between the differential editing sites in Fragile X patients and editing sites in the turquoise module of ASD for data from one of the two brain banks (Supplementary Fig. 26). Overall, these results again support our hypothesis that the turquoise module encapsulates a subset of dysregulated editing sites in ASD that are under regulation by FMRP.

Altogether, the analysis of editing profiles in Fragile X patient brain provides a strong independent line of evidence showing convergence of dysregulated RNA editing between Fragile X syndrome and ASD through a common mechanism involving FMRP regulation of RNA editing.

2.2.11 Consistent hypoediting patterns observed for different brain regions of ASD patients

Here we investigated whether other brain regions share similar editing patterns with the frontal cortex. In temporal cortex and cerebellum, we also observed global down-regulation of RNA editing and enrichment in synapses, cellular junctions, and ionic channels (Fig. 5a and Supplementary Fig. 27). Overall, differential editing sites shared between brain regions showed significant correlation in levels of dysregulation (Fig. 5b). Likewise, WGCNA, performed on the editing sites identified in temporal cortex and cerebellum, identified downregulated modules (colored turquoise by WGCNA convention) strongly associated with ASD in these brain regions respectively (Fig. 5c, Supplementary Table 5). The turquoise modules of the three brain regions shared many editing sites (Fig. 5d). Overall, these results demonstrate that the global patterns of dysregulated editing are common across implicated brain regions in ASD.

A small set of 65 and 66 genes were, however, exclusively differentially edited in cortex and cerebellum respectively (Fig. 5e, Supplementary Table 8). They exhibited significant cortex- and cerebellum-specific expression patterns (Fig. 5f), suggesting that the region-specific differential editing may be explained by higher expression in their respective brain regions. It is likely that these region-specific genes have distinct functional roles in ASD.

We also examined 59 editing sites conserved across multiple phylogenetic taxa, likely serving as functionally paramount RNA editing sites in human[38]. Thirteen were identified as differentially edited in at least one brain region. Strikingly, they all exhibited hypoediting in ASD, 6 of which were recoding sites (Fig. 5g). Four of the recoding sites are located in glutamate receptors: *GRIA2* (R764G), *GRIA4* (R765G), *GRIK1* (Q621R) and *GRIK2* (Y571C)[13]. Additionally, another recoding site was found in the *NOVA1* gene (Fig. 5g), which codes for a brain-specific splicing factor that reportedly may cause down-regulated splicing in ASD[39]. This recoding site (S363G) stabilizes protein half-life of NOVA1[39], suggesting that the down-regulated editing may be an upstream causal factor of down-regulated splicing in ASD[9]. Overall these findings strengthen the association between RNA editing and aberrant synaptic signaling in ASD.

2.2.12 Common and brain region-specific mechanisms of RNA editing regulation in ASD

Next, we examined the prospective regulation of hypoediting in the other brain regions. The eigen-editing sites of the turquoise modules in the other two brain regions also displayed correlations with both *FMR1* and *FXR1* expression (Fig. 5c, although the correlation for *FXR1* in cerebellum was not statistically significant, p = 0.07), suggesting that regulation of RNA editing by FMRP and FXR1P may be a common mechanism for multiple afflicted brain regions in ASD.

Correlation of the expression levels of the *ADAR* (1, 2 and 3) and Fragile X-related genes with the 1st principal component (PC) of all differential editing sites (Supplementary Fig. 28) also recapitulated many of the turquoise module associations: *FMR1* significantly associated with the 1st PC in both frontal cortex and cerebellum, and *FXR1* negatively correlated in all 3 brain regions, corroborating their roles in positive and negative regulation of RNA editing respectively. Although we did not observe significant changes of the *ADAR* mRNAs between ASD and control groups in any brain region (Supplementary Fig. 28), *ADAR2* was significantly associated with the 1st PC of differential editing in temporal cortex (Supplementary Fig. 28) and validated by Western blot analysis showing a possible trend of downregulated ADAR2 protein in the temporal cortex of ASD (Fig. 5h,i). Lastly, *FXR2*, though not associated with the turquoise module in frontal cortex, showed significantly positive correlation with the turquoise module in temporal cortex (Fig. 5c) and with the PC of differential editing in cerebellum (Supplementary Fig. 28). Future studies are needed to examine the roles of FXR2 and ADAR2 in these brain regions.

2.2.13 Exacerbated severity of hypoediting patterns in dup15q patients

Duplication of chromosome 15q11.2-q13.1 (i.e., dup15q), accounting for 0.25-3% ASD diagnoses[40], clinically manifests with more severe motor impairments and intellectual disability than idiopathic ASD[40,41], along with greater magnitude and homogeneous dysregulation of gene expression and splicing[9]. We analyzed RNA editing in dup15q from frontal cortex (8 samples), temporal cortex (9 samples), and cerebellum (5 samples) against covariate matched controls (Supplementary Fig. 29). Dup15q patients exhibited more profound hypoediting (Fig. 6a) than idiopathic ASD (Fig. 1c, 5a). Correlation between differential editing levels and the Intelligent Quotient (IQ) scores of the idiopathic ASD individuals (Supplementary Fig. 30) was also very high, though not significant because only a handful of ASD subjects had IQ information: temporal cortex (R^2=0.64), cerebellum (R^2=0.36), and most prominently in frontal cortex (R^2=0.80), the region considered most strongly associated with cognitive function[42]. These results suggest that editing dysregulation could be related to the severity of cognitive deficits.

The landscape of editing in dup15q recapitulated the trends in idiopathic ASD. Differential editing levels in dup15q significantly correlated with those in idiopathic ASD (Fig. 6b), and were enriched in the turquoise modules observed in the idiopathic subjects (Fig. 6c), and showed greater concordance and magnitude of hypoediting (Fig. 6d). We found hypoediting at nearly all the testable (Methods) 59 conserved sites, including replicated differential editing at the glutamate receptors *GRIA2* (R764G), *GRIA4* (R765G), *GRIK1* (Q621R), *GRIK2* (Y571C), and *NOVA1* (Fig. 6e). Overall, these results not only strongly validate the hypoediting landscape identified across the 3 brain regions of ASD but also reveal an exacerbated hypoediting bias in a subset of ASD patients with severe clinical phenotypes.

2.3 Discussion

Here we performed the first global investigation of RNA editing in ASD and uncovered a common trend of hypoediting in ASD patients across different brain regions and different patient cohorts. Furthermore, we showed correlation between the hypoediting and *FMR1* and *FXR1* genes, which we validated as direct regulators of multiple and diverse sites in human. Consistent with these roles, we demonstrated convergent RNA editing patterns between ASD and Fragile X syndrome, revealing a shared molecular deficit in these closely related neurodevelopmental disorders.

As the cause of the Fragile X syndrome and as a syndromic ASD, FMRP has been subject to a myriad of ASD studies: 1) genes with rare de novo mutations[17], common variation[43], and copy number variants[44] in ASD are enriched in FMRP target genes[30]. 2) Multiple transcriptome analyses identified significant correlation between FMRP expression and ASD-associated gene modules[8,18]. 3) Many similar cognitive and behavioral symptoms manifest in both ASD and Fragile X syndrome[36]. 4) The protein level of FMRP has been shown to be downregulated in ASD patients[45]. The plethora of related literature supports the involvement of FMRP in the pathogenesis of ASD and highlights the need to elucidate its potential molecular mechanisms. Our study addresses this question by showing that RNA editing may be strongly associated with the molecular pathology via which FMRP contributes to the molecular abnormalities observed in ASD.

Our data supports a model where FMRP directly mediates the interaction between ADAR and the RNA substrates to promote editing, which advances previous studies of FMRP and RNA editing in Mouse, Drosophila, and Zebrafish[29]. The involvement of FXR1P in RNA editing regulation was unknown, and intriguingly, we observed that FXR1P, likely through a similar model, represses editing. Additionally, FMRP and FXR1P showed distinct features among the validated regulatory targets, where FXR1P acted on promiscuous sites and FMRP on site-selective editing sites. Nevertheless, the two proteins also shared common validated target sites, suggesting they could have some synergistic regulation of RNA editing, as they do in other biological processes relevant to neurodevelopmental disorders, such as neurogenesis[29,46].

Our study revealed substantial similarities and highly reproducible patterns in global editing changes in ASD across the three brain regions we profiled, indicating it may affect molecular pathways in general neurological function. Nevertheless, our data also allude to some region-specific editing regulation, such as a downregulation trend of ADAR2 protein in the temporal cortex, but not in the frontal cortex or cerebellum. Expression levels of the gene *FXR2*, a homolog of *FMR1*, demonstrated strong correlation with RNA editing levels, which is again a temporal cortex-specific observation (Fig. 5c). Future studies aimed at studying region-specific RNA editing will further elucidate these and other region-specific regulatory mechanisms.

Individuals with ASD frequently score lower in IQ testing than neurotypicals[47]. Our analyses, although based on a small data set, showed a high correlation between differential editing and IQ scores in all 3 brain regions. Additionally, dup15q patients, generally known to manifest more severe motor impairments and intellectual disability than idiopathic ASD, showed nearly unidirectional and greater severity of hypoediting than idiopathic patients in all 3 brain regions. These findings support an association between intellectual disability and RNA editing in ASD, which awaits confirmation in subsequent cohorts.

RNA editing alterations occurred in genes of critical neurological relevance (Supplementary Fig. 11), including contactins (*CNTNAP2*, *CNTNAP4*), neurexins (*NRXN1*, *NRXN3*), ankyrins (*ANK2*), and neuronal splicing factors (*NOVA1* and *RBFOX1*), which all harbor genetic mutations associated with ASD[25]. Although causality here is indeterminable, the occurrence of aberrant RNA hypoediting in known ASD risk genes suggests these changes contribute to disease risk. They certainly contribute to the disorder’s molecular pathology. Additionally, the differential editing sites significantly overlapped with developmentally regulated editing sites, suggesting that hypoediting may disrupt editing dependent functions during cortical developmental and further accentuates the potentiating role of early-onset molecular pathologies in ASD. Furthermore, some differential editing sites showed correlated editing levels with expression levels of their host genes, which may indicate a functional relationship. Together, this current work indicates that it will be important to further explore the role of RNA editing in ASD pathophysiology, so as to determine whether these changes are causal, or reflect homeostatic or dyshomeostatic responses.

2.4 Methods

2.4.1 RNA-Seq data sets of ASD and control brain samples

We obtained RNA-Seq data sets of three brain regions of ASD and control subjects from our previous study[9]. For idiopathic ASD, we used all data sets except (1) samples from subjects < 7 years old (which showed outlying expression patterns compared to all other samples), and (2) samples containing a 15q duplication (dup15q), an established genetic cause of syndromic ASD[48]. Note that the dup15q samples were analyzed separately as described below. We confirmed that ASD diagnosis was not confounded by age, batch, and other biological and technical variables (Supplementary Fig. 2). The final sample set consisted of an approximately equal number of controls and ASD samples totaling 62 samples in frontal cortex, 57 samples in temporal cortex, and 60 samples in cerebellum (Supplementary Table 1).

2.4.2 Dup15q dataset

A total of 5, 8, and 9 RNA-Seq datasets of dup15q patients were obtained from cerebellum, frontal cortex, and temporal cortex respectively (Supplementary Table 1). In addition, 11, 14, and 13 controls were chosen respectively from the above idiopathic dataset to balance covariates (Supplementary Fig. 29), except batch and brain bank, as there were nominally significant (although not passing Bonferroni correction) confounding effects between batch, brain bank, and dup15q diagnosis for this subset of data[9].

2.4.3 Frontal cortex replication dataset

For replication of idiopathic results, we downloaded previously published RNA-Seq data that were obtained from frontal cortex of 63 ASD and control subjects[27]. After balancing confounding variables, 22 ASD and 23 control datasets remained, none of which overlapped subjects from our original dataset.

2.4.4 RNA-Seq data sets of Fragile X patients and carriers/controls

Postmortem frontal cortex samples of Fragile X patients and Carriers were obtained from the University of Maryland Brain and Tissue Bank and the University of California at Davis FXTAS Brain Repository (Supplementary Fig. 25). Total RNA was extracted using TRIzol (Thermo Fisher Scientific, 15596018). RNA-Seq libraries were prepared using NEBNext Poly (A) mRNA magnetic isolation module (NEB, E7490) followed by NEBNext Ultra Directional RNA library prep kit for Illumina according to manufacturer's instruction. RNA-Seq data were collected on an Illumina HiSeq 2000 sequencer.

2.4.5 RNA-Seq read mapping and RNA editing identification

RNA-Seq reads were mapped using RASER[49], an aligner optimized for detecting RNA editing sites, using parameters m = 0.05 and b = 0.03. Uniquely mapped read pairs were retained for further analysis. Unmapped reads were extracted and processed to identify "hyperediting" sites. A recent study showed that previous RNA editing identification methods failed to detect editing sites in hyperedited regions due to existence of a high number of mismatches in the reads[20]. Our implementation of the hyperediting pipeline closely followed a strategy used by a previous study[20]. In brief, all adenosines in unmapped reads were converted into guanosines. These reads were aligned to a modified hg19 genome where adenosines were also substituted by guanosines. Uniquely mapped read pairs were obtained from this alignment step, and previously converted adenosines were reinstituted. We then combined these hyperedited reads with the originally uniquely mapped reads to identify RNA editing sites.

The procedures described in our previous studies were used to identify RNA editing sites[19,50]. First, RNA editing sites were identified as mismatches between the reads and the human reference genome. A log-likelihood test was carried out to determine whether an RNA editing site is likely resulted from a sequence error[19]. A number of posterior filters were then applied to remove RNA editing sites that were likely caused by technical artifacts in sequencing or read mapping[50].

Due to limited sequencing depth and the inherent nature of random sampling in RNA-Seq, some editing sites are observable in only a small number of subjects within a population cohort. Editing sites with low apparent prevalence lack sufficient sample size to enable a comparison between ASD and control groups. Therefore, we applied the following filters to retain a subset of editing sites: (1) in each sample, an editing site was required to have at least 5 total reads among which at least 2 reads were edited; (2) the editing site should satisfy filter (1) in at least 5 samples. We applied these filters to editing sites called within each brain region separately.

2.4.6 Identification of differential RNA editing sites

We define differential RNA editing sites as those: (1) that had significantly different average editing levels between ASD and controls, or (2) that were observed at significantly different population frequencies. A challenge with statistical testing for differential editing levels is that editing level estimation has a larger uncertainty at lower read coverage. More accurate calculations could be obtained by setting a high threshold for read coverage. However, this remedy leads to fewer samples or reduced power per editing site. We developed a strategy that attempts to optimize the trade-off between read coverage requirement and sensitivity in detecting differential editing.

Specifically, the following procedures were implemented for each editing site e_i. (1) we first identified the highest total read coverage requirement for e_i at which there were at least 5 samples in both control and ASD groups. The following read coverage was considered: 20, 15 and 10, in the order of high to low. (2) If a read coverage requirement C was reached in (1), we calculated the average editing level of e_i among the ASD and control samples (M_{ai}, M_{ci}), respectively, that satisfied C. (3) We then considered samples where e_i did not have at least C reads, but satisfied a lower read coverage cutoff (15, 10, or 5). These samples were included if their inclusion did not alter M_{ai} and M_{ci} by more than 0.03. (4) We carried out Wilcoxon rank-sum test between editing levels of the above samples in ASD and control groups. (5) If a read coverage requirement C was not reached in (1), then we tested all samples where e_i had ≥ 5 read coverage so long as there were at least 10 ASD and 10 control samples. Differential editing sites were defined as those with a p value < 0.05 and an effect size > 5%, in lieu that an editing change of approximately this magnitude was sufficient to cause dendritic deficits in mice[51].

Another type of differential editing was defined as editing sites that have different prevalence between ASD and controls. For each editing site, a Fisher's Exact test was carried out to compare the numbers of ASD and control samples with nonzero editing levels, with the background being the numbers of ASD and control samples with zero editing level. The minimum read coverage requirement per site was obtained using the same adaptive procedure as described above for the first type of differential editing sites. Differential editing sites were defined as those with p < 0.05 and an effect size > 5%. Differential editing sites identified via the above two methods overlapped significantly (Supplementary Fig. 7).

Differential editing sites detection in the replication ASD dataset[27] was performed similarly. However, because only 4952 editing sites had sufficient coverage, we eschewed effect size cutoffs and considered sites differential in which $p < 0.05$.

The dup15q subset sample size was too small to leverage the adaptive coverage model. Instead, we only tested editing sites where ≥ 5 dup15q and ≥ 5 control samples had ≥ 5 read coverage (defined as tesSupplementary Table ites). Differential editing sites had $p < 0.05$ and effect size > 5% from either Wilcoxon rank-sum test or Fisher's Exact test.

2.4.7 Computational comparison of methods and parameters for differential editing identification

Another *de facto* method for conducting differential testing in postmortem brain studies is to leverage a multilinear regression model to correct for potential technical confounders. We compared the results of our methods against those of a multilinear regression model including diagnosis, sex, age, and RIN as independent factors against RNA editing level. The set of differential editing sites strongly and significantly overlapped across all brain regions and available sample sizes (Supplementary Fig. 8, odds ratio 7-139), suggesting that *a priori* balancing of ASD and control groups was sufficient to obviate technical conflation. An additional issue with multilinear regression is a propensity for spuriously introducing noise at editing sites with smaller training sizes. Indeed, we found that the differential editing sets at smaller sample sizes (0-10 and 10-20) had more disparate calls between the two methods than the larger samples sizes (20-60).

We also tested whether the particular choice of parameters chosen for M_{ai} and M_{ci} significantly altered the differential editing values. We performed differential editing analysis with varying values of M_{ai} and $M_{ci,}$ and juxtaposed the differential editing values with the originally called values (M_{ai} and M_{ci} = 0.03*)* (Supplementary Fig. 9). The correlation remained nearly at 1, which shows that the differential editing values are robust to the choice of M_{ai} and $M_{ci.}$

2.4.8 Identification of genes enriched with differential editing

This analysis aims to identify genes that are enriched with differential editing sites. One might consider the top differentially edited genes as those with the largest number of differential editing sites. However, as expected, there exists a positive correlation between gene length and the number of differential editing sites (Supplementary Fig. 11). Therefore, we used a linear model to construct a regression between these two variables. We defined genes as enriched with differential editing if they had more differential editing sites than expected (beyond 95% confidence interval of the expected mean).

2.4.9 Differential editing sites associated with gene expression

To examine the association between differential editing and gene expression, we screened for significant correlations between editing level of each differential editing site and the FPKM value of its host gene. Specifically, the correlation coefficient between editing level and FPKM had to pass nominal significance ($P < 0.05$) within a multilinear regression: FPKM ~ age, sex, batch, RIN, brain bank, seqStatPC1, seqStatPC2, editing level. seqStatPC1 and seqStatPC2 are the first and second principal components encompassing 99% of variance of technical variables as described in our previous work[9].

2.4.10 Enrichment of editing sites in developmentally distinct editing clusters

Editing sites identified in 33 postmortem frontal cortex samples spanning the human lifespan (fetal, infant, child, teen, middle, and old age) were obtained from a previous study[26]. The original study classified editing sites into 3 developmental trajectories (constantly lowly edited sites, perpetually highly edited sites, and developmentally increasing sites). We recapitulated the 3 developmental trajectories on editing sites residing in *all* genomic regions using similar clustering criteria as in the original study. Briefly, editing sites with a median coverage < 20 reads across all samples were discarded. Then, we performed one-way ANOVA on each editing site across the six age groups. We considered editing sites passing FDR < 0.05 as developmentally increasing sites. Amongst the remaining sites, those with median editing level > 0.5 were categorized as perpetually highly edited sites; those with median < 0.5, as constantly lowly edited sites. Enrichment of editing sites in ASD within these 3 developmental clusters was performed using Fisher's Exact test.

2.4.11 Annotation of editing sites and heatmap generation

Editing sites were annotated using ANNOVAR[52]. Heatmaps throughout this study were generated using circlize[53].

2.4.12 Principal components analysis (PCA)

PCA was conducted on differential editing sites in order to examine associations between PCs and potential confounding covariates. The R function prcomp was used for this purpose. Missing values in the editing level matrix were imputed using the missMDA package[54]. The PCs were then correlated against technical and biological covariates such as age and gender (Supplementary Fig. 10). The first PC was predominantly associated with ASD diagnosis, and was thus used as the PC for differential editing.

2.4.13 Weighted gene co-expression network analysis (WGCNA)

The WGCNA package[28] in R was used to find modules of correlated editing sites. In multi-sample analysis, it is typical that some editing sites have no available values (missing data) in certain samples that lack read coverage at those sites. To preclude inaccurate calculations due to samples with too much missing data, we used the following requirements for editing sites to be included in WGCNA: (1) with ≥5 reads in ≥ 90% of samples and (2) with nonzero editing levels in ≥ 10% samples. In addition, to detect variation in the data, we further required that the included editing sites had a standard deviation ≥ 0.1 in their editing levels across samples. A soft threshold power of 10 was used to fit scale-free topology. To avoid obtaining modules driven by outlier samples, we followed our previous bootstrapping strategy[9,11]. One hundred bootstraps of the data set were carried out to compute the topological overlap matrix of each resampled network. Co-editing modules were obtained using the consensus topological overlap matrix of the 100 bootstraps.

WGCNA offers a dynamic tree-cutting algorithm, which enables identification of modules at various dendrogram heights and allows delineation of nested modules[55]. However, upon examination of the WGCNA dendrogram (Fig. 2d), we observed only one pronounced module of editing sites. Furthermore, most modules, identified through dynamic tree cutting, were generally unstable, highly dependent on tree cutting parameters. Therefore, we used the traditional constant height tree cutting, provided by WGCNA as cutreeStaticColor, with cutHeight set to 0.9965, which produced the single turquoise module. This is the largest module that is most likely biologically relevant and technically robust. In addition, this module is conserved across brain regions (Fig. 5c).

2.4.14 Association of modules with ASD diagnosis and RNA binding proteins

To test the association of the turquoise module with diagnosis, we first defined eigen-editing sites as the first principal component of the module, according to WGCNA recommendation[56]. A linear regression model was constructed between the eigen-editing sites and diagnosis, in addition to biological and technical covariates including

RIN, age, gender, sequencing batch, PMI, brain bank, 5’ to 3’ RNA bias, AT dropout rate, GC dropout rate, mapped bases in intergenic regions, uniquely mapped reads. The linear model was fit with backwards selection, and the module was deemed as associated with ASD diagnosis if $p \leq 0.05$ for the coefficient of this variable.

We tested if a module was enriched with differential editing sites using Fisher’s Exact test. In addition, we tested the association between modules and potential regulatory genes by examining the correlation between the eigen-editing sites and mRNA expression levels of the genes. It should be noted that the mRNA expression levels were corrected values following removal of variability contributed by technical covariates[9].

2.4.15 eCLIP-Seq experiment and data analysis

The eCLIP experimental procedure is detailed in our previous studies[32,57]. The antibodies used for this experiment are: FMRP antibody (MBL, RN016P) and FXR1 antibody (Bethyl Laboratories, A303-892A). Flash-frozen brain tissue was cryo-ground in pestles pre-chilled with liquid nitrogen, spread out in standard tissue culture plates pre-chilled to -80°C, and UV crosslinked twice at 254 nM (400 mJ/cm^2). 50 mg of crosslinked tissue was then used for each eCLIP experiment, performed as previously described[32,57]. As controls, we sampled 2% of the pre-immunoprecipitation (post-lysis and fragmentation) sample and prepared libraries identically to the FMRP or FXR1P eCLIP (including the membrane size selection step). These libraries served as “size-matched input (SMInput)” to minimize non-specific background signal in the identical size range on the membrane as well as any inherent biases in ligations, RT-PCR, gel migration and transfer steps.

eCLIP-Seq data were analyzed using the CLIPper software[32] that generated a list of predicted binding peaks of the corresponding protein. In each replicate, peaks were further filtered to retain those whose abundance was at least 2 fold of that in the SMInput sample.

To examine the FMPR or FXR1P binding relative to RNA editing sites, we compared the distances from eCLIP peaks to turquoise editing sites compared to gene-matched random adenosines. Only editing sites residing in genes containing at least 1 eCLIP peak were considered. The closest distance between an editing site or random adenosine and eCLIP peaks were calculated. A total of 10,000 sets of controls were generated using this procedure. To determine a P value, we first plotted the cumulative distribution of the distances between editing sites or controls and the eCLIP sites. The area under the curve (AUC) of this distribution was calculated for the set of editing sites and each set of controls. The AUC calculation was constricted to the distance interval [0,100,000 kb]. AUC values of the 10,000 sets of controls were modeled by a Gaussian distribution, which was used to calculate a P value for the AUC of the set of editing sites. Density plots were generated using the geom_density function in the ggplot2 package in R. To avoid overplotting, we randomly selected and plotted ten of the control sets for visualization. Note that the observed linear distance between protein-RNA binding and the regulated target sites may be larger than the actual proximity of the protein and its targets, due to limited sensitivity of CLIP or existence of secondary or tertiary RNA structures.

To identify the motifs enriched in eCLIP peaks, we used two alternative methods: HOMER[58], and DREME[59]. We ran DREME with all eCLIP peaks of each protein using default parameters, which creates control sequences through dinucleotide shuffling. HOMER was run with the findMotifsGenome.pl program (-p 4 -rna -S 10 -len 5,6,7,8,9). Background controls were defined as randomly chosen sequences in the same type of genic region as the true peaks. The control sequences have one-to-one match in length with the actual peaks. Three sets of random controls were constructed. Homopolymer or dinucleotide repeats were discarded. We required the final consensus motif to be the most enriched motif identified by HOMER that was also one of the most enriched motifs resulting from DREME.

2.4.16 RNA editing analysis of Fragile X samples

The RNA-Seq data derived from Fragile X patients and carriers were analyzed similarly as those of the ASD cohorts. Fisher's Exact test was used to identify differentially edited sites using pooled patient and carrier data sets ($p \leq 0.05$ and effect size > 5%).

2.4.17 Gene ontology enrichment analysis

Gene ontology (GO) terms were downloaded from Ensembl[60]. For each query gene, a random control gene was chosen to match gene expression level and gene length (±10% relative to that of query gene). GO terms present in the sets of query genes and control genes were collected respectively. A total of 10,000 sets of control genes were obtained. For each GO term, a Gaussian distribution was fit to the number of control genes containing this GO term. The enrichment p value of the GO term among the query genes was calculated using this distribution.

2.4.18 Validation of RNA editing levels

RNA Extraction. Brain tissues were homogenized in RNA TRIzol reagent (Thermo Fisher Scientific, 15596018). Mixture was incubated on ice for 15 min. Chloroform was added to the mixture and incubated at room temperature for 10 min. The mixture was centrifuged at 12000g for 15 min, and the top layer was carefully extracted. Equal volume of 200-proof ethanol was added to the top chloroform layer and mixed thoroughly. RNA was further purified using Direct-zol™ RNA MiniPrep Plus kit (Zymo Research, R2072) following the manufacture's protocol.

cDNA synthesis and PCR. cDNA synthesis was carried out using random hexamers, 1 µg total RNA, and the SuperScript IV Reverse Transcriptase (Thermo Fisher Scientific, 18090050) as described in the manufacturer's protocol. Next, 2uL cDNA (corresponding to 1/10th of the original RNA) was used as template for PCR reactions using the DreamTaq PCR Master Mix (2X) (Thermo Fisher Scientific, k1082). PCR was performed on an Eppendorf thermal cycler using the following thermal cycle conditions for all candidate sites (5 min, 95°C for hot start followed by 30 cycles of 15 s, 95°C; 15 s, 55°C and 1min/kb, 72°C).

Topo Cloning and Clonal Sequencing. PCR products were run on 1% agarose gel and visualized under UV light. The correct size band was cut and digested by Zymoclean™ Gel DNA Recovery Kit (Zymo Research, D4002) following the manufacturer's protocol. PCR product was inserted into kanamycin resistant pCR 2.1-

TOPO vector (Thermo Fisher Scientific, 450641). Ligated clones were transformed into One Shot TOP10 Chemically Competent E. coli (Thermo Fisher Scientific, C404003). Transformed cells were streaked on LB/Agar plates containing kanamycin and X-Gal as selection markers and incubated overnight at 37°C. Each plate was divided into 4 quadrants and 6 white clones were randomly selected from each quadrant (total of 24 clones per patient sample per editing site). Each clone was inoculated overnight in LB/Kanamycin. Plasmid was extracted using Plasmid DNA Miniprep Kits (Thermo Fisher Scientific, K210011). Miniprep samples were sequenced using Genewiz Sanger sequencing. The number of the clones presenting G peak at the editing site of interest was counted to determine the estimated editing ratio.

2.4.19 Co-immunoprecipitation

HeLa cells were maintained with DMEM supplemented with 10% FBS and 100 U ml-1 penicillin/ streptomycin at 37 °C and 5% CO2. Ten million HeLa cells were collected and lysed in 1 ml non-denaturing lysis buffer at pH 8.0, containing 20 mM Tris-HCl, 137 mM NaCl, 1% NP-40, and 2 mM EDTA supplemented with complete protease inhibitor cocktail. Extracted proteins were incubated overnight with ADAR1 antibody (Santa Cruz, sc-271854) or FMRP antibody (Millipore, MAB2160) at 4 °C; precipitation of the immune complexes was performed with Dynabeads Protein G (Thermo Fisher Scientific, 1003D) for 4 h at 4 °C, according to the manufacturer’s instructions. For experiments involving Flag-ADAR2, the supernatant derived from Flag-tagged hADAR2 overexpressing cells was incubated for 3 h at 4 °C with Flag M2 antibody (Sigma, F1804). 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), Flag antibody (sc-807), FMRP antibodies (Millipore, MAB2160 and Abcam, ab17722), FXR1P antibody (Bethyl Laboratories, A303-892A), and FXR2 antibody (Sigma-Aldrich, F1554). The HRP-linked secondary antibodies were used and the blots were visualized with the ECL kit (GE, RPN2232).

2.4.20 Subcellular fractionation

Cells were fractionated following a previously published protocol with some modifications[61]. Briefly, monolayers of cells in 10-cm plates were washed twice with ice-cold PBS, followed by gentle scraping of cells. Cells were resuspended with the ice-cold HLB+N buffer (10 mM Tris-HCl, at pH 7.5, 10 mM NaCl, 2.5 mM MgCl2 and 0.5% NP-40) on ice for 5 min. Lysates were layered over a chilled 10% sucrose cushion made in the ice-cold HLB+NS buffer (10 mM Tris-HCl, at pH 7.5, 10 mM NaCl, 2.5 mM MgCl2, 0.5% NP-40 and 10% sucrose) and centrifuged for 5 min at 4 °C at 420g. After centrifugation, the supernatant was collected and served as the cytoplasmic fraction. The nuclear pellet was then treated with the ice-cold nuclei lysis buffer (10 mM HEPES, at pH 7.6, 300 mM NaCl, 7.5 mM MgCl2, 0.2 mM EDTA, 1 mM DTT, 1 M Urea, and 1% NP-40) after washing. Fractionation efficiency was validated by Western blot using antibody specific to the marker for each fraction: β-tubulin (Sigma, T8328) for the cytoplasmic fraction and rabbit polyclonal U1-70k (Santa Cruz, sc-390899) for the nucleoplasmic fraction.

2.4.21 Construction of minigenes and site-directed mutagenesis

Partial 3’ UTRs (EEF2K and TEAD1) and intronic (CNTNAP4, NLGN1, and TENM2) regions were restriction digested and inserted between the SacII/XhoI sites in the pEGFP-C1 vector. Overlapping oligonucleotide primers containing the desired mutations were used to amplify mutation-containing fragments from the wild-type minigene plasmid, using Q5 High-Fidelity DNA polymerase (New England Biolabs, M0491L). All resulting amplification products were confirmed by sequencing.

2.4.22 Transfection, RNA isolation, RT-PCR amplification, and analysis of RNA editing

HeLa cells were grown on 6-well plates under standard conditions at 37 °C with 5% CO2. Cells were grown to 70% confluence, and transfection was performed using Lipofectamine 3000 (Thermo Fisher Scientific, L3000015) with 100 ng of minigene plasmid. For editing validation of endogenous substrate, two neuroblastoma cell lines, SK-N-BE(2) and KELLY, were grown on 6-well plates without transfection of a minigene. Cells were harvested after 24 h. Total RNA was extracted using TRIzol reagent (Thermo Fisher Scientific, 15596018), followed by treatment with 1 U of DNase I (Zymo Research, E1011-A). RNA was further purified using Direct-zol RNA MiniPrep kit following the manufacture’s instruction (Zymo Research, R2072). Reverse transcription (RT) was performed on 1 μg total RNA for 1 h at 42 °C using random hexamer primer and SuperScript IV (Thermo Fisher Scientific, 18090050). The cDNA products derived from the expressed minigenes were detected by PCR using the pEGFP-C1-specific forward primer and a gene-specific reverse primer. On the other hand, cDNA products for the endogenous substrate were amplified with gene-specific primer set. Amplification was performed for 30 cycles, consisting of 30 s at 95 °C, 30 s at 55 °C, and 2 minutes at 72 °C. The products from RT-PCR were resolved on 0.8% agarose gels. The appropriate PCR product was excised and the DNA was extracted, purified, and analyzed by Sanger sequencing. A-to-I editing levels were calculated as relative peak heights (that is, ratio between the G peak height and the combined height of A and G peaks, height G / height A + height G).

2.4.23 Production of lentivirus and cell transduction for protein knockdown

pLKO1 non-target control-shRNA (SHC016), FMR1-targeting shRNA (TRCN0000059758) or FXR1-targeting shRNA (TRCN0000159153) constructs were used. We produced lentiviruses via co-transfection of pCMV-d8.91, pVSV-G and pLKO1 into HEK293T cells using Lipofectamine 3000 (Thermo Fisher Scientific, L3000015). Transduction was carried out according to the standard protocol of the ENCODE consortium[62]. Briefly, viruses were collected from conditioned media after 48 h co-transfection. Lentivirus-containing media was mixed with the same volume of DMEM media that contain polybrene (8 μg/ml), which was used to infect HeLa, SK-N-BE(2), and KELLY cells. After 24 h, cells were incubated with puromycin (2 μg/ml for HeLa and 1 μg/ml for SK-N-BE(2) and KELLY) for 3-7 days. Knockdown efficiency was evaluated by Western blot. Cells were lysed in RIPA containing complete protease inhibitor cocktail. Cell lysates were then resolved through 8% SDS-PAGE and probed by ADAR1 antibody (Santa Cruz, sc-271854), ADAR2 antibody (Santa Cruz, sc-73409), FMRP antibody (Millipore, MAB2160), FXR1P antibody (Bethyl Laboratories, A303-892A), and FXR2 antibody (Sigma-Aldrich, F1554).

2.4.24 Western Blot in ASD and Fragile-X brain samples

Brain tissues were homogenized in RIPA lysis and extraction buffer containing protease inhibitor (Thermo Scientific, 88866). Mixture was then incubated on ice for 30 minutes, sonicated, and spun down. Crude protein concentration was obtained using Pierce BCA Protein Assay Kit (Thermo Fisher Scientific, 23225). Equal amount of protein was separated using 8% SDS–PAGE and then transferred onto nitrocellulose membrane. The membrane was blocked with 5% non-fat milk (Genesee Scientific, 20-241) and 0.1% Tween 20 in tris-buffered saline. The blot was incubated in primary antibody solution against the protein of interest with 5% non-fat milk and 0.1% Tween 20 in TBS overnight at 4°C on shaker. Antibodies used in this experiment include ADAR1 antibody (Santa Cruz, sc-271854), ADAR2 antibody (Santa Cruz, sc-73409), ADAR3 antibody (Santa Cruz, sc-73410), FMRP antibody (Millipore, MAB2160). Secondary antibody containing goat anti-mouse IgG-HRP (sc-2005, Santa Cruz Biotechnology) or goat anti-rabbit IgG-HRP (sc-2004, Santa Cruz Biotechnology) was used to label the corresponding primary antibody. The blot was developed using Amersham ECL Prime

Western Blotting Detection Reagent (GE Healthcare Life Sciences, RNP2232) and imaged with the Syngene PXi immunoblot imaging system. Beta Actin was used as a loading control. Western blot images were analyzed using ImageJ. All uncropped images are included in Supplementary Fig. 31.

2.4.25 RNA immunoprecipitation (RIP)–PCR

RIP was performed according to previously published protocols with some modifications[63]. Cells were harvested on the second day of minigene transfection in RIP buffer (25 mM Tris-HCl, at pH 7.4, 150 mM KCl, 5 mM EDTA, 0.5% NP-40 and 0.5 mM DTT supplemented with complete protease inhibitor cocktail and 100 U ml-1 RNase OUT), sonicated (10 s three times with 1 min intervals) and centrifuged at 13,000 rpm for 10 min at 4 °C. Supernatant was treated with 100 U RNase-free DNase I (Zymo Research, E1011-A) at 37 °C for 30 min and then centrifuged again at 13,000 rpm for 10 min at 4 °C. For immunoprecipitation, lysates were incubated with FXR1P antibody (Santa Cruz, sc-374148) or anti-mouse IgG (Santa Cruz, sc-2025) as a negative control overnight at 4 °C. The Dynabeads were washed three times with the RIP buffer and bound RNA was isolated using TRIzol (Thermo Fisher Scientific, 15596018), according to the manufacturer’s instructions. Eluted RNA was reverse-transcribed using SuperScript IV (Thermo Fisher Scientific, 18090050) with random hexamer primers. PCR was carried out for 30 cycles, consisting of 30 s at 95 °C, 30 s at 55 °C, and 30 s at 72 °C. PCR products were analyzed by agarose gel electrophoresis.

2.4.26 Immunofluorescence

HeLa cells were seeded on Millicell EZ Slide 8-well glass (Millipore, PEZGS0816) and incubated overnight in DMEM with 10% FBS to obtain 60% monolayer cell confluency. Each chamber was carefully rinsed with ice-cold PBS. Cells were fixed in 4% paraformaldehyde at room temperature for 10 min and washed with ice-cold 0.1% PBS-T three times for total of 15 min. Cells were permeabilized with either 0.1% Tween-20 or Triton X-100 in PBS for 5 min. Block solution containing 5% normal donkey serum and 1% BSA in 0.3% PBS-T was applied for 1 h at room temperature on shaker. Cells were incubated in primary antibody solution of mouse anti-ADAR1 (1: 100; sc-271854, Santa Cruz Biotechnology) and rabbit anti-FMR1 (1: 100; ab17722, Abcam) in 0.3% PBS-T containing 1% NDS and 1% BSA for overnight at 4°C. Cells were washed three times with ice-cold 0.1% PBS-T for 5 min. Cells were then incubated in a secondary cocktail containing Highly Cross-Adsorbed AlexaFluor® 488-conjugated Donkey anti-Mouse IgG (1: 200; A-21202, Thermo Fisher Scientific), and AlexaFluor® 488-conjugated Donkey anti-Rabbit IgG (1: 200; ab150074, Abcam) in 0.3% PBS-T containing 1% NDS and 1% BSA. Chamber was disassembled to expose the slide. Vectashield Anti-fade mounting medium containing 4',6-Diamidino-2-Phenylindole, Dihydrochloride (DAPI) stain was applied to the slide and covered with a coverslip. Cells were examined and imaged at 63x oil-immersion objective using Zeiss LSM 780 confocal microscope with ZEN 2011 (Black edition) software and post-processed with ImageJ. All images were taken under identical setting and conditions.

2.4.27 Statistics

Differential editing sites were obtained using a two-tailed Wilcoxon signed-rank test under an adapting scheme (see previous section). Ascertaining bias for hypoediting was performed using a Chi-square test under the null hypothesis of equal numbers of up- and down-regulated editing sites. Significance of gene set and editing set overlaps were determined using a two-tailed Fisher’s exact test. Significance of minigene reporter assays were summarized using one-way ANOVA and a Student’s t-test against proper controls, where data distributions were assumed to be normal, but this was not formally tested. Data generated in this study was not randomized according to experimental conditions or stimulus presentations, and data collection and analyses were not performed blind to the conditions of the experiments. For statistics of more specific analyses, see the appropriate sections in Methods and Figure legends (also refer to the online "Life Sciences Reporting Summary").

2.4.28 Sample size selection

No statistical methods were used to pre-determine sample sizes, but our samples sizes are similar to those reported in previous publications.[8,9]

2.5 Acknowledgements

Postmortem brain samples of used in this study were obtained from the University of Maryland Brain and Tissue Bank, which is a component of the NIH NeuroBioBank. We are grateful to the patients and families who participate in the tissue donation programs. This work was funded by grants from the National Institute of Health to X.X. (HG009417, HG006264), G.W.Y. (HG004659, HG009417, HG007005 and MH107367), S.T. (T32HG002536), E.L.V.N. (HG009530), V.M.C. (MH094681), R.J.H. (HD 036071). S.T is supported by the UCLA Eureka Scholarship. E.L.V.N. is a Merck Fellow of the Damon Runyon Cancer Research Foundation (DRG-2172-13). G.A.P. is supported by the National Science Foundation Graduate Research Fellowship. G.R. is supported by NIH fellowship 1F32MH114620.

2.6 Figures

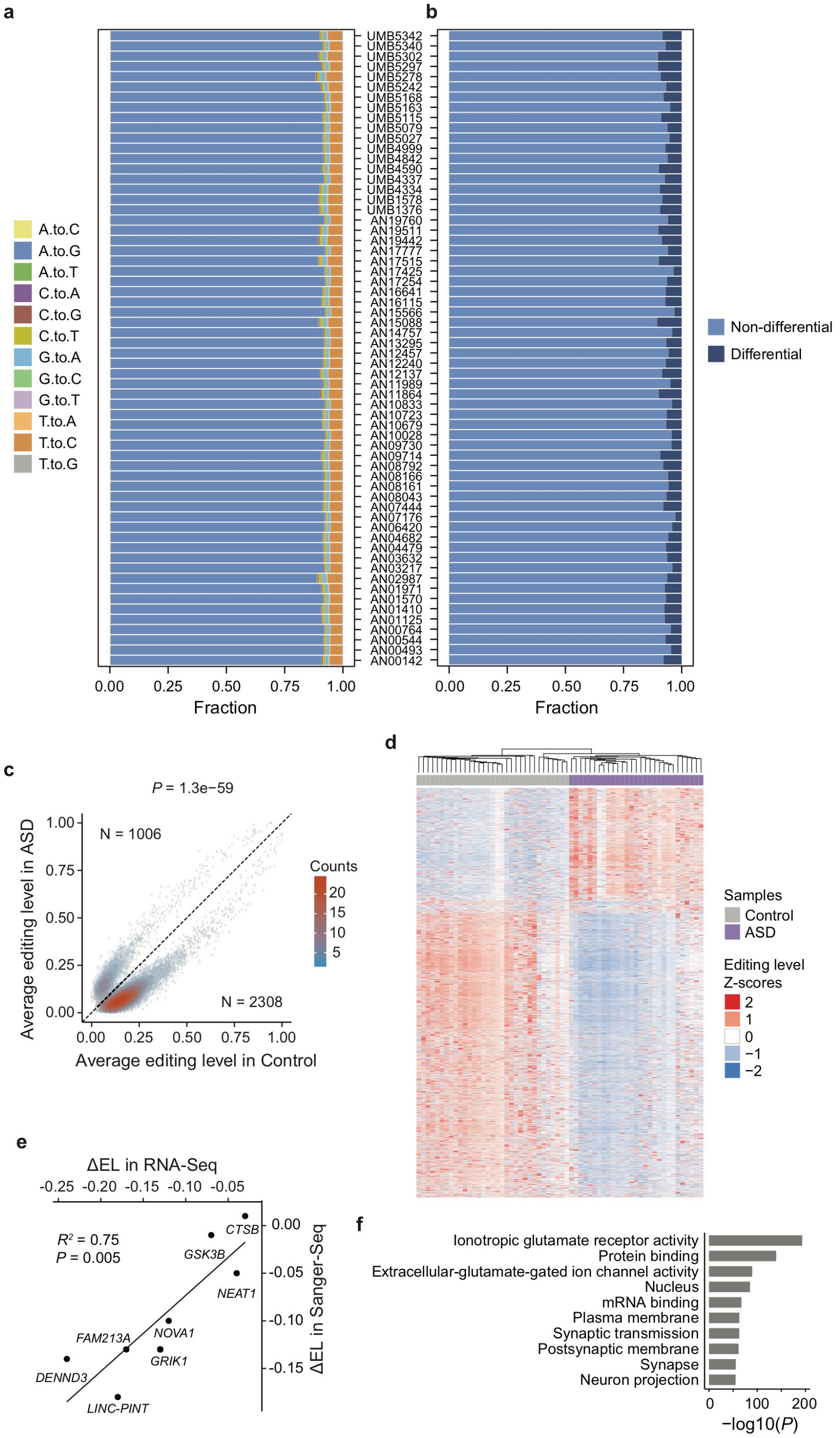

a
b
A.to.C
A.to.G
A.to.T
C.to.A
C.to.G
C.to.T
G.to.A
G.to.C
G.to.T
T.to.A
T.to.C
T.to.G
UMB5342
UMB5340
UMB5302
UMB5297
UMB5278
UMB5242
UMB5168
UMB5163
UMB5115
UMB5079
UMB5027
UMB4999
UMB4842
UMB4590
UMB4337
UMB4334
UMB1578
UMB1376
AN19760
AN19511
AN19442
AN17777
AN17515
AN17425
AN17254
AN16641
AN16115
AN15566
AN15088
AN14757
AN13295
AN12457
AN12240
AN12137
AN11989
AN11864
AN10833
AN10723
AN10679
AN10028
AN09730
AN09714
AN08792
AN08166
AN08161
AN08043
AN07444
AN07176
AN06420
AN04682
AN04479
AN03632
AN03217
AN02987
AN01971
AN01570
AN01410
AN01125
AN00764
AN00544
AN00493
AN00142
0.00
0.25
0.50
0.75
1.00
Fraction
0.00
0.25
0.50
0.75
1.00
Fraction
Non-differential
Differential
c
P = 1.3e−59
N = 1006
N = 2308
Average editing level in ASD
Average editing level in Control
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Counts
20
15
10
5
d
Samples
Control
ASD
Editing level
Z-scores
2
1
0
−1
−2
e
ΔEL in RNA-Seq
-0.25
-0.20
-0.15
-0.10
-0.05
R² = 0.75
P = 0.005
CTSB
GSK3B
NEAT1
NOVA1
FAM213A
GRIK1
DENND3
LINC-PINT
0.00
-0.05
-0.10
-0.15
ΔEL in Sanger-Seq
f
Ionotropic glutamate receptor activity
Protein binding
Extracellular-glutamate-gated ion channel activity
Nucleus
mRNA binding
Plasma membrane
Synaptic transmission
Postsynaptic membrane
Synapse
Neuron projection
0
100
200
−log10(P)

Figure 2.1 Transcriptome-wide differential editing in the frontal cortex of ASD.

a, Fraction of all types of RNA-DNA differences (RDDs) identified in the RNA-Seq data of each subject. **b**, Fraction of differential and non-differential editing sites for each subject. **c**, Average editing levels of differential editing sites in ASD and controls. Numbers (N) of editing sites that were up- or down-regulated in ASD are shown, which were compared via Chi-squared test (P value shown above plot). **d**, Differential editing sites segregate ASD and control samples. Normalized editing levels (z-scores) were used in hierarchical clustering. Each row corresponds to one editing site. Each column represents a sample. **e**, Experimental validation of differential editing levels using Sanger sequencing. The frontal cortex samples used in this experiment are shown in Supplementary Table 1. ΔEL: change in editing level (ASD - control), n=8 editing sites. **f**, GO enrichment analysis of genes harboring differential editing sites (n=1,189 genes, p-values determined by one-sided Gaussian test, see methods).

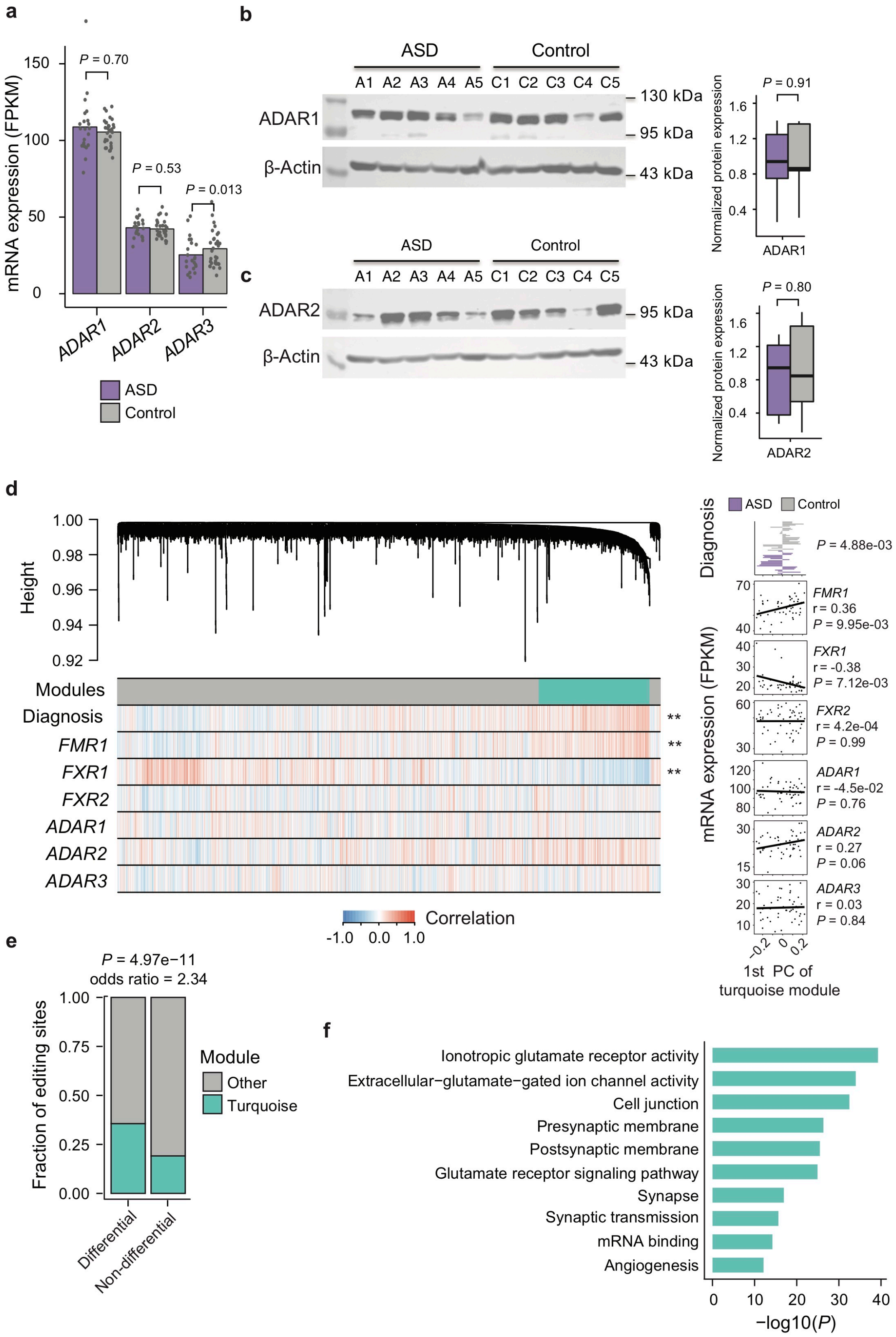

a
P = 0.70
P = 0.53
P = 0.013
150
100
50
0
mRNA expression (FPKM)
ADAR1
ADAR2
ADAR3
ASD
Control
b
ASD
Control
A1 A2 A3 A4 A5 C1 C2 C3 C4 C5
ADAR1
β-Actin
130 kDa
95 kDa
43 kDa
P = 0.91
1.6
1.2
0.8
0.4
Normalized protein expression
ADAR1
c
ASD
Control
A1 A2 A3 A4 A5 C1 C2 C3 C4 C5
ADAR2
β-Actin
95 kDa
43 kDa
P = 0.80
1.6
1.2
0.8
0.4
Normalized protein expression
ADAR2
d
1.00
0.98
0.96
0.94
0.92
Height
Modules
Diagnosis
FMR1
FXR1
FXR2
ADAR1
ADAR2
ADAR3
**
**
**
Correlation
-1.0 0.0 1.0
Diagnosis
ASD
Control
P = 4.88e-03
mRNA expression (FPKM)
70
40
FMR1
r = 0.36
P = 9.95e-03
40
30
20
FXR1
r = -0.38
P = 7.12e-03
60
30
FXR2
r = 4.2e-04
P = 0.99
120
100
80
ADAR1
r = -4.5e-02
P = 0.76
30
15
ADAR2
r = 0.27
P = 0.06
30
20
10
ADAR3
r = 0.03
P = 0.84
-0.2 0 0.2
1st PC of turquoise module
e
P = 4.97e−11
odds ratio = 2.34
1.00
0.75
0.50
0.25
0.00
Fraction of editing sites
Differential
Non-differential
Module
Other
Turquoise
f
Ionotropic glutamate receptor activity
Extracellular−glutamate−gated ion channel activity
Cell junction
Presynaptic membrane
Postsynaptic membrane
Glutamate receptor signaling pathway
Synapse
Synaptic transmission
mRNA binding
Angiogenesis
0 10 20 30 40
−log10(P)

Figure 2.2 Global analysis reveals potential regulators of differential editing in the frontal cortex of ASD.

a, mRNA expression levels (FPKM) of ADAR1, ADAR2, and ADAR3 estimated from RNA-Seq data (n=62 samples). P values were calculated using a regression approach where covariates were accounted for[15]. Dots show individual sample FPKMs. **b**, Western blot of ADAR1 protein in ASD and control samples. (Note the images are cropped and uncropped images are in Supplementary Fig. 31, same for all Western blot images hereafter.) Protein level was normalized against that of β-actin. Samples used in this experiment are shown in Supplementary Table 1 (chosen based on availability). A1-A5: ASD samples. C1-C5: control samples. P value was calculated via two-tailed Student's t-test. Boxplot definition: center=median, lower hinge=25th percentile, upper hinge=75th percentile, min and max extend to observations at most 1.5 * inter quartile range (IQR) . **c**, Similar as **b**, for ADAR2 protein. **d**, WGCNA analysis of RNA editing in frontal cortex (n=51 samples). Dendrogram of RNA editing sites is shown. The turquoise module is indicated by the turquoise color. Correlation of editing sites with diagnosis (ASD or control) and mRNA expression levels of a few genes is shown in the Heatmap. **: P < 0.01. Right panels: Bar graph and scatter plots represent association between diagnosis or mRNA expression levels and the first principal component (PC) of the turquoise module. P values of Pearson's correlation are shown. **e**, Overlap between the turquoise sites and differential editing sites. P value was calculated via Fisher's Exact test (n=4061 editing sites, two-tailed). **f**, GO enrichment analysis of genes harboring the turquoise sites (n=846 genes, one-tailed Gaussian test, see methods).

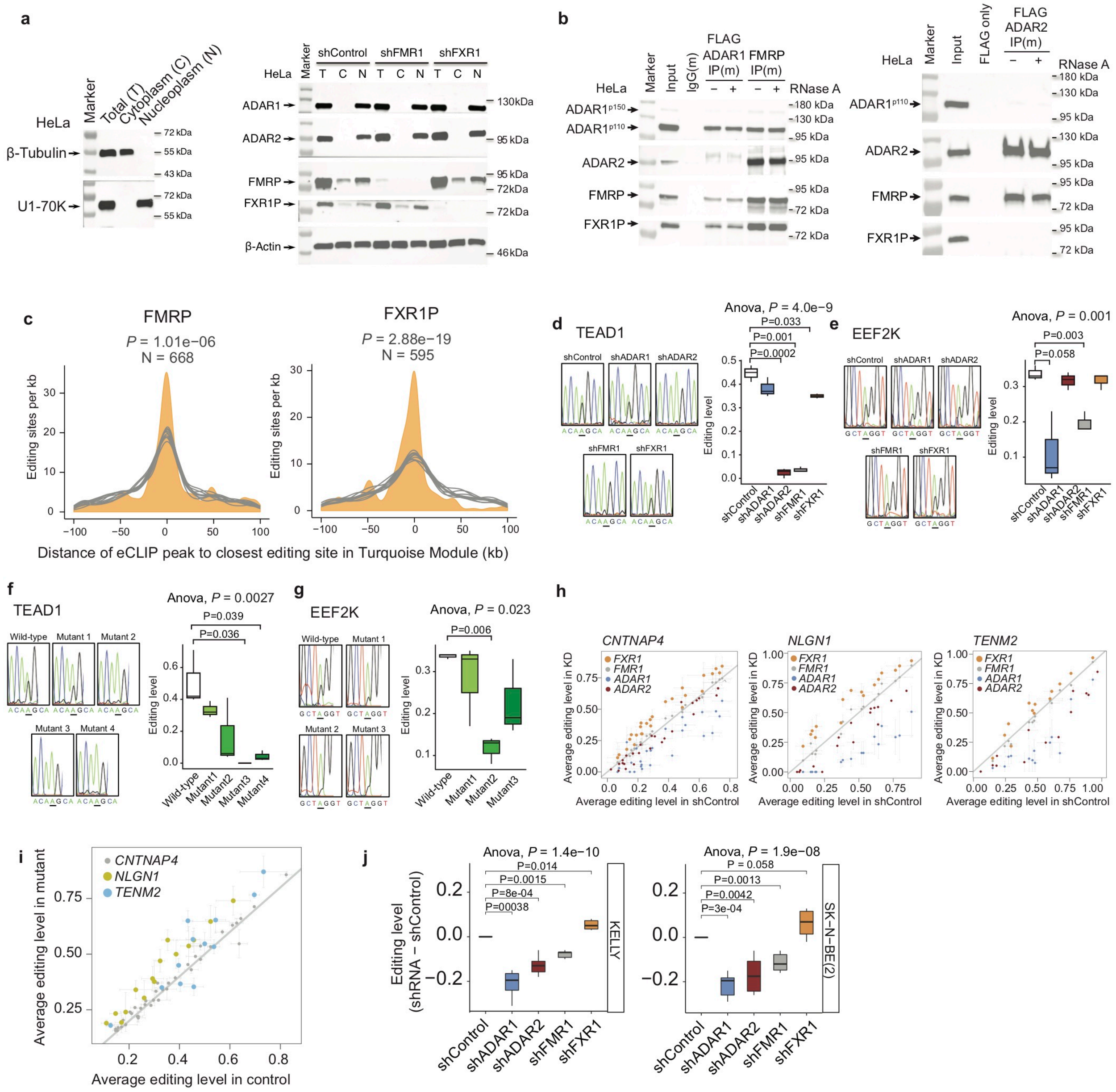


Figure 2.3 FMRP and FXR1P regulate RNA editing.

a, Western blot of ADAR1, ADAR2, FMRP and FXR1P proteins in the nuclear and cytoplasmic fractions of HeLa cells. Cell fractionation was confirmed by Western blotting of β-tubulin and U1-70K as marker proteins. Control cells and cells with stable knockdown of FMR1 or FXR1 were used. Experiment was repeated twice with similar results. **b**, Co-IP experiments with or without RNase A in HeLa cells between ADAR1, ADAR2, FMRP and FXR1P. Endogenous proteins were targeted except for ADAR2 (where a FLAG-tagged ADAR2 was expressed). Experiment was repeated 3 times with similar results. **c**, Shortest distance between FMRP or FXR1P eCLIP peaks and turquoise editing sites resulted from the WGCNA analysis (orange). Ten sets of random control sites (gray) are depicted for comparison (see Methods). Number of editing sites (N) is shown (see Methods for P value calculation). **d**, Experimental testing of an RNA editing site in the *TEAD1* gene for its dependence on ADAR1, ADAR2, FMRP or FXR1P. Control HeLa cells or cells with stable knockdown of one of these proteins were used to express a minigene that contains the editing site. Editing levels were measured by Sanger sequencing. Example sequencing traces are shown for each sample with the targeted editing site underlined. Boxplots include three biological replicates. Overall P value (shown above plot) was calculated by one-way ANOVA. Individual comparison P values were calculated by one-tailed Student’s t-test. **e**, Similar as **d**, for an editing site in the *EEF2K* gene. **f**. similar as **d**, but displaying editing levels of the TEAD1 editing site in minigenes with the wild-type sequence or different versions of mutants introduced to predicted FMRP binding motifs (see Supplementary Fig. 20). Wild-type HeLa cells were used to express these minigenes. **g**, Similar as **f**, for the editing site in the EEF2K gene (see Supplementary Fig. 20). **h**, RNA editing levels in control HeLa cells and cells with stable knockdown of ADAR1, ADAR2, FMR1 or FXR1. Hyper-editing sites in three genes were tested. Error bars represent standard errors of three biological replicates. **i**, Editing levels in control HeLa cells in the same hyper-edited genes as in **h**, but with mutations introduced at predicted FXR1 binding motifs (see Supplementary Fig. 23c-e). Error bars are standard errors of three biological replicates. **j**, Editing changes at six differential editing sites in ASD induced by control shRNA (shControl) or shRNA knockdown of ADAR1, ADAR2, FMR1, and FXR1 in two neuroblastoma cell lines, KELLY and SK-N-BE(2) (see Supplementary Fig. 24). Boxplots show editing changes against shRNA control over the six editing sites (n=6 editing sites). P-values calculated using two-tailed t-test. Boxplot definitions for d-g, j: center=median, lower hinge=25th percentile, upper hinge=75th percentile, min and max extend to observations at most 1.5 * IQR.

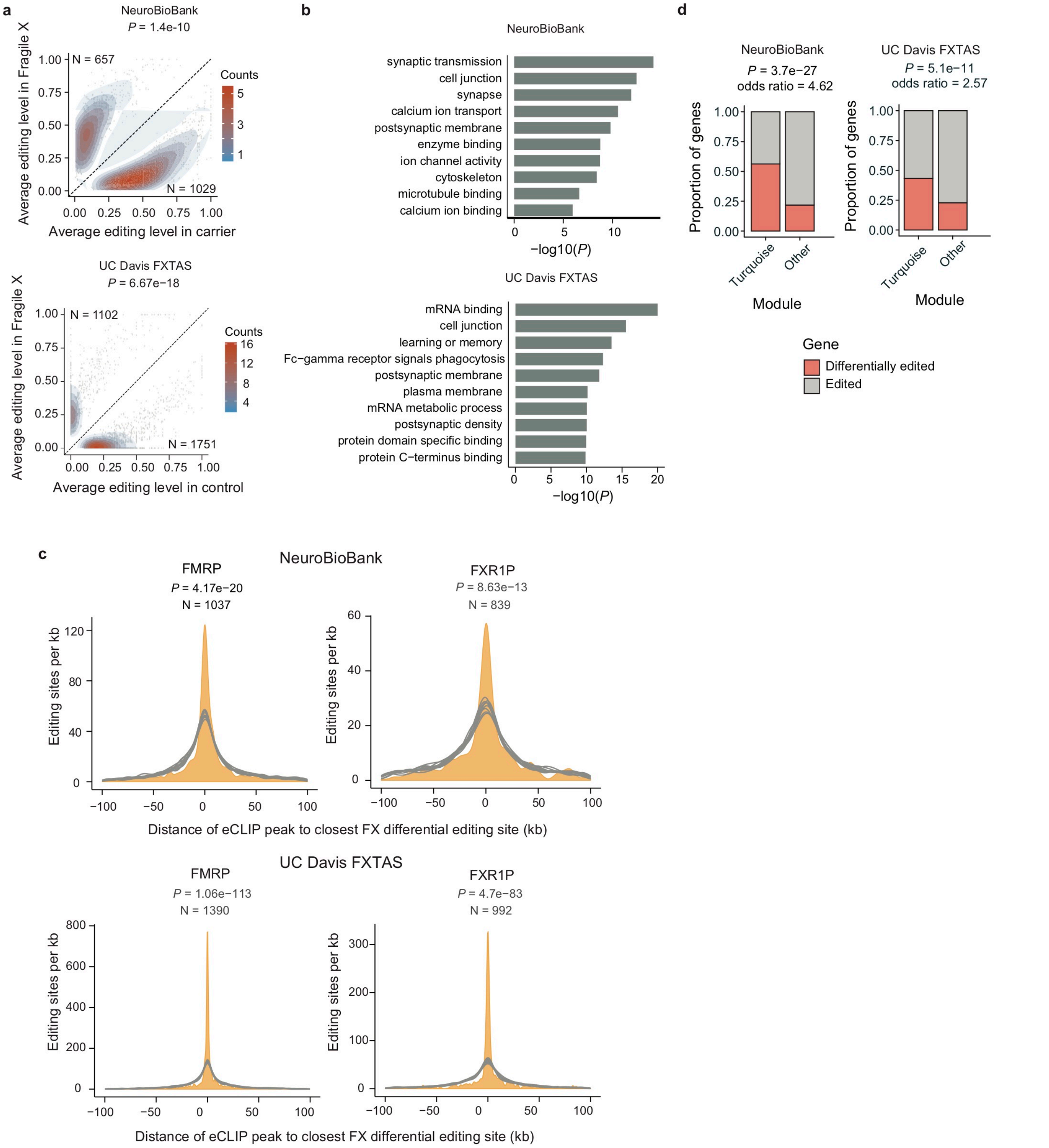

a
NeuroBioBank
P = 1.4e-10
N = 657
N = 1029
Counts
5
3
1
1.00
0.75
0.50
0.25
0.00
0.00 0.25 0.50 0.75 1.00
Average editing level in Fragile X
Average editing level in carrier
UC Davis FXTAS
P = 6.67e-18
N = 1102
N = 1751
Counts
16
12
8
4
Average editing level in Fragile X
Average editing level in control
b
NeuroBioBank
synaptic transmission
cell junction
synapse
calcium ion transport
postsynaptic membrane
enzyme binding
ion channel activity
cytoskeleton
microtubule binding
calcium ion binding
0 5 10
−log10(P)
UC Davis FXTAS
mRNA binding
cell junction
learning or memory
Fc−gamma receptor signals phagocytosis
postsynaptic membrane
plasma membrane
mRNA metabolic process
postsynaptic density
protein domain specific binding
protein C−terminus binding
0 5 10 15 20
−log10(P)
d
NeuroBioBank
P = 3.7e−27
odds ratio = 4.62
UC Davis FXTAS
P = 5.1e−11
odds ratio = 2.57
Proportion of genes
1.00
0.75
0.50
0.25
0.00
Turquoise
Other
Module
Gene
Differentially edited
Edited
c
NeuroBioBank
FMRP
P = 4.17e−20
N = 1037
FXR1P
P = 8.63e−13
N = 839
Editing sites per kb
120
80
40
0
60
40
20
0
−100 −50 0 50 100
Distance of eCLIP peak to closest FX differential editing site (kb)
UC Davis FXTAS
FMRP
P = 1.06e−113
N = 1390
FXR1P
P = 4.7e−83
N = 992
Editing sites per kb
800
600
400
200
0
300
200
100
0
−100 −50 0 50 100
Distance of eCLIP peak to closest FX differential editing site (kb)

Figure 2.4 Transcriptome-wide differential editing in the frontal cortex of Fragile X patients and controls.

a, Average editing levels of differential editing sites in Fragile X patients compared to carriers (left; NeuroBiobank dataset) or controls (right; UC Davis FXTAS dataset). Numbers of editing sites (N) that were up- or down-regulated in the patients are shown, which were compared via Chi-squared test (P value shown above plot). **b**, Gene ontology enrichment of genes harboring differential editing sites (n=961 and 1914 genes for Neurobiobank and UC Davis FXTAS respectively; one-tailed Gaussian test, see methods). **c**, Similar as Fig. **3c**, shortest distance between FMRP or FXR1P eCLIP peaks and differential editing sites in **a** (orange) (n=number of differential editing sites overlapping eCLIP genes; P-value from one-tailed Gaussian test, see methods). **d**, Overlap between the genes harboring WGCNA turquoise sites of ASD frontal cortex and those harboring differential editing sites in the Fragile X patients. P value was calculated via two-tailed Fisher's Exact test (N = 4051 and 7915 genes in Neurobiobank and UC Davis FXTAS respectively).

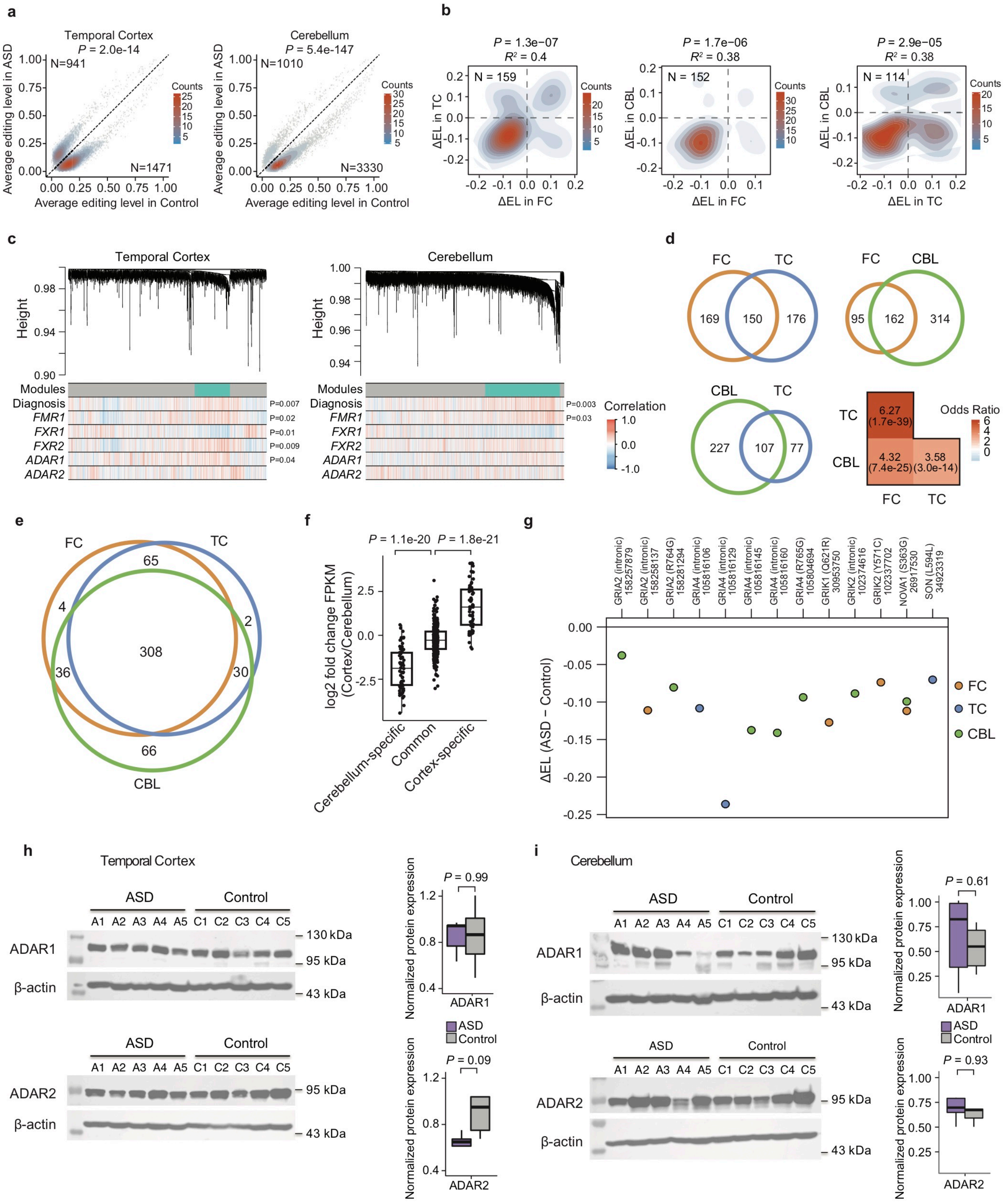

a
Temporal Cortex
P = 2.0e-14
N=941
N=1471
Average editing level in ASD
Average editing level in Control
Counts 25 20 15 10 5
Cerebellum
P = 5.4e-147
N=1010
N=3330
Counts 30 25 20 15 10 5
b
P = 1.3e−07
R² = 0.4
N = 159
ΔEL in TC
ΔEL in FC
P = 1.7e−06
R² = 0.38
N = 152
ΔEL in CBL
ΔEL in FC
P = 2.9e−05
R² = 0.38
N = 114
ΔEL in CBL
ΔEL in TC
c
Temporal Cortex
Cerebellum
Height
Modules
Diagnosis
FMR1
FXR1
FXR2
ADAR1
ADAR2
P=0.007
P=0.02
P=0.01
P=0.009
P=0.04
P=0.003
P=0.03
Correlation 1.0 0.0 -1.0
d
FC TC 169 150 176
FC CBL 95 162 314
CBL TC 227 107 77
Odds Ratio
TC 6.27 (1.7e-39)
CBL 4.32 (7.4e-25) 3.58 (3.0e-14)
FC TC
e
FC TC CBL
65 4 2 308 36 30 66
f
P = 1.1e-20 P = 1.8e-21
log2 fold change FPKM (Cortex/Cerebellum)
Cerebellum-specific
Common
Cortex-specific
g
GRIA2 (intronic) 158257879
GRIA2 (intronic) 158258137
GRIA2 (R764G) 158281294
GRIA4 (intronic) 105816106
GRIA4 (intronic) 105816129
GRIA4 (intronic) 105816145
GRIA4 (intronic) 105816160
GRIA4 (R765G) 105804694
GRIK1 (Q621R) 30953750
GRIK2 (intronic) 102374616
GRIK2 (Y571C) 102337702
NOVA1 (S363G) 26917530
SON (L594L) 34923319
ΔEL (ASD − Control)
FC
TC
CBL
h
Temporal Cortex
ASD Control
A1 A2 A3 A4 A5 C1 C2 C3 C4 C5
ADAR1
β-actin
ADAR2
130 kDa
95 kDa
43 kDa
P = 0.99
P = 0.09
Normalized protein expression
ADAR1
ADAR2
ASD
Control
i
Cerebellum
P = 0.61
P = 0.93

Figure 2.5 RNA editing dysregulation in different brain regions.

a, Average editing levels of differential editing sites in ASD and controls. Similar as Fig. **1c**, but data for temporal cortex and cerebellum are shown respectively. N=number differential editing sites, P-value calculated by Chi-square test. **b**, Changes in editing levels (ΔEL) between ASD and control (ASD-control) of differential editing sites shared across brain regions (N=number of differential editing sites shared). Abbreviations (same below): TC: temporal cortex; FC: frontal cortex; CBL: cerebellum. Pearson's correlation P and R^2 values are shown. **c**, Similar as Fig. **2d**, for WGCNA analysis of editomes in the temporal cortex and cerebellum regions. P-values calculated from linear regression (Methods). N = 46 and 47 samples in temporal cortex and cerebellum respectively. **d**, Overlap of editing sites in the turquoise modules of pairs of brain regions. Only editing sites with sufficient read coverage in both brain regions for WGCNA analysis are included. Odds ratios and P values (in parenthesis) for the overlaps are shown in the heatmap (two-tailed Fisher's Exact test). **e**, Overlap of genes harboring differential editing sites across brain regions. **f**, Relative FPKM values (log2 fold change) between cortex and cerebellum samples for genes that harbor differential editing sites only in cerebellum, only in cortex or in both types of regions. P values were calculated by two-tailed Student's t-test. N=66 cerebellum, 301 common, and 59 cortex specific genes respectively. **g**, Editing level difference (ΔEL, ASD-control) for a small number of literature-reported evolutionarily conserved RNA editing sites that showed differential editing between ASD and control groups in at least one brain region. **h**, Similar as Fig. **2b**, Western blot of ADAR1 and ADAR2 proteins in temporal cortex samples. Samples used in this experiment are shown in Supplementary Table 1. N=10 samples. **i**, Similar as **h**, Western blot of ADAR1 and ADAR2 proteins in cerebellum samples. N = 10 samples. Boxplot definitions: center=median, lower hinge=25th percentile, upper hinge=75th percentile, min and max extend to observations at most 1.5 * IQR.

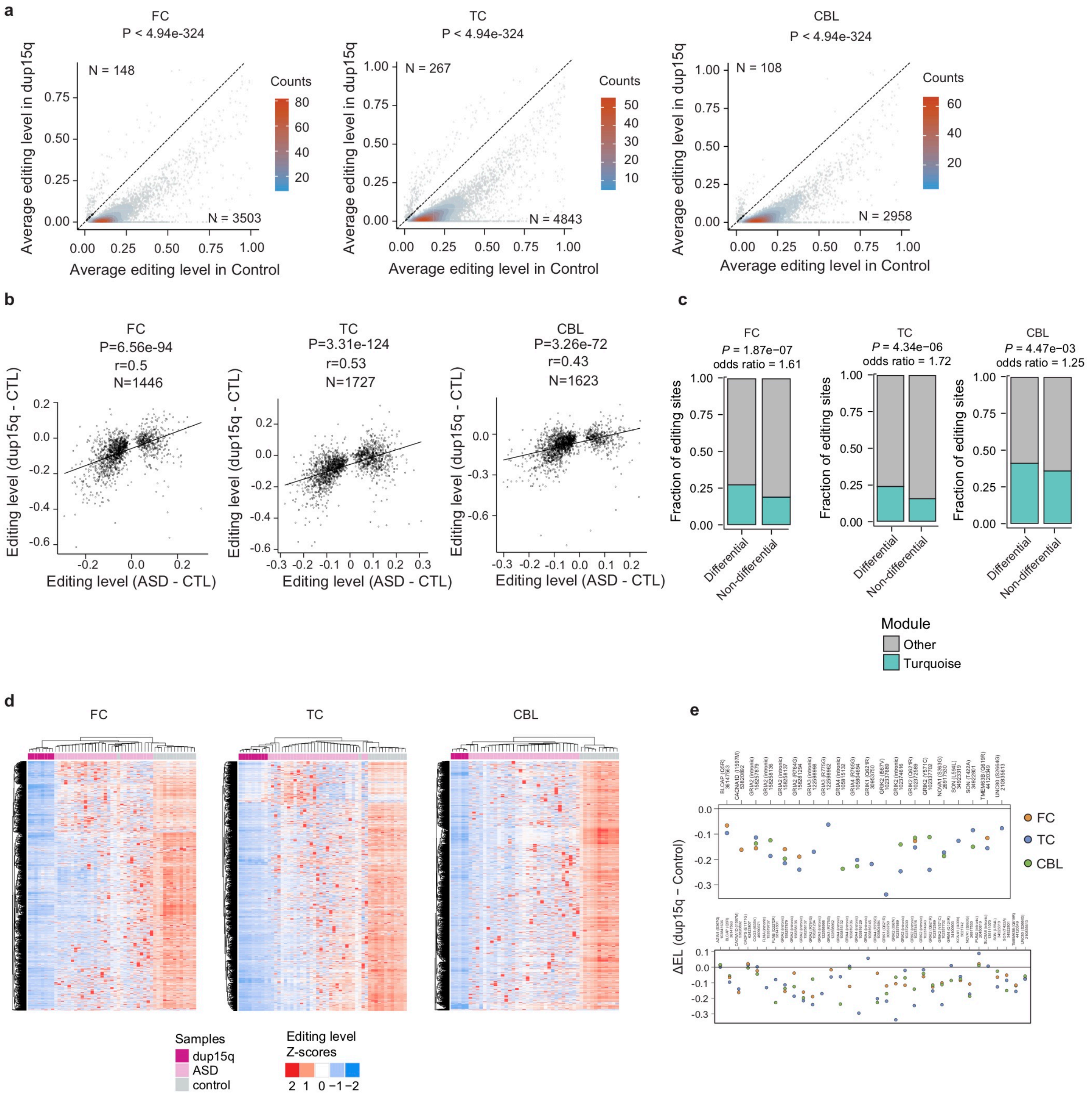


Figure 2.6 Hypo-editing in 3 brain regions of dup15q patients.

a, Contour scatterplot of differential editing sites in 3 brain regions of dup15q patients vs matched controls, similar as Fig. **1c**. N=number of differential editing sites hypo and hyperedited. P-value calculated using Chi-square test. **b**, Comparison of editing level difference in idiopathic ASD and dup15q patients relative to their respective controls (CTL). Pearson correlation P values and correlation coefficient (r) are shown. N represents the number of editing sites that are differential in idiopathic ASD and testable (see Methods) in the dup15q patients. **c**, Overlap between differential editing sites in dup15q patients and the turquoise modules of the respective brain regions of idiopathic ASD. P values were calculated via two-tailed Fisher's Exact test, n=3411, 2224, and 3834 in FC, TC, and CBL respectively. **d**, Heatmaps (similar as Fig. **1d**) of editing sites shown in **b**, including dup15q patients, matched controls, and the entire idiopathic ASD cohort. **e**, Editing level difference (ΔEL, dup15q-control) for a small number of literature-reported evolutionarily conserved RNA editing sites. Top panel: differential sites in at least one brain region of dup15q patients. Bottom panel: all testable editing sites.

2.7 References

1 Association, A. P. *Diagnostic and statistical manual of mental disorders (4th ed., text rev.)*. 4th edn, (2000).

2 Rojas, D. C. The role of glutamate and its receptors in autism and the use of glutamate receptor antagonists in treatment. *Journal of neural transmission (Vienna, Austria : 1996)* **121**, 891-905, doi:10.1007/s00702-014-1216-0 (2014).

3 Guo, Y. P. & Commons, K. G. Serotonin neuron abnormalities in the BTBR mouse model of autism. *Autism research : official journal of the International Society for Autism Research* **10**, 66-77, doi:10.1002/aur.1665 (2017).

4 Ha, S., Sohn, I. J., Kim, N., Sim, H. J. & Cheon, K. A. Characteristics of Brains in Autism Spectrum Disorder: Structure, Function and Connectivity across the Lifespan. *Experimental neurobiology* **24**, 273-284, doi:10.5607/en.2015.24.4.273 (2015).

5 Nelson, S. B. & Valakh, V. Excitatory/Inhibitory Balance and Circuit Homeostasis in Autism Spectrum Disorders. *Neuron* **87**, 684-698, doi:10.1016/j.neuron.2015.07.033 (2015).

6 Marchetto, M. C. *et al.* Altered proliferation and networks in neural cells derived from idiopathic autistic individuals. *Mol Psychiatry*, doi:10.1038/mp.2016.95 (2016).

7 de la Torre-Ubieta, L., Won, H., Stein, J. L. & Geschwind, D. H. Advancing the understanding of autism disease mechanisms through genetics. *Nature medicine* **22**, 345-361 (2016).

8 Gupta, S. *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat Commun* **5**, 5748, doi:10.1038/ncomms6748 (2014).

9 Parikshak, N. N. *et al.* Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* **540**, 423-427, doi:10.1038/nature20612 (2016).

10 Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**, 1511-1523, doi:10.1016/j.cell.2014.11.035 (2014).

11 Wu, Y. E., Parikshak, N. N., Belgard, T. G. & Geschwind, D. H. Genome-wide, integrative analysis implicates microRNA dysregulation in autism spectrum disorder. *Nat Neurosci*, doi:10.1038/nn.4373 (2016).

12 Nishikura, K. A-to-I editing of coding and non-coding RNAs by ADARs. *Nature reviews. Molecular cell biology* **17**, 83-96, doi:10.1038/nrm.2015.4 (2016).

13 Behm, M. & Ohman, M. RNA Editing: A Contributor to Neuronal Dynamics in the Mammalian Brain. *Trends in genetics : TIG* **32**, 165-175, doi:10.1016/j.tig.2015.12.005 (2016).

14 Slotkin, W. & Nishikura, K. Adenosine-to-inosine RNA editing and human disease. *Genome medicine* **5**, 105, doi:10.1186/gm508 (2013).

15 Khermesh, K. *et al.* Reduced levels of protein recoding by A-to-I RNA editing in Alzheimer's disease. *RNA (New York, N.Y.)* **22**, 290-302, doi:10.1261/rna.054627.115 (2016).

16 Eran, A. *et al.* Comparative RNA editing in autistic and neurotypical cerebella. *Mol Psychiatry* **18**, 1041-1048, doi:10.1038/mp.2012.118 (2013).

17 Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-299, doi:10.1016/j.neuron.2012.04.009 (2012).

18 Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008-1021, doi:10.1016/j.cell.2013.10.031 (2013).

19 Bahn, J. H. *et al.* Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome research* **22**, 142-150, doi:10.1101/gr.124107.111 (2012).

20 Porath, H. T., Carmi, S. & Levanon, E. Y. A genome-wide map of hyper-edited RNA reveals numerous new sites. *Nat Commun* **5**, 4726, doi:10.1038/ncomms5726 (2014).

21 Picardi, E., D'Erchia, A. M., Lo Giudice, C. & Pesole, G. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic acids research* **45**, D750-d757, doi:10.1093/nar/gkw767 (2017).

22 Levanon, E. Y. *et al.* Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nature biotechnology* **22**, 1001-1005, doi:10.1038/nbt996 (2004).

23 Oakes, E., Anderson, A., Cohen-Gadol, A. & Hundley, H. A. Adenosine Deaminase That Acts on RNA 3 (ADAR3) Binding to Glutamate Receptor Subunit B Pre-mRNA Inhibits RNA Editing in Glioblastoma. *The Journal of biological chemistry* **292**, 4326-4335, doi:10.1074/jbc.M117.779868 (2017).

24 Rakic, P. Evolution of the neocortex: a perspective from developmental biology. *Nature reviews. Neuroscience* **10**, 724-735, doi:10.1038/nrn2719 (2009).

25 Abrahams, B. S. *et al.* SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Molecular autism* **4**, 36, doi:10.1186/2040-2392-4-36 (2013).

26 Hwang, T. *et al.* Dynamic regulation of RNA editing in human brain development and disease. *Nat Neurosci* **19**, 1093-1099, doi:10.1038/nn.4337 (2016).

27 Liu, X. *et al.* Disruption of an Evolutionarily Novel Synaptic Expression Pattern in Autism. *PLoS biology* **14**, e1002558, doi:10.1371/journal.pbio.1002558 (2016).

28 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **9**, 559, doi:10.1186/1471-2105-9-559 (2008).

29 Davis, J. K. & Broadie, K. Multifarious Functions of the Fragile X Mental Retardation Protein. *Trends in genetics : TIG*, doi:10.1016/j.tig.2017.07.008 (2017).

30 Darnell, J. C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247-261, doi:10.1016/j.cell.2011.06.013 (2011).

31 Zhang, Y. *et al.* The fragile X mental retardation syndrome protein interacts with novel homologs FXR1 and FXR2. *The EMBO journal* **14**, 5358-5366 (1995).

32 Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature methods* **13**, 508-514, doi:10.1038/nmeth.3810 (2016).

33 Ascano, M., Jr. *et al.* FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature* **492**, 382-386, doi:10.1038/nature11737 (2012).

34 Vasudevan, S. & Steitz, J. A. AU-rich-element-mediated upregulation of translation by FXR1 and Argonaute 2. *Cell* **128**, 1105-1118, doi:10.1016/j.cell.2007.01.038 (2007).

35 Van Nostrand, E. L. *et al.* A Large-Scale Binding and Functional Map of Human RNA Binding Proteins. *bioRxiv*, doi:10.1101/179648 (2017).

36 Hagerman, R., Hoem, G. & Hagerman, P. Fragile X and autism: Intertwined at the molecular level leading to targeted treatments. *Molecular autism* **1**, 12, doi:10.1186/2040-2392-1-12 (2010).

37 Abbeduto, L., McDuffie, A. & Thurman, A. J. The fragile X syndrome-autism comorbidity: what do we really know? *Front Genet* **5**, 355, doi:10.3389/fgene.2014.00355 (2014).

38 Pinto, Y., Cohen, H. Y. & Levanon, E. Y. Mammalian conserved ADAR targets comprise only a small fragment of the human editosome. *Genome biology* **15**, R5, doi:10.1186/gb-2014-15-1-r5 (2014).

39 Irimia, M. *et al.* Evolutionarily conserved A-to-I editing increases protein stability of the alternative splicing factor Nova1. *RNA biology* **9**, 12-21, doi:10.4161/rna.9.1.18387 (2012).

40 DiStefano, C. *et al.* Identification of a distinct developmental and behavioral profile in children with Dup15q syndrome. *Journal of neurodevelopmental disorders* **8**, 19, doi:10.1186/s11689-016-9152-y (2016).

41 Battaglia, A. *et al.* The inv dup(15) syndrome: a clinically recognizable syndrome with altered behavior, mental retardation, and epilepsy. *Neurology* **48**, 1081-1086 (1997).

42 Frith, C. & Dolan, R. The role of the prefrontal cortex in higher cognitive functions. *Brain research. Cognitive brain research* **5**, 175-181 (1996).

43 Jansen, A. *et al.* Gene-set analysis shows association between FMRP targets and autism spectrum disorder. *European journal of human genetics : EJHG* **25**, 863-868, doi:10.1038/ejhg.2017.55 (2017).

44 Pinto, D. *et al.* Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *American journal of human genetics* **94**, 677-694, doi:10.1016/j.ajhg.2014.03.018 (2014).

45 Fatemi, S. H. & Folsom, T. D. Dysregulation of fragile x mental retardation protein and metabotropic glutamate receptor 5 in superior frontal cortex of individuals with autism: a postmortem brain study. *Molecular autism* **2**, 6, doi:10.1186/2040-2392-2-6 (2011).

46 Patzlaff, N. E., Nemec, K. M., Malone, S. G., Li, Y. & Zhao, X. Fragile X related protein 1 (FXR1P) regulates proliferation of adult neural stem cells. *Human molecular genetics* **26**, 1340-1352, doi:10.1093/hmg/ddx034 (2017).

47 Charman, T. *et al.* IQ in children with autism spectrum disorders: data from the Special Needs and Autism Project (SNAP). *Psychological medicine* **41**, 619-627, doi:10.1017/s0033291710000991 (2011).

48 Finucane, B. M. *et al.* in *GeneReviews(R)* (eds M. P. Adam *et al.*) (University of Washington, Seattle

University of Washington, Seattle. GeneReviews is a registered trademark of the University of Washington, Seattle. All rights reserved., 1993).

49 Ahn, J. & Xiao, X. RASER: reads aligner for SNPs and editing sites of RNA. *Bioinformatics (Oxford, England)* **31**, 3906-3913, doi:10.1093/bioinformatics/btv505 (2015).

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

51 Feldmeyer, D. *et al.* Neurological dysfunctions in mice expressing different levels of the Q/R site-unedited AMPAR subunit GluR-B. *Nat Neurosci* **2**, 57-64, doi:10.1038/4561 (1999).

52 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164, doi:10.1093/nar/gkq603 (2010).

53 Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize Implements and enhances circular visualization in R. *Bioinformatics (Oxford, England)* **30**, 2811-2812, doi:10.1093/bioinformatics/btu393 (2014).

54 Josse, J. & Husson, F. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. *2016* **70**, 31, doi:10.18637/jss.v070.i01 (2016).

55 Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics (Oxford, England)* **24**, 719-720, doi:10.1093/bioinformatics/btm563 (2008).

56 Langfelder, P. & Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *BMC systems biology* **1**, 54, doi:10.1186/1752-0509-1-54 (2007).

57 Wheeler, E. C., Van Nostrand, E. L. & Yeo, G. W. Advances and challenges in the detection of transcriptome-wide protein-RNA interactions. *Wiley interdisciplinary reviews. RNA*, doi:10.1002/wrna.1436 (2017).

58 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).

59 Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic acids research* **43**, W39-49, doi:10.1093/nar/gkv416 (2015).

60 Aken, B. L. *et al.* The Ensembl gene annotation system. *Database : the journal of biological databases and curation* **2016**, doi:10.1093/database/baw093 (2016).

61 Nojima, T., Gomes, T., Carmo-Fonseca, M. & Proudfoot, N. J. Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide. *Nature protocols* **11**, 413-428, doi:10.1038/nprot.2016.012 (2016).

62 Sundararaman, B. *et al.* Resources for the Comprehensive Discovery of Functional RNA Elements. *Mol Cell* **61**, 903-913, doi:10.1016/j.molcel.2016.02.012 (2016).

63 Bahn, J. H. *et al.* Genomic analysis of ADAR1 binding and its involvement in multiple RNA processing pathways. *Nat Commun* **6**, 6355, doi:10.1038/ncomms7355 (2015).

Chapter 3

Patient derived cortical spheroids uncover dysregulated RNA editing over fetal progression of Autism spectrum disorder

3.1 Introduction

Autism spectrum disorder (ASD) encompasses neurodevelopmental deficits in communicative skills, reciprocal interactions, and repetitive and stereotyped behaviors[1]. The diagnosis of ASD happens as early as infancy usually via clinical observation aided by diagnostic tests such as the Autism Diagnostic Observation Schedule[2] or Autism Diagnostic Interview-Revised[3]. The pathogenesis of ASD, however, likely stems earlier at fetal development. Multiple studies of early autism pathogenesis, including iPSC derived neurons and mouse models, have observed aberrations in morphology, electrophysiology, and molecular processes[4,5]. Unfortunately, characterization of pathogenic alterations in actual fetal brain is impossible since clinical diagnosis cannot be made that early.

The invention of organoids has provided powerful in vitro models to study early ASD brain development[6]. Organoids are organ-like aggregates originating from stem cells that can form 3-dimensional cellular structures that recapitulate organ structure, physiology, and cellular organization and composition[6]. A few studies have modelled early autism using organoids of the cerebral cortex[6]. However, they had small sample size (e.g. < 5 disease samples)[6] and limited protocol reproducibility[7].

Recently, a differentiation method for generating 3D cerebral organoids, called cortical spheroids, was shown to have strong reproducibility across multiple donors and time points[8]. Leveraging this method, the Geschwind and Pasca labs have generated RNA-sequencing (RNA-seq) data of 553 organoids across multiple time points encompassing both idiopathic ASD and 7 syndromic ASD with known genetic mutations (unpublished work). Being the first large-scale organoid study of ASD, this dataset enables elucidation of the transcriptomic landscape across fetal progression of core ASD pathology. Our focus here is to examine the profiles and dynamic changes of RNA editing in this dataset.

Our group recently found widespread dysregulation of RNA editing in a large cohort of postmortem brains of autistic individuals[9]. Genes harboring differential editing sites in ASD were significantly enriched in functional categories related to synaptic maintenance and transmission[9]. Interestingly, many of these differential editing sites demonstrated substantial increase in their editing levels during fetal to infant transition[9]. Previous transcriptome studies identified modules of genes that exclusively express during fetal development and harbor enrichment of autism-related rare genetic variants[10,11], suggesting the importance of distinct transcriptomic changes in fetal progression of ASD. Based on the above findings, we hypothesize that RNA editing is significantly altered over fetal development in ASD, which is distinct from the dysregulation observed in postnatal brain. Therefore, in this study we analyzed the above organoid dataset to characterize the RNA editing landscape over modelled fetal brain development. We uncovered significant dysregulated trends across idiopathic and syndromic autism disorders. We observed that RNA editing levels globally increase over cortical spheroid development and are hypoedited in ASD. Functionally these hypoedited sites likely affect cellular development and proliferation specific to radial glia, neural progenitors, and early born neurons.

3.2 Results

3.2.1 Overview of the organoid dataset

RNA-seq data were generated as part of a larger study examining aberrations in neurobiology, morphology, and transcriptomics of cortical spheroids modelling autistic cortex (unpublished work). Specifically, a directed differentiation method[12,13] was applied to generate 553 cortical spheroid datasets from neurotypical controls and patients with ASD encompassing both idiopathic ASD and syndromic ASD in 8 susceptibility loci including 15q13 duplication, 16p11 deletion, 16p11 duplication, 22q11 deletion, 22q13 deletion, mutations in PCDH19, SHANK3, and Timothy Syndrome (Supplemental Table 1). Spheroids were differentiated for 25, 50, 75, or 100 days, followed by RNA-seq (rRNA depleted, strand specific, paired-end) with an average of 25.5 (95% CI, 8.4-39) million read pairs. An average of 16.4 million (95% CI, 6.3-27.8) read pairs were uniquely mapped to the human genome (Figure 1a).

Potential confounding covariates, such as batch, sex, and the first two principal components summarizing sequencing biases and mapping quality (PCSeq1 and

PCSeq2, Methods) spread roughly evenly across ASD and control groups, permitting downstream covariate adjustment (Figure 1b-c). Additionally, RNA integrity number measured from a subset of samples was consistently high (median 8.95; 95% CI, 7.8-9.7; Figure 1d), indicating that RNA quality remained high throughout differentiation and RNA extraction.

3.2.2 RNA editing sites in cortical spheroids

Next we applied our previously developed bioinformatic methods[14,15] to globally identify RNA editing sites in the cortical spheroid samples. Similar to other studies[9,16], the number of detectable editing sites per sample (range: 25,00-290,000 editing sites) varied depending on sequencing depth (Figure 2a). On average, 97% (95% CI, 95-98%) of the detected RNA editing sites were consistent with the canonical A-to-G and C-to-T editing types, indicating high detection specificity (Figure 2b). A-to-G constituted the majority of all detected editing sites, consistent with the known rarity of C-to-U editing in human[17] (Figure 2b). On average, 76% of editing sites fell in Alu regions (Figure 2c; 95% CI, 66-83%), commensurate with the known binding preference of ADAR proteins for Alu structures[18].

3.2.3 Hypoediting in ASD across cortical spheroid development

In a previous study we found that differential RNA editing sites in ASD were enriched in a subset that dramatically increase during the transition from fetal to infant stages[9]. However, the actual developmental trajectory of RNA editing across ASD fetal development remains unknown. The span of differentiation time points of our cortical spheroids enabled us to model this and concomitantly investigate aberrations in the developmental trajectory in ASD. Over fetal development, RNA editing could have multiple modules that project towards distinct trajectories. Thus, we first searched for modules of RNA editing in our cortical spheroids using weighted gene co-expression analysis (WGCNA). Interestingly, almost all RNA editing sites clustered into a single module (labelled as the turquoise module) (Figure 3, except a small blue module reflecting confounding variables). This observation indicates that RNA editing globally follows a single trajectory over organoid development. Using the eigengene (1st principal component) to view the projection of the turquoise module over differentiation points, we found that RNA editing generally increased over organoid development (Figure 4a). Interestingly, RNA editing of the ASD cohort followed the same increasing trajectory, but exhibited lower editing levels compared to the controls (i.e., hypoediting) across all time points (Figure 4a). The same trends also appeared in many of the individual ASD genetic susceptibility mutations (Figure 4b), suggesting that hypoediting is likely implicated in both core ASD etiology and distinct syndromic phenotypes.

3.2.4 Potential RNA editing regulators in cortical spheroid development

To investigate potential regulatory mechanisms of the turquoise module, we compiled from previous studies a list of RNA binding proteins (RBPs) that regulate RNA editing[16,19]. We hypothesized that potential regulators would show strong correlations between their gene expression and the 1st principal component of the turquoise module. Interestingly, we identified many RBPs with strong positive and negative correlations (Figure 5). Two members of the ADAR family, *ADAR2* and *ADAR3* showed positive correlations (Figure 5). *FXR1* displayed strong negative correlation, consistent with its previously discovered role suppressing RNA editing in postnatal Autism brain[9] (Figure 5). Out of all the candidate RBPs, *ILF3*, a repressor of editing[16], had the strongest (negative) correlation (Figure 5). Overall, our results suggest that RNA binding proteins may have prominent roles in regulating RNA editing over cortical spheroid development.

3.2.5 Hypoediting affects neuronal propagation

To investigate the possible functional roles of hypoediting in ASD fetal development, we first identified differential editing sites between ASD and control cohorts within each differentiation time point. We leveraged a mixed effects model that could adjust for technical and biological covariates (e.g. batch and sex) and handle the hierarchical dependencies between samples (e.g. multiple samples derived from differentiations from the same donor) (Methods). We found 495, 343, 659, and 459 differential editing sites at day 25, 50, 75 and 100 respectively (FDR < 0.1 and difference in editing ≥ 5%). Consistent with the WGCNA analysis, the differential editing sites displayed a hypoediting bias at all time points (Figure 6a).

Next, we examined differential editing sites for cell type enrichment and found that they resided over abundantly within genes specific to dividing radial glia, intermediate progenitor cells, and newborn neurons (Figure 6b). Gene ontology analysis found that the differential editing sites were located in genes involved in transcriptional regulation and processing and cell growth (Figure 6c). These results indicate that hypoediting is likely involved in cellular development and proliferation of maturing neurons.

3.3 Discussion

Transcriptomic studies in postmortem autistic brains have vastly improved our understanding of the molecular etiology of ASD. However, because the diagnosis of ASD occurs at early infancy, the study of transcriptomics in autistic brain is limited to postnatal time periods. In this study, we overcame this issue by modeling cortical development of both idiopathic and multiple syndromic forms of ASD using cortical spheroids. In particular, we examined RNA editing over hundreds of cortical spheroids spanning a comprehensive developmental range. RNA editing globally increased over spheroid development, and was hypoedited in the ASD cohort, specifically affecting genes relevant to radial glia, neuroprogenitor cells, and newborn neurons in functional categories related to cellular proliferation and transcription regulation.

Interestingly, the functional pathways targeted by dysregulated RNA editing differs between fetal and postnatal ages. RNA editing in postnatal human brain targeted neuronal transmission genes[9] whereas, in cortical spheroids, it primarily targeted cellular development and proliferation genes. Interestingly, these results are consistent with studies of gene expression in prenatal vs postnatal ASD. Dysregulated genes[10,11] and ASD susceptibility genetic mutations[20,21] consistently converge onto either transcriptional regulation genes or synaptic genes; the transcriptional regulation genes express specifically during fetal development, whereas, the synaptic genes express primarily postnatally. Thus, the changing pathways harboring dysregulated RNA editing could serve a convergent purpose with the changing function of dysregulated gene expression modules over ASD disease progression.

3.4 Methods

3.4.1 Cortical spheroid generation

Skin or blood samples were obtained from 45 ASD and 18 control human subjects and transformed using a previous developed differentiation method for specifying 3D cortical spheroids[12,13]. The 45 ASD patients and 18 control individuals begot 62 ASD and 21 CTL cell lines which were differentiated across 86 ASD and 53 CTL inductions. The multiple inductions were grown and harvested across 25, 50, 75, and 100 days to produce 82 ASD and 52 CTL at day 25, 81 ASD and 52 CTL at day 50, 80 ASD and 52 CTL at day 75, and 80 ASD and 49 CTL at day 100 for a combined total of 323 ASD and 205 CTL cortical spheroids. The avoid batch effects we sequenced some of the cortical spheroids across multiple batches, generating 82 ASD and 52 CTL RNA-sequencing datasets at day 25, 81 ASD and 53 CTL at day 50, 90 ASD and 64 CTL at day 75, and 80 ASD and 51 CTL at day 100 for a combined total of 333 ASD and 220 CTL cortical spheroid RNA-sequencing datasets.

3.4.2 High throughput sequencing

Cortical spheroid RNA was harvested and RNA sequenced using rRNA depletion, paired end, 2nd read sense, 101 nucleotide long reads to an average coverage of 25.5 million read pairs. Reads were aligned with Hisat2[22] to the hg19 genome and transcriptome from Ensembl gene annotations. Picard tools was run on mapped read files to obtain sources of mapping quality and sequencing bias including duplicate reads, GC dropout, AT dropout, 5' to 3' bias, high quality aligned reads, and ribosomal, intronic, intergenic, coding, and UTR bases.

These metrics were summarized using the first two principal components henceforth called PCSeq1 and PCSeq2.

3.4.3 Identifying RNA editing sites

To our knowledge, no studies have identified RNA editing sites in organoids of human brain. Therefore, we performed de novo RNA editing detection using our previous methods[14,15]. First, to rescue reads that were unmappable due to harboring too many editing sites[23], we re-mapped unmapped reads onto an hg19 genome with all adenosines converted to guanosines. Then, RNA editing sites were identified as mismatches between mapped reads and the human reference genome. A log-likelihood test and posterior filters were then applied to remove sequencing errors and technical artifacts caused by difficult to map genomic regions. The RNA editing level per sample was estimated as the number of reads harboring the edited allele divided by total reads covering the site. Editing level was only estimated for sites having at least 5 reads coverage.

3.4.4 Calculating gene expression

Gene expressions measured by transcripts per million (tpm) were calculated from in-house scripts using the exon union of RefSeq transcripts obtained from the UCSC genome browser. TPMs were adjusted for potential confounding variables (e.g. sex, batch, and PCseq1-2) using a hierarchical linear mixed effects regression modelling confounding variables as fixed effects and cell line, induction, and individual donor as nested random effects. Specifically we implemented the model using the R package lme4 as lmer(tpm ~ sex + disease condition + differentiation day + batch2 + batch3 + batch3 redo + batch4 + PCseq1 + PCseq2 + (1|individual/cell line/induction)). P-values for differential gene expression between ASD and CTL samples were obtained by fitting tpm to a reduced model without disease condition, and testing significance of the maximum likelihood fit of the full model against the reduced model.

3.4.5 Detection differential RNA editing sites

Per differential time point (25, 50, 75, and 100 days), we identified differential editing sites as those with significant differences in editing level between ASD and control samples. To account for confounding variables (e.g. sex, batch) and to handle the hierarchical design of organoid generation (i.e. multiple cell lines and inductions from same individual), we used the lme4 package[24] in R to fit a linear mixed effect model to editing level against sex, disease condition, batch, and PCseq metrics from Picard tools. The command was: lmer(editing levels ~ sex + disease condition + batch2 + batch3 + batch3_redo + batch4 + PCseq1 + PCseq2 + (1|individual/cell line/induction)), where all fixed variables except PCseq1 and PCseq2 were encoded as 0 or 1. If the data was insufficient to fit the nested random effect term, nested levels were iteratively removed until convergence was attained. To follow the 10:1 rule, we filtered out editing sites that had fewer than 80 samples with at least 5 read coverage, leaving 27201, 27521, 30958, and 27109 editing sites for 25, 50, 75, and 100 days respectively. P-values were obtained by fitting a reduced linear mixed effects model without disease condition and then testing significance of the maximum likelihood fit of the full model against the reduced model. Editing sites were considered differential if Benjamin Hochberg adjusted FDR < 0.1 and average difference in editing level between ASD and controls > 0.05.

3.4.6 Weighted gene co-expression network analysis for RNA editing sites

To find modules of RNA editing sites over organoid development, we performed weighted gene co-expression network analysis (WGCNA)[25]. WGCNA creates modules through hierarchical clustering of topological overlap metrics, which are based on correlations between editing sites. To suppress confounders (e.g. sex, batch) on module formation, we first adjusted RNA editing levels using the lme4 package[24] in R to run mixed effects linear regression with sex, batch, ASD, differentiation day, PCSeq1, and PCSeq2 as fixed effects and individual, cell line, and induction as nested random effects (Supplemental Table 1). To ensure calculation of all pairwise correlations with at least 20 samples, we next filtered out editing sites that had fewer than 300 samples with at least 5 read coverage and nonzero editing. We also ran WGCNA goodSamplesGenes function to remove samples with too many missing editing level estimates. These filtering steps left 540 samples and 5,770 editing sites for further WGCNA steps. We used a soft power threshold of 10 to achieve scale free topology and spearman correlation to calculate topological overlap. To prevent confounding modules with outlier samples, we applied a robust WGCNA methodological as described in previous work[26], running 200 randomizations over all covariates except ASD condition and differentiation day. The final topological overlap matrix was hierarchically clustered and cut using various tree cutting parameters, all of which gave similar module partitions (Figure 3). The final module clusterings found only 233 (4%) out of the 5770 editing sites confounded by covariates (Figure 3), indicating that our adjustment procedures effectively removed most technical artifacts from the RNA editing data.

3.4.7 Significance of hypoediting trajectory in ASD

To test significance in ASD of the observed hypoedited trajectory over cortical spheroid development, we randomized the ASD and CTL labels over 10,000 permutations and shifted the minimum value of the 1st principal component of the turquoise module to zero. For each permutation, loess curves were fitted over development time from 25 days to 100 days. The difference in area under the loess curves for nonrandomized CTL samples versus ASD samples was used as the test statistic. The p-value was then calculated against a null distribution of the differences in areas from the 10,000 permutations.

3.4.8 Enrichment analyses

Gene ontology enrichment of differential editing sites was performed as described in our previous work[9]. Briefly, genes harboring one or more differential editing sites were agglomerated into a unique query list. Each gene in the list was matched by a random gene that had same gene length ($\pm 10\%$) and same tpm averaged across all samples ($\pm 10\%$). This randomization process was performed for 10,000 permutations, and the number of random genes having a given GO term was fit to a Gaussian distribution. The number of query genes having the GO term was tested for enrichment using a one-tailed Gaussian test.

Cell type enrichment of differential editing sites was performed similarly as gene ontology enrichment, except that cell type marker genes were used instead of GO terms. We ascertained the cell type marker genes from a previous single cell sequencing study of fetal brain[27].

3.5 Figures

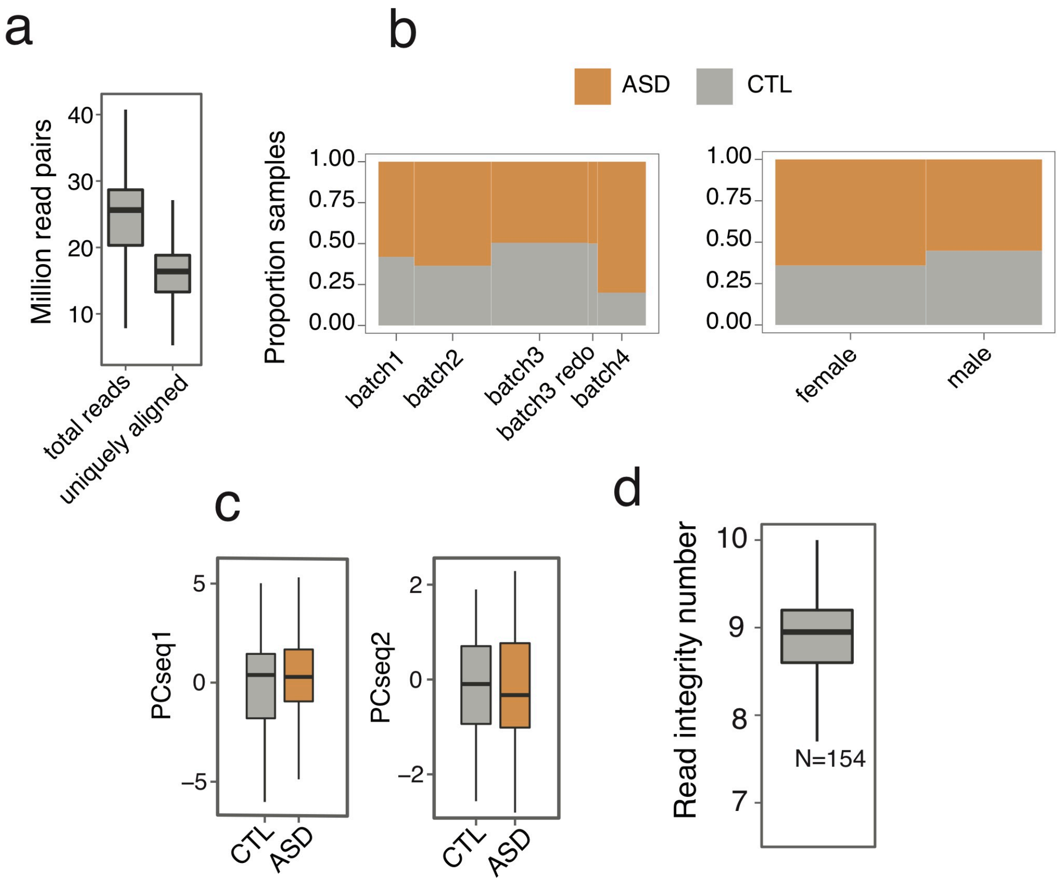


Figure 3.1 Overview of cortical spheroid data

a) Boxplots summarizing the sequencing coverage over all samples (n=553) in terms of total read coverage and number of uniquely aligned reads. b) Distribution plots of batch (left) and sex (right) over ASD (autism) and CTL (control) samples. The height of bars shows proportion of samples. Width of bars is proportional to the number of samples out of total (n=553). c) Boxplots summarizing distribution of PCseq1 and PCseq2 over CTL and ASD samples. PCseq1 and PCseq2 are the first two principal components of metrics from Picard tools summarizing mapping bias and sequencing quality (Methods). d) Summary of read integrity number measured for a subset of samples (n=154) shows that RNA quality is high for all samples. Boxplot definition: center, median; lower hinge, 25th percentile; upper hinge, 75th percentile; minimum and maximum extend to at most 1.5 X IQR.

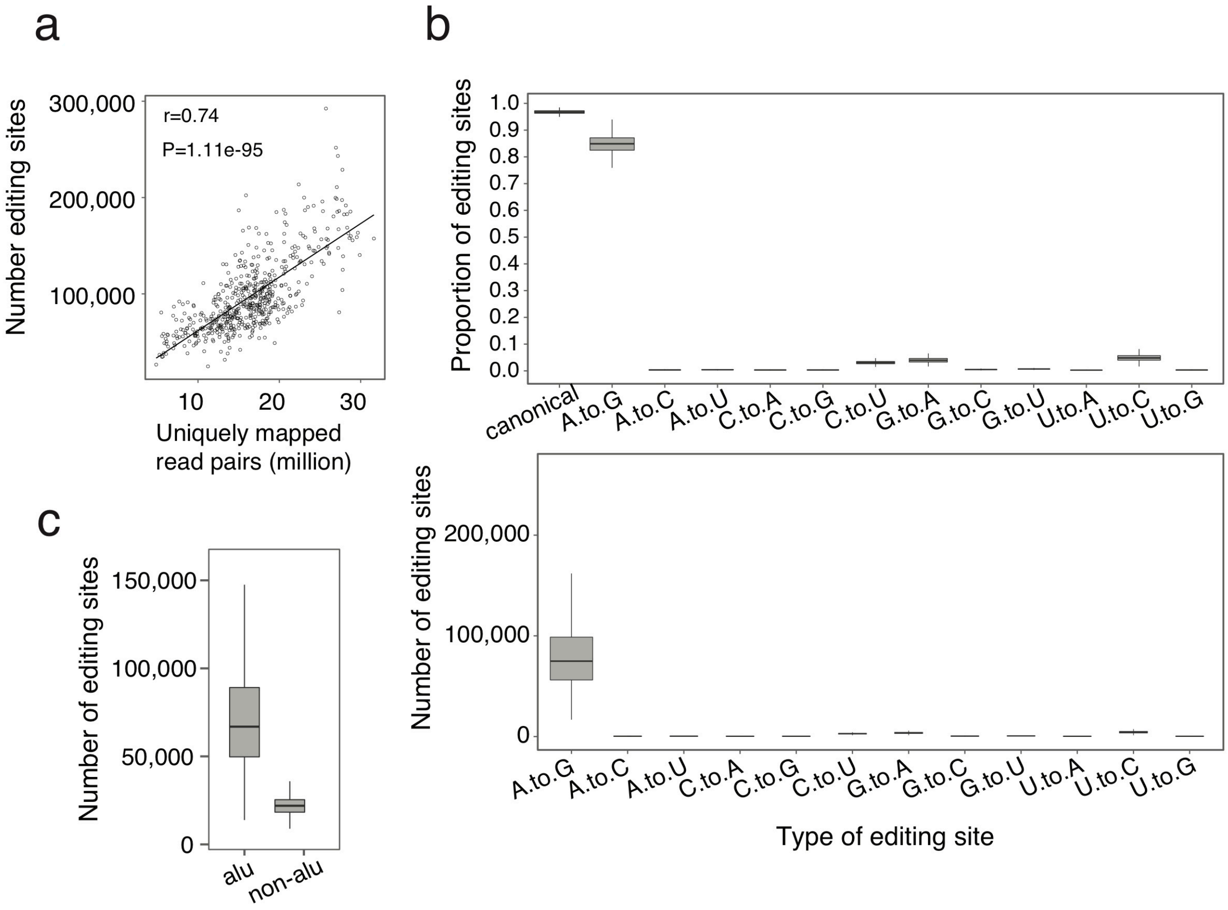
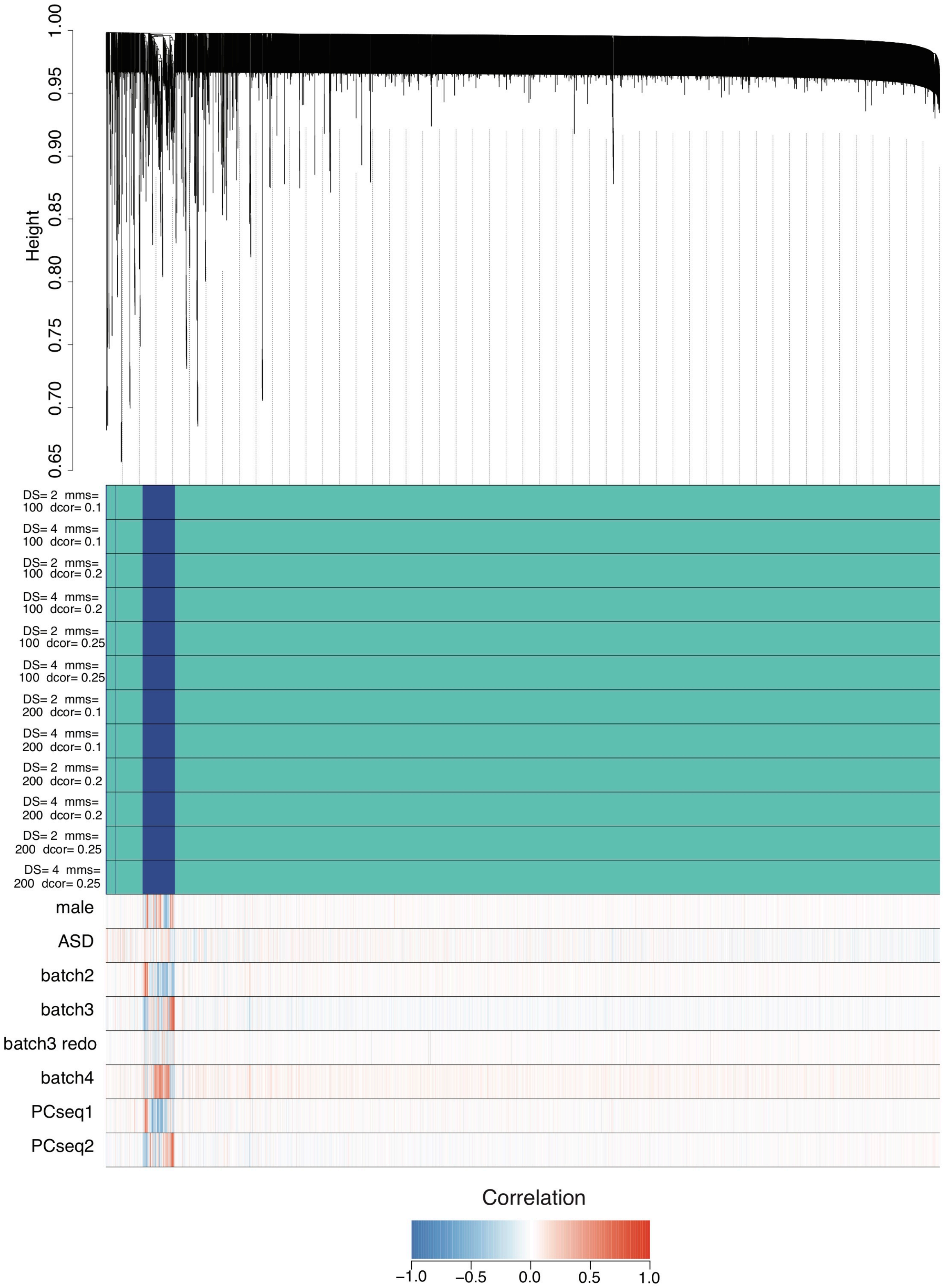


Figure 3.2 Quality of RNA editing site detection

Graphs show various properties of the RNA editing sites detected from the RNA-sequencing data. a) Corrrelation between number of editing sites detected and number of uniquely mapped reads per sample. The number of detectable editing sites strongly correlates with coverage. b) Types of editing sites detected over all samples. Boxplots show the proportion (top) and number (bottom) of types of editing detected over all samples. Canonical editing sites include the known A-to-G and C-to-U and their reverse complements U-to-C and G-to-A. The predominant A.to.G proportion of editing indicates high specificity of RNA editing detection. c) Number of editing sites over all samples falling in alu or non-alu genomic regions. Boxplot defiinition: center, median; lower hinge, 25th percentile; upper hinge, 75th percentile; minimum and maximum extend to at most 1.5 X IQR. n=553 samples for all graphs.



Height
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
DS= 2 mms= 100 dcor= 0.1
DS= 4 mms= 100 dcor= 0.1
DS= 2 mms= 100 dcor= 0.2
DS= 4 mms= 100 dcor= 0.2
DS= 2 mms= 100 dcor= 0.25
DS= 4 mms= 100 dcor= 0.25
DS= 2 mms= 200 dcor= 0.1
DS= 4 mms= 200 dcor= 0.1
DS= 2 mms= 200 dcor= 0.2
DS= 4 mms= 200 dcor= 0.2
DS= 2 mms= 200 dcor= 0.25
DS= 4 mms= 200 dcor= 0.25
male
ASD
batch2
batch3
batch3 redo
batch4
PCseq1
PCseq2
Correlation
−1.0
−0.5
0.0
0.5
1.0

Figure 3.3 Modules of RNA editing sites over cortical spheroid development

Top dendrogram shows topological overlap of editing sites (n=5770). Middle colored maps show module member assignment of each editing sites after applying various tree cutting parameters. DS=deep splitting parameter, mms=minimal number of editing sites required to form a module, dcor=1-minimum correlation used to fuse smaller modules together. Bottom heatmaps show spearman correlation of each editing site with various biological and confounding variables. Module assignment is highly robust to choice of tree cutting parameters and identifies 2 modules. A small subset of sites is strongly confounded by technical variables (blue module). All other sites fall in the turquoise module.

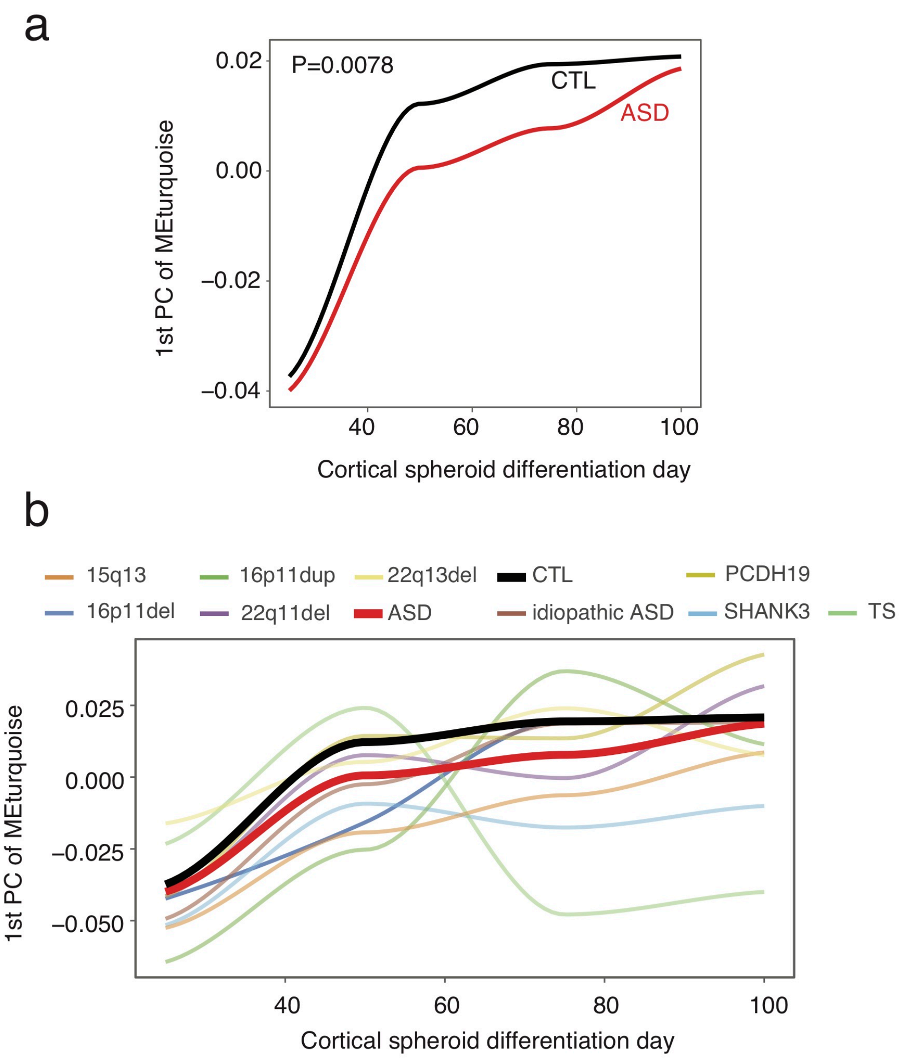


Figure 3.4 Hypoediting in ASD across cortical spheroid development

Plots show the projection of the turquoise module over cortical spheroid differentiation days. a) Loess fitted curve to the 1st principal component of the turqouise module over differentiation days partitioned by control (CTL) and autism (ASD) samples. P-value shows significance of the observed hypoediting pattern in ASD through permutation label swapping. b) Similar to (a) but including partitions over the various autism spectrum disorders. 15q13=15q13 duplication, 16p11del =16p11 deletion, 16p11dup=16p11 duplication, 22q11del=22q11 deletion, 22q13del=22q13 deletion, ASD=all the Autism spectrum disorders combined, CTL=control, idiopathic ASD=idiopathic Autism, PCDH19 and SHANK3=mutations in the respective genes, TS=Timothy syndrome.

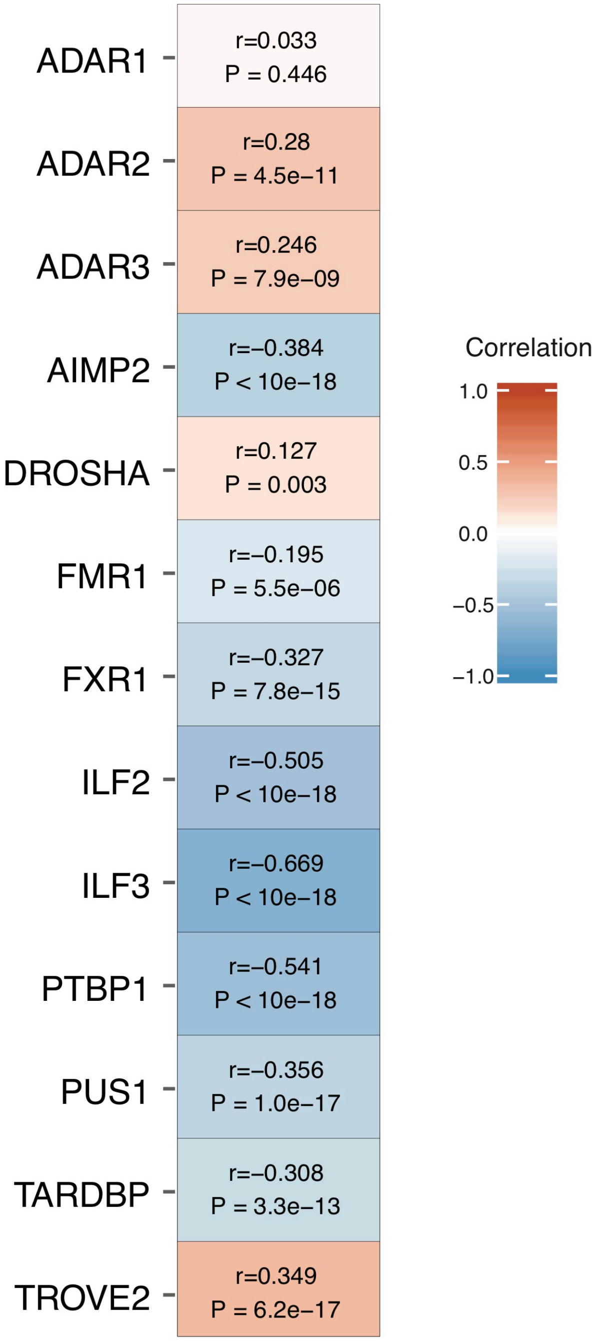

ADAR1
r=0.033
P = 0.446
ADAR2
r=0.28
P = 4.5e−11
ADAR3
r=0.246
P = 7.9e−09
AIMP2
r=−0.384
P < 10e−18
DROSHA
r=0.127
P = 0.003
FMR1
r=−0.195
P = 5.5e−06
FXR1
r=−0.327
P = 7.8e−15
ILF2
r=−0.505
P < 10e−18
ILF3
r=−0.669
P < 10e−18
PTBP1
r=−0.541
P < 10e−18
PUS1
r=−0.356
P = 1.0e−17
TARDBP
r=−0.308
P = 3.3e−13
TROVE2
r=0.349
P = 6.2e−17
Correlation
1.0
0.5
0.0
−0.5
−1.0

Figure 3.5 Potential regulators of RNA editing in cortical spheroid development

Heatmap of the spearman correlation between the 1st principal component of the turquoise module and RNA binding proteins known to regulate RNA editing. Annotations include the spearman correlation coefficient and corresponding p-value, adjusted using the Benjamin Hochberg method.

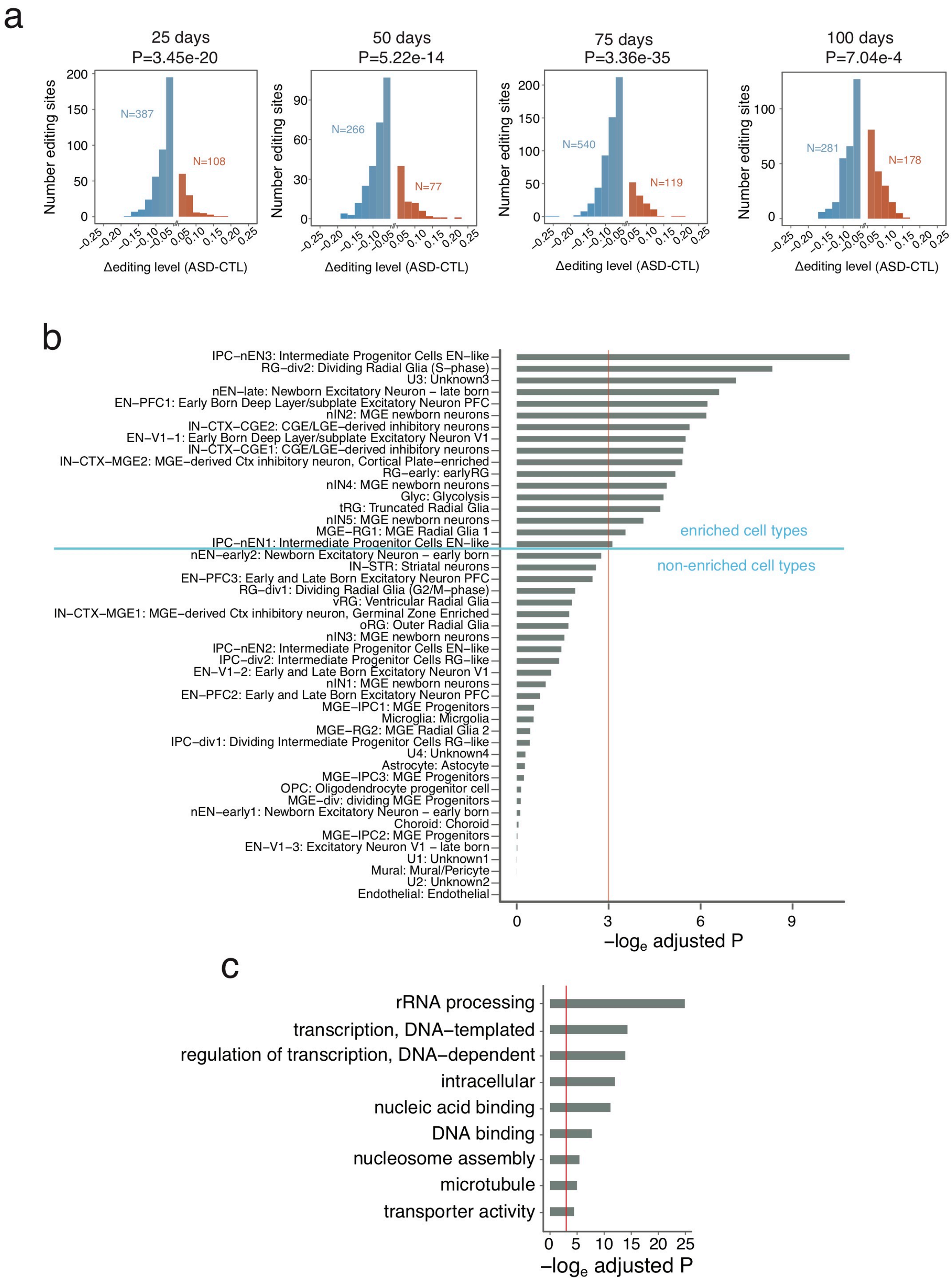

a
25 days
P=3.45e-20
50 days
P=5.22e-14
75 days
P=3.36e-35
100 days
P=7.04e-4
Number editing sites
N=387
N=108
N=266
N=77
N=540
N=119
N=281
N=178
0
50
100
150
200
30
60
90
−0.25
−0.20
−0.15
−0.10
−0.05
0.05
0.10
0.15
0.20
0.25
Δediting level (ASD-CTL)
b
IPC−nEN3: Intermediate Progenitor Cells EN−like
RG−div2: Dividing Radial Glia (S−phase)
U3: Unknown3
nEN−late: Newborn Excitatory Neuron − late born
EN−PFC1: Early Born Deep Layer/subplate Excitatory Neuron PFC
nIN2: MGE newborn neurons
IN−CTX−CGE2: CGE/LGE−derived inhibitory neurons
EN−V1−1: Early Born Deep Layer/subplate Excitatory Neuron V1
IN−CTX−CGE1: CGE/LGE−derived inhibitory neurons
IN−CTX−MGE2: MGE−derived Ctx inhibitory neuron, Cortical Plate−enriched
RG−early: earlyRG
nIN4: MGE newborn neurons
Glyc: Glycolysis
tRG: Truncated Radial Glia
nIN5: MGE newborn neurons
MGE−RG1: MGE Radial Glia 1
IPC−nEN1: Intermediate Progenitor Cells EN−like
nEN−early2: Newborn Excitatory Neuron − early born
IN−STR: Striatal neurons
EN−PFC3: Early and Late Born Excitatory Neuron PFC
RG−div1: Dividing Radial Glia (G2/M−phase)
vRG: Ventricular Radial Glia
IN−CTX−MGE1: MGE−derived Ctx inhibitory neuron, Germinal Zone Enriched
oRG: Outer Radial Glia
nIN3: MGE newborn neurons
IPC−nEN2: Intermediate Progenitor Cells EN−like
IPC−div2: Intermediate Progenitor Cells RG−like
EN−V1−2: Early and Late Born Excitatory Neuron V1
nIN1: MGE newborn neurons
EN−PFC2: Early and Late Born Excitatory Neuron PFC
MGE−IPC1: MGE Progenitors
Microglia: Microglia
MGE−RG2: MGE Radial Glia 2
IPC−div1: Dividing Intermediate Progenitor Cells RG−like
U4: Unknown4
Astrocyte: Astocyte
MGE−IPC3: MGE Progenitors
OPC: Oligodendrocyte progenitor cell
MGE−div: dividing MGE Progenitors
nEN−early1: Newborn Excitatory Neuron − early born
Choroid: Choroid
MGE−IPC2: MGE Progenitors
EN−V1−3: Excitatory Neuron V1 − late born
U1: Unknown1
Mural: Mural/Pericyte
U2: Unknown2
Endothelial: Endothelial
enriched cell types
non-enriched cell types
0
3
6
9
$-\log_e$ adjusted P
c
rRNA processing
transcription, DNA−templated
regulation of transcription, DNA−dependent
intracellular
nucleic acid binding
DNA binding
nucleosome assembly
microtubule
transporter activity
0
5
10
15
20
25
$-\log_e$ adjusted P

Figure 3.6 Functional enrichment of differential RNA editing in ASD cortical spheroids

a) Histograms show distributions of difference in editing levels (average ASD - average CTL) of differential editing sites identfied per differentiation timepoint. Blue bars show number of sites downregulated in ASD. Red bars show sites upregulated in ASD. P-value shows significance of downregulated bias of RNA editing (two-tailed Fisher's Exact test). b) Enrichment of differential editing sites in fetal brain cell types. Red line denotes FDR < 0.05 cutoff. P-value shows Benjamin Hochberg adjusted FDR. Differential editing sites show enrichment in cell types related to radial glia, neuroprogenitor cells, and newborn neurons. c) Gene ontology enrichment of differential editing sites. Red line denotes FDR < 0.05 cutoff. P-value shows Benjamin Hochberg adjusted FDR.

3.6 References

1 Association, A. P. *Diagnostic and statistical manual of mental disorders (4th ed., text rev.)*. 4th edn, (2000).

2 Lord, C. *et al.* The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders* **30**, 205-223 (2000).

3 Lord, C., Rutter, M. & Le Couteur, A. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with

possible pervasive developmental disorders. *Journal of autism and developmental disorders* **24**, 659-685 (1994).

4 Vitrac, A. & Cloez-Tayarani, I. Induced pluripotent stem cells as a tool to study brain circuits in autism-related disorders. *Stem cell research & therapy* **9**, 226, doi:10.1186/s13287-018-0966-2 (2018).

5 Shinoda, Y., Sadakata, T. & Furuichi, T. Animal models of autism spectrum disorder (ASD): a synaptic-level approach to autistic-like behavior in mice. *Experimental animals* **62**, 71-78, doi:10.1538/expanim.62.71 (2013).

6 Wang, H. Modeling Neurological Diseases With Human Brain Organoids. *Frontiers in synaptic neuroscience* **10**, 15, doi:10.3389/fnsyn.2018.00015 (2018).

7 Di Lullo, E. & Kriegstein, A. R. The use of brain organoids to investigate neural development and disease. *Nature reviews. Neuroscience* **18**, 573-584, doi:10.1038/nrn.2017.107 (2017).

8 Yoon, S. J. *et al.* Reliability of human cortical organoid generation. *Nature methods* **16**, 75-78, doi:10.1038/s41592-018-0255-0 (2019).

9 Tran, S. S. *et al.* Widespread RNA editing dysregulation in brains from autistic individuals. *Nat Neurosci* **22**, 25-36, doi:10.1038/s41593-018-0287-x (2019).

10 Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008-1021, doi:10.1016/j.cell.2013.10.031 (2013).

11 Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997-1007, doi:10.1016/j.cell.2013.10.020 (2013).

12 Pasca, A. M. *et al.* Functional cortical neurons and astrocytes from human pluripotent stem cells in 3D culture. *Nature methods* **12**, 671-678, doi:10.1038/nmeth.3415 (2015).

13 Sloan, S. A., Andersen, J., Pasca, A. M., Birey, F. & Pasca, S. P. Generation and assembly of human brain region-specific three-dimensional cultures. *Nature protocols* **13**, 2062-2085, doi:10.1038/s41596-018-0032-7 (2018).

14 Bahn, J. H. *et al.* Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome research* **22**, 142-150, doi:10.1101/gr.124107.111 (2012).

15 Hsiao, Y. E. *et al.* RNA editing in nascent RNA affects pre-mRNA splicing. *Genome research* **28**, 812-823, doi:10.1101/gr.231209.117 (2018).

16 Quinones-Valdez, G. *et al.* Regulation of RNA editing by RNA-binding proteins in human cells. *Communications biology* **2**, 19, doi:10.1038/s42003-018-0271-8 (2019).

17 Liu, Z. & Zhang, J. Human C-to-U Coding RNA Editing Is Largely Nonadaptive. *Molecular biology and evolution* **35**, 963-969, doi:10.1093/molbev/msy011 (2018).

18 Nishikura, K. A-to-I editing of coding and non-coding RNAs by ADARs. *Nature reviews. Molecular cell biology* **17**, 83-96, doi:10.1038/nrm.2015.4 (2016).

19 Tan, M. H. *et al.* Dynamic landscape and regulation of RNA editing in mammals. *Nature* **550**, 249-254, doi:10.1038/nature24041 (2017).

20 Ruzzo, E. K. *et al.* Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks. *Cell* **178**, 850-866.e826, doi:10.1016/j.cell.2019.07.015 (2019).

21 Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215-1233, doi:10.1016/j.neuron.2015.09.016 (2015).

22 Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nature methods* **12**, 357-360, doi:10.1038/nmeth.3317 (2015).

23 Porath, H. T., Carmi, S. & Levanon, E. Y. A genome-wide map of hyper-edited RNA reveals numerous new sites. *Nat Commun* **5**, 4726, doi:10.1038/ncomms5726 (2014).

24 Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823* (2014).

25 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **9**, 559, doi:10.1186/1471-2105-9-559 (2008).

26 Wu, Y. E., Parikshak, N. N., Belgard, T. G. & Geschwind, D. H. Genome-wide, integrative analysis implicates microRNA dysregulation in autism spectrum disorder. *Nat Neurosci* **19**, 1463-1476, doi:10.1038/nn.4373 (2016).

27 Nowakowski, T. J. *et al.* Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science (New York, N.Y.)* **358**, 1318-1323, doi:10.1126/science.aap8809 (2017

Chapter 4

Statistical inference of differential RNA editing sites from RNA-sequencing data by hierarchical modeling

4.1 Introduction

RNA editing alters RNA sequences by base modifications, insertions, and deletions, which in metazoans[1], predominantly consists of adenosine to inosine (A-to-I) changes catalyzed by the ADAR proteins. A plethora of studies have demonstrated diverse biological roles for A-to-I editing. The most-studied type of RNA editing consists of recoding sites. Because inosines are recognized as guanosines by the cellular machinery, such editing sites can lead to amino acid substitutions in proteins. Functional studies have implicated many of these protein-recoding sites in modulating neuronal function, synaptic permeability and emissions[2]. In contrast, RNA editing in non-coding regions is less well understood, despite the vast number of sites. Nonetheless, the functional relevance of such editing sites is starting to be elucidated, such as in splicing regulation[3-5], microRNA targeting and polyadenylation of 3' UTRs[6-8], and modulation of double-stranded RNA-related immunity[9]. Importantly, widespread aberrant RNA editing patterns have been reported across a large number of diseases, including neurological diseases[10,11], atherosclerosis[12], cancer[13-16], and autoimmune disorders[17].

Recent studies of RNA editing were greatly facilitated by the RNA sequencing (RNA-seq) technology and related bioinformatic tools, which allow comprehensive delineation of global editomes in diverse biological processes. In editome profiling, a critical task is to identify editing sites that are statistically different in their quantitative levels between two groups of samples (such as disease vs. controls). Alternatively, editing sites whose levels are statistically associated with certain variables, such as age, are sought after. Most previous studies utilized the classic t-test[18-20], Wilcoxon rank-sum test[11,13,17,18,21], Fisher's Exact test[11,22-24], or linear regression-based test[8,11,20,23,25,26] for these purposes. However, these tests are limited due to the lack of consideration of uncertainty in the read counts or variability in editing quantification between biological replicates.

Here, we develop and evaluate a suite of tools, REDITs (RNA editing tests), that are built upon beta-binomial-based models to carry out differential editing analyses. Specifically, REDITs consist of two methods. The first method handles the classic case of identifying editing sites that possess differential editing levels between two conditions (i.e. case vs. control). The second method carries out statistical inference of categorical or quantitative variables that covary with editing levels (e.g. age-correlated RNA editing). Beta-binomial models have been applied to analyze DNA methylation[27-31]. However, these methods are not directly applicable to RNA editing studies due to their methylation-specific aspects (e.g. methylome-wide priors).

We show via simulated and actual data that REDITs have improved sensitivity and mitigated false-positive rate compared to commonly used alternatives. We applied REDITs to human tissue editomes (GTEx) and examined the association between RNA editing and age or gender. Interestingly, we observed that editing shows increasing or decreasing trajectories over age in many human tissues and exhibits gender-associations in some tissue types.

4.2 Methods

4.2.1 Beta-binomial model underlying REDITs

We first consider an A-to-G editing site measured using RNA-seq across m samples from two conditions (Fig. 1a). For each sample m_i, we denote the total read coverage as n_i and the number of reads harboring the edited nucleotide as k_i. As in most previous studies, the observed editing level can be calculated as $\frac{k_i}{n_i}$. If we assume the true underlying editing level is θ_i, then k_i follows a binomial distribution:

$$k_i|\theta_i \sim Binomial(n_i, \theta_i),$$

$$P(k_i|\theta_i) = \binom{n_i}{k_i}\theta_i^{k_i}(1-\theta_i)^{n_i-k_i}.$$

The value of θ_i, is expected to vary amongst samples due to biological variability. Thus, we model θ_i using a beta distribution for samples from the same condition. More specifically, for condition 1:

$$\theta_i \sim Beta(\alpha_1, \beta_1)$$

$$P(\theta_i) = \frac{\theta_i^{\alpha_1 - 1}\,(1-\theta_i)^{\beta_1 - 1}}{B(\alpha_1, \beta_1)},\ i = 1 \ldots j;$$

For condition 2:

$$\theta_i \sim Beta(\alpha_2, \beta_2)$$

$$P(\theta_i) = \frac{\theta_i^{\alpha_2 - 1}\,(1-\theta_i)^{\beta_2 - 1}}{B(\alpha_2, \beta_2)}, i = j + 1 \ldots m,$$

where α, β are hyper parameters and $B(\alpha, \beta)$ is the beta function. We choose the beta distribution because it adheres to the restriction that editing levels must fall in a continuum between (0,1). Additionally, the beta distribution is conjugate to the binomial distribution (below), which eases our inference procedure. Given the inordinate flexibility of both concave and convex shapes possessed by the beta distribution, we restricted $\alpha \geq 1$ and $\beta \geq 1$ to enforce that it partake only uniform or unimodal shapes and eschew U-shapes. Intuitively, this restriction presumes that the distribution of true editing levels per condition has measures of centrality and dispersion that approximately correspond to the peak and width of the distribution respectively. In addition, these parameter restrictions were observed for all editing sites analyzed in the GTEx dataset (Results, supplemental Fig. 1a). In totality, each sample follows a generative model whereby its true editing level θ_i is an observation from a uniform or unimodal beta distribution characteristic of its condition, and the random sampling of edited reads or non-edited reads from RNA-seq follows a binomial distribution (Fig. 1b). The entire generative model is a beta-binomial distribution:

$$P(k_i, \theta_i |\, n_i, \alpha_l, \beta_l) = \ P(k_i | \theta_i, n_i)\ \cdot P(\theta_i | \alpha_l, \beta_l), \text{ where } l = 1 \text{ for } i \leq j \text{ and } l = 2 \text{ for } i \geq j + 1$$

$$= \binom{n_i}{k_i}\theta_i^{k_i}(1-\theta_i)^{n_i-k_i} \cdot \frac{\theta_i^{\alpha_l-1}\,(1-\theta_i)^{\beta_l-1}}{B(\alpha_l,\beta_l)}.$$

Integrating over θ_i yields the marginal likelihood of k_i given the hyper-parameter (α,β):

$$P(k_i|n_i,\alpha_l,\beta_l) = \int_0^1 \binom{n_i}{k_i}\theta_i^{k_i}(1-\theta_i)^{n_i-k_i} \cdot \frac{\theta_i^{\alpha_l-1}\,(1-\theta_i)^{\beta_l-1}}{B(\alpha_l,\beta_l)}\,d\theta_i$$

$$= \binom{n_i}{k_i}/B(\alpha_l,\beta_l)\int_0^1 \theta_i^{k_i}(1-\theta_i)^{n_i-k_i} \cdot \theta_i^{\alpha_l-1}\,(1-\theta_i)^{\beta_l-1}\,d\theta_i$$

$$= \binom{n_i}{k_i}/B(\alpha_l,\beta_l)\int_0^1 \theta_i^{k_i+\alpha_l-1}\,(1-\theta_i)^{n_i-k_i+\beta_l-1}\,d\theta_i$$

$$= \binom{n_i}{k_i} \cdot \frac{B(k_i+\alpha_l,n_i-k_i+\beta_l)}{B(\alpha_l,\beta_l)}\ . \qquad (1)$$

4.2.2 Statistical inference of differential editing between two groups

Given two groups of samples (e.g., cases and controls), under the null hypothesis of no between-group difference on editing (i.e. $\alpha_1 = \alpha_2 = \alpha_0$ and $\beta_1 = \beta_2 = \beta_0$), the likelihood of the data is given by:

$$L_0 = \prod_1^m P(k_i|n_i,\alpha_0,\beta_0) = \prod_1^m \binom{n_i}{k_i} \cdot \frac{B(k_i+\alpha_0,n_i-k_i+\beta_0)}{B(\alpha_0,\beta_0)}\ .$$

The likelihood of the alternative model where significant difference exists between the two groups is given by:

$$L_A = \prod_{i=1}^{j} P(k_i|n_i,\alpha_1,\beta_1) \prod_{i=j+1}^{m} P(k_i|n_i,\alpha_2,\beta_2)$$

$$= \prod_{i=1}^{j} \binom{n_i}{k_i} \cdot \frac{B(k_i+\alpha_1,n_i-k_i+\beta_1)}{B(\alpha_1,\beta_1)} \ \prod_{i=J=1}^{m} \binom{n_i}{k_i} \cdot \frac{B(k_i+\alpha_2,n_i-k_i+\beta_2)}{B(\alpha_2,\beta_2)}\ .$$

Wilk’s theorem states that the statistical significance of differential editing is given by:

$$-2 \cdot \log\left(\frac{L_0}{L_A}\right) \sim \chi^2, \text{ with 2 degrees of freedom}$$

where L_0 and L_A are evaluated at the maximum likelihood estimates (MLEs) of α’s and β’s. Thus, we call this method REDIT-LLR henceforth.

4.2.3 Statistical inference of editing sites that covary with quantitative variables

The beta-binomial model can be expanded to handle statistical inference under the regression scenarios (REDIT-Regression). For convenience, we describe this method using the example of identification of editing sites that covary with age (A_i). For a specific editing site, we assume the underlying true editing levels per sample, ($\frac{k_i}{n_i} \sim \theta_i$) ($i = 1 \dots m$), follow a beta distribution with constant dispersion (σ) but with mean (μ) linearly dependent on age (Fig. 1c). The assumption of a constant σ and the dependency of μ on age is analogous to that of linear regression. We re-parameterize the beta-binomial model as follows,

$$\mu = \frac{\alpha}{\alpha + \beta}$$

$$\sigma = \frac{1}{\alpha + \beta}.$$

Then the marginal likelihood of k_i in Eq. (1) becomes

$$P(k_i|n_i, u_i, \sigma) = \binom{n_i}{k_i} \frac{B\left(k_i + \frac{\mu_i}{\sigma}, n_i - k_i + \frac{1-\mu_i}{\sigma}\right)}{B\left(\frac{\mu_i}{\sigma}, \frac{(1-\mu_i)}{\sigma}\right)}.$$

The dependency of μ on age is linear:

$$\mu_i = \beta_{age} \cdot A_i + \beta_{0.}$$

Under the above re-parameterization, μ must fall between (0,1) and σ must be > 0, which we enforce during maximum likelihood estimation. The likelihood of the data is given by:

$$L_{data} = \prod_{i=1}^{m} P(k_i|n_i, \mu_i, \sigma) = \prod_{i=1}^{m} \binom{n_i}{k_i} \frac{B\left(\frac{k_i + \mu_i}{\sigma}, n_i - k_i + \frac{1 - \mu_i}{\sigma}\right)}{B\left(\frac{\mu_i}{\sigma}, \frac{(1 - \mu_i)}{\sigma}\right)} \ .$$

The null hypothesis H_0 is that age does not impact editing, i.e., $\beta_{age} = 0$ or equivalently $\mu_i = \beta_0$. Based on Wilk's theorem, the statistical significance of the alternative model (H_A: $\beta_{age} \neq 0$) is given by:

$$-2 \cdot \log\left(\frac{L_0}{L_A}\right) \sim \chi^2, \text{with 1 degree of freedom}$$

where L_0 and L_A are maximum likelihood under H_0 and H_A respectively. General inference of multiple covariates ($\beta_1, \beta_2, \ldots$) with respective observations ($X_{1i}, X_{2i}, \ldots$) can be carried out by comparing maximum likelihood under the alternative model, $\mu_i = \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \cdots + \beta_0$, to maximum likelihood under null models with $\beta_j = 0$ to determine statistical association of covariate j with editing for each j=1,2,…. Categorical variables (e.g. gender, ethnicity) can also be included by encoding them as 0 and 1. Regression for one or more quantitative and/or categorical covariates is handled by our provided code (see code availability).

For regressions of proportions (values restricted between 0 and 1), the μ term is conventionally transformed using a logistic link function: $\mu_i = \frac{1}{1+e^{-(\beta_{age} \cdot A_i + \beta_0)}}$. However, the logistic transformation can over-fit the data and lead to inflated maximum likelihood ratios. Thus, we chose not to use the logistic link function and instead opted to constrict regression coefficients within the maximum likelihood estimation so that editing levels would never fall below zero or above 1.

4.2.4 Simulations to evaluate REDIT-LLR

To simulate RNA editing data, we extracted RNA editing sites from the REDIportal database[32] derived from 2660 GTEx samples and other sources. Using these data, we simulated realistic read coverages and hyper-parameter distributions reflecting biological variance of editing levels. First, we used maximum likelihood estimation to fit beta distributions to the editing levels of each editing site in brain samples of GTEx. Brain was chosen since it has the largest sample size among all histological types (Supplemental Table 1a). Furthermore, to acquire highly accurate parameters, we required the editing sites to have ≥ 20 reads in ≥ 250 brain samples. A total of 1206 editing sites were retained. The α and β parameters were then clustered (k-means) into 10 clusters, yielding 10 representative parameter values (Supplemental Fig. 1a-b, Supplemental Table 1b). As an alternative, we repeated the simulations using a truncated Gaussian distribution instead of beta distribution. The mean and variance parameters were converted directly from the mean and variance of the beta distributions (Supplemental Fig. 1d, Supplemental Table 1b).

Editing levels were sampled from the beta or truncated Gaussian distributions, and the numbers of edited reads for each sample were simulated using the corresponding binomial distribution. To simulate read coverages, we used maximum likelihood estimation to fit negative binomial distributions to the coverage data of the above editing sites from 10 random GTEx brain samples (Supplemental Table 1c, Supplemental Fig. 1c). A total of 100 independent simulations of 1000 editing sites were created for each group, with 2, 3 or 5 samples per group.

4.2.5 Evaluating sensitivity and false-positive rates of REDIT-LLR

The false-positive rate and sensitivity of REDIT-LLR was evaluated by simulating case-control scenarios where the case and control groups were each characterized by 1 of the 10 beta (or truncated Gaussian) distributions (Supplemental Fig. 1). Sensitivity was evaluated where the cases and controls were simulated using different underlying distributions, and false-positive rate was evaluated using identical distributions to generate cases and controls. A $p < 0.05$ was imposed to call significant comparisons. We piloted evaluation of sensitivity and false-positive rate on a single set of parameters for a case-control comparison of 3 replicates, where the parameters for sensitivity were beta(α=13.52, β=11.95) versus beta(α =43.64, β=0.23), and parameters for false-positives were beta(α=13.52, β=11.95) for both groups (Figure 2a,b); simulations were then expanded to include all combinations of parameters and sample sizes (Supplemental Figure 2,4).

Pooled Fisher's Exact test, t-tests, and Wilcoxon Rank-sum tests were also applied on the above simulated editing sites. Pooled Fisher's Exact test was carried out by pooling reads from replicates and testing the resulting 2x2 contingency table. The t-test and Wilcoxon rank sum tests were performed in 2 ways, respectively, (1) using editing levels estimated without minimal read coverage requirement, and (2) filtering out, for each editing site, any sample with read coverage < 10 (i.e., thresholded t-test or Wilcoxon rank sum test).

4.2.6 Evaluating false-positive rates of REDIT-Regression using simulated data

To evaluate REDIT-Regression, we simulated editing sites that covary with age. We based the simulations on a previous dataset of 33 postmortem human brains spanning fetal stages to old age[20]. A total of 267,766 editing sites were reported by this study.

To test the false-positive rates, we simulated editing sites where age had no effect on editing level. For each editing site, we extracted its read coverage in each sample of the original dataset. The editing level and number of edited reads were simulated similarly as described for REDIT-LLR, using the beta or truncated Gaussian distributions (Supplemental Fig. 1). For each editing site, one of the 10 beta or truncated Gaussian distributions, respectively, was randomly selected to simulate edited reads. Each simulation included 3, 5, 7, and 33 samples and 267,766 editing sites, with the age values of the samples unaltered. For sample sizes of 3, 5, and 7 we chose to use samples (R5805, R3523, R3990), (R5805, R3591, R3497, R4371, R3990), and (R5805, R5815, R3552, R3497, R4054, R3539, R3990), respectively, as these samples represented the age range of the dataset. Each simulation included editing sites where median coverage was at least 5 (93,437 sites for n=3, 86,290 for n=5, 68,235 for n=7, and 90,919 for n=33), and 100 independent simulations were carried out per sample size. For each simulation, the false-positive rate was calculated as the fraction of sites with significant age associations (REDIT-Regression $p < 0.05$) among all sites tested. We piloted the simulations on the sample size of 3 (Figure 2d) and then expanded to the other sample sizes (Supplemental Figure 7b, 8b).

Binomial regression, linear regression, and thresholded linear regression (using samples with read coverage ≥ 10) were also performed for the above simulated data.

4.2.7 Evaluating sensitivity of REDIT-Regression using simulated data

To test sensitivity, we simulated various correlations between age and editing levels. First, we estimated representative correlations of these two variables using the original data of the 33 postmortem samples. We used editing sites where ≥20 samples had ≥ 20 read coverage. The observed editing levels (calculated as the number of edited reads divided by read coverage) were then regressed against age using linear regression. To tractably limit the number of simulated age-associations, we used a stricter p-value threshold of $p<0.005$ to deem editing sites significant. These sites were used to derive 5 representative slope and intercept values to simulate linear relationships between age and editing levels (Supplemental Fig. 6a-b, Supplemental Table 1d).

Using the above relationships, we simulated true editing levels of each editing site by randomly sampling from a beta or truncated Gaussian distribution whose mean was set as $\mu_i = \beta_{age} \cdot A_i + \beta_0$, where A_i was the unaltered age of the sample. The standard deviation for each editing site was randomly selected from the standard deviations of the 10 distributions from the GTEx data described above. Other aspects of the simulations are similar as described for false-positive evaluation. Sensitivity was calculated as the fraction of editing sites with significant ($p < 0.05$) age association among all sites tested. We piloted sensitivity evaluation on sample size of 3 and where true editing level was a beta distribution with mean set as $\mu_i = 0.005 \cdot A_i + 0.33$ (Figure 2c) and then expanded to the other 4 slope and intercept values and sample sizes (Supplemental Figure 7a, 8a).

4.2.8 Evaluating false-positive rates of REDITs on actual data

To test the false-positive rate of REDIT-LLR on real data, we obtained 6814 editing sites in 18 control samples from a previous study of system lupus erythematosus[24,33]. We randomly permuted the samples and formed two groups (n=2, 3, 5, 10, 15, or 18 per group). Thus, any editing site called with $p < 0.05$ is deemed a false-positive.

For REDIT-Regression, we used 267,766 editing sites from the 33 postmortem brains[20] as described above and randomly grouped samples with replacement to achieve various sample sizes (n=5, 10, 15, 20, 25, 30, 33). Each sample was then randomly assigned a covariate value (1 to n). Each random sampling and covariate assignment were repeated 100 times.

4.2.9 Application of REDIT-Regression to identify age and gender associated RNA editing

We applied REDIT-Regression to the GTEx dataset to investigate how RNA editing varies with human age (4,668,508 editing sites obtained from the REDIportal database).

To expedite run-time, we removed GTEx tissues that had fewer than 10 samples, and required editing sites to have ≥ 1 read coverage in $\geq$ 10 samples. Partitioning samples per body site and histological type, we identified sites that significantly associated with age ($|\text{beta}_{\text{age}}| > 0.01$ and FDR < 0.1) through REDIT-Regression using age as the only covariate. Tissues with an increasing trajectory in editing over age were defined as those where the number of editing sites demonstrating an increasing trend is at least twice of that with a decreasing trend (and Fisher's Exact test FDR < 0.1). Tissues with a decreasing trajectory were defined similarly. The same criteria were used when finding editing sites associated with age in a dataset of 33 postmortem frontal cortex samples[20]. The subset of samples used to match the age range of GTEx frontal cortex samples were R4054 age: 40.60548, R2897 age: 41.04109, R4049 age: 41.20274, R4371 age: 41.77808, R3791 age: 42.06575, R2826 age: 42.83836, R3539 age: 57.48219, R3479 age: 58.60548, R3766 age: 59.26027, R3445 age: 61.16712, R4038 age: 67.86575, R3990 age: 71.10959.

To identify editing sites that significantly associate with gender, we ran REDIT-Regression using both gender and age as covariates, since age is already a known variable correlated with editing[20,34-36]. $|\text{Beta}_{\text{sex}}| > 0.05$ and FDR < 0.1 was used to call significant associations.

4.2.10 Implementing statistical tests

The t-test, Wilcoxon rank-sum test, Fisher's Exact test, and linear regression were performed using corresponding base functions in R. Binomial regression was run in R using the gamlss package (version 5.1-2) using default arguments.

4.2.11 Running-time performance evaluation

The running-time performances of the REDIT-LLR and REDIT-Regression were evaluated by running various numbers of editing sites from real datasets. REDIT-LLR was run on 29 Autism and 33 controls from frontal cortex[11], and REDIT-Regression was run on the 33 frontal cortex samples spanning human development[20]. We calculated the average running-time (average time required to run 1 editing site) by fitting a least-squares linear regression on running time (minutes) vs number of editing sites tested. No computational parallelization was used for these evaluations.

4.3 Results

4.3.1 Overview of REDITs

REDITs model read counts in RNA editing using a beta-binomial distribution, where read coverage is modelled using the binomial component, and biological variance between replicates is concurrently modelled with the beta component. For differential editing tests between groups (e.g. cases vs. controls), REDITs evaluate a null model assuming no between-group difference, compared to an alternative model that includes two distinct groups. It then determines differential editing based on the significance of the likelihood ratio of the two models (thus, called REDIT-LLR, Fig. 1a-b). To test the correlation of an editing site with one or multiple biological factors (e.g. covariance of editing with age), REDITs model the covariates as a linear combination of regressors that together constitute the mean of the beta component. It then tests whether inclusion of each covariate significantly improves the maximum likelihood ratio (thus, called REDIT-Regression, Fig. 1a,c).

4.3.2 Evaluation of the REDIT-LLR method via simulated data

To evaluate the REDIT-LLR method, we simulated read counts of editing sites using beta distributions estimated from GTEx brain tissues (Supplemental Fig. 1a-c, Supplemental Table 1b-c). We simulated 2, 3, or 5 biological replicates per group, which are typical sample sizes in case-control studies. The significance level was chosen to be 0.05. REDIT-LLR had much greater sensitivity (true positive rate) than the t-test or Wilcoxon rank-sum test, particularly for smaller sample sizes (Figure 2a, Supplemental Fig. 2), which is consistent with the lack of depreciation of sites with inadequate coverages by the latter two methods. Also, this problem of the two methods is not alleviated using thresholds to impose a minimal coverage requirement (Figure 2a, Supplemental Fig. 2), due to loss of sample size.

Although the Fisher's Exact test has comparable sensitivity as REDIT-LLR, its false-positive rate is much higher than the nominal level of 0.05 and that of REDIT-LLR (Figure 2b, Supplemental Fig. 2). This limitation of Fisher's Exact test is consistent with its theoretical flaw of neglecting biological variability. Notably, this problem of Fisher's Exact test exacerbates as sequencing coverage increases (Supplemental Fig. 3).

As an alternative method, we simulated editing sites using a different hyper-parameter distribution (truncated normal instead of beta distribution) (Supplemental Fig. 1d). The REDIT-LLR method still outperformed the other methods (Supplemental Fig. 4), indicating that this method is robust to the underlying distribution of editing levels.

4.3.3 Evaluation of the REDIT-LLR method using actual RNA-seq data

We evaluated the false positive rates of different methods by randomly grouping control samples of a previous study[33] into two groups (Methods). Editing sites identified with $p < 0.05$ were considered as false-positive predictions. REDIT-LLR yielded the lowest false positive rates in the majority of comparisons (Supplemental Fig. 5a). All methods except Fisher's exact test had false-positive rates < 5%. In particular, Fisher's exact test performed poorly at editing sites with highly variable editing levels between samples, which are enriched in Alu and intronic regions (Supplemental Fig. 5b-c). In contrast, for sites with lower variability, which were enriched in exonic and non-Alu regions, Fisher's exact test performed adequately (false-positive rates < 5%). The assumption of low variance between biological replicates likely holds for these editing sites where precise regulation of editing level may be critical for homeostasis[37].

4.3.4 Evaluation of the REDIT-Regression method

To evaluate the REDIT-Regression method, we simulated editing sites based on RNA-seq data of 33 postmortem frontal cortex samples used in a study of RNA editing in human development[20]. The simulations incorporated unaltered read coverages per editing site from the actual data. To evaluate sensitivity, we simulated editing sites that covary with age ($\beta_{age}$and β_0 parameters)(Supplemental Table 1d). For all simulations, the REDIT-Regression method had higher sensitivity (proportion of sites with $p < 0.05$) than the linear regressions, though lower sensitivity than binomial regression (Figure 2c, Supplemental Fig. 7a). Similar trends were observed using simulated data generated from a truncated normal instead of beta distribution (Supplemental Fig. 8a).

We also used the same 33 samples as described above, but did not impose any correlation between RNA editing and age. Thus, the prediction of a significant association between editing and age is a false positive. Based on these simulations, the false-positive rate of REDIT-Regression remains at or below 5%, whereas binomial regression yielded much higher false-positive rates (Figure 2d, Supplemental Fig. 7b, Supplemental Fig. 8b).

We next evaluated the false-positive rate of REDIT-Regression using actual RNA-seq data by bootstrapping samples and shuffling the associated random covariate values (Methods). Similar to the simulation results, all methods except the binomial regression maintained false-positive rates below 5% (Supplemental Fig. 9).

Overall, out of the methods that mitigate false-positives to 5%, REDIT-Regression had the highest sensitivity, demonstrating the optimal balance between false and true positive rate.

4.3.5 REDIT-Regression on GTEx data uncovers association of RNA editing with human aging

Multiple studies indicate that RNA editing levels in the brain increase over age[20,34-36]. However, this trend has not been evaluated across many samples for the panoply of human tissues. We applied REDIT-Regression to the GTEx dataset, to comprehensively investigate the trajectory of editing variations over human aging. Overall in most tissue types, we observed hundreds of editing sites associated with age

(Supplemental Table 2a-b, FDR < 0.1 Methods). Interestingly, many also exhibited homogenously increasing or decreasing trajectories of editing (Supplemental Table 2a-b, Fig. 3, Supplemental Fig. 10).

One striking observation in our results is that editing decreased with age across brain regions (Fig. 3, Supplemental Fig. 10). A previous study reported an increasing trend of brain editing with age (frontal cortex)[20]. However, this increasing trend was predominantly driven by the fetal to infant transition, which was replicated in another study[11]. The ages of GTEx subjects ranged from 20 years to 70 years (Supplemental Fig. 11a-b), which only encapsulates the period of human adulthood to older age. Thus, we hypothesized that the disparity in the age-editing association between ours and previous studies[20,34-36] were attributable to differences in the ages of the respective cohorts. To level the comparison, we performed REDIT-Regression in two ways using the previous dataset[20], with the entire cohort and with a subset of frontal cortex samples aged $\geq$ 20 years, respectively. REDIT-Regression on the entire cohort recapitulated that editing levels predominantly increased during development (762 sites increasing vs 148 sites decreasing, Chi-Square p-value =$3.1e^{-54}$, odds ratio=5.14). However, in the subset of samples aged $\geq$ 20 years, REDIT-Regression identified no editing sites associated with age, which is similar to that observed in the GTEx samples where only 24 sites were associated with age (Supplemental Table 2b, brain-frontal cortex). Overall, our findings underscore that the trends of RNA editing changes differ between early development versus aging.

4.3.6 REDIT-Regression on GTEx data reveals gender-biased RNA editing

Although humans display sexual dimorphism in morphology and physiology, RNA editing differences between genders is largely unknown. Recent studies implicated RNA editing in gender-specific stratification of glioblastoma survival[38]. However, comprehensive investigation of gender-biased editing across human tissues has not been reported. We carried out REDIT-Regression analysis on the GTEx dataset using both age and gender as covariates. Strikingly, we observed gender-biased editing across diverse tissue types (Figure 4, Supplemental Fig. 12, FDR < 0.1 Methods). The two tissues with the greatest number of gender-associated sites were thyroid and adipose tissues. Interestingly these tissues also have morphological or physiological differences between the genders, such as higher thyroid-stimulated hormone levels in females[39] and lower subcutaneous fat levels in males[40]. Our observation suggests that gender-specific RNA editing may be involved in sexual dimorphism of different aspects of physiology.

4.3.7 Computational speed of REDITs

We evaluated REDITs run time using data from two previous studies (62 samples in the LLR analysis, and 33 samples in the regression analysis)[11,20]. The REDIT-LLR method processed 100,000 editing sites in about 14 minutes (Fig. 5a). REDIT-Regression ran 36 minutes for 100,000 editing sites (Fig. 5b). For the preponderance of RNA editing studies, this level of speed is efficient and should obviate the need for parallelization, permitting application of the methods given the most basic computational resources. Nevertheless, we provide an example of how to include parallelization in the REDITs codes (see code availability).

4.4 Discussion

In this work we introduce REDITs which leverage beta-binomial models to detect editing differences between groups or editing association with covariates. Compared to nominal methods used in previous studies, REDITs proffers the advantage of handling the uncertainty in RNA editing levels calculated from limited sequencing depth in RNA-seq data, while still maintaining biological variance modelling. Using both simulated and actual data, we demonstrated that REDITs have superior performance than commonly utilized tests in RNA editing studies. Since REDITs consider biological replicates to model the variability across data sets, they are particularly suitable for handling data with sparse and limited counts. If the sequencing depth is very high, REDITs mathematically simplify to a likelihood ratio test of beta distributions for case-control studies, or a beta regression for regression analyses. Since most RNA-seq data have limited coverage at the single-nucleotide level, REDITs serve a widespread utility.

In this study we clarified the trajectory of editing changes associated with age in human tissues. Specifically, we found hundreds of editing sites associated with age in most of the GTEx tissues and that many tissues exhibited homogenously increasing or decreasing trajectories of editing over age. Our analyses also clarified that, in brain, editing level increases during early development, and then decreases from adulthood to old age. Interestingly, some tissues also have many RNA editing sites with gender bias, which may contribute to sexual differences in physiology and anatomy. The functional importance and mechanistic underpinnings of these trends on human aging and sexual dimorphism merit further examination. Overall, REDITs should serve prodigiously to expand our understanding of how RNA editing undergirds molecular systems, biological phenotypes, and disease.

4.5 Acknowledgements

We thank John Krusckche for references, including his book *Doing Bayesian Analysis*, on how to diagram generative models. We would also like to thank the members of the Xiao laboratory for their helpful discussions and comments on this work.

4.6 Code availability

R scripts for running the REDITs are available on GitHub (https://github.com/gxiaolab/REDITs). The repository also provides a code example for leveraging parallelization using multiple cores.

4.7 Figures

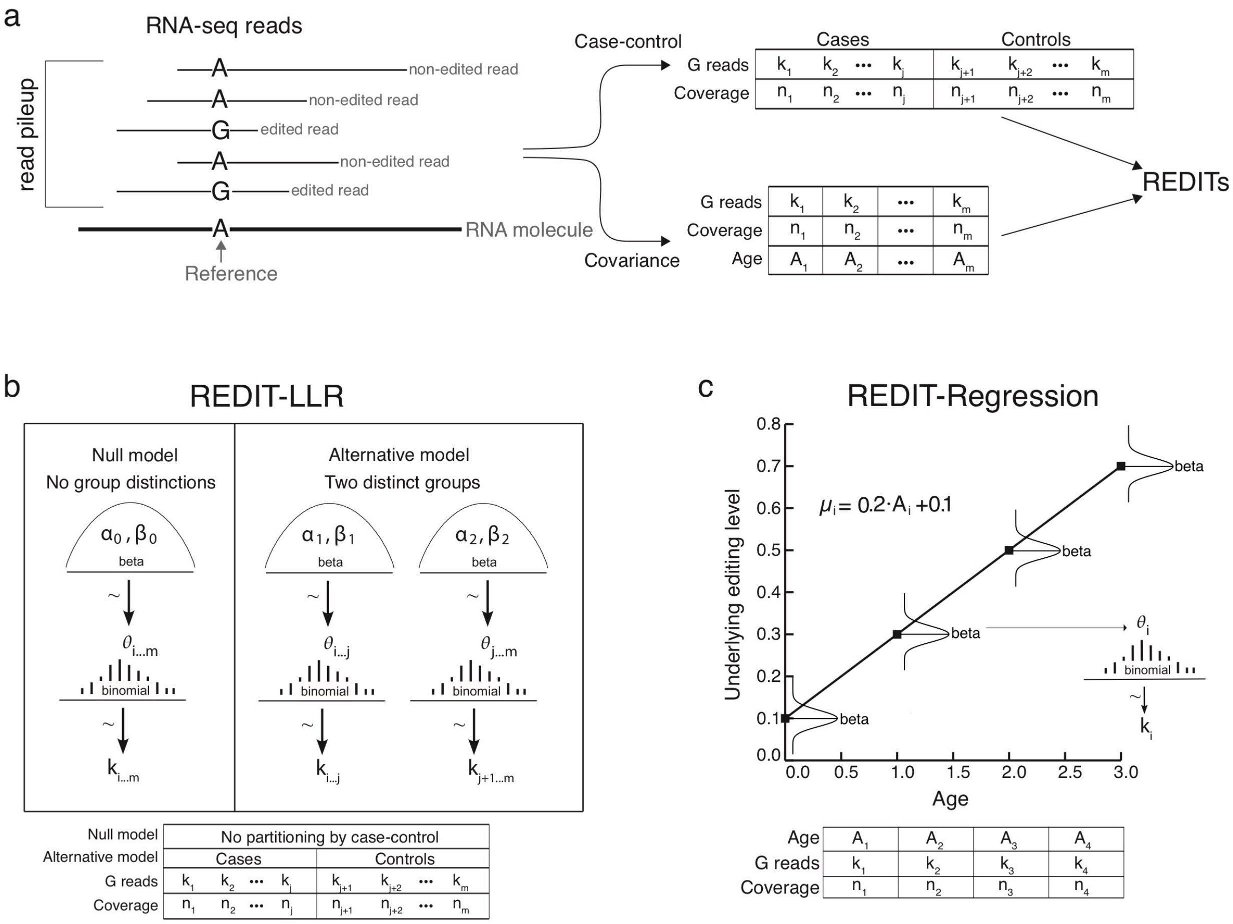


Figure 4.7.1 Overview of REDITs

a) Context for usage of REDITs within RNA editing studies. RNA-seq is generated on multiple samples (i=1...m), and editing sites are quantified by piling reads and counting number of matches/edited reads (A) to the reference and mismatches/non-edited reads (G) to the reference. k_i = # of G reads and n_i = # of A reads + # of G reads. Depending on the type of inference sought, samples are partitioned into tables, which are input into

REDITs for statistical inference. b) Overview of REDITs-LLR for case-control inference. Under the null model, a single beta distribution characterizes the underlying editing levels $\theta_{1...m}$ of all samples, whereas the alternative model posits distinct beta distributions characterizing condition 1 ($\theta_{1...j}$) against condition 2 ($\theta_{j+1...m}$) . A binomial distribution characterizes editing reads per sample ($k_{1...m}$). The coverage acquired via RNA-seq directly determines $n_{1...m}$. c) Overview of REDITs Regression for inference of covariance with RNA editing. For simplicity, the model is depicted showing covariance of editing with age in 4 samples. The underlying editing level of each sample is characterized by a distinct beta distribution where mean (u_i) is linearly dependent on age through u_i = 0.2 Ai + 0.1, and dispersion (σ) is constant. Points along regression line show locations of the means (u_i) of beta distributions. The number of edited reads (k_i) then follow a binomial(k_i | n_i, θ_i) distribution where n_i is determined by sequencing coverage per sample at this editing site, and θ_i is an observation from the respective beta distribution per sample.

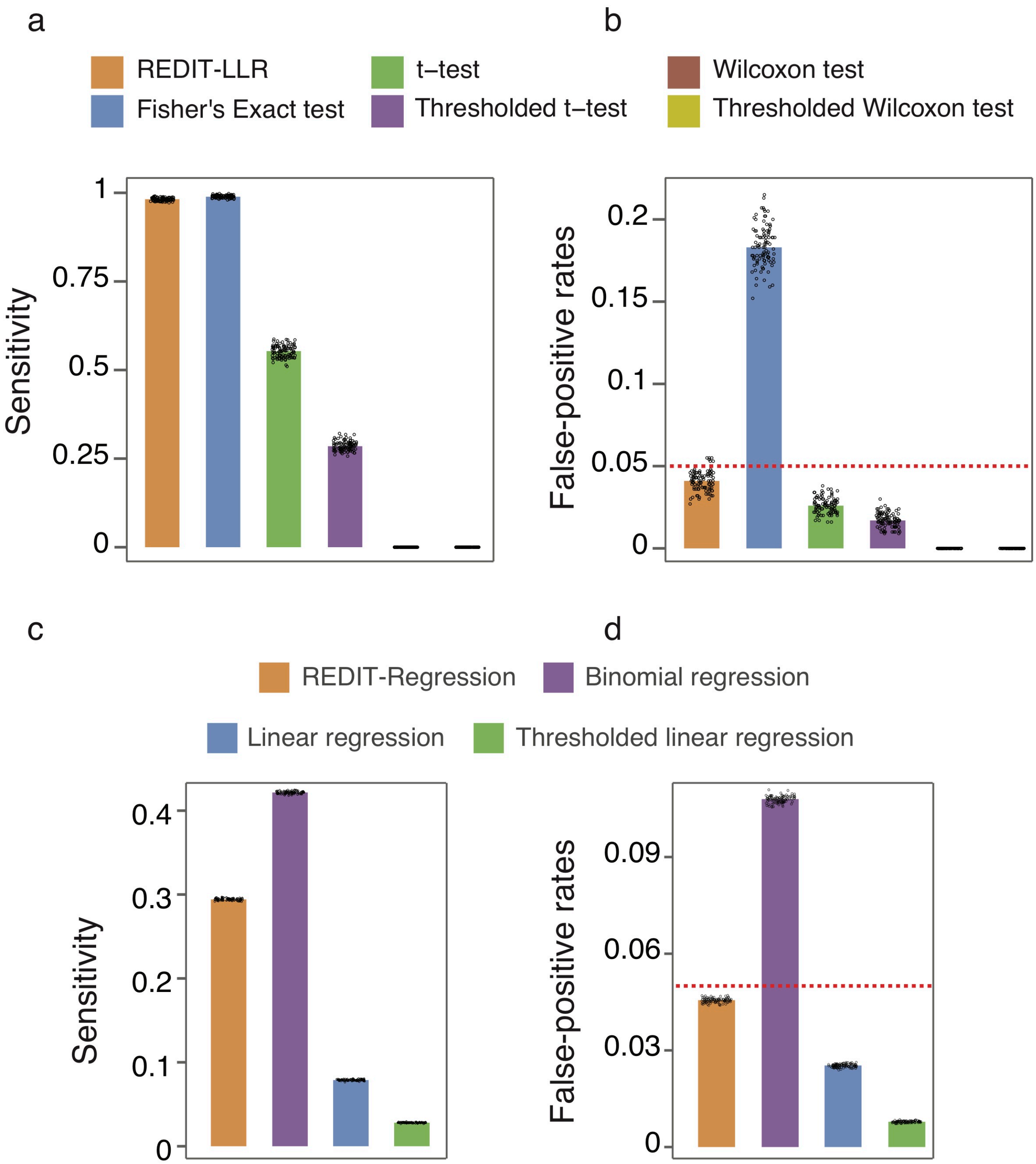

a
b
REDIT-LLR
Fisher's Exact test
t–test
Thresholded t–test
Wilcoxon test
Thresholded Wilcoxon test
Sensitivity
1
0.75
0.5
0.25
0
False-positive rates
0.2
0.15
0.1
0.05
0
c
d
REDIT-Regression
Binomial regression
Linear regression
Thresholded linear regression
Sensitivity
0.4
0.3
0.2
0.1
0
False-positive rates
0.09
0.06
0.03
0

Figure 4.7.2 Evaluation of REDITs using simulated data

a) Sensitivity of REDIT-LLR was evaluated using simulations where group 1 editing level was characterized by beta(α=13.52, β=11.95) and group 2 as beta(α=43.64, β=0.23) with 3 samples per group. b) False-positive rate was evaluated where both group1 and group2 were simulated as beta(α=13.52, β=11.95). c) Sensitivity of REDIT-Regression was evaluated where editing level was characterized as a beta distribution with mean simulated as μ_i= 0.005*A_i+ 0.33 over 3 samples. d) False-positive rate was evaluated where age had no simulated effect on editing levels. Individual points show 100 replicate simulation results. Red dotted lines show the 5% false-positive threshold. Thresholded t-test and thresholded Wilcoxon test = t-test and Wilcoxon rank-sum test run on samples with minimal 10 read coverage. Thresholded linear regression = linear regression on only samples with minimal 10 read coverage. REDITs have highest sensitivities out of all tests that remain within a 5% false-positive rate.

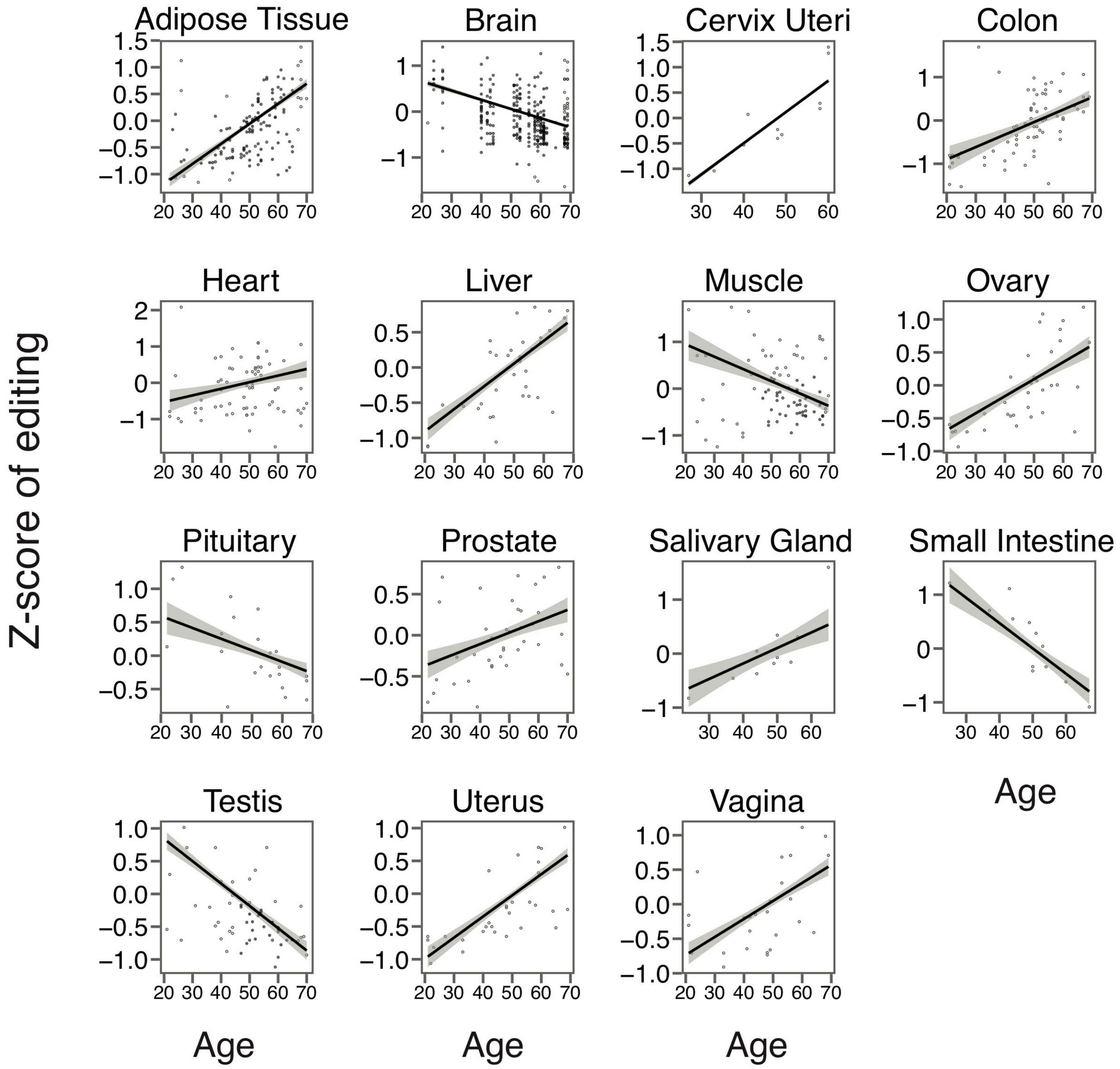


Figure 4.7.3 REDIT-Regression uncovers overall trend of increasing RNA editing over human aging

REDIT-Regression was performed to find RNA editing levels that statistically associate with age. Plots show linear regression lines fit to z-scores of all RNA editing sites that were found significantly associated with age in samples partitioned by GTEx histological type. Only histological types exhibiting homogenous trajectories of editing over age are plotted (Supplemental Table 2a). Grey shading shows the 99% confidence interval from regression. Points show the median z-score per sample.

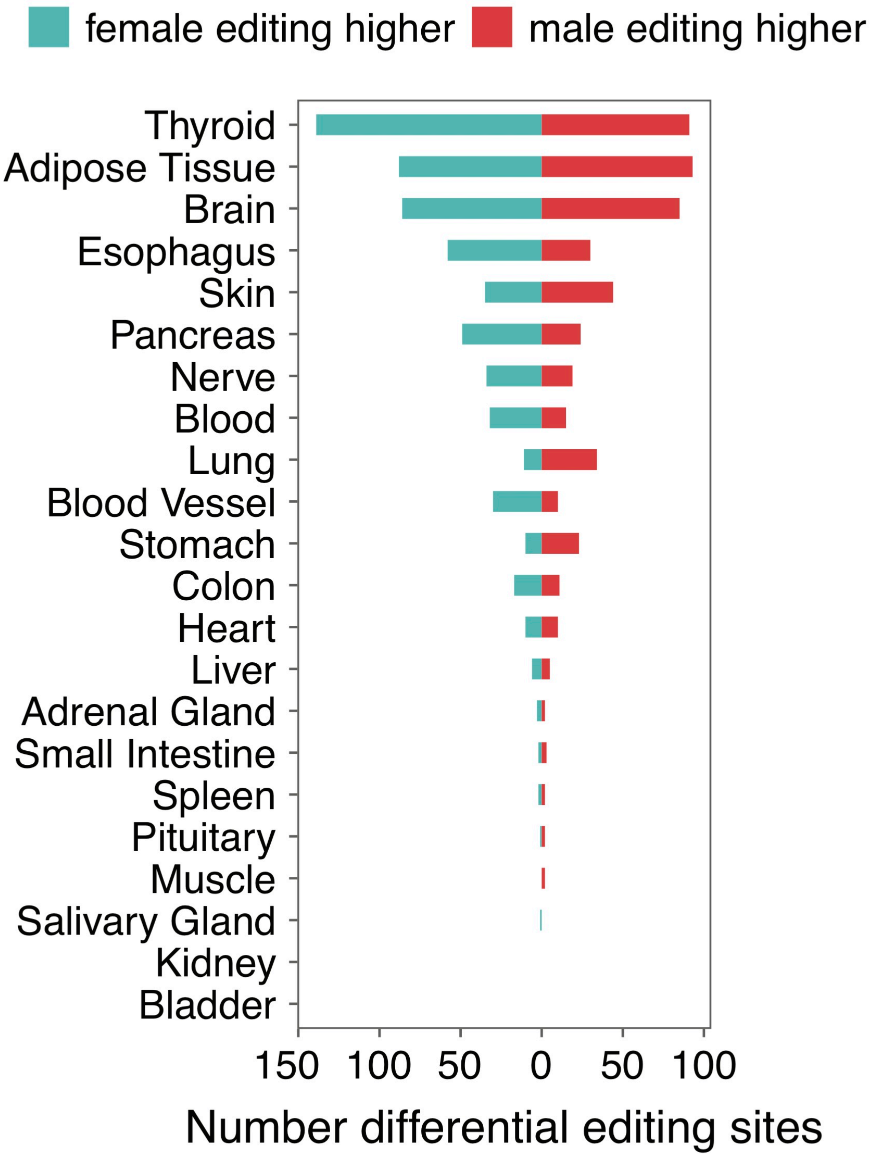


Figure 4.7.4 Gender-biased RNA editing in various tissues

REDIT-Regression was performed across GTEx tissues partitioned by histological type to find editing sites associated with gender. Bargraphs show number of editing sites found significantly more highly edited in males (red) or females (blue). Significant sites defined with FDR < 0.1 and mean difference between genders > 0.05.

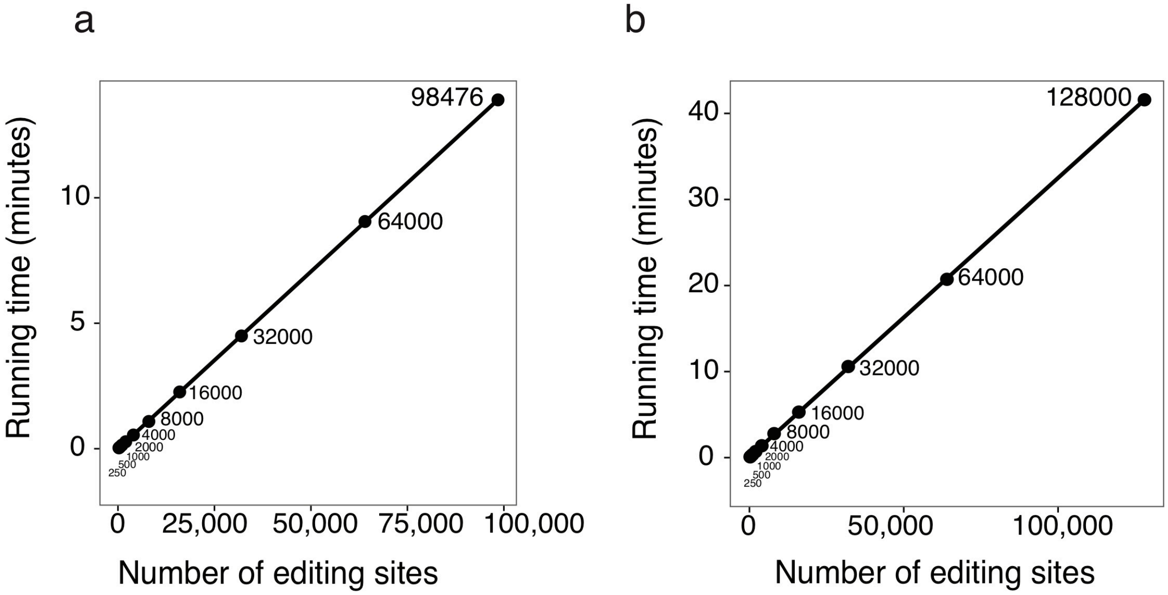


Figure 4.7.5 Computational running-time of REDITs

The points in the scatterplots show the running time (minutes) in analyzing various numbers of editing sites from real datasets. a) Computational running-time of REDIT-LLR. b) Computational running-time of REDIT-Regression. Annotations specify the exact numbers of editing sites tested. The fitted lines were derived by least squares regression.

4.8 References

1 Yablonovitch, A. L., Deng, P., Jacobson, D. & Li, J. B. The evolution and adaptation of A-to-I RNA editing. *PLoS Genet* **13**, e1007064, doi:10.1371/journal.pgen.1007064 (2017).

2 Behm, M. & Ohman, M. RNA Editing: A Contributor to Neuronal Dynamics in the Mammalian Brain. *Trends in genetics : TIG* **32**, 165-175, doi:10.1016/j.tig.2015.12.005 (2016).

3 Rueter, S. M., Dawson, T. R. & Emeson, R. B. Regulation of alternative splicing by RNA editing. *Nature* **399**, 75-80, doi:10.1038/19992 (1999).

4 Feng, Y., Sansam, C. L., Singh, M. & Emeson, R. B. Altered RNA editing in mice lacking ADAR2 autoregulation. *Molecular and cellular biology* **26**, 480-488, doi:10.1128/mcb.26.2.480-488.2006 (2006).

5 Hsiao, Y. E. *et al.* RNA editing in nascent RNA affects pre-mRNA splicing. *Genome research* **28**, 812-823, doi:10.1101/gr.231209.117 (2018).

6 Bahn, J. H. *et al.* Genomic analysis of ADAR1 binding and its involvement in multiple RNA processing pathways. *Nat Commun* **6**, 6355, doi:10.1038/ncomms7355 (2015).

7 Nishikura, K. A-to-I editing of coding and non-coding RNAs by ADARs. *Nature reviews. Molecular cell biology* **17**, 83-96, doi:10.1038/nrm.2015.4 (2016).

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

9 Liddicoat, B. J. *et al.* RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. *Science (New York, N.Y.)* **349**, 1115-1120, doi:10.1126/science.aac7049 (2015).

10 Hideyama, T. *et al.* Profound downregulation of the RNA editing enzyme ADAR2 in ALS spinal motor neurons. *Neurobiology of disease* **45**, 1121-1128, doi:10.1016/j.nbd.2011.12.033 (2012).

11 Tran, S. S. *et al.* Widespread RNA editing dysregulation in brains from autistic individuals. *Nat Neurosci* **22**, 25-36, doi:10.1038/s41593-018-0287-x (2019).

12 Stellos, K. *et al.* Adenosine-to-inosine RNA editing controls cathepsin S expression in atherosclerosis by enabling HuR-mediated post-transcriptional regulation. *Nat Med* **22**, 1140-1150, doi:10.1038/nm.4172 (2016).

13 Han, L. *et al.* The Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in Human Cancers. *Cancer cell* **28**, 515-528, doi:10.1016/j.ccell.2015.08.013 (2015).

14 Paz-Yaacov, N. *et al.* Elevated RNA Editing Activity Is a Major Contributor to Transcriptomic Diversity in Tumors. *Cell reports* **13**, 267-276, doi:10.1016/j.celrep.2015.08.080 (2015).

15 Fumagalli, D. *et al.* Principles Governing A-to-I RNA Editing in the Breast Cancer Transcriptome. *Cell reports* **13**, 277-289, doi:10.1016/j.celrep.2015.09.032 (2015).

16 Ishizuka, J. J. *et al.* Loss of ADAR1 in tumours overcomes resistance to immune checkpoint blockade. *Nature* **565**, 43-48, doi:10.1038/s41586-018-0768-9 (2019).

17 Roth, S. H. *et al.* Increased RNA Editing May Provide a Source for Autoantigens in Systemic Lupus Erythematosus. *Cell reports* **23**, 50-57, doi:10.1016/j.celrep.2018.03.036 (2018).

18 Kang, L. *et al.* Genome-wide identification of RNA editing in hepatocellular carcinoma. *Genomics* **105**, 76-82, doi:10.1016/j.ygeno.2014.11.005 (2015).

19 Qin, Y. R. *et al.* Adenosine-to-inosine RNA editing mediated by ADARs in esophageal squamous cell carcinoma. *Cancer research* **74**, 840-851, doi:10.1158/0008-5472.can-13-2545 (2014).

20 Hwang, T. *et al.* Dynamic regulation of RNA editing in human brain development and disease. *Nat Neurosci* **19**, 1093-1099, doi:10.1038/nn.4337 (2016).

21 Srivastava, P. K. *et al.* Genome-wide analysis of differential RNA editing in epilepsy. *Genome research* **27**, 440-450, doi:10.1101/gr.210740.116 (2017).

22 Paz, N. *et al.* Altered adenosine-to-inosine RNA editing in human cancer. *Genome research* **17**, 1586-1595, doi:10.1101/gr.6493107 (2007).

23 Tan, M. H. *et al.* Dynamic landscape and regulation of RNA editing in mammals. *Nature* **550**, 249-254, doi:10.1038/nature24041 (2017).

24 Quinones-Valdez, G. *et al.* Regulation of RNA editing by RNA-binding proteins in human cells. *Communications biology* **2**, 19, doi:10.1038/s42003-018-0271-8 (2019).

25 Chen, L. *et al.* Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma. *Nat Med* **19**, 209-216, doi:10.1038/nm.3043 (2013).

26 Picardi, E. *et al.* Profiling RNA editing in human tissues: towards the inosinome Atlas. *Sci Rep* **5**, 14941, doi:10.1038/srep14941 (2015).

27 Sun, D. *et al.* MOABS: model based analysis of bisulfite sequencing data. *Genome biology* **15**, R38, doi:10.1186/gb-2014-15-2-r38 (2014).

28 Feng, H., Conneely, K. N. & Wu, H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic acids research* **42**, e69, doi:10.1093/nar/gku154 (2014).

29 Dolzhenko, E. & Smith, A. D. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC bioinformatics* **15**, 215, doi:10.1186/1471-2105-15-215 (2014).

30 Hebestreit, K., Dugas, M. & Klein, H. U. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics (Oxford, England)* **29**, 1647-1653, doi:10.1093/bioinformatics/btt263 (2013).

31 Park, Y., Figueroa, M. E., Rozek, L. S. & Sartor, M. A. MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics (Oxford, England)* **30**, 2414-2422, doi:10.1093/bioinformatics/btu339 (2014).

32 Picardi, E., D'Erchia, A. M., Lo Giudice, C. & Pesole, G. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic acids research* **45**, D750-d757, doi:10.1093/nar/gkw767 (2017).

33 Hung, T. *et al.* The Ro60 autoantigen binds endogenous retroelements and regulates inflammatory gene expression. *Science (New York, N.Y.)* **350**, 455-459, doi:10.1126/science.aac7442 (2015).

34 Li, Z. *et al.* Evolutionary and ontogenetic changes in RNA editing in human, chimpanzee, and macaque brains. *RNA (New York, N.Y.)* **19**, 1693-1702, doi:10.1261/rna.039206.113 (2013).

35 Dillman, A. A. *et al.* mRNA expression, splicing and editing in the embryonic and adult mouse cerebral cortex. *Nat Neurosci* **16**, 499-506, doi:10.1038/nn.3332 (2013).

36 Wahlstedt, H., Daniel, C., Enstero, M. & Ohman, M. Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. *Genome research* **19**, 978-986, doi:10.1101/gr.089409.108 (2009).

37 Pinto, Y., Cohen, H. Y. & Levanon, E. Y. Mammalian conserved ADAR targets comprise only a small fragment of the human editosome. *Genome biology* **15**, R5, doi:10.1186/gb-2014-15-1-r5 (2014).

38 Silvestris, D. A. *et al.* Dynamic inosinome profiles reveal novel patient stratification and gender-specific differences in glioblastoma. *Genome biology* **20**, 33, doi:10.1186/s13059-019-1647-x (2019).

39 Suzuki, S., Nishio, S., Takeda, T. & Komatsu, M. Gender-specific regulation of response to thyroid hormone in aging. *Thyroid research* **5**, 1, doi:10.1186/1756-6614-5-1 (2012).

40 Fuente-Martin, E., Argente-Arizon, P., Ros, P., Argente, J. & Chowen, J. A. Sex differences in adipose tissue: It is not only a question of quantity and distribution. *Adipocyte* **2**, 128-134, doi:10.4161/adip.24075 (2013).

Chapter 5

Concluding remarks

5.1. Summary

RNA editing has paramount function in neuronal development and synaptic transmission[1]. However, no studies have yet globally examined the landscape of editing in Autism spectrum disorders, a prevalent disease affecting 1 in 68 individuals and widely considered a disorder of synaptic transmission[2]. Large scale global studies of RNA editing in disease also lack tools to handle unique difficulties with RNA editing data when calling statistical associations. In this work, we conducted the first global studies of RNA editing in autistic postmortem brains and cortical spheroids to better understand potential roles of RNA editing across critical periods of ASD etiology and multiple implicated brain regions, and to identify convergent trends across the heterogenic landscape of autistic genetic aberrations. In addition, we developed advanced methodologies for RNA editing data to handle commonly desired statistical analyses.

In chapter 2, we examined the global landscape of RNA editing in postmortem brain of ASD patients. The ASD cohort displayed a global trend of downregulated RNA editing within synaptic transmission genes. To ascertain prospective regulators of this trend, we performed the first massive screen for RNA editing regulators from knockdown datasets of RNA editing regulators from ENCODE[3]; from this screen and subsequent experimental validation, we identified FMRP and FXR1P as two novel RNA editing regulators in human and which likely contribute to the hypoediting trend. Strikingly, these findings were convergent across multiple brain regions and across multiple syndromic forms of ASD, implicating RNA editing contributions to the core Autism phenotype. This work has been published in *Nature Neuroscience*, 2019, S.S. Tran, H. Jun, J.H. Bahn, A. Azghadi, G. Ramaswami, E.L. Van Nostrand, T. B. Nguyen, Y.H. Hsiao, C. Lee, G. A. Pratt, V. M. Cerdeno, R. J. Hagerman, G. W. Yeo, D. H. Geschwind, X. Xiao. “Widespread RNA editing dysregulation in brains from autistic individuals.”

In chapter 3, we studied the landscape of RNA editing over fetal development through hundreds of cortical spheroids generated from individuals with idiopathic Autism and individuals harboring a myriad of penetrant Autism-susceptibility genes. RNA editing globally increased as development progressed, and at all time points we found hypoediting of editing within the ASD cohort. Interestingly, the differential RNA editing sites were enriched genes from radial glia, intermediate progenitor cells, and newborn neurons and cellular pathways important for cell maintenance and proliferation. Overall these results suggest that RNA editing could contribute to the abnormal neuronal development in ASD. This work is part of a larger project comparing morphological, immunohistological, and transcriptomic differences between control and ASD cortical spheroids with the labs of Daniel Geschwind and Sergiu Pasca.

In chapter 4, we advanced statistical methodologies to develop the method REDITs (RNA editing tests) which handles unique challenges in RNA editing data. REDITs leverages a beta-binomial model which can compute common statistical associations such as case-control associations or regression association with covariates, while considering limited and variable coverage of individual editing sites. REDITs had higher sensitivity and lower rates of false-positives than the commonly used methods across most RNA editing studies. Applying REDITs to the GTEx dataset, we found editing sites displaying significant male-female sex bias and clarified the overall trajectories of RNA editing over human aging across multiple tissues. This work has been submitted to *Oxford Bioinformatics*, 2019, S. S. Tran, Q. Zhou, X. Xiao. “Statistical inference of differential RNA editing sites from RNA-sequencing data by hierarchical modeling.”

5.2. Conclusions

Our contributions towards RNA editing in ASD and development of methodological tools for RNA editing analysis motivate many important directions for future work.

First, our findings of RNA editing in ASD provide an important high-level view of the aberrant landscape; but they are constrained in resolution. In particular, the findings of RNA editing from bulk tissue and organoids represent an average across a myriad of cell types, brain layers, and intra-cellular locations. Yet, it remains unknown how RNA editing would present in the discrete functional units of brain such as at a single-cell resolution or across different intra-cellular compartments. Single-cell sequencing across multiple ASD and control brains would resolve which cell types experience dysregulated RNA editing and whether the hypoediting trend observed in bulk tissue and organoids occurs uniquely within one cell type or multiple. Unfortunately, commonly used single-cell sequencing technologies only capture 3’ ends of transcripts and have limited coverage hundreds of thousands to a couple of million reads per cell[4,5]. A comprehensive single-cell study of RNA editing in ASD will require full length transcript coverage and tens of millions of reads sequenced per cell.

RNA-sequencing in bulk tissue and organoids also does not inform which cellular compartments experience dysregulated RNA editing. Brain cells have compartmentalized milieus of gene expression pools[6], and likely also tailor their RNA editomes to suit diverse functional purposes in nucleus, mitochondria, cellular junctions, synapses, trans-vesicles, or other cellular compartments. Any observed dysregulated RNA editing specific to cellular compartments would greatly pinpoint the etiological contribution to ASD. Excitingly, a technology for accomplishing this, APEX-seq, was just recently developed[7]. APEX-seq utilizes APEX2 fusion proteins for precise labelling, isolation, and sequencing of transcriptomes within specific subcellular compartments[7]. The application of APEX-seq within iPSC models, organoids, or animal models would greatly accelerate understanding RNA editing in ASD at a subcellular resolution.

Expanding upon with previous literature, we observed that RNA editing in brain generally increased over age stemming from fetal development to adulthood, and then remains relatively stationary or slightly decreases from adulthood onwards. It remains unknown which brain cell types harbor this increasing trajectory. Interestingly, the slope of increasing RNA editing strongly correlated with major periods of human brain maturation, particularly fetal development[2]. Furthermore, out of all brain cells, neurons have the highest levels of RNA editing[8]. This evidence suggests the observed increase of RNA editing in bulk tissue is driven by neurons and is reflecting proliferation and maturation of neurons over these time periods. Single cell sequencing of brain over multiple developmental time points would clarify this and could potentially uncover distinct trajectories of editing in other cell types. Any observations of trends distinct from neurons would open new fields for eliciting the role of editing in other novel non-neuronal brain functions. Even if editing is primarily a consequence of developing neurons, future work could try attenuating global or subsets of neuronal editing sites to decipher importance for various facets of neuronal development, which is also relevant to ASD given the observed hypoediting at all timepoints of ASD included in our findings.

Future studies should elaborate on the functional relevance of hypoediting observed across ASD brains and organoids. Only a couple dozen editing sites have been found evolutionarily conserved between human and mouse[9]. Most of these sites reside in synaptic genes, and, coincidentally, included all of the dozen of studied recoding sites found in human brain from previous literature[9] (Chapter 1 introduction). Interestingly, many of these sites also exhibited significant downregulation in brains of Autistic patients, suggesting a direct effect by RNA editing on aberrant synaptic transmission. In addition to these conserved sites, we also found a dozen non-conserved recoding sites downregulated in the ASD brains. Non-conserved RNA editing sites have been shown to have critical roles in human physiology. For example, non-conserved editing in the 3'UTR of in cathepsin S mRNA regulates its post-transcriptional stability and is associated with changes in cathepsin S levels in patients with atherosclerotic vascular diseases[10]. Therefore, the isoforms generated by these novel recoding sites could serve important roles in brain and contribute to human-specific facets of ASD disease etiology.

The predominate majority of hypoediting, however, occurred in noncoding regions of RNA, particularly introns and 3'UTRs. The noncoding editing sites resided mostly in genes responsible for synaptic development and transmission, suggesting they might serve complementary roles to recoding sites towards synaptic function. It is generally unknown how noncoding RNA editing affects neurons, but previous studies have found evidence for a role of noncoding RNA editing in neuronal RNA processing and regulation such as microRNA targeting[11], circular RNA formation[12,13], and alternative splicing[14,15]. Dysfunctional relationships between immune cells and neurons in synaptic pruning is increasing considered as an integral part of ASD etiology[2]. Though not demonstrated yet in neuronal tissue, RNA editing was found to modulate the activation of immune cells in embryonic mouse and cancer cells[16-18], and thus might mechanistically constitute part of the aberrant immune states in ASD. Future studies should also consider how noncoding RNA editing changes in response to depolarizing or desensitizing responses to various external stimuli and neurotransmitters. A study of CA1 neurons in hippocampus found that RNA editing affecting flip/flop exon usages in *GRIA2* glutamate receptor gene varied depending on cell state[19]. Elevated activity led to increased editing and flip inclusion which promoted slower receptor desensitization[20]. Future experiments coupling inducement of various neuronal states with RNA sequencing could address the role of these noncoding RNA editing sites in promoting neuronal excitability, inhibition, and plasticity on a global scale.

The mechanistic underpinnings of the observed hypoediting in ASD also remains incompletely understood. The FMRP and FXR1P proteins, which were involved in the regulation of hypoediting in ASD, explained 13% and 14% percent of editing variance across samples, indicating that other factors remain undiscovered. We initially searched for prospective RNA editing regulators in ASD by screening hundreds of RNA binding protein gene knockdowns from the ENCODE consortium for changes in RNA editing[3]. Unfortunately, the ENCODE consortium used HepG2 (liver) and K562 (lung) cells, which do not express brain specific proteins, many of which are relevant to ASD[21,22]. An important extension of this work would be to expand the gene knockdown screens to brain-specific proteins in neuronal cultures.

The observed aberrations of RNA editing in ASD prompt prospective directions for novel therapeutic development. As Fragile X proteins seem important contributors towards hypoediting, one might restore normal editing levels by therapeutically raising expression of *FMR1* and *FXR1*, or any other relevant RNA binding proteins that will be discovered. A recent study developed a therapeutic using CRISPRa that enables increasing expression of target genes through adenovirus injection[23]. Using this technology to increase expression of *SIM1* reverted obesity in mice down to normal weights[23]. Another therapeutic option is antisense oligonucleotides (ASO); FDA approval in 2016 for ASO treatment of Spinal muscular dystrophy[24] has opened the possibility of applying similar oligo treatments in ASD if repressors of RNA editing get discovered. Recent technological advances also make directly restoring editing levels of target hypoedited sites possible. The Zhang lab in MIT developed catalytically inactive Cas13 fused to ADAR2 to precisely edit target adenosines on RNA[25]. An alternative approach, “leveraging endogenous ADAR for programmable editing of RNA” (LEAPER), used synthetic RNAs to recruit endogenous ADARs to target adenosines with high specificity[26]. Future studies of RNA editing may enable prioritization of the most relevant sites to ASD pathology.

Lastly, in Chapter 4 we generated more advanced statistical methodologies for performing statistical associations in RNA editing studies. Our proposed beta-binomial statistical model, which handles both uncertainty from read coverage and variance between multiple samples, should propitiously function as a framework to develop additional methodologies for RNA editing. We used our statistical model to handle RNA editing associations between case-control groups and for regression against fixed covariates such as age. Yet, many other common analyses still need methodologies better suited for RNA editing data. For example, many gene expression studies of ASD applied Principal component analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) as methods for data dimensionality reduction to summarize overall trends of groups of genes[5,27,28]. As another example, many large datasets such as single cell datasets require mixed-effects modelling because multiple samples can be classified by hierarchical categories[5] (e.g. same patient, same cell line, same induction). Unfortunately, PCA and t-SNE and mixed-effects all disallow missing data and assume accurate data measurements; in contrast, RNA editing data is sparse and contains varying accuracies of editing level measurements depending on coverage. Future work could adapt our beta-binomial model to handle data reduction, mixed-effects modelling, and other analyses for RNA editing data. In larger scope, these models could also be used in other biological datasets that suffer from similar constraints, such as massive parallel reporter assays and RNA splicing measured from RNA-sequencing.

5.3 References

1 Behm, M. & Ohman, M. RNA Editing: A Contributor to Neuronal Dynamics in the Mammalian Brain. *Trends in genetics : TIG* **32**, 165-175, doi:10.1016/j.tig.2015.12.005 (2016).

2 de la Torre-Ubieta, L., Won, H., Stein, J. L. & Geschwind, D. H. Advancing the understanding of autism disease mechanisms through genetics. *Nat Med* **22**, 345-361, doi:10.1038/nm.4071 (2016).

3 Quinones-Valdez, G. *et al.* Regulation of RNA editing by RNA-binding proteins in human cells. *Communications biology* **2**, 19, doi:10.1038/s42003-018-0271-8 (2019).

4 Habib, N. *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature methods* **14**, 955-958, doi:10.1038/nmeth.4407 (2017).

5 Velmeshev, D. *et al.* Single-cell genomics identifies cell type-specific molecular changes in autism. *Science (New York, N.Y.)* **364**, 685-689, doi:10.1126/science.aav8130 (2019).

6 Holt, C. E., Martin, K. C. & Schuman, E. M. Local translation in neurons: visualization and function. *Nature structural & molecular biology* **26**, 557-566, doi:10.1038/s41594-019-0263-5 (2019).

7 Fazal, F. M. *et al.* Atlas of Subcellular RNA Localization Revealed by APEX-Seq. *Cell* **178**, 473-490.e426, doi:10.1016/j.cell.2019.05.027 (2019).

8 Picardi, E., Horner, D. S. & Pesole, G. Single-cell transcriptomics reveals specific RNA editing signatures in the human brain. *RNA (New York, N.Y.)* **23**, 860-865, doi:10.1261/rna.058271.116 (2017).

9 Pinto, Y., Cohen, H. Y. & Levanon, E. Y. Mammalian conserved ADAR targets comprise only a small fragment of the human editosome. *Genome biology* **15**, R5, doi:10.1186/gb-2014-15-1-r5 (2014).

10 Stellos, K. *et al.* Adenosine-to-inosine RNA editing controls cathepsin S expression in atherosclerosis by enabling HuR-mediated post-transcriptional regulation. *Nat Med* **22**, 1140-1150, doi:10.1038/nm.4172 (2016).

11 Ekdahl, Y., Farahani, H. S., Behm, M., Lagergren, J. & Ohman, M. A-to-I editing of microRNAs in the mammalian brain increases during development. *Genome research* **22**, 1477-1487, doi:10.1101/gr.131912.111 (2012).

12 Ivanov, A. *et al.* Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals. *Cell reports* **10**, 170-177, doi:10.1016/j.celrep.2014.12.019 (2015).

13 Jeck, W. R. *et al.* Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA (New York, N.Y.)* **19**, 141-157, doi:10.1261/rna.035667.112 (2013).

14 Hsiao, Y. E. *et al.* RNA editing in nascent RNA affects pre-mRNA splicing. *Genome research* **28**, 812-823, doi:10.1101/gr.231209.117 (2018).

15 Rueter, S. M., Dawson, T. R. & Emeson, R. B. Regulation of alternative splicing by RNA editing. *Nature* **399**, 75-80, doi:10.1038/19992 (1999).

16 Mannion, N. M. *et al.* The RNA-editing enzyme ADAR1 controls innate immune responses to RNA. *Cell reports* **9**, 1482-1494, doi:10.1016/j.celrep.2014.10.041 (2014).

17 Vitali, P. & Scadden, A. D. Double-stranded RNAs containing multiple IU pairs are sufficient to suppress interferon induction and apoptosis. *Nature structural & molecular biology* **17**, 1043-1050, doi:10.1038/nsmb.1864 (2010).

18 Liddicoat, B. J. *et al.* RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. *Science (New York, N.Y.)* **349**, 1115-1120, doi:10.1126/science.aac7049 (2015).

19 Balik, A., Penn, A. C., Nemoda, Z. & Greger, I. H. Activity-regulated RNA editing in select neuronal subfields in hippocampus. *Nucleic acids research* **41**, 1124-1134, doi:10.1093/nar/gks1045 (2013).

20 Grosskreutz, J. *et al.* Kinetic properties of human AMPA-type glutamate receptors expressed in HEK293 cells. *The European journal of neuroscience* **17**, 1173-1178, doi:10.1046/j.1460-9568.2003.02531.x (2003).

21 Ruzzo, E. K. *et al.* Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks. *Cell* **178**, 850-866.e826, doi:10.1016/j.cell.2019.07.015 (2019).

22 Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215-1233, doi:10.1016/j.neuron.2015.09.016 (2015).

23 Matharu, N. *et al.* CRISPR-mediated activation of a promoter or enhancer rescues obesity caused by haploinsufficiency. *Science (New York, N.Y.)* **363**, doi:10.1126/science.aau0629 (2019).

24 Wood, M. J. A., Talbot, K. & Bowerman, M. Spinal muscular atrophy: antisense oligonucleotide therapy opens the door to an integrated therapeutic landscape. *Human molecular genetics* **26**, R151-r159, doi:10.1093/hmg/ddx215 (2017).

25 Cox, D. B. T. *et al.* RNA editing with CRISPR-Cas13. *Science (New York, N.Y.)* **358**, 1019-1027, doi:10.1126/science.aaq0180 (2017).

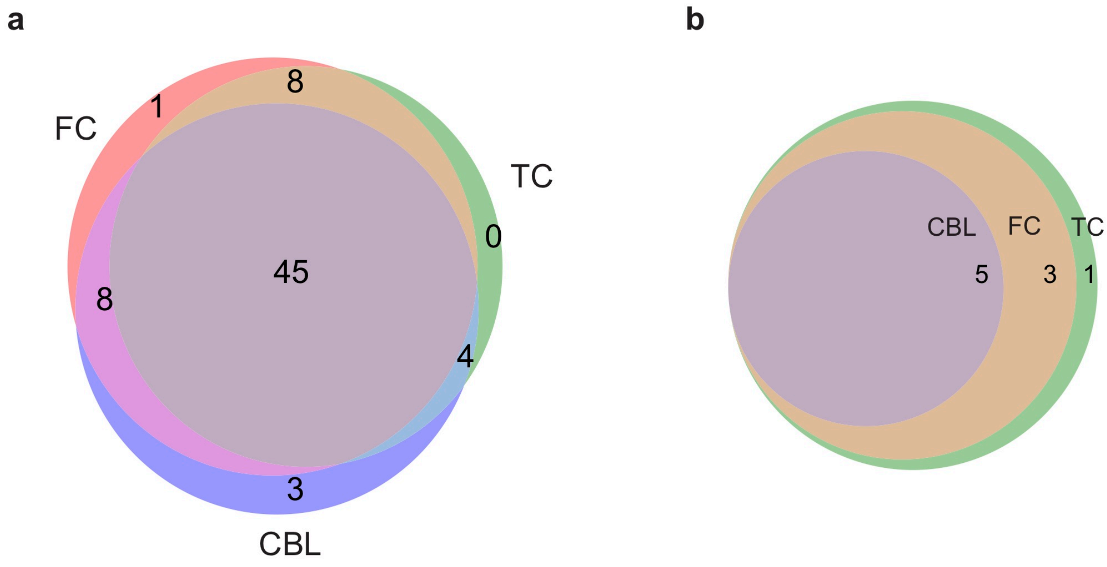
26 Qu, L. *et al.* Programmable RNA editing by recruiting endogenous ADAR using engineered RNAs. *Nature biotechnology*, doi:10.1038/s41587-019-0178-z (2019).

27 Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008-1021, doi:10.1016/j.cell.2013.10.031 (2013).

28 Parikshak, N. N. *et al.* Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* **540**, 423-427, doi:10.1038/nature20612 (2016).

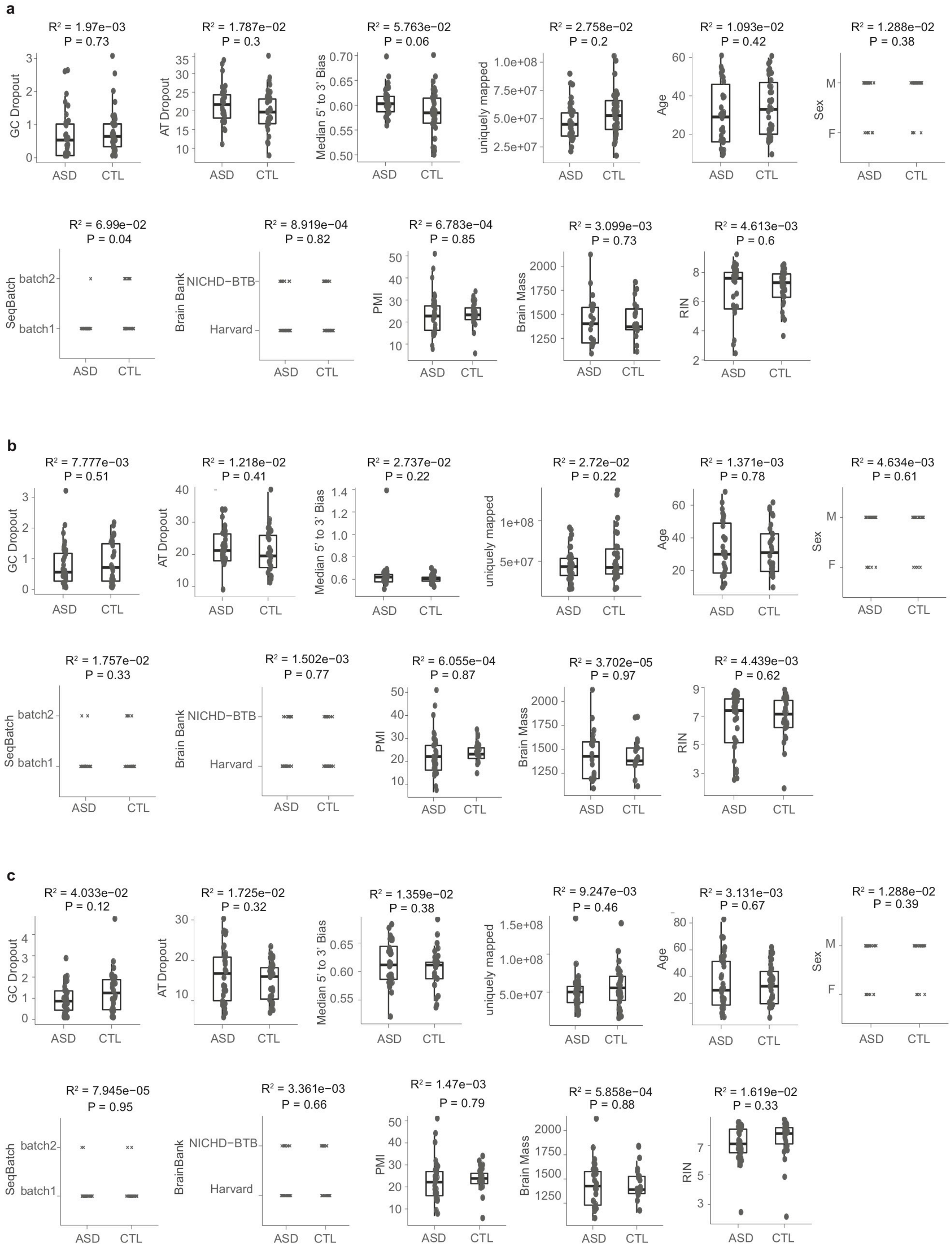
Appendix

Supplemental Figures for Chapter 2



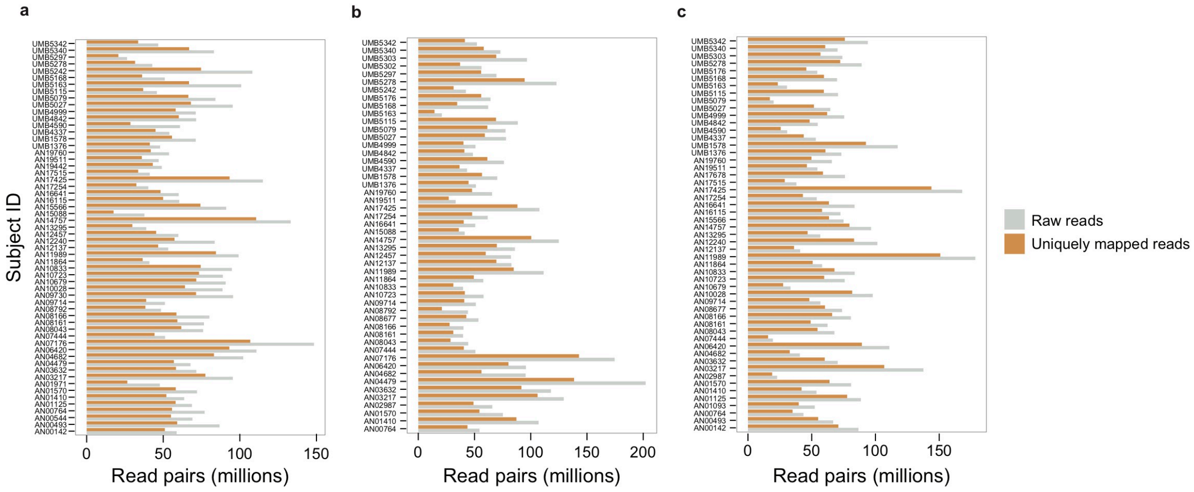
Supplemental Figure 2.1 Unique subjects in idiopathic ASD, control, dup15q samples

a, Venn diagram showing idiopathic ASD and control samples derived per unique subject across the 3 brain regions. b, Venn diagram showing dup15q samples derived per unique subject. FC: frontal cortex, TC, temporal cortex, CBL, cerebellum.


a
R² = 1.97e−03
P = 0.73
GC Dropout
ASD
CTL
R² = 1.787e−02
P = 0.3
AT Dropout
R² = 5.763e−02
P = 0.06
Median 5' to 3' Bias
R² = 2.758e−02
P = 0.2
uniquely mapped
R² = 1.093e−02
P = 0.42
Age
R² = 1.288e−02
P = 0.38
Sex
M
F
R² = 6.99e−02
P = 0.04
SeqBatch
batch2
batch1
R² = 8.919e−04
P = 0.82
Brain Bank
NICHD−BTB
Harvard
R² = 6.783e−04
P = 0.85
PMI
R² = 3.099e−03
P = 0.73
Brain Mass
R² = 4.613e−03
P = 0.6
RIN
b
R² = 7.777e−03
P = 0.51
R² = 1.218e−02
P = 0.41
R² = 2.737e−02
P = 0.22
R² = 2.72e−02
P = 0.22
R² = 1.371e−03
P = 0.78
R² = 4.634e−03
P = 0.61
R² = 1.757e−02
P = 0.33
R² = 1.502e−03
P = 0.77
R² = 6.055e−04
P = 0.87
R² = 3.702e−05
P = 0.97
R² = 4.439e−03
P = 0.62
c
R² = 4.033e−02
P = 0.12
R² = 1.725e−02
P = 0.32
R² = 1.359e−02
P = 0.38
R² = 9.247e−03
P = 0.46
R² = 3.131e−03
P = 0.67
R² = 1.288e−02
P = 0.39
R² = 7.945e−05
P = 0.95
R² = 3.361e−03
P = 0.66
BrainBank
R² = 1.47e−03
P = 0.79
R² = 5.858e−04
P = 0.88
R² = 1.619e−02
P = 0.33

Supplemental Figure 2.2 Comparison of technical covariates between ASD and control (CTL) samples.

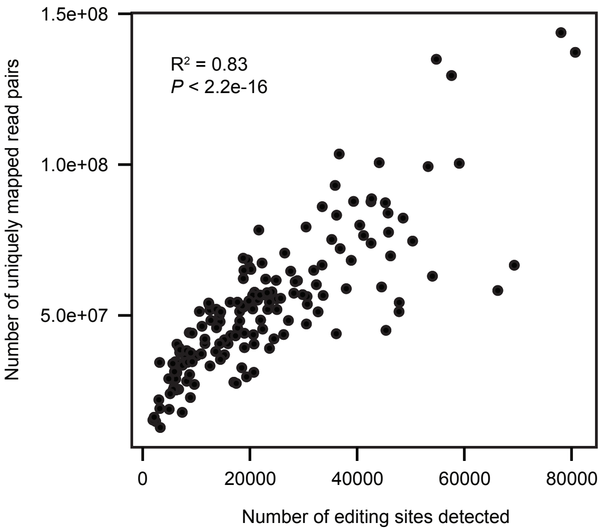
Pearson correlation was calculated. No p values passed the Bonferroni corrected cutoff 0.0045. a, frontal cortex b, temporal cortex, and c, cerebellum. N=45 samples in each comparison. Boxplot definitions: center=median, lower hinge=25th percentile, upper hinge=75th percentile, min and max extend to observations at most 1.5 * IQR



Supplemental Figure 2.3

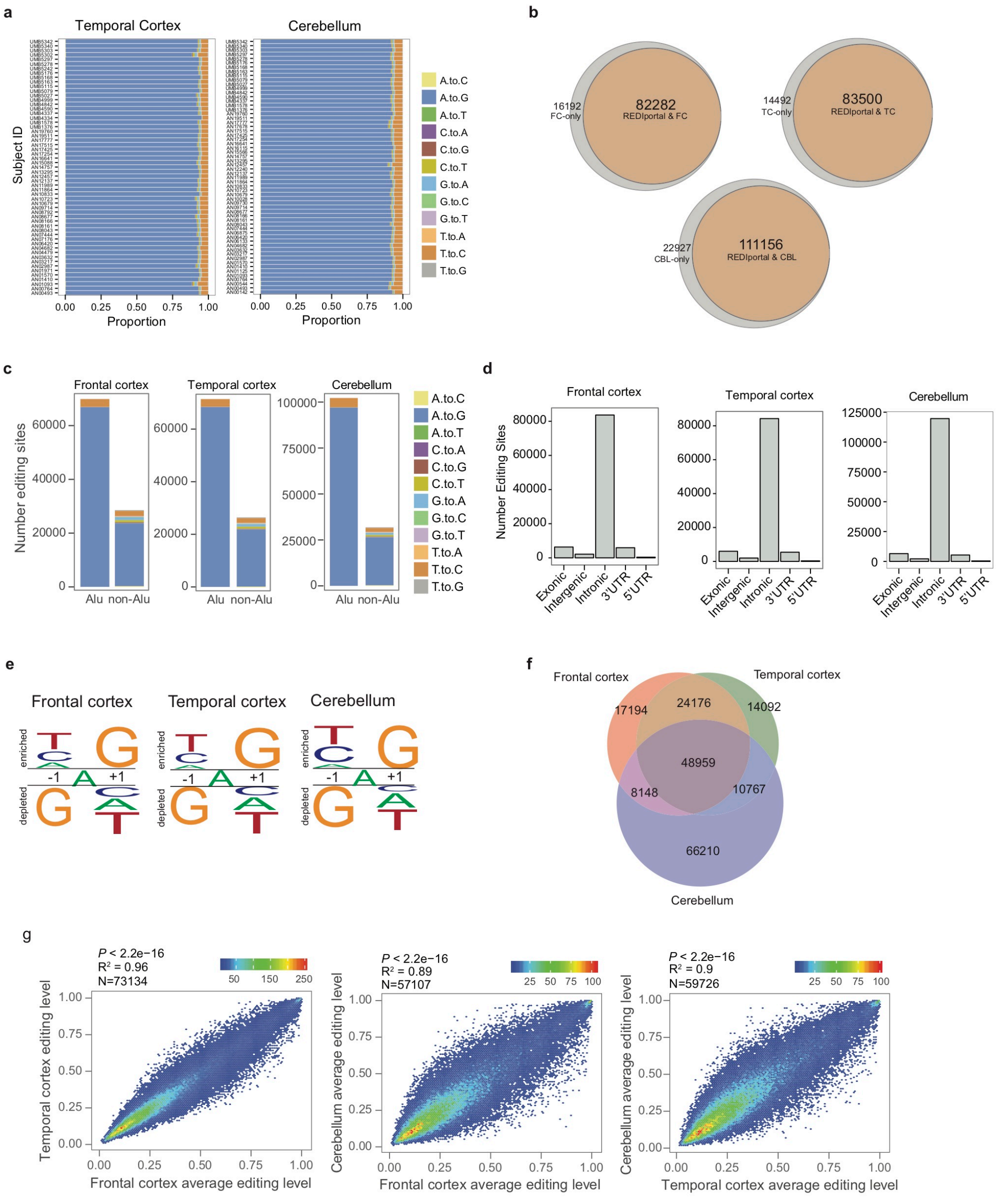
Numbers of raw read pairs (gray) and uniquely mapped read pairs (orange) per sample.

a, frontal cortex. b, temporal cortex. c, cerebellum.



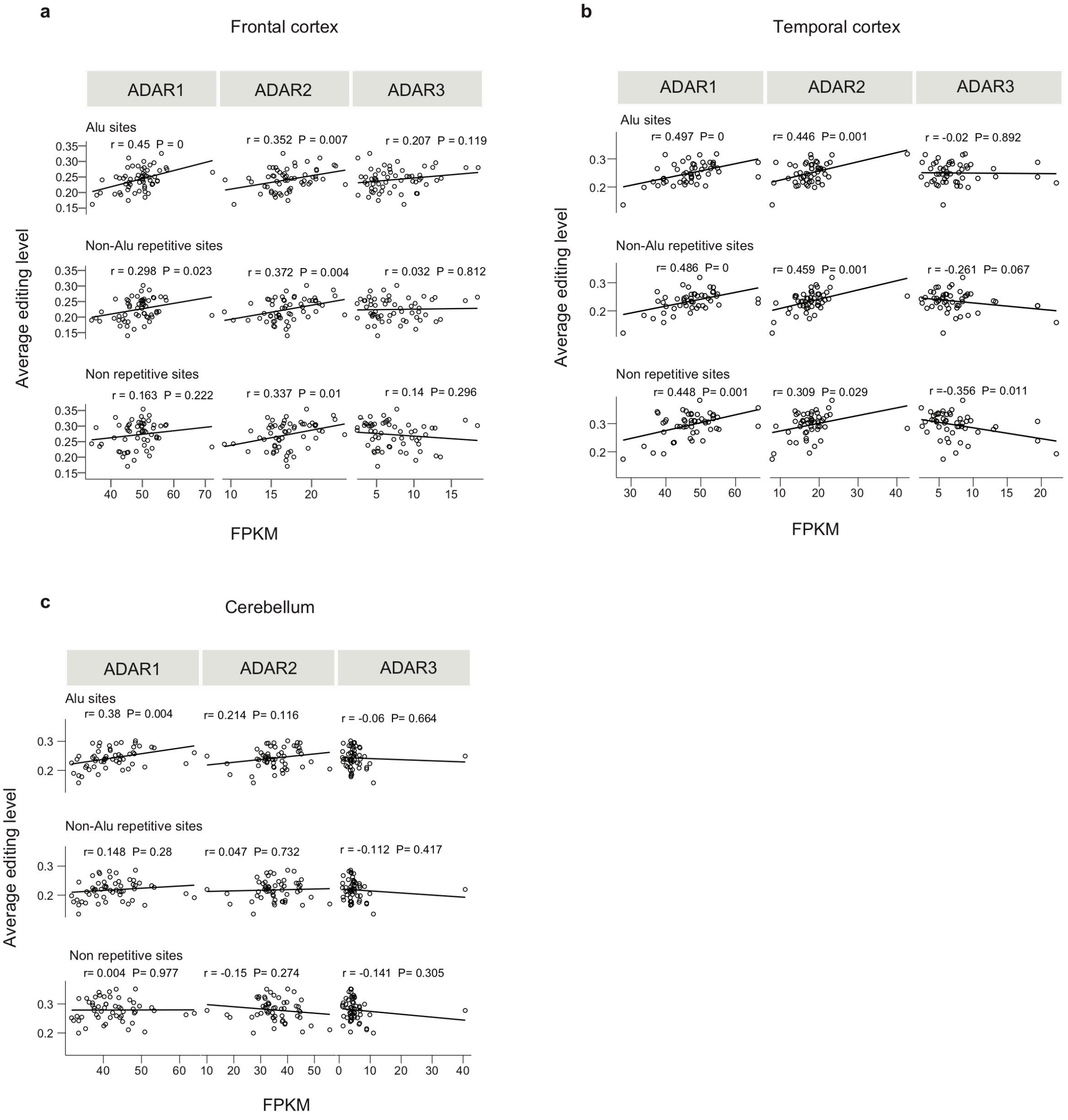
Supplemental Figure 2.4

Number of editing sites detected per brain sample is highy correlated with sequencing depth. All samples from all three brain regions are shown. Pearson correlation was calculated. N= 179 total samples.



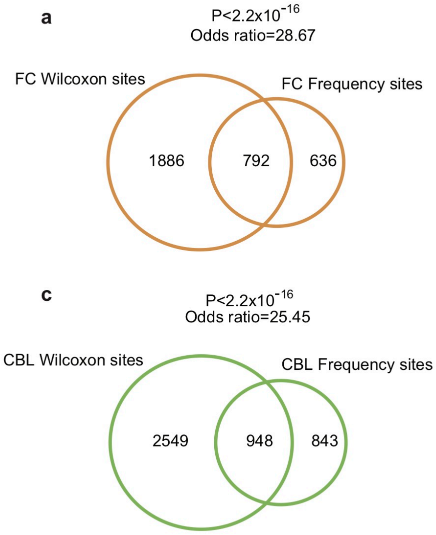
Supplemental Figure 2.5 Global RNA editomes in different brain regions and their comparisons.

a, Fraction of all types of predicted RNA editing sites identified in the RNA-Seq data of temporal cortex and cerebellum of each subject. b, Overlap of RNA editing sites identified in this study with those in the REDIportal database. FC: frontal cortex. TC: temporal cortex. CBL: cerebellum. c, Number of different types of predicted editing sites in alu and non-alu regions. d, Distribution of common editing sites in different types of genomic regions. e, Sequence consensus in the immediate upstream and downstream positions of common RNA editing sites. f, Overlap of common editing sites across brain regions. g, Correlation density graphs of average editing levels of common editing sites between brain regions. N=number of editing sites.



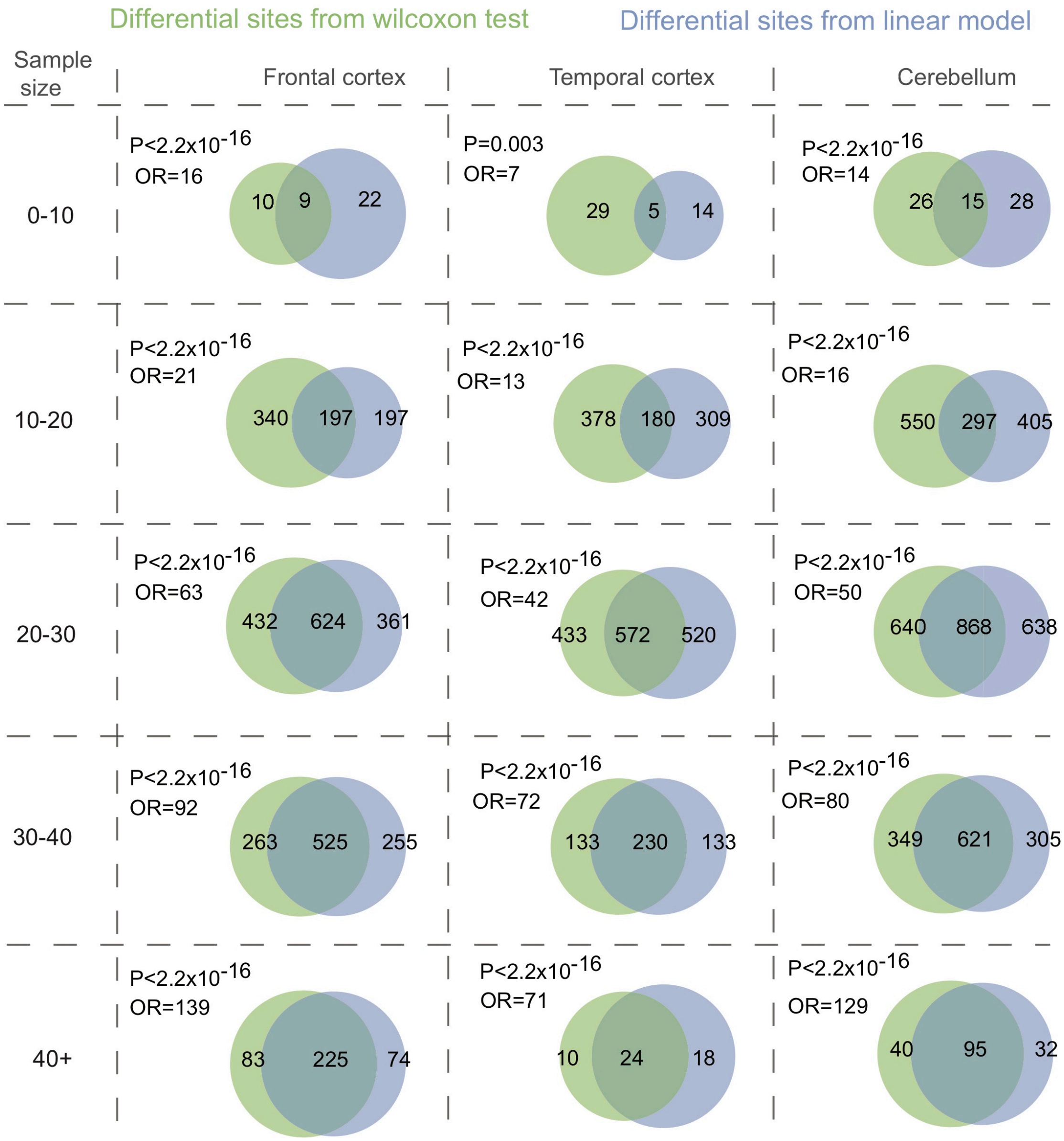
Supplemental Figure 2.6

Correlation between ADAR mRNA expression (FPKM) and average editing levels of different categories of editing sites. a. Alu sites, b.non-Alu repetitive sites, c.non repetitive sites. All correlations use Pearson's method. N=62, 57, and 60 samples in frontal cortex, temporal cortex, and cerebellum respectively.



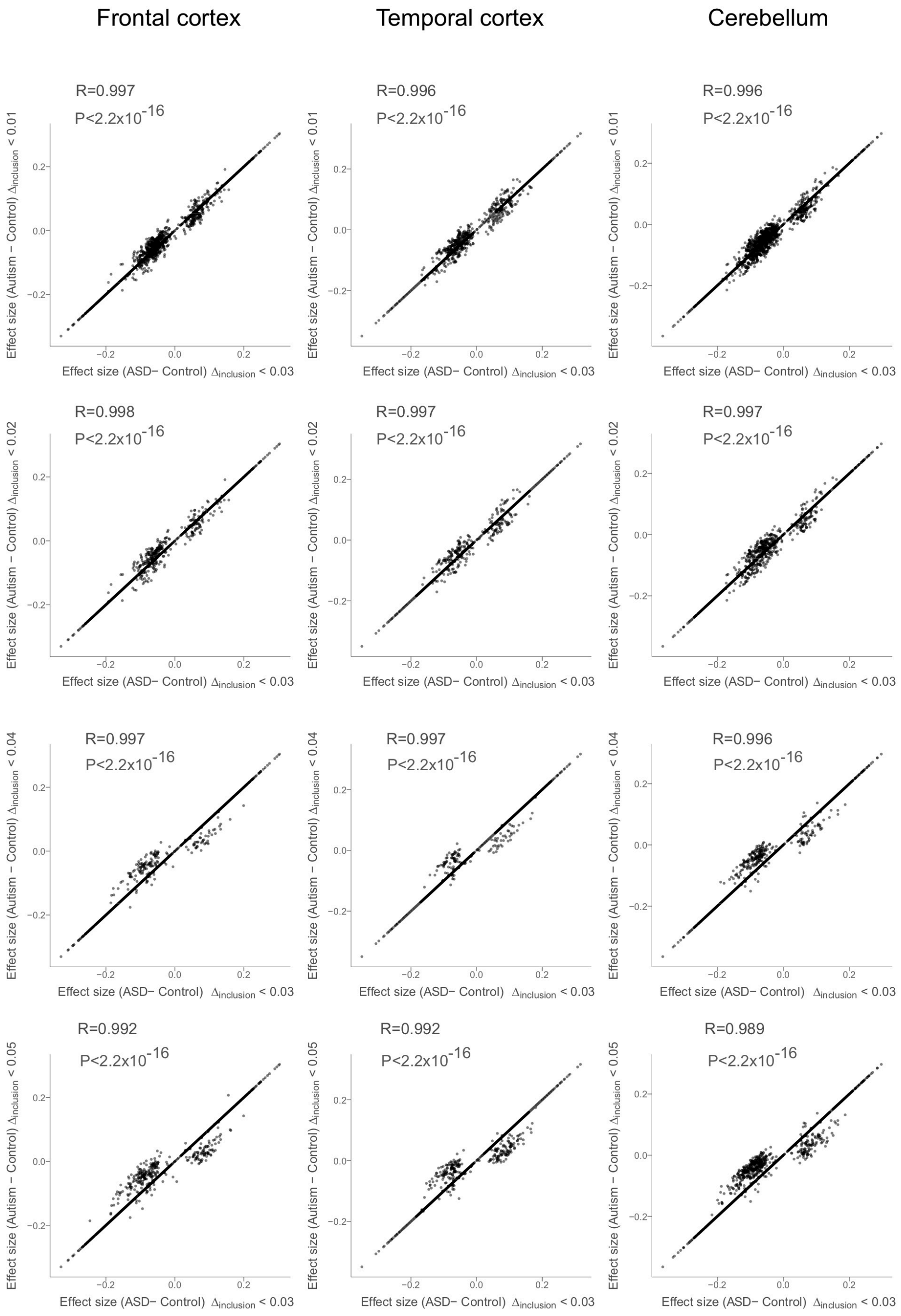
Supplemental Figure 2.7

Overlap between wilcoxon differential editing sites and population frequency differential sites Left circle: Differential editing sites identified from wilcoxon test on editing level disparities. Right circle: Differential editing sites identified from differences in population frequencies. a, frontal cortex, b, temporal cortex, c, cerebellum. P values calculated via two-tailed Fisher's Exact test.



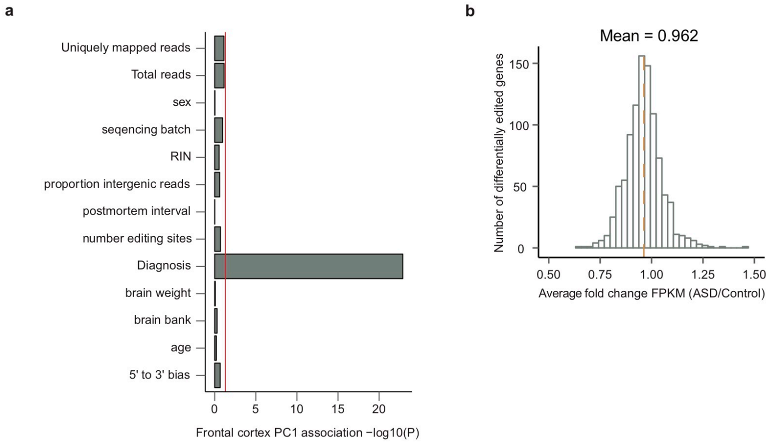
Supplemental Figure 2.8 Robustness of differential editing sites to variations in statistical modeling.

Green circles: Sets of differential editing sites identified through adaptive wilcoxon rank-sum test used in this study. Blue circles: Sets of differnetial editing sites identified through multivariable linear model, y=editing level, x=read integrity number, sex, age, and diagnosis (i.e. ASD, control). Effect of N (sample size) on robustness of linear model was evaluated by binning editing sites by sample size (N = 0-10, 10-20, 20-30, 30-40, 40+). Across frontal cortex, temporal cortex, and cerebellum, differential sites identified are significantly overlapping (P=two-tailed Fisher's Exact test, OR=odds ratio) and converge greatest at larger sample sizes.


Frontal cortex
Temporal cortex
Cerebellum
R=0.997
P<2.2x10^{-16}
R=0.996
P<2.2x10^{-16}
R=0.996
P<2.2x10^{-16}
Effect size (Autism − Control) $\Delta_{\text{inclusion}}$ < 0.01
Effect size (ASD− Control) $\Delta_{\text{inclusion}}$ < 0.03
R=0.998
P<2.2x10^{-16}
R=0.997
P<2.2x10^{-16}
R=0.997
P<2.2x10^{-16}
Effect size (Autism − Control) $\Delta_{\text{inclusion}}$ < 0.02
Effect size (ASD− Control) $\Delta_{\text{inclusion}}$ < 0.03
R=0.997
P<2.2x10^{-16}
R=0.997
P<2.2x10^{-16}
R=0.996
P<2.2x10^{-16}
Effect size (Autism − Control) $\Delta_{\text{inclusion}}$ < 0.04
Effect size (ASD− Control) $\Delta_{\text{inclusion}}$ < 0.03
R=0.992
P<2.2x10^{-16}
R=0.992
P<2.2x10^{-16}
R=0.989
P<2.2x10^{-16}
Effect size (Autism − Control) $\Delta_{\text{inclusion}}$ < 0.05
Effect size (ASD− Control) $\Delta_{\text{inclusion}}$ < 0.03
−0.2
0.0
0.2

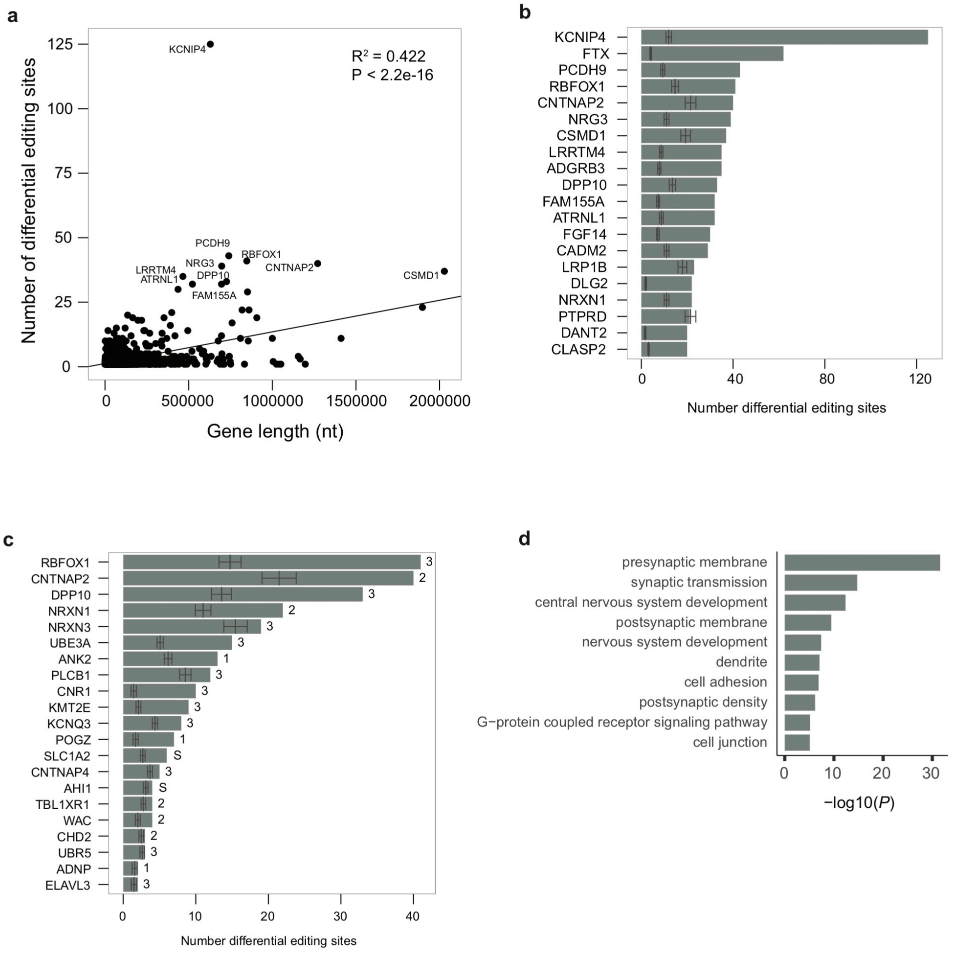
Supplemental Figure 2.9 Impact of M_{ai} and M_{ci} on differential editing levels.

M_{ai} and M_{ci} (see Methods) presented here as $\Delta_{inclusion}$. The panels show correlation of differential editing levels (ASD- Control) between $\Delta_{inclusion} < 0.03$ (this study) and alternative $\Delta_{inclusion}$ thresholds ($\Delta_{inclusion} < 0.01$, 0.02, 0.03, and 0.04). Differential editing levels remain unchanged upon varying $\Delta_{inclusion}$. Graph annotations: R (correlation coefficient) and P value via Pearson correlation. N=3314, 2412, and 4340 editing sites in frontal cortex, temporal cortex, and cerebellum respectively.



Supplemental Figure 2.10 Examination of differential editing relative to potential confounding variables.

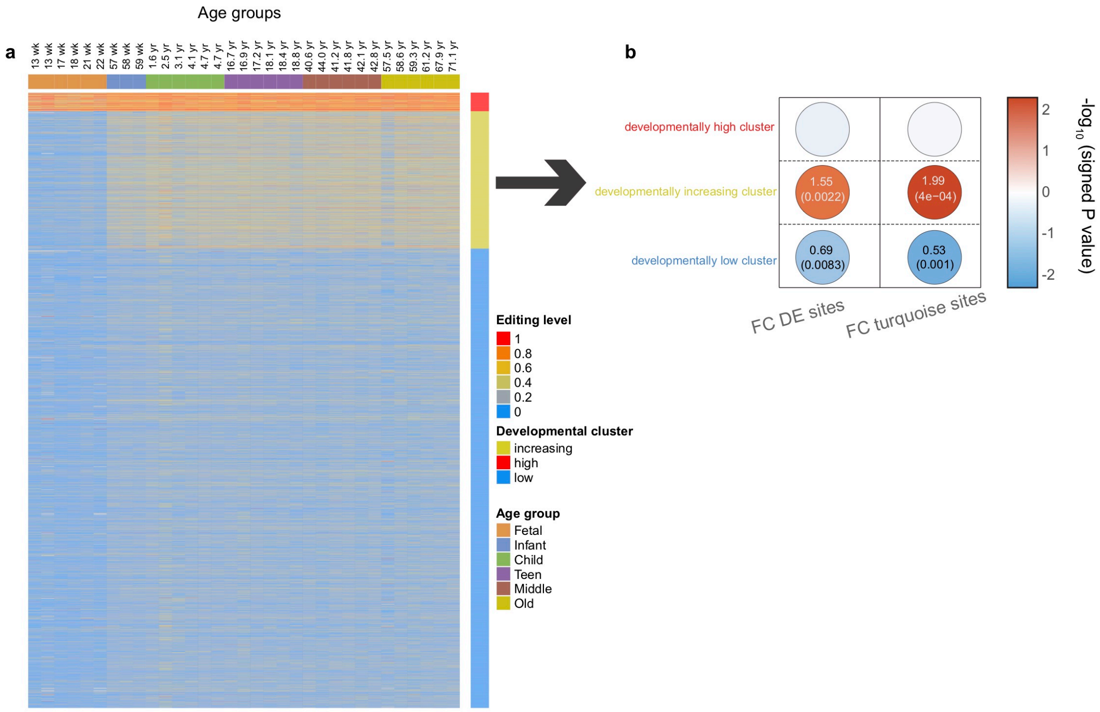
a, Correlation between the first principal component of differential editing sites from frontal cortex and confounding variables and diagnosis; N=62 samples. b, Fold change of gene expression levels (ASD/control) for genes harboring differential editing sites in frontal cortex. The average fold change shown above the plot.



Supplemental Figure 2.11 Genes enriched with differential editing sites.

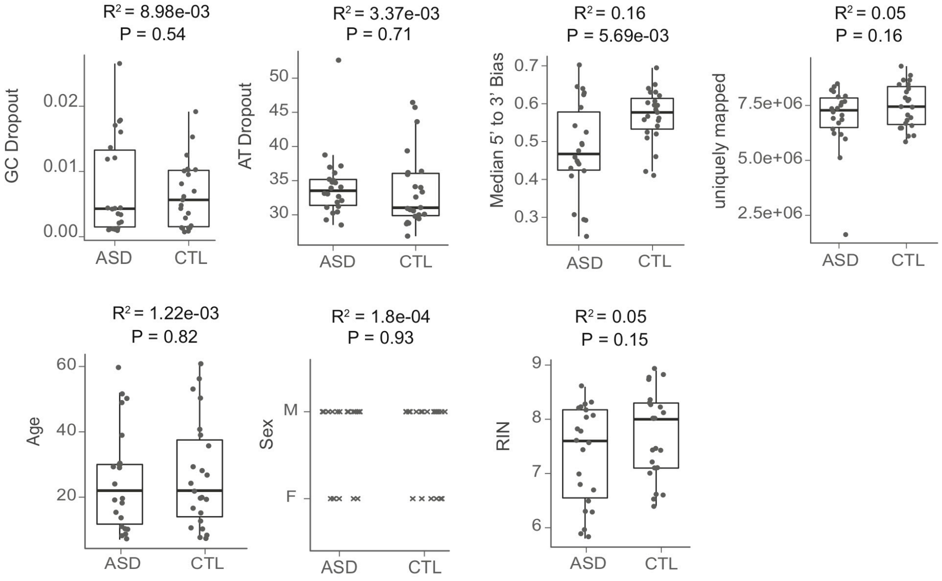
a, The number of differential editing sites per gene correlates with gene length. Pearson correlation was calculated. b, Genes with the largest numbers of differential editing sites. In these genes, enrichment of differential editing sites is not explained by gene length (except for PTPRD). Error bars show the 95% confidence interval of expected number of editing sites calculated via a linear model trained on gene length (see

Methods). c, Genes with differential editing sites that are also known ASD susceptibility genes. ASD susceptibility genes were required to have a SFARI score < 4 (scores indicated to the right of the bars). S = syndromic, 1 = high confidence, 2 = strong candidate, 3 = suggestive evidence. a-c, N=1189 genes in lnear model. d, Gene ontology of differential editing sites correlated with gene expression. Top 10 GO terms are shown. P-value calculated using one-tailed Gaussian test (Methods); N=148 genes.



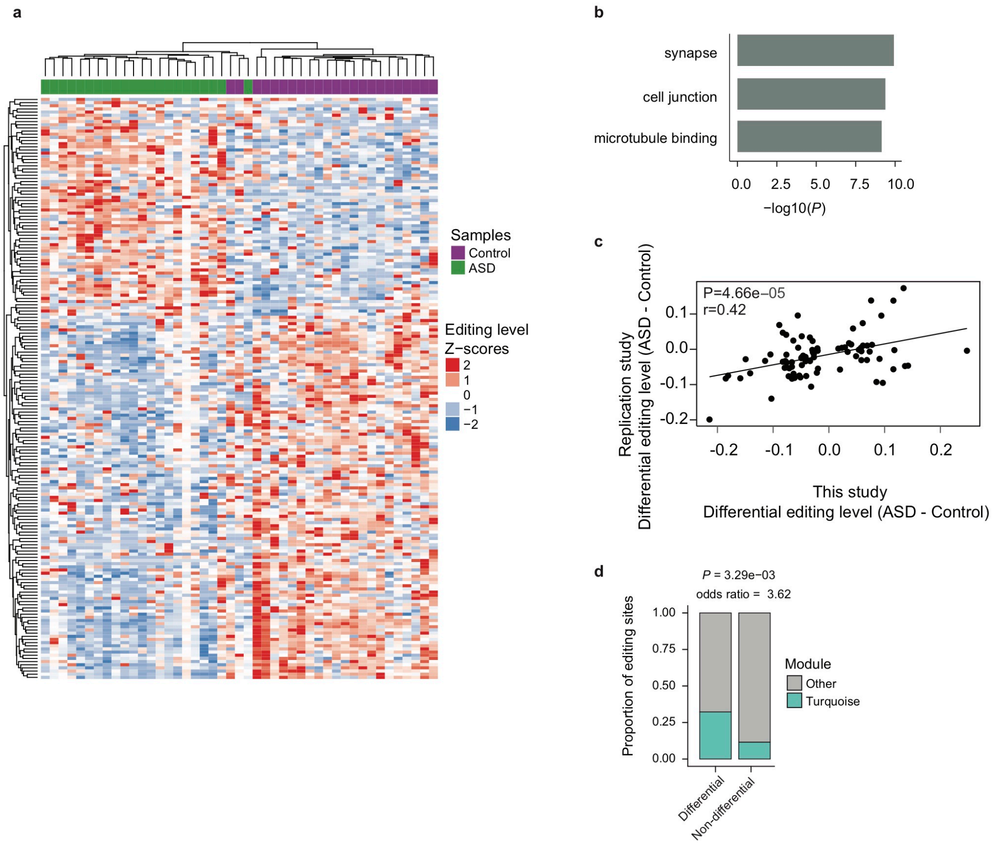
Supplemental Figure 2.12

Global enrichment of differential editing in ASD among developmentally regulated brain-specific editing sites. a. Recapitulation of 3 distinct trajectories for all editing sites in dorsalateral prefrontal cortex over human lifespan. Editing sites were obtained from a previous study. Editing sites are partitioned into clusters based on similar criteria as in the original study: Red cluster: Editing sites constitutively highly edited across lifespan, anova fdr > 0.05 and median editing level > 0.5. Yellow cluster: Editing sites increasing over development, particularly within the fetal-infant stages, anova fdr < 0.05. Blue cluster: Editing sites with perpetually low editing levels, anova fdr > 0.05 and median editing level < 0.5. Top axis shows individual ages and age groupings. b. Differential editing (FC DE sites) and turquoise module sites (FC turquoise sites) from frontal cortex are enriched in the fetal-infant developmentally increasing yellow cluster and depleted of sites in the blue developmentally low cluster. P values, two-tailed Fisher's Exact test. Odds ratio < 1 colored blue; > 1 colored red. N=3355 and 1116 editing sites for the DE and turquoise site comparisons respectively.



Supplemental Figure 2.13

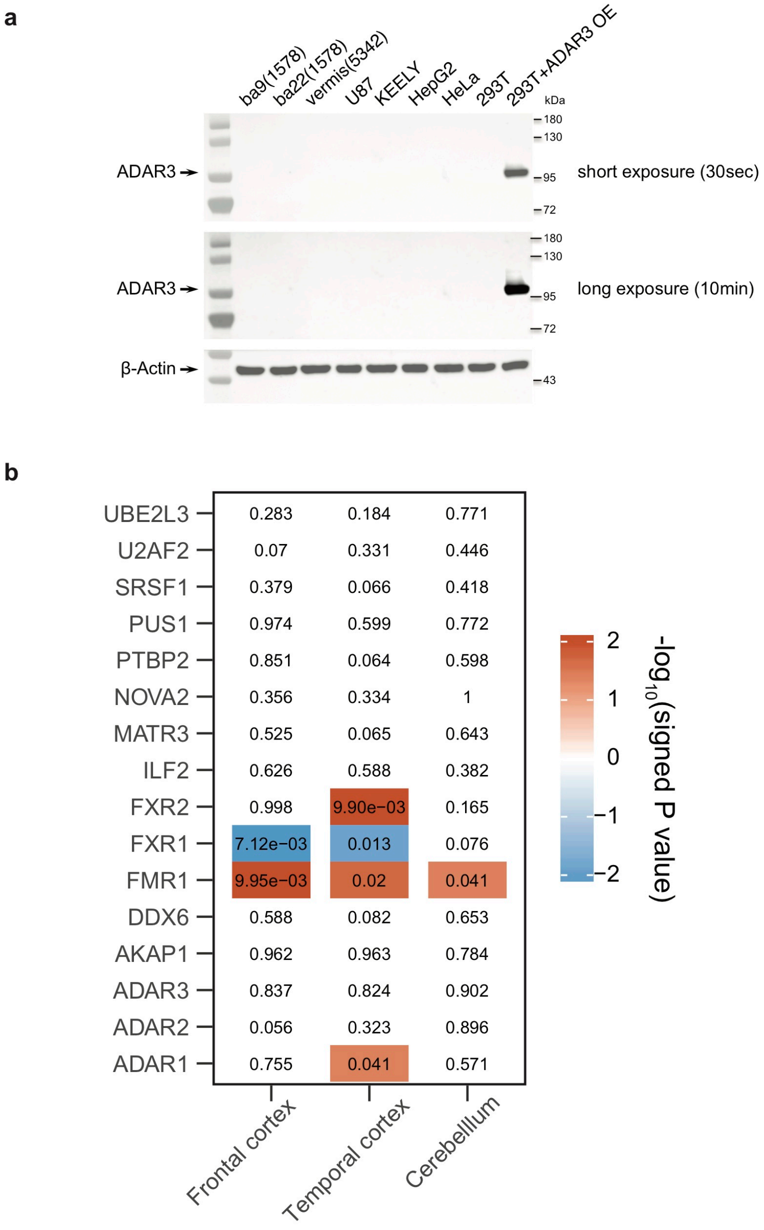
Comparison of technical covariates in replicate dataset between ASD and control (CTL) samples. Pearson correlation was calculated. N=45 samples in each covariate comparison.



Supplemental Figure 2.14 Global replication of differential editing in the frontal cortex of ASD.

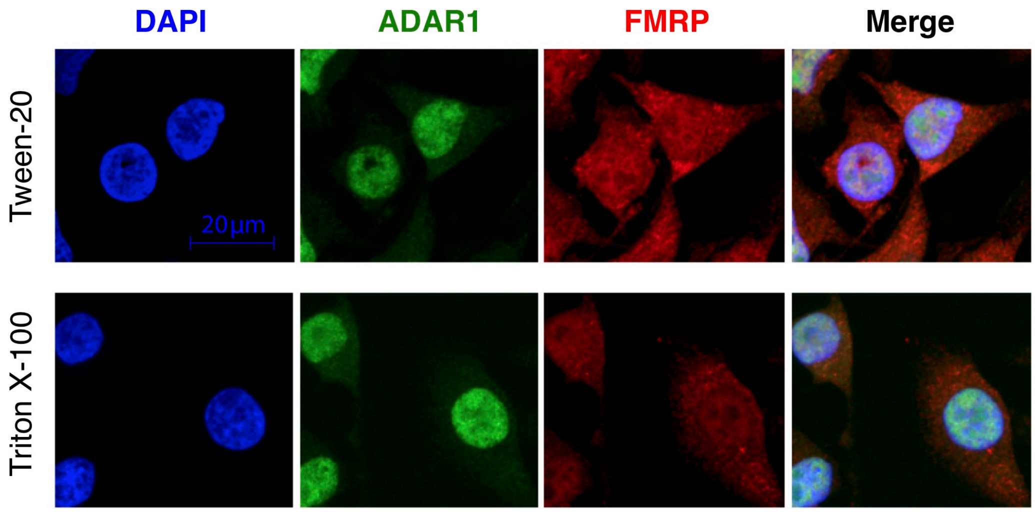
ASD and Control RNA-seq samples obtained from Liu et al[46]. a, Heatmap of differnetial editing sites recapitulates segregation of samples by ASD and Controls similar as Fig. 1d. About 65% of differential editing sites are hypoedited in ASD, reproducing the hypoediting trend observed in Fig. 1c. b, Gene Ontology of genes harboring differential editing sites show enrichment in synaptic function; P-value calculated from one-tailed Gaussian test (Methods), N=129 genes. c, Pearson correlation between differential editing sites in our study and those testable from replication cohort; N=86 editing sites.

d, Overlap between the differential editing sites in the replication cohort and editing sites in the turquoise module of frontal cortex in our study. P value calculated via two-tailed Fisher's Exact test, n=428 editing sites.


a
ba9(1578)
ba22(1578)
vermis(5342)
U87
KEELY
HepG2
HeLa
293T
293T+ADAR3 OE
kDa
180
130
95
72
ADAR3
short exposure (30sec)
180
130
95
72
ADAR3
long exposure (10min)
β-Actin
43
b
UBE2L3 0.283 0.184 0.771
U2AF2 0.07 0.331 0.446
SRSF1 0.379 0.066 0.418
PUS1 0.974 0.599 0.772
PTBP2 0.851 0.064 0.598
NOVA2 0.356 0.334 1
MATR3 0.525 0.065 0.643
ILF2 0.626 0.588 0.382
FXR2 0.998 9.90e−03 0.165
FXR1 7.12e−03 0.013 0.076
FMR1 9.95e−03 0.02 0.041
DDX6 0.588 0.082 0.653
AKAP1 0.962 0.963 0.784
ADAR3 0.837 0.824 0.902
ADAR2 0.056 0.323 0.896
ADAR1 0.755 0.041 0.571
Frontal cortex
Temporal cortex
Cerebellum
2
1
0
−1
−2
-log10(signed P value)

Supplemental Figure 2.15 Examination of ADAR3 and other RNA binding proteins.

a, Western blot of ADAR3 across a myriad of tissues and cell types (Supplementary Fig. 31). Right, we confirmed the quality of ADAR3 antibody (Santa Cruz, sc-73410) for detecting ADAR3 protein expression in 293T ADAR3 overexpression. ADAR3 protein was undetected in cell lines U87, KELLY, HepG2, HeLa, and 293T. Left, ADAR3 was also undetectable in human postmortemtissue samples (Table S1b). The experiment was repeated twice independently with similar results. b, Heatmap showing Pearson correlation of FPKM of RNA binding proteins with “eigengene” of turquoise modules. RNA binding proteins were chosen from a parallel screen of potential RNA editing regulators. Significantly associated RBPs are colored red or blue corresponding to positive or negative correlation. N=51 samples per correlation.

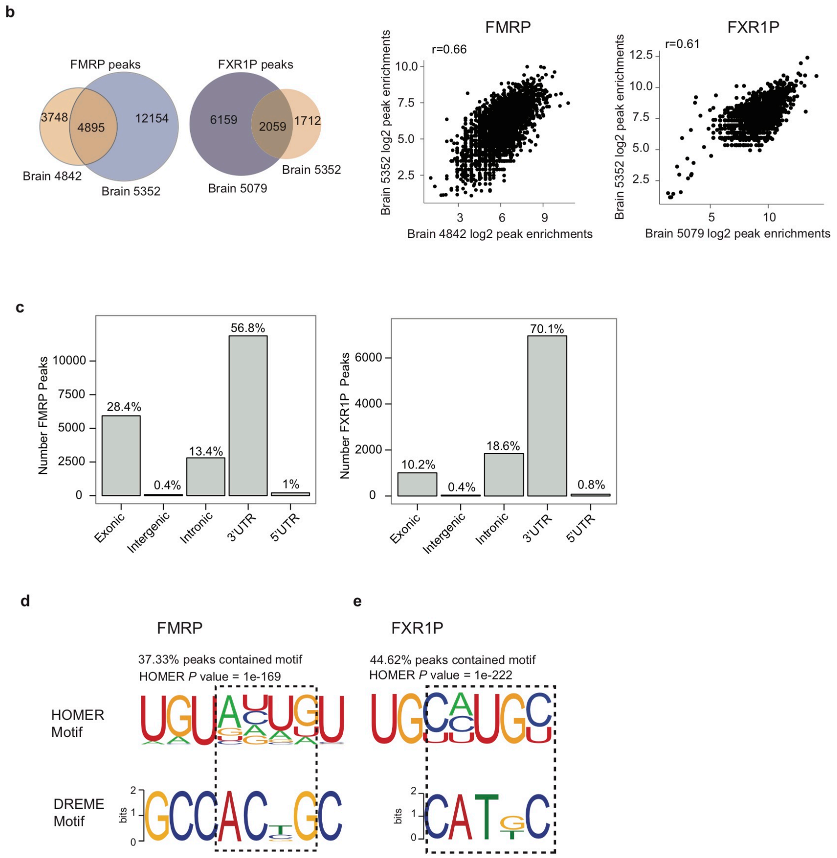


Supplemental Figure 2.16 Immunofluorescence of ADAR1 and FMRP in HeLa.

Confocal images of immunofluorescence staining of ADAR1 (green), FMRP (red), and DAPI (blue) in HeLa cells. Cells were permeabilized with either 0.1% Tween-20 (upper) orTriton X-100 (bottom). Experiment was repeated 3 times independently with similar results.

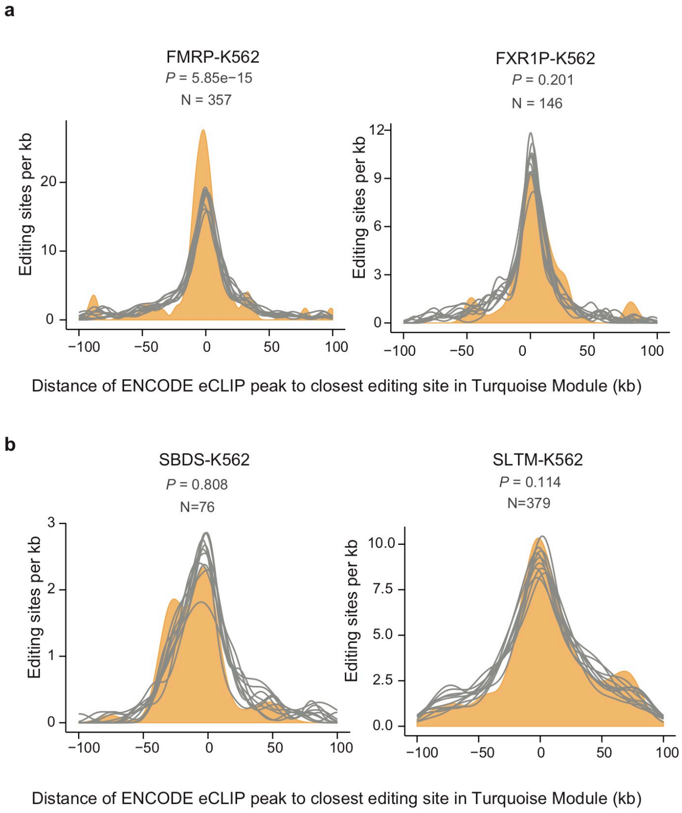
a

Protein	Maryland Brain Bank ID	IP raw reads	Control IP raw reads	Usable reads	Number peaks identified
FMR1P	4842	3342402	27744098	177375	9778
FMR1P	5352	4949318	26925538	339489	19277
FXR1P	5079	14757848	32020356	173021	8706
FXR1P	5352	14226896	36237522	159985	4352



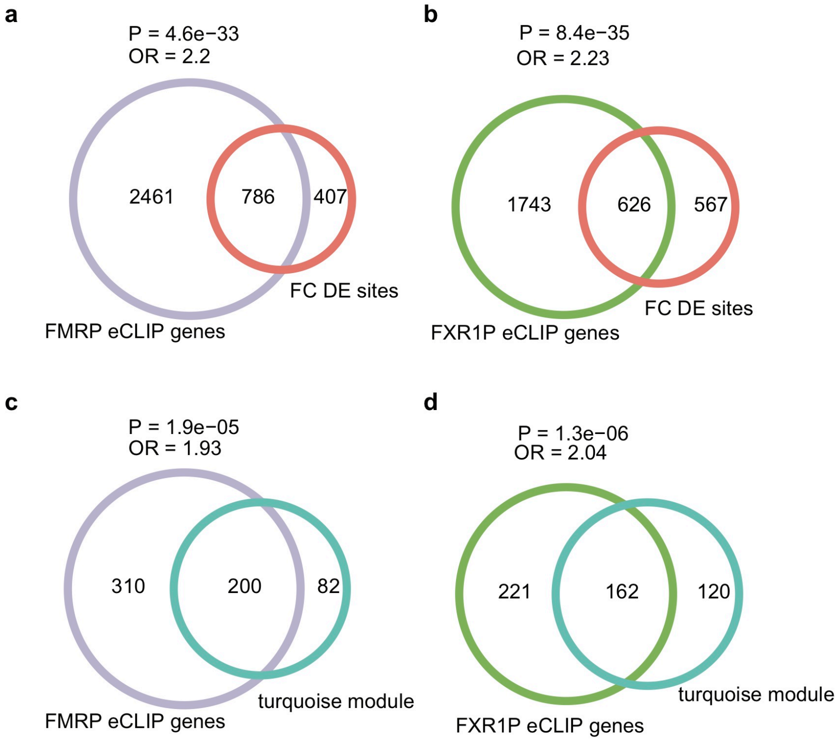
Supplemental Figure 2.17 Global RNA binding patterns of FMRP and FXR1P in human frontal cortex.

a, Human frontal cortex samples used for eCLIP experiments and the numbers of raw reads, usable reads (after quality checks and removal of PCR duplicates) and final eCLIP peaks. b, Left: Venn diagrams showing overlap of eCLIP peaks between two replicates of each protein. Right: scatter plots showing Pearson’s correlation of log2 fold enrichment (Number of eCLIP reads/number of control reads) of overlapping peaks for each protein. N=4870 and 2055 peaks correlated between FMRP and FXR1P replicates respectively. c, Distribution of FMRP and FXR1P eCLIP peaks in different types of genomic regions. d. Top motifs identified by HOMER and DREME packages in eCLIP peaks of FMRP. e, Same as d, but for FXR1P. For details of P-value calculations see documentation of respective packages. N=29055 and 13058 combined peaks for FMRP and FXR1P respectively.



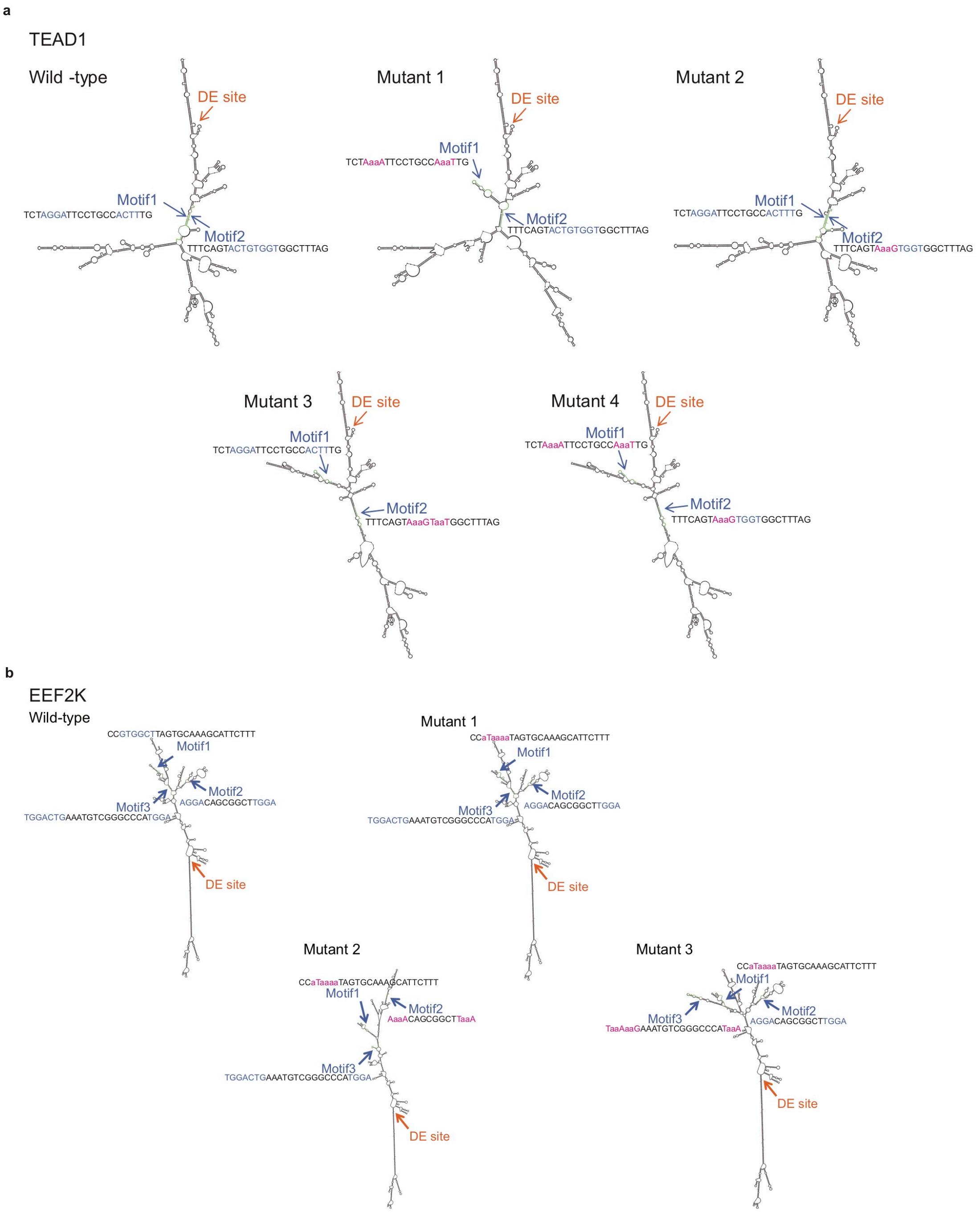
Supplemental Figure 2.18

Distances between turquoise editing sites in frontal cortex and eCLIP peaks generated in cancer cell lines, similar as Fig. 3c. a, Shortest distance between turquoise editing sites and FMRP or FXR1P eCLIP peaks from K562 cells generated by the ENCODE consortium. b, Shortest distances between turquoise editing sites and eCLIP peaks of two negative control RNA binding proteins that do not regulate RNA editing. ENCODE accession numbers are (FMR1: ENCSR331VNX, FXR1: ENCSR774RFN, SBDS: ENCSR059CWF, SLTM: ENCSR000SSH). P-values from one-tailed Gaussian test (Methods), N=number of editing sites.



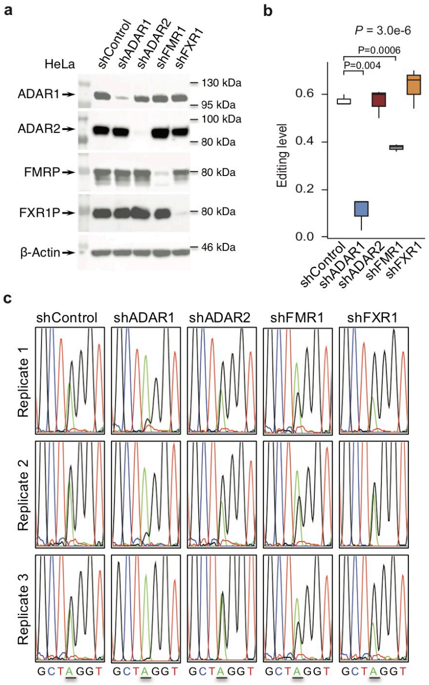
Supplemental Figure 2.19

Overlap between FMRP and FXR1P eCLIP-bound genes and differentially editied genes. a, Overlap between FMRP gene targets (FMRP eCLIP genes) and differential editing sites in frontal cortex (FC DE sites). P values calculated via two-tailed Fisher's Exact test. OR: odds ratio. b, Similar as a, for FXR1P eCLIP genes. c, similar as a, but using editing sites in the turquoise module of frontal cortex. d, similar as c, for FXR1P eCLIP genes.


a
TEAD1
Wild -type
DE site
Motif1
TCTAGGATTCCTGCCACTTTG
Motif2
TTTCAGTACTGTGGTGGCTTTAG
Mutant 1
DE site
Motif1
TCTAaaATTCCTGCCAaaTTG
Motif2
TTTCAGTACTGTGGTGGCTTTAG
Mutant 2
DE site
Motif1
TCTAGGATTCCTGCCACTTTG
Motif2
TTTCAGTAaaGTGGTGGCTTTAG
Mutant 3
DE site
Motif1
TCTAGGATTCCTGCCACTTTG
Motif2
TTTCAGTAaaGTaaTGGCTTTAG
Mutant 4
DE site
Motif1
TCTAaaATTCCTGCCAaaTTG
Motif2
TTTCAGTAaaGTGGTGGCTTTAG
b
EEF2K
Wild-type
CCGTGGCTTAGTGCAAAGCATTCTTT
Motif1
Motif2
AGGACAGCGGCTTGGA
Motif3
TGGACTGAAATGTCGGGCCCATGGA
DE site
Mutant 1
CCaTaaaaTAGTGCAAAGCATTCTTT
Motif1
Motif2
AGGACAGCGGCTTGGA
Motif3
TGGACTGAAATGTCGGGCCCATGGA
DE site
Mutant 2
CCaTaaaaTAGTGCAAAGCATTCTTT
Motif1
Motif2
AaaACAGCGGCTTaaA
Motif3
TGGACTGAAATGTCGGGCCCATGGA
DE site
Mutant 3
CCaTaaaaTAGTGCAAAGCATTCTTT
Motif1
Motif2
AGGACAGCGGCTTGGA
Motif3
TaaAaaGAAATGTCGGGCCCATaaA
DE site

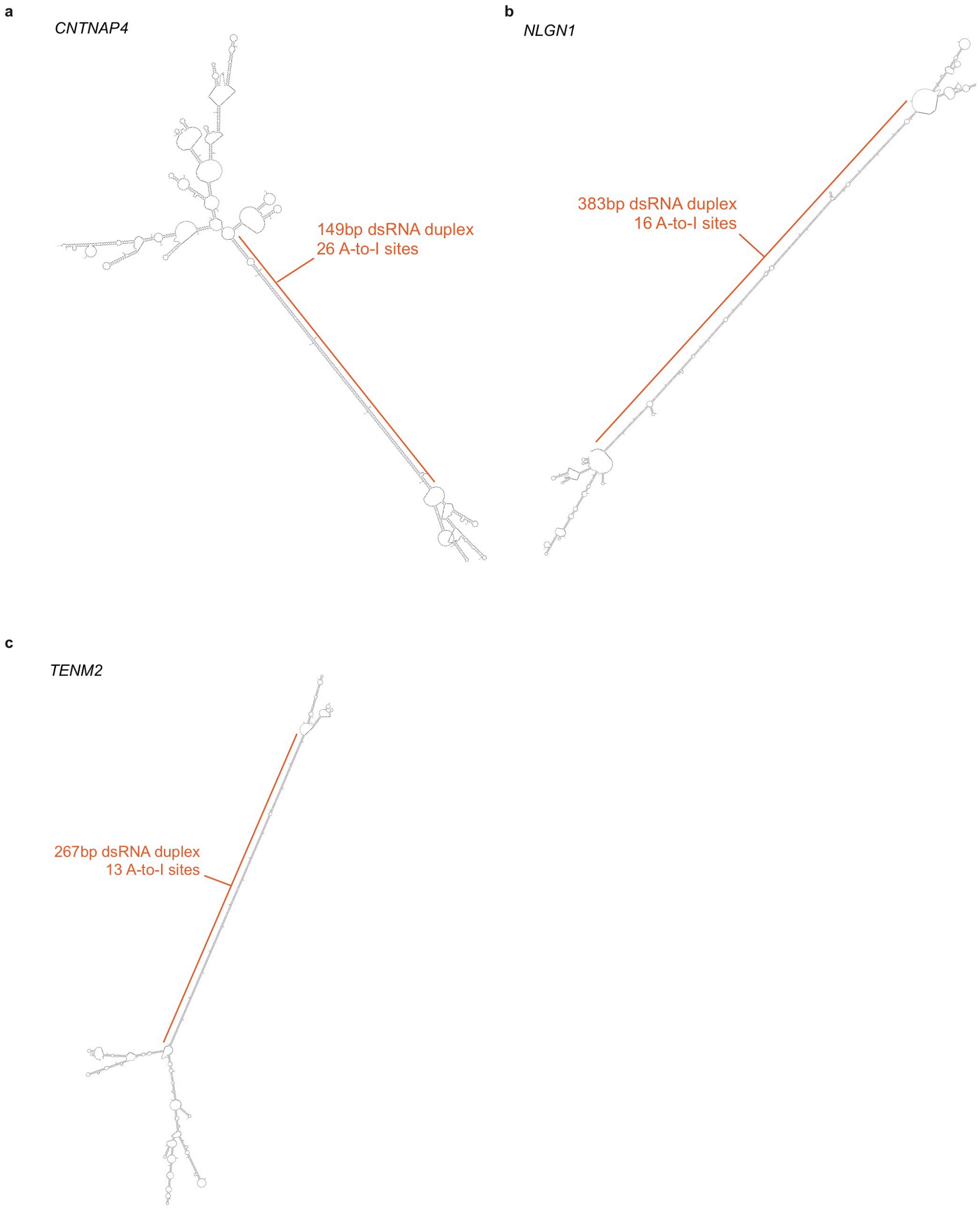
Supplemental Figure 2.20

RNA secondary structures predicted by mFold. a, Differential editing (DE) site in the TEAD1 gene. Wild-type: wild-type sequence; Mutant 1-4: mutant sequences designed to disrupt FMRP binding motifs. Lower case letters (in pink) represents mutations introduced to each motif. b, Similar as a, for a DE site in the EEF2K gene.



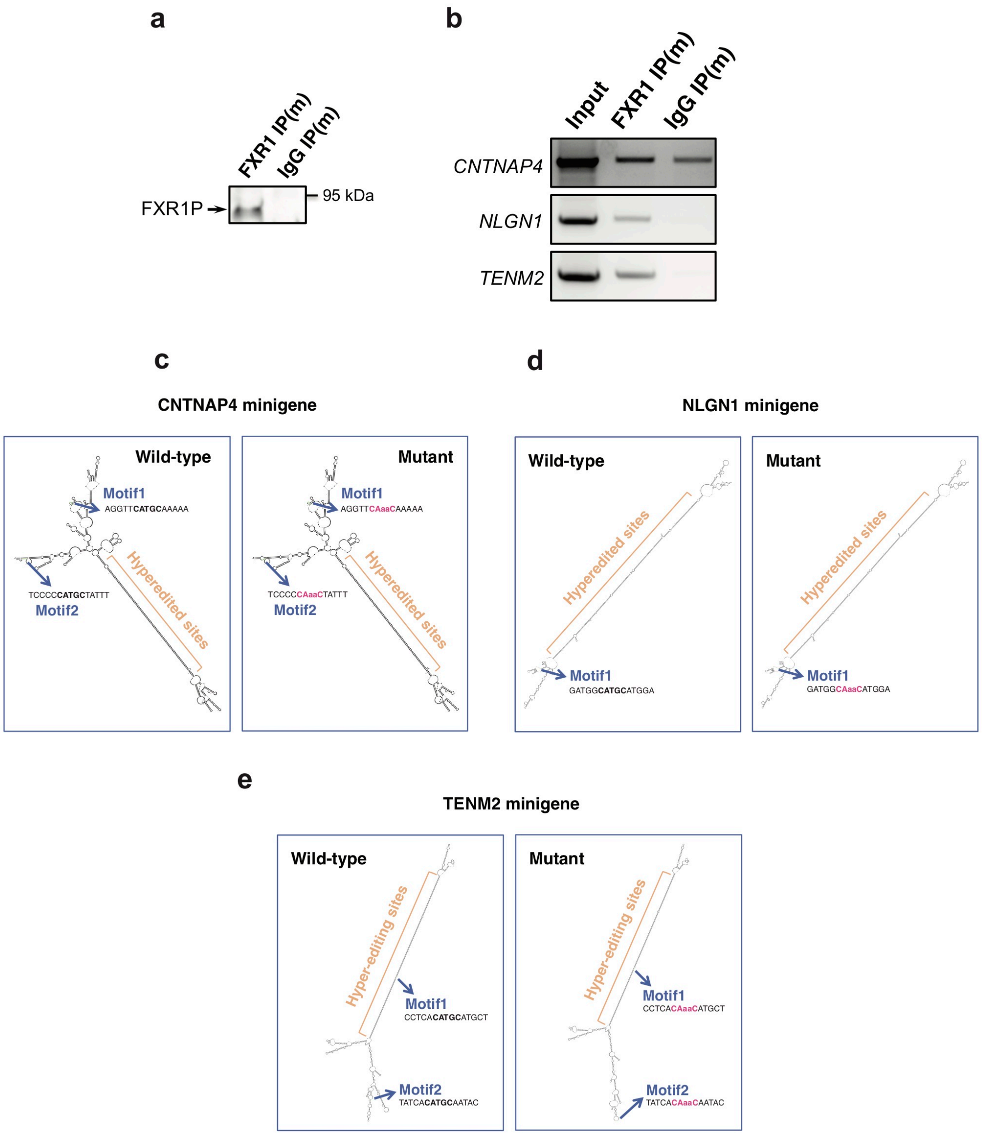
Supplemental Figure 2.21

Validation of gene knockdown and endogenous editing. a, Western blot (Supplementary Fig. 31) of FMRP, FXR1P, and ADAR proteins in HeLa cells with stable shRNA knockdown. All knockdowns are confirmed with high efficiency. Experiment was repeated 3 times with similar results. b, Endogenous editing levels of the RNA editing site in the EEF2K gene in cells shown in a. Boxplots were derived from three biological replicates. Overall P value calculated by one-way ANOVA. Individual comparison P values were calculated by two-tailed Student's t-test. Boxplot definition: center=median, lower hinge=25th percentile, upper hinge=75th percentile, min and max extend to observations at most 1.5 * IQR c, Sanger sequencing traces of the results in b. The edited site is underlined.


a
CNTNAP4
149bp dsRNA duplex
26 A-to-I sites
b
NLGN1
383bp dsRNA duplex
16 A-to-I sites
c
TENM2
267bp dsRNA duplex
13 A-to-I sites

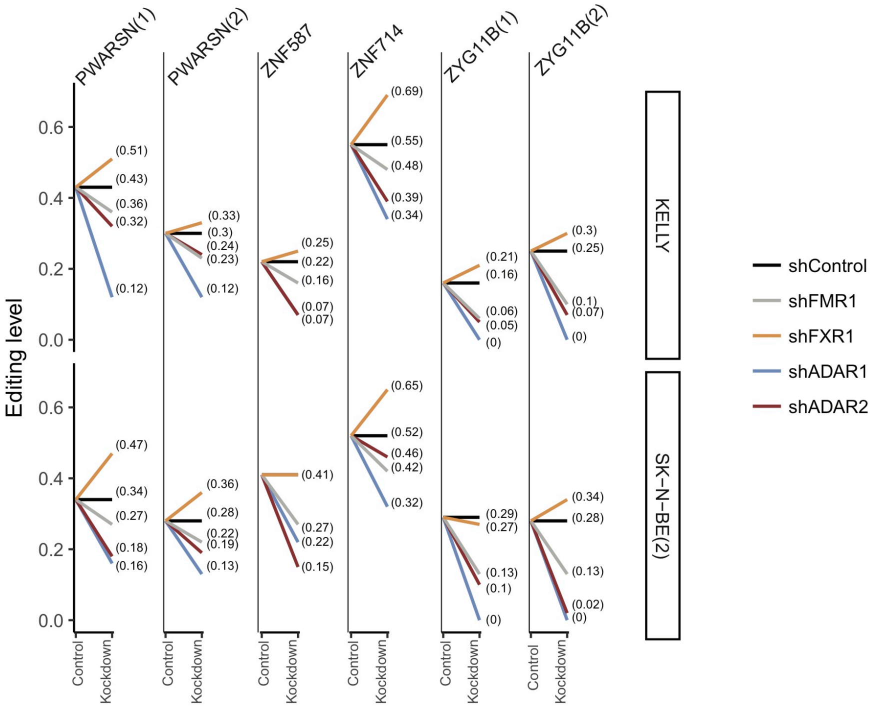
Supplemental Figure 2.22

mRNA Secondary structure prediction via mFold for regions harboring hyper-editing sites. a, CNTNAP4, b, NLGN1, and c, TENM2. Double-stranded regions are illustrated by the orange line. The number of editing sites within each double-stranded region is listed.



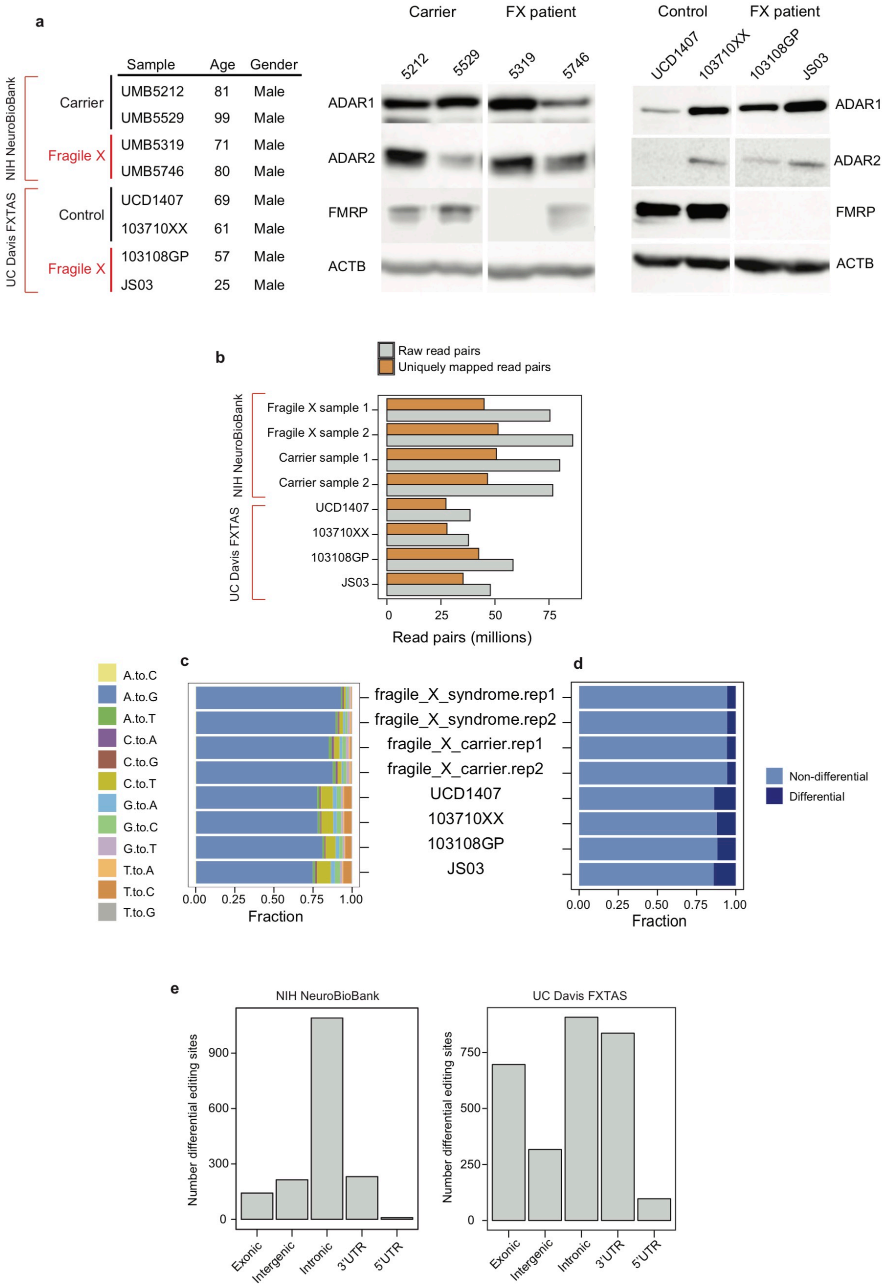
Supplemental Figure 2.23

Three target genes with hyper-editing sites that are dependent on FXR1P. a, Efficiency of anti-FXR1P immunoprecipitation was validated by Western blot. Experiment was performed once b, Semi-quantitative RT–PCR followed by agarose gel electrophoresis verified the presence of intronic target RNA in anti-FXR1P RIP. Non-specific binding of CNTNAP4 RNA to IgG negative control was detected. However, immunoprecipitation with the FXR1P antibody pulled down more CNTNAP4 RNA than the non-specific IgG antibody. c-e, related to Fig. 3i, RNA secondary structures predicted by mFold. in c, CNTNAP4, d, NLGN1, e, TENM2. Wild-type (left in each panel) shows wild-type sequence. Mutant (right in each pane), shows mutations introduced to each FXR1P motif.



Supplemental Figure 2.24

Validation of the dependency of differential editing sites on Fragile X proteins and ADARs. Endogenous editing levels of 6 differential editing sites (columns) measured in shControl (black) cells or cells with shRNA knockdown of FMR1 (grey), FXR1 (orange), ADAR1 (blue), and ADAR2 (red) in two neuroblastoma cell lines (KELLY and SK-N-BE(2)). Genomic coordinates (hg19) are: PWARSN(1)(chr15:25227816), PWARSN(2) (chr15:25227838), ZNF587 (chr19:58372723), SNF714 (chr19:21303017), ZYG11B(1) (chr1:53289547), ZYG11B(2) (chr1:53291420).

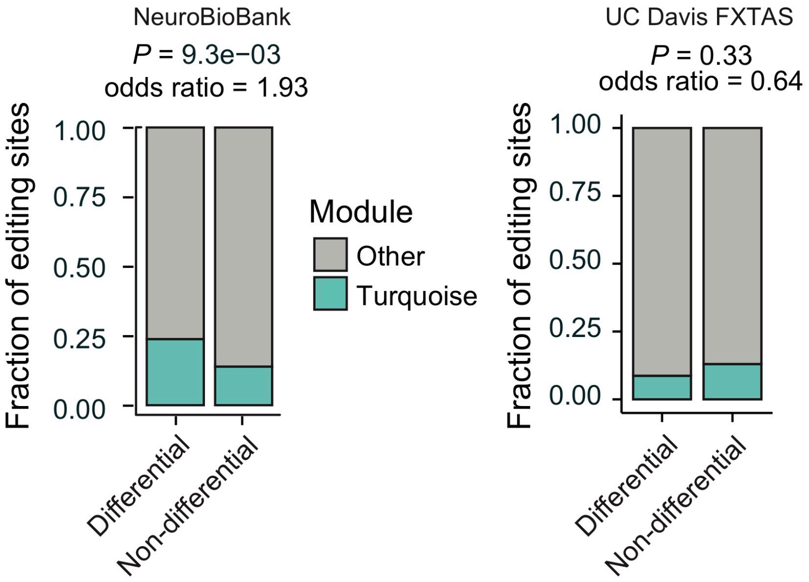


		Sample	Age	Gender
NIH NeuroBioBank	Carrier	UMB5212	81	Male
		UMB5529	99	Male
	Fragile X	UMB5319	71	Male
		UMB5746	80	Male
UC Davis FXTAS	Control	UCD1407	69	Male
		103710XX	61	Male
	Fragile X	103108GP	57	Male
		JS03	25	Male

Supplemental Figure 2.25

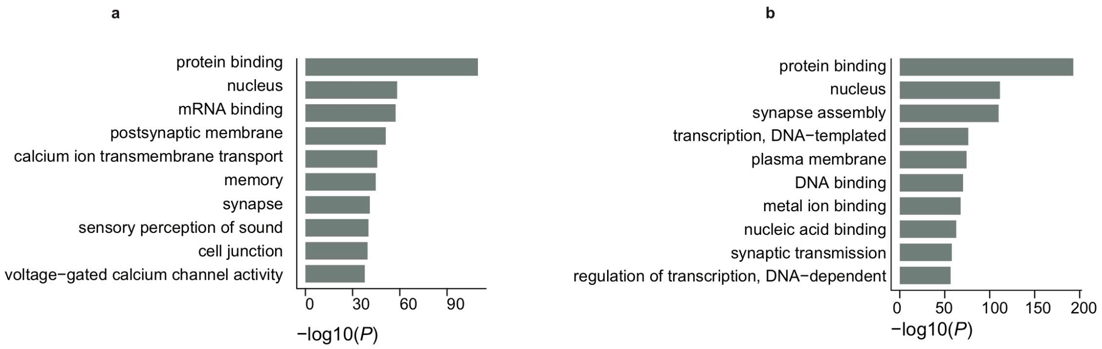
Global RNA editing analysis of RNA-Seq data obtained from frontal cortex of Fragile X patients and carriers/controls. Dataset analyzed from NIH NeuroBioBank and UC Davis

FXTAS brain repository are delineated within the plots. a, Western blot of ADARs and FMRP in the frontal cortex of patients, carriers, and controls from NIH NeuroBioBank (middle) and FXTAS (right). b, Number of raw and uniquely mapped read pairs for each sample. c, Fraction of all types of RNA editing sites identified in each sample. d, Fraction of differential and non-differential editing sites identified in each sample. e, Distribution of differential editing sites in different types of genomic regions.



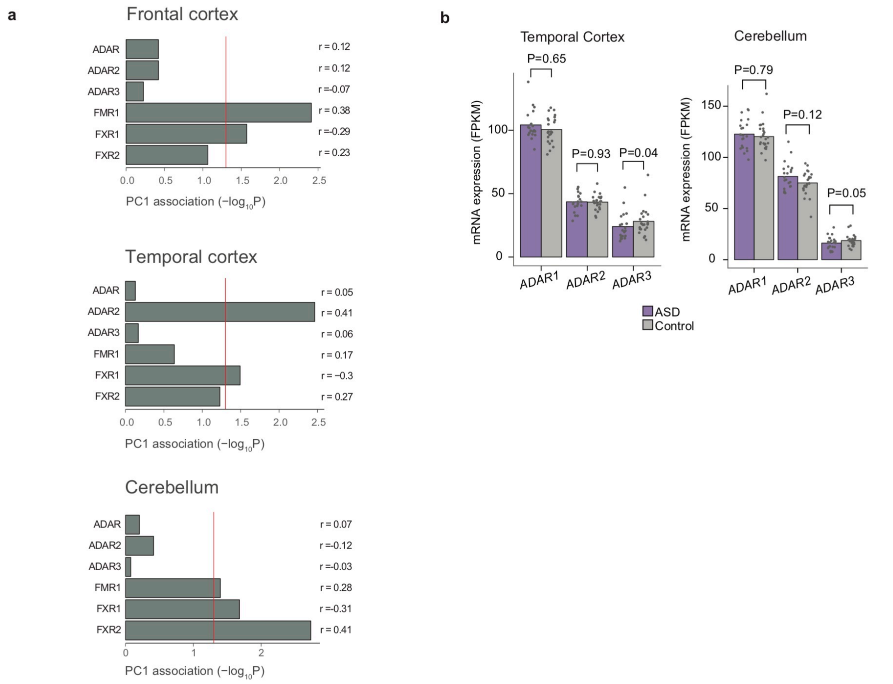
Supplemental Figure 2.26

Overlap between differential editing sites in Fragile X patients and those in the turquoise module of ASD frontal cortex. Left shows Fragile X vs carriers from NeuroBioBank. Right shows Fragile X vs controls from UC Davis FXTAS brain repository. P values were calculated using two-tailed Fisher's Exact test. N= 1679 and 1206 editing sites for NeuroBioBank and UC Davis FXTAS respectively.



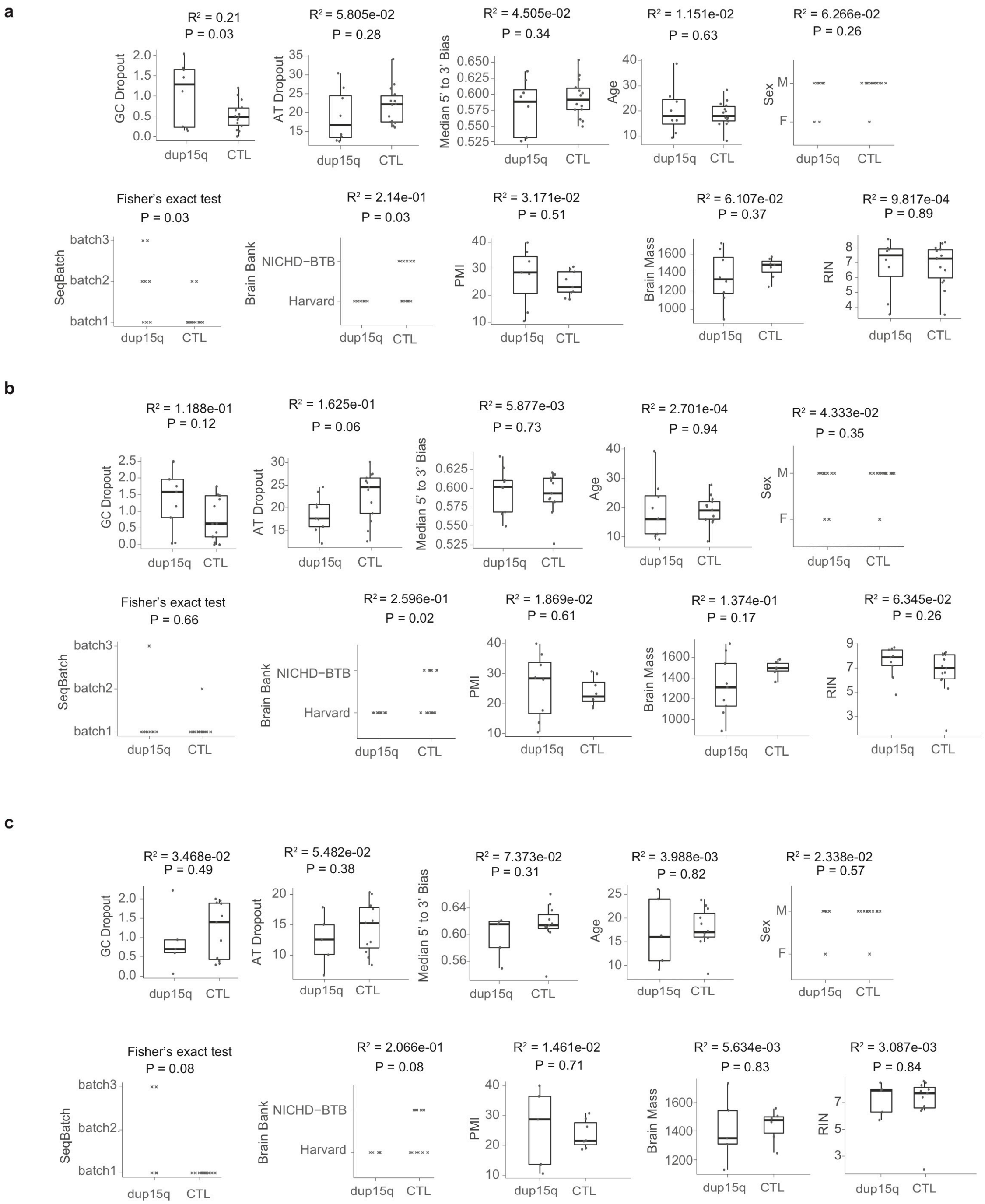
Supplemental Figure 2.27

Gene ontology of genes harboring differential editing sites in ASD. a. temporal cortex (N=1048 genes); b. cerebellum (N=1437 genes). P-value calculated using one-tailed Gaussian test (Methods).



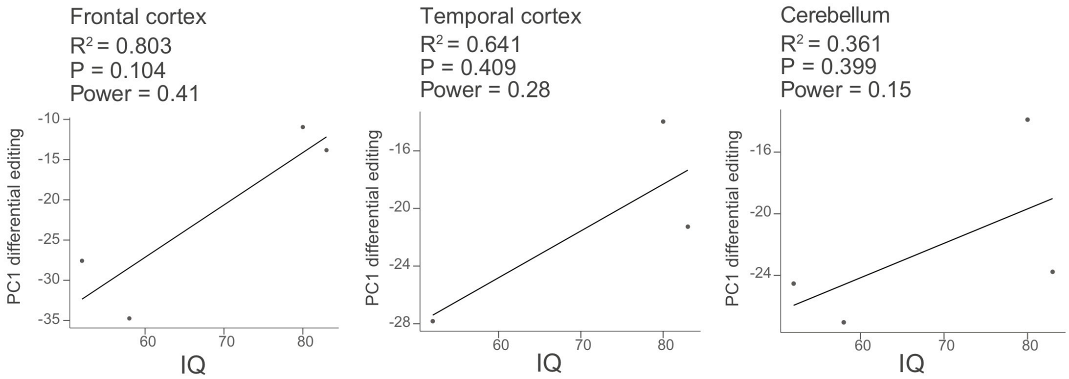
Supplemental Figure 2.28

Global regulatory profile of ADARs and Fragile X proteins across brain regions. a, Pearson correlation between mRNA levels of ADAR1, ADAR2, ADAR3, FMR1, FXR1, and FXR2, and the first principal component of differential RNA editing sites in each brain region of ASD. b, ADAR1, ADAR2 and ADAR3 mRNA expression in different brain regions of ASD and control samples. P values were calculated similarly as in Fig. 2a, using a regression approach where covariates were accounted for9. (a-b) N=62, 57, and 60 samples for frontal cortex, temporal cortex, and cerebellum respectively.


a
R² = 0.21
P = 0.03
GC Dropout
2.0
1.5
1.0
0.5
0.0
dup15q
CTL
R² = 5.805e-02
P = 0.28
AT Dropout
35
30
25
20
15
dup15q
CTL
R² = 4.505e-02
P = 0.34
Median 5' to 3' Bias
0.650
0.625
0.600
0.575
0.550
0.525
dup15q
CTL
R² = 1.151e-02
P = 0.63
Age
40
30
20
10
dup15q
CTL
R² = 6.266e-02
P = 0.26
Sex
M
F
dup15q
CTL
Fisher's exact test
P = 0.03
SeqBatch
batch3
batch2
batch1
dup15q
CTL
R² = 2.14e-01
P = 0.03
Brain Bank
NICHD−BTB
Harvard
dup15q
CTL
R² = 3.171e-02
P = 0.51
PMI
40
30
20
10
dup15q
CTL
R² = 6.107e-02
P = 0.37
Brain Mass
1600
1400
1200
1000
dup15q
CTL
R² = 9.817e-04
P = 0.89
RIN
8
7
6
5
4
dup15q
CTL
b
R² = 1.188e-01
P = 0.12
GC Dropout
2.5
2.0
1.5
1.0
0.5
0.0
dup15q
CTL
R² = 1.625e-01
P = 0.06
AT Dropout
30
25
20
15
dup15q
CTL
R² = 5.877e-03
P = 0.73
Median 5' to 3' Bias
0.625
0.600
0.575
0.550
0.525
dup15q
CTL
R² = 2.701e-04
P = 0.94
Age
40
30
20
10
dup15q
CTL
R² = 4.333e-02
P = 0.35
Sex
M
F
dup15q
CTL
Fisher's exact test
P = 0.66
SeqBatch
batch3
batch2
batch1
dup15q
CTL
R² = 2.596e-01
P = 0.02
Brain Bank
NICHD−BTB
Harvard
dup15q
CTL
R² = 1.869e-02
P = 0.61
PMI
40
30
20
10
dup15q
CTL
R² = 1.374e-01
P = 0.17
Brain Mass
1600
1400
1200
1000
dup15q
CTL
R² = 6.345e-02
P = 0.26
RIN
9
7
5
3
dup15q
CTL
c
R² = 3.468e-02
P = 0.49
GC Dropout
2.0
1.5
1.0
0.5
0.0
dup15q
CTL
R² = 5.482e-02
P = 0.38
AT Dropout
20
15
10
dup15q
CTL
R² = 7.373e-02
P = 0.31
Median 5' to 3' Bias
0.64
0.60
0.56
dup15q
CTL
R² = 3.988e-03
P = 0.82
Age
25
20
15
10
dup15q
CTL
R² = 2.338e-02
P = 0.57
Sex
M
F
dup15q
CTL
Fisher's exact test
P = 0.08
SeqBatch
batch3
batch2
batch1
dup15q
CTL
R² = 2.066e-01
P = 0.08
Brain Bank
NICHD−BTB
Harvard
dup15q
CTL
R² = 1.461e-02
P = 0.71
PMI
40
30
20
10
dup15q
CTL
R² = 5.634e-03
P = 0.83
Brain Mass
1600
1400
1200
dup15q
CTL
R² = 3.087e-03
P = 0.84
RIN
7
5
3
dup15q
CTL

Supplemental Figure 2.29

Comparison of technical covariates between dup15q and control (CTL) samples in a, frontal cortex b, temporal cortex, and c, cerebellum. Pearson correlation was calculated; two-tailed Fisher's Exact test was alternatively used when present with more than 2 categories. Uncorrected P values are depicted. The bonferroni significance cut-off is $P < 0.005$. N=22, 22, and 16 samples from frontal cortex, temporal cortex, and cerebellum respectively. Boxplot definitions: center=median, lower hinge=25th percentile, upper hinge=75th percentile, min and max extend to observations at most 1.5 * IQR



Supplemental Figure 2.30

Correlation between IQ and differential RNA editing. Pearson correlation between the 1st principal component of differential RNA editing and samples with measured IQ in frontal cortex (left), temporal cortex (middle), and cerebellum (right). Power calculated for a significance level of 0.05 and n = 4 (the minimum requisite sample size for power calculation).

Supplemental Figure 2.31

Uncropped western blot images

Uncropped western blots in the main text are collected. Some of western blot membranes were cut to use for multiple antibodies. The protein standards are represented on the left of the blots.

Antibody information

ADAR1 (Santa Cruz, sc-73408, Lot # F1417, 1:200)

ADAR2 (Santa Cruz, sc-73409, Lot # K0917, 1:200)

ADAR3 (Santa Cruz, sc-73410, Lot # B2316, 1:200)

FMRP (Millipore, MAB2160, Lot # 2984225, 1:500)

FMRP (Abcam, ab17722, Lot # GR272723-1, 1:1000)

FXR1P (Bethyl Laboratories, A303-892A, 1:2000)

FLAG (Sigma, F7425, Lot # 018M4828V, 1:1000)

β-Actin (Santa Cruz, sc-47778, Lot # J2915, 1:500)

β-Tubulin (Santa Cruz, sc-23949, Lot # C0718, 1:200)

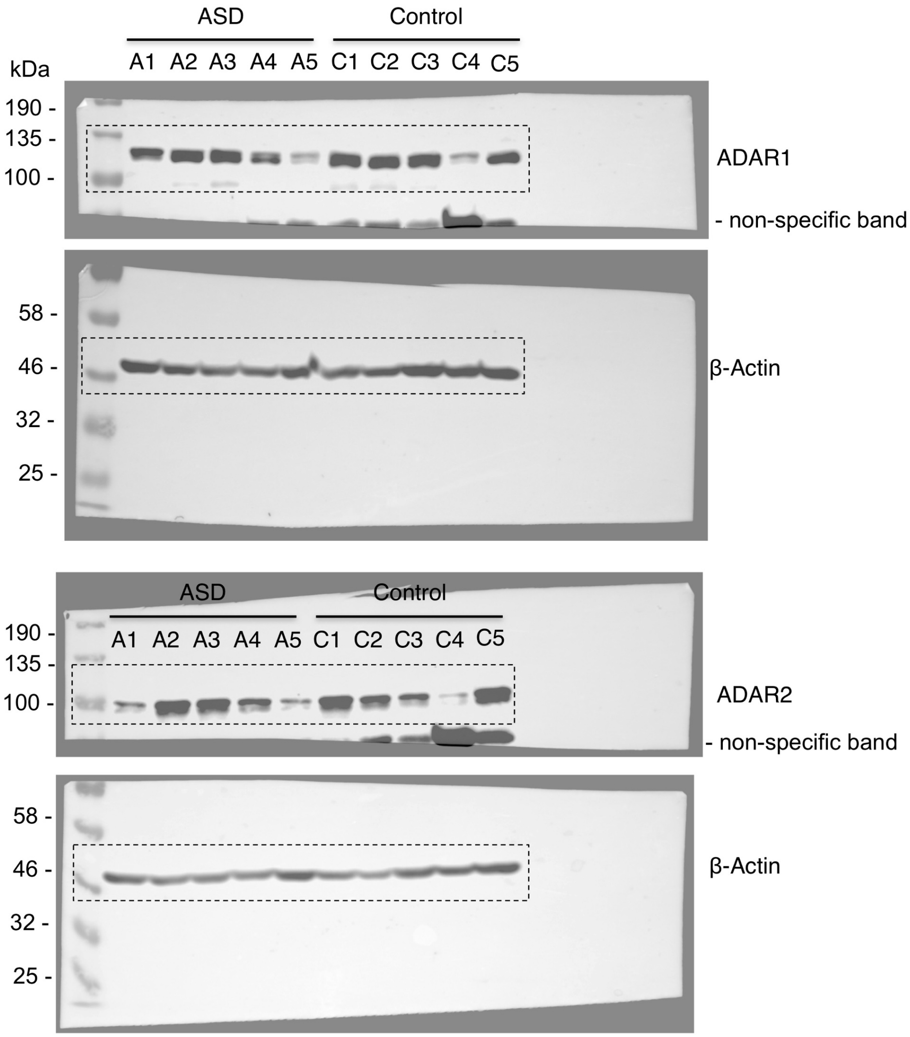
U1-70K (Santa Cruz, sc-390899, Lot # G0616, 1:100)

ADAR1 (Santa Cruz, sc-271854, Lot # F1616, 5ug/each IP sample)

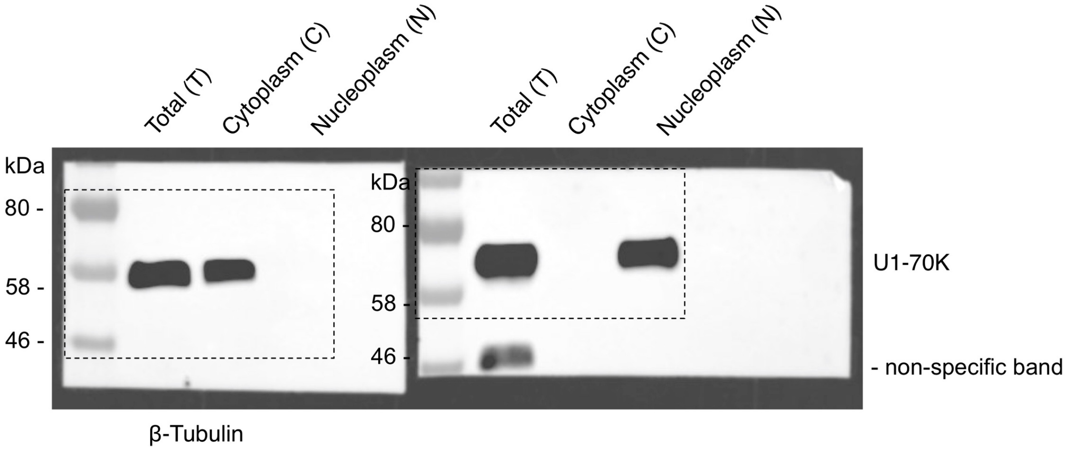
FMRP (Millipore, MAB2160, Lot # 2984225, 5ug/each IP sample)

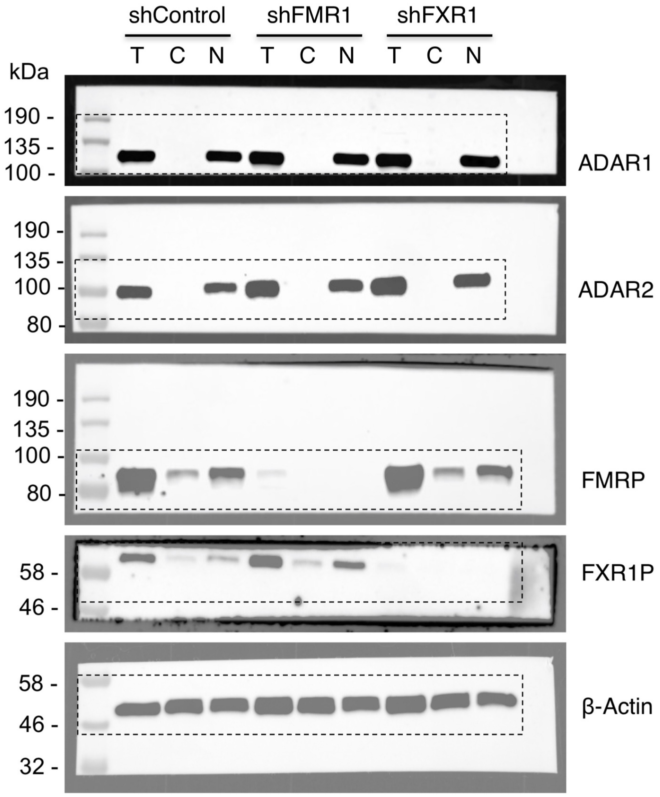
FLAG (Sigma, F1804, Lot # SLBT7654, 2.5ug/each IP sample)

Supplementary Figure 31 related to Fig. 2b in main text.

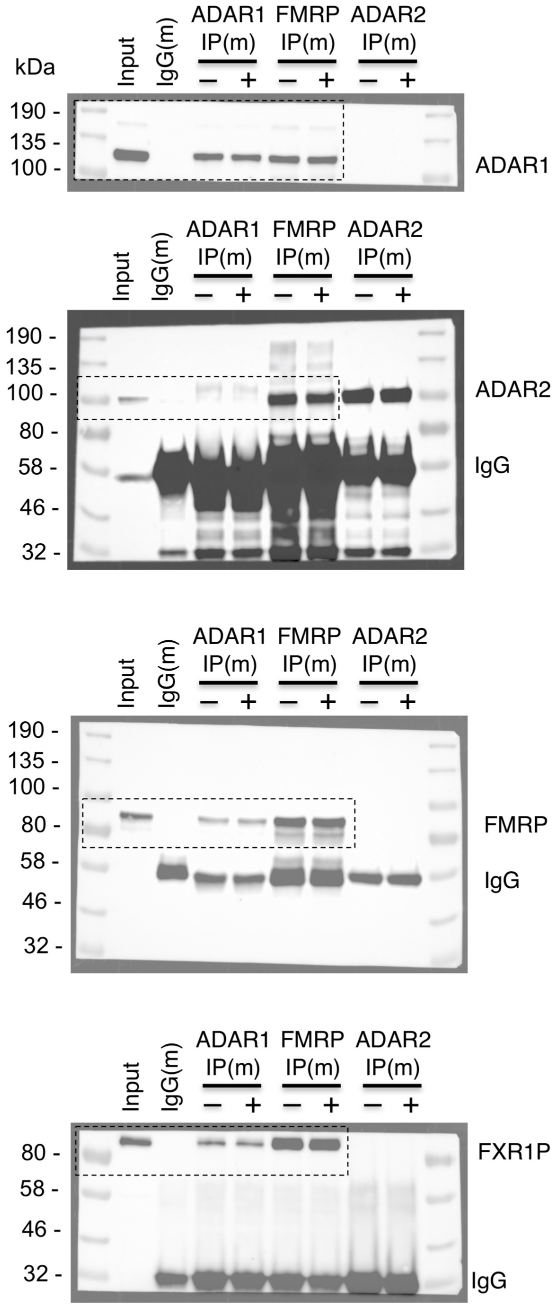


Supplementary Figure 31 related to Fig. 3a in main text.

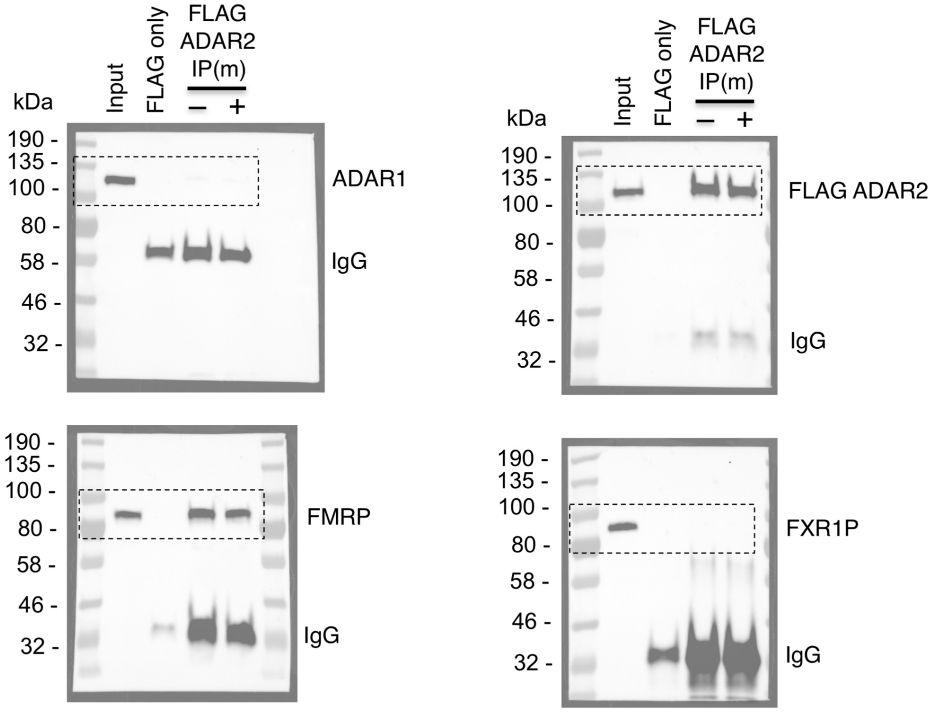




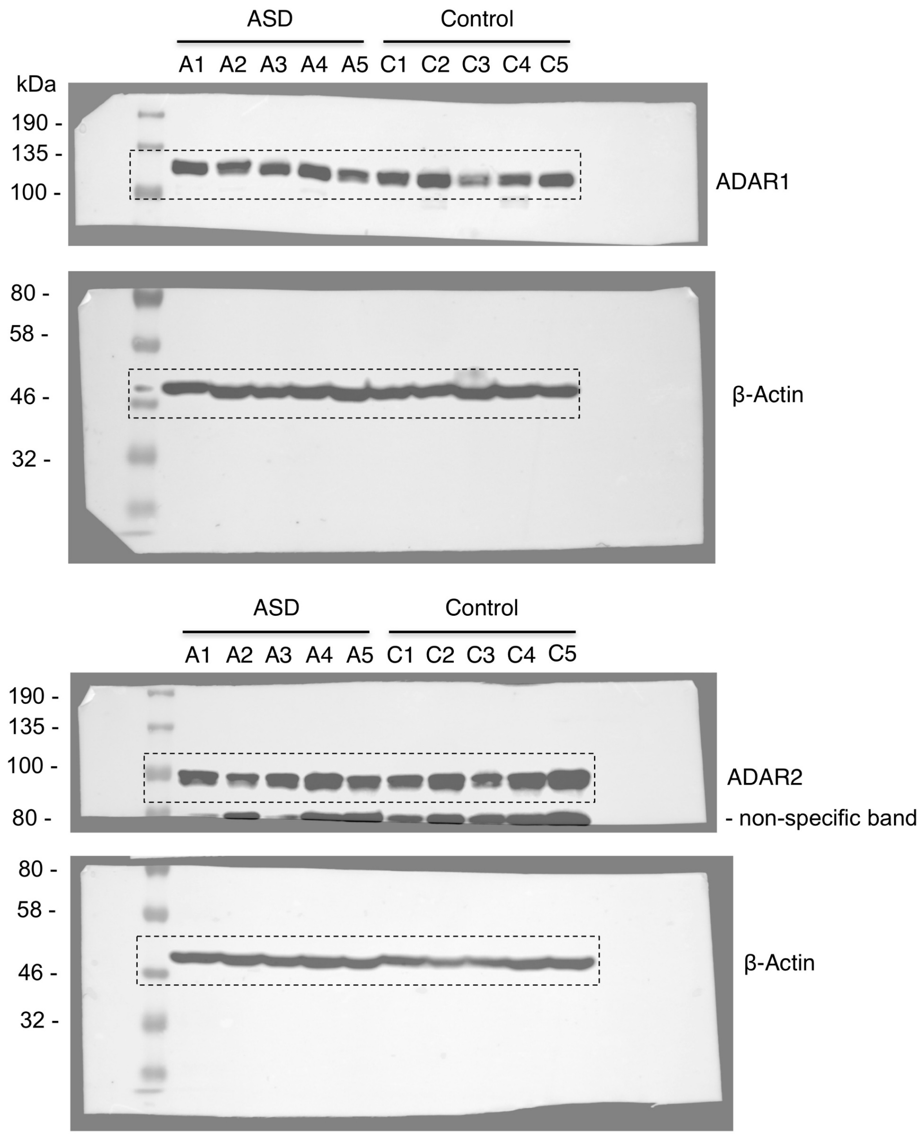
Supplementary Figure 31 related to Fig. 3b in main text.



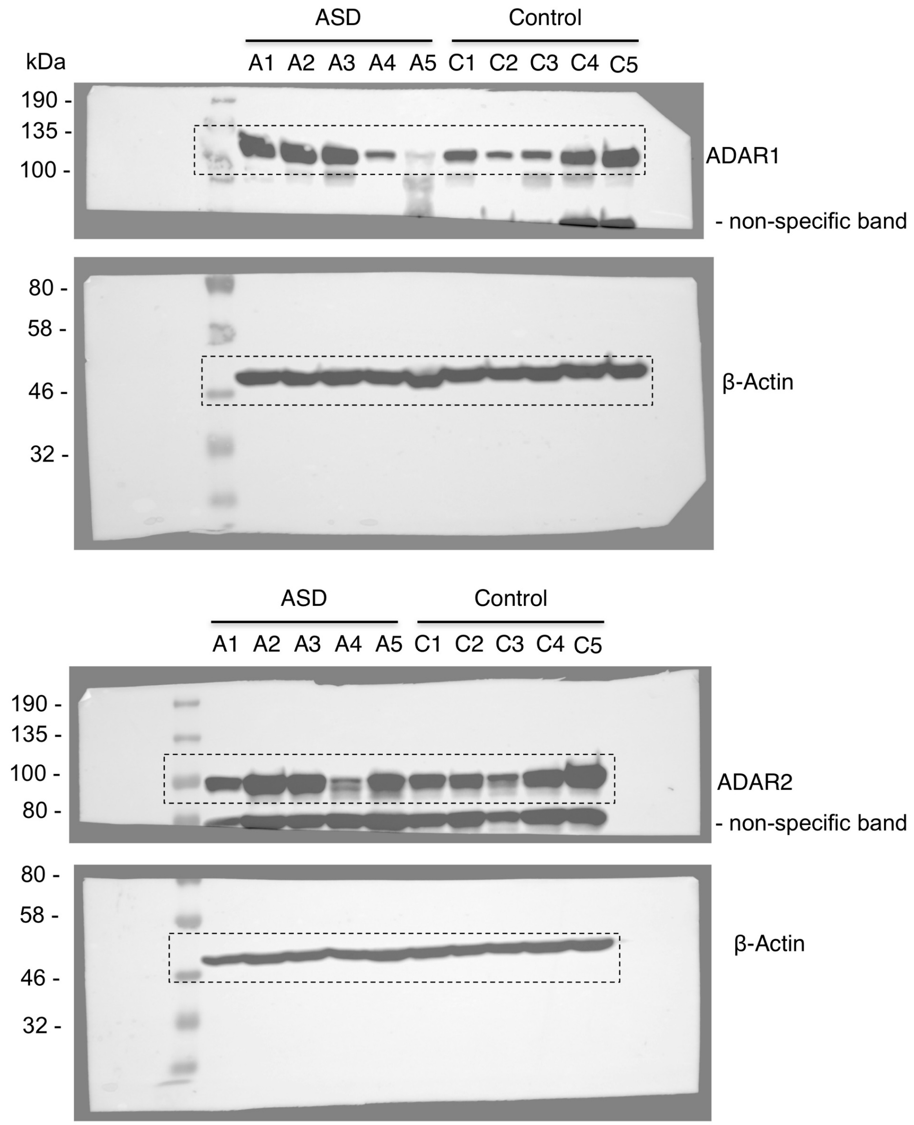
Supplementary Figure 31 related to Fig. 3b in main text.



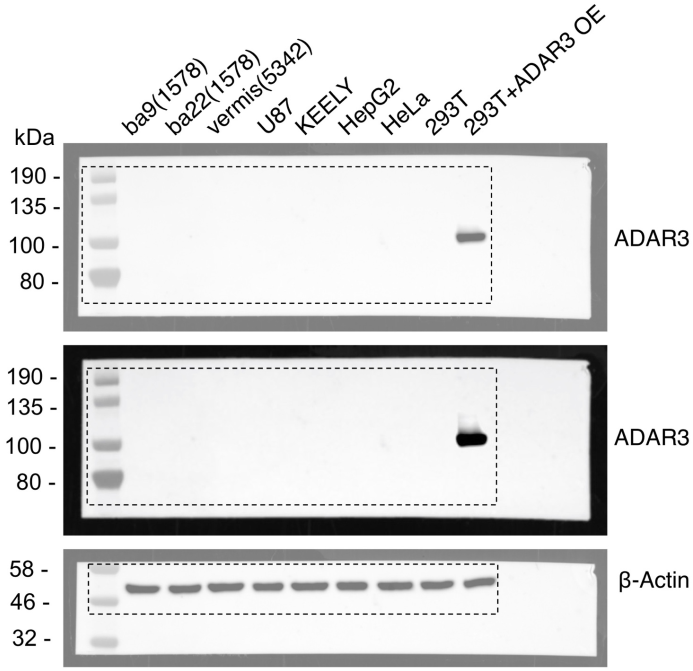
Supplementary Figure 31 related to Fig. 5h in main text.



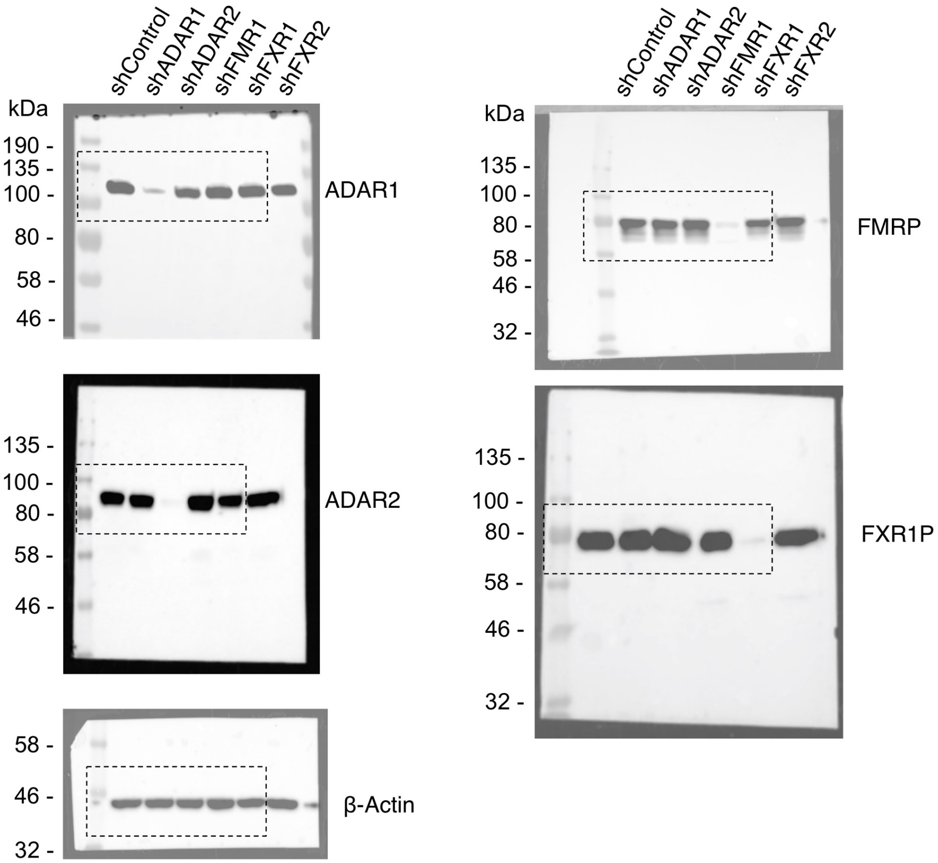
Supplementary Figure 31 related to Fig. 5i in main text.



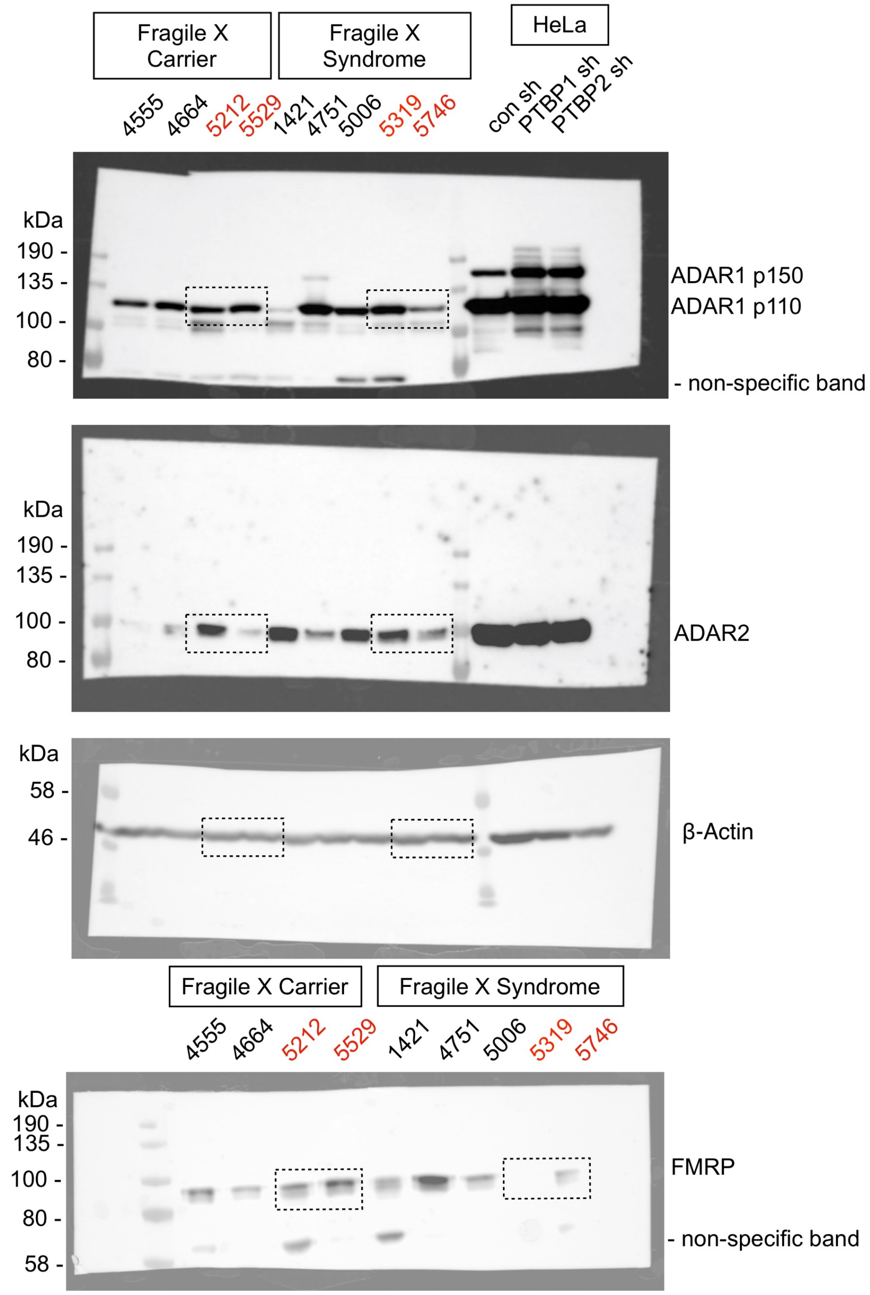
Supplementary Figure 31 related to Supplementary Fig.15



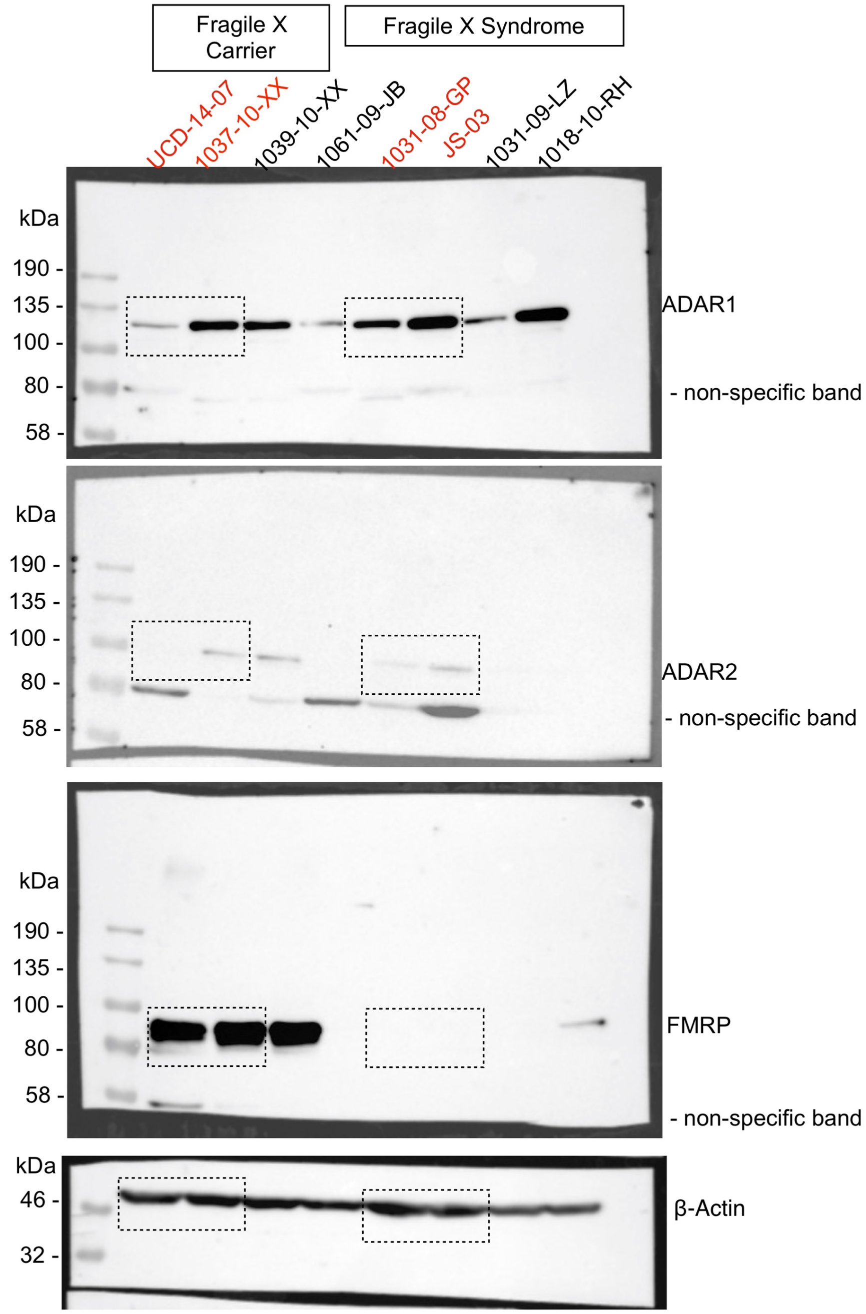
Supplementary Figure 31 related to Supplementary Fig. 21a



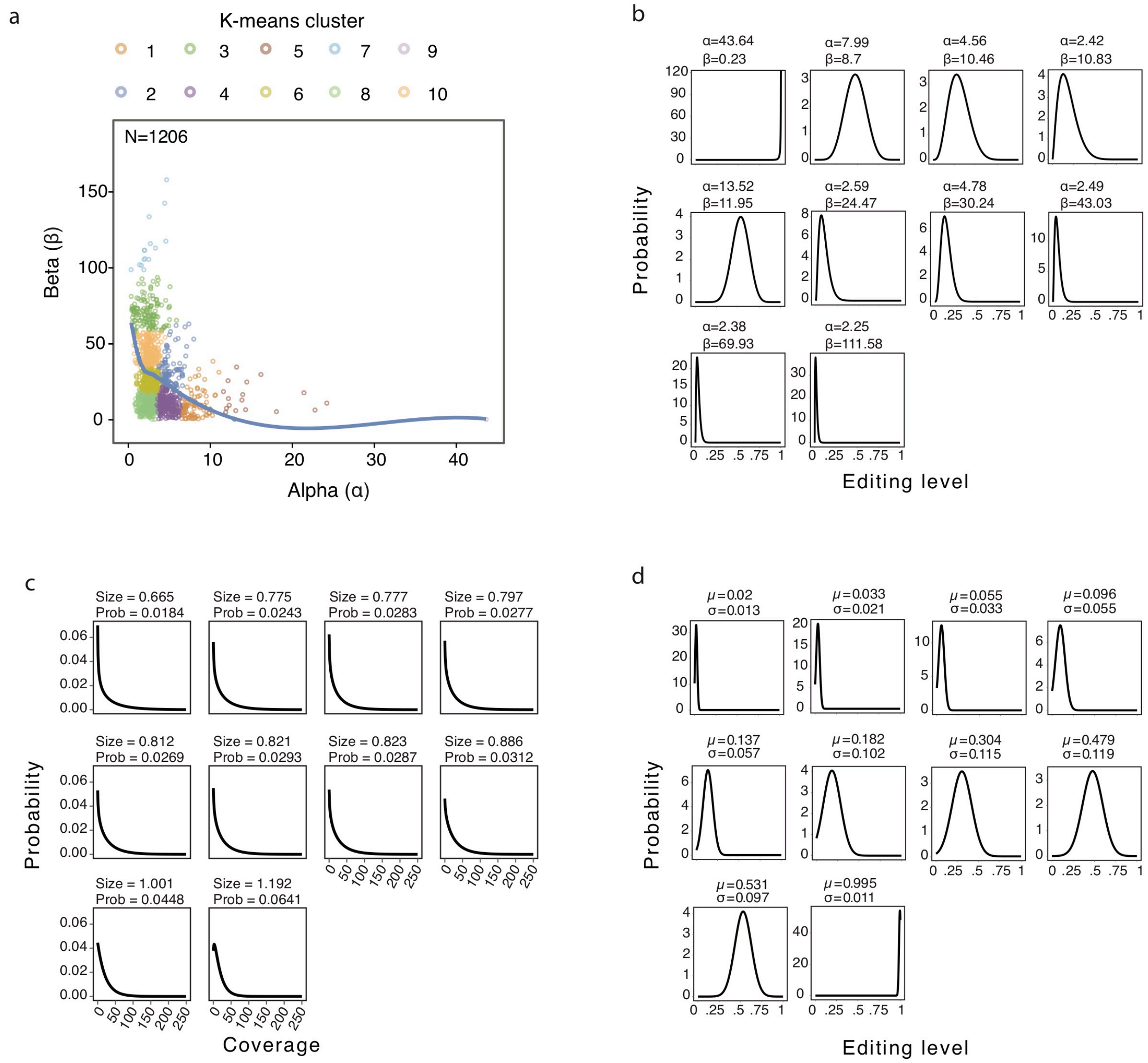
Supplementary Figure 31 related to Supplementary Fig. 25a



Supplementary Figure 31 related to Supplementary Fig. 25a

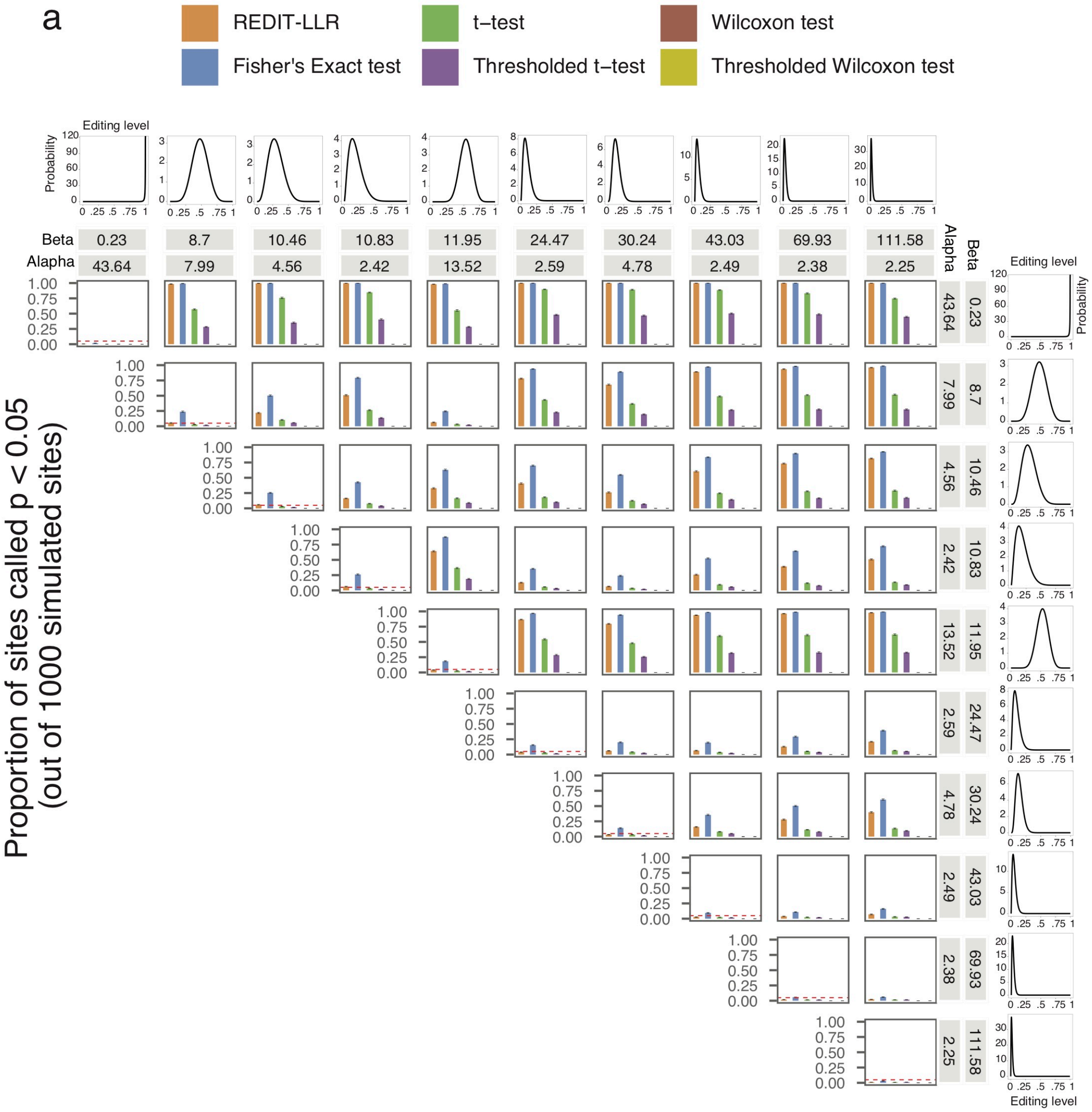


Supplemental Figures for Chapter 4



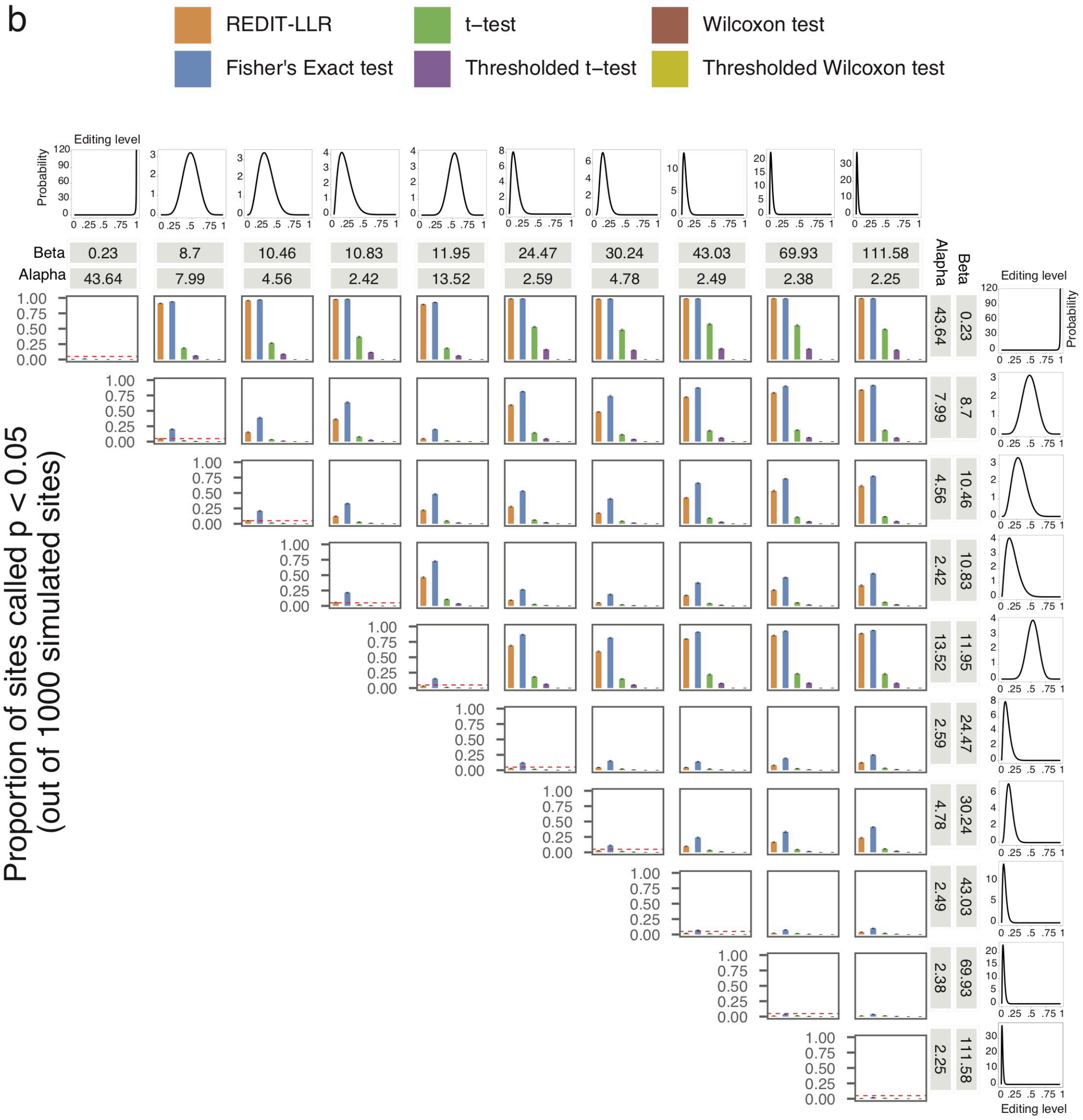
Supplemental Figure 4.1 Simulations for REDIT-LLR

Realistic distributions were simulated from GTEx to test REDIT-LLR. a) To obtain realistic editing level distributions, K-means clustering was run on alpha (α) and beta (β) parameters for beta distributions fit by maximum likelihood to editing sites from brain GTEx samples. N : number of editing sites that had ⩾ 250 samples each with ⩾ 20 reads. Blue line: loess regression line showing distribution of data point density. b) Editing level densities used in simulations were modeled with beta distributions and constructed using the medians of alpha and beta parameters from the k-means clusters from (a). c) Realistic coverage distributions per sample were modelled using negative binomial distributions. Distributions were fit by maximum likelihood from the coverages of all editing sites in 10 samples sampled from GTEx. Size and probability (Prob) parameters are annotated. d) Similar to (b) but using truncated normal distribution to model editing level distributions. Truncated normal distributions were translated by method of moments estimation from the beta distributions from (b); μ = mean, σ = variance

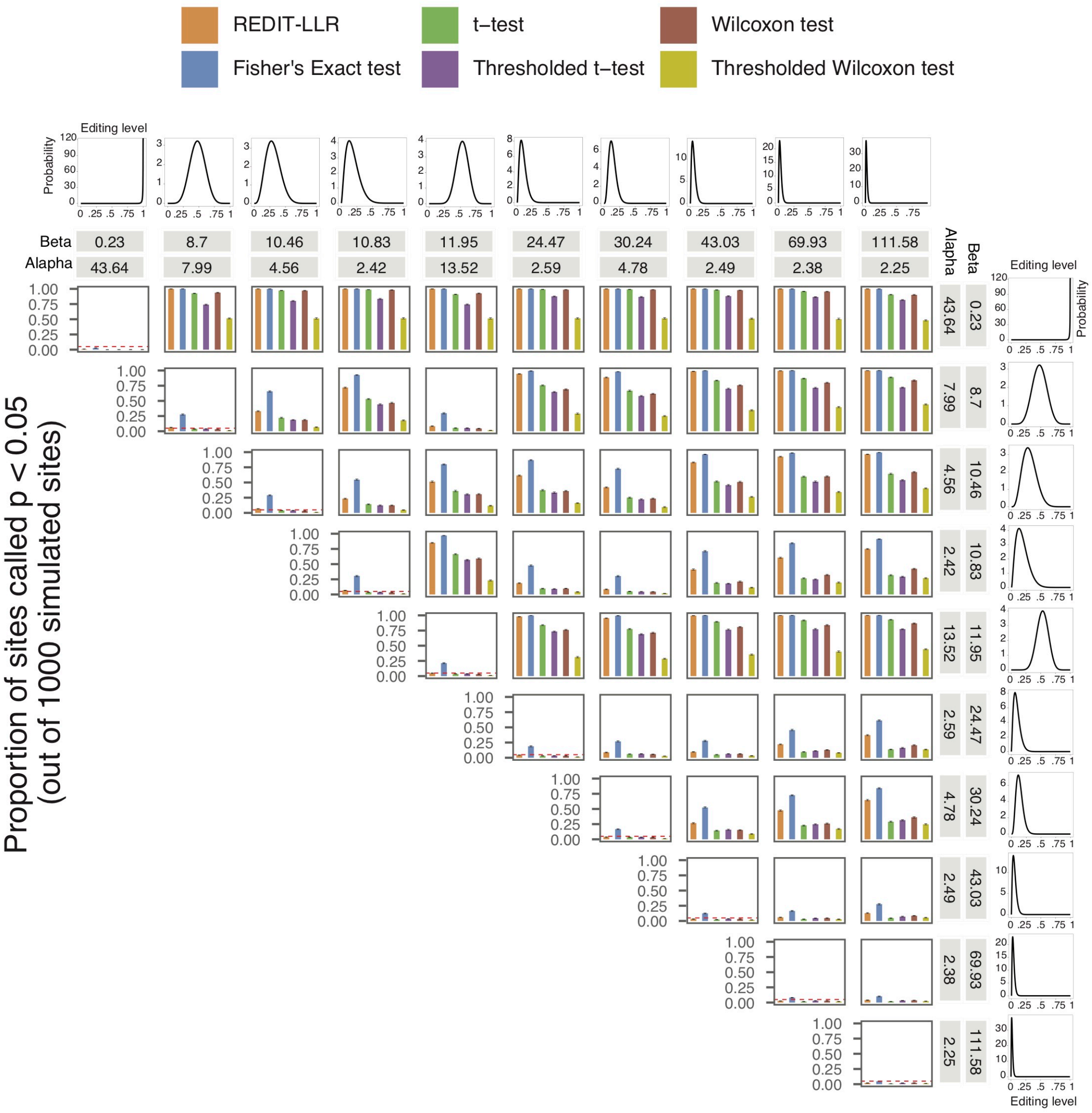

a
REDIT-LLR
Fisher's Exact test
t−test
Thresholded t−test
Wilcoxon test
Thresholded Wilcoxon test
Editing level
Probability
Beta
0.23
8.7
10.46
10.83
11.95
24.47
30.24
43.03
69.93
111.58
Alapha
43.64
7.99
4.56
2.42
13.52
2.59
4.78
2.49
2.38
2.25
0 .25 .5 .75 1
1.00
0.75
0.50
0.25
0.00
Proportion of sites called p < 0.05
(out of 1000 simulated sites)

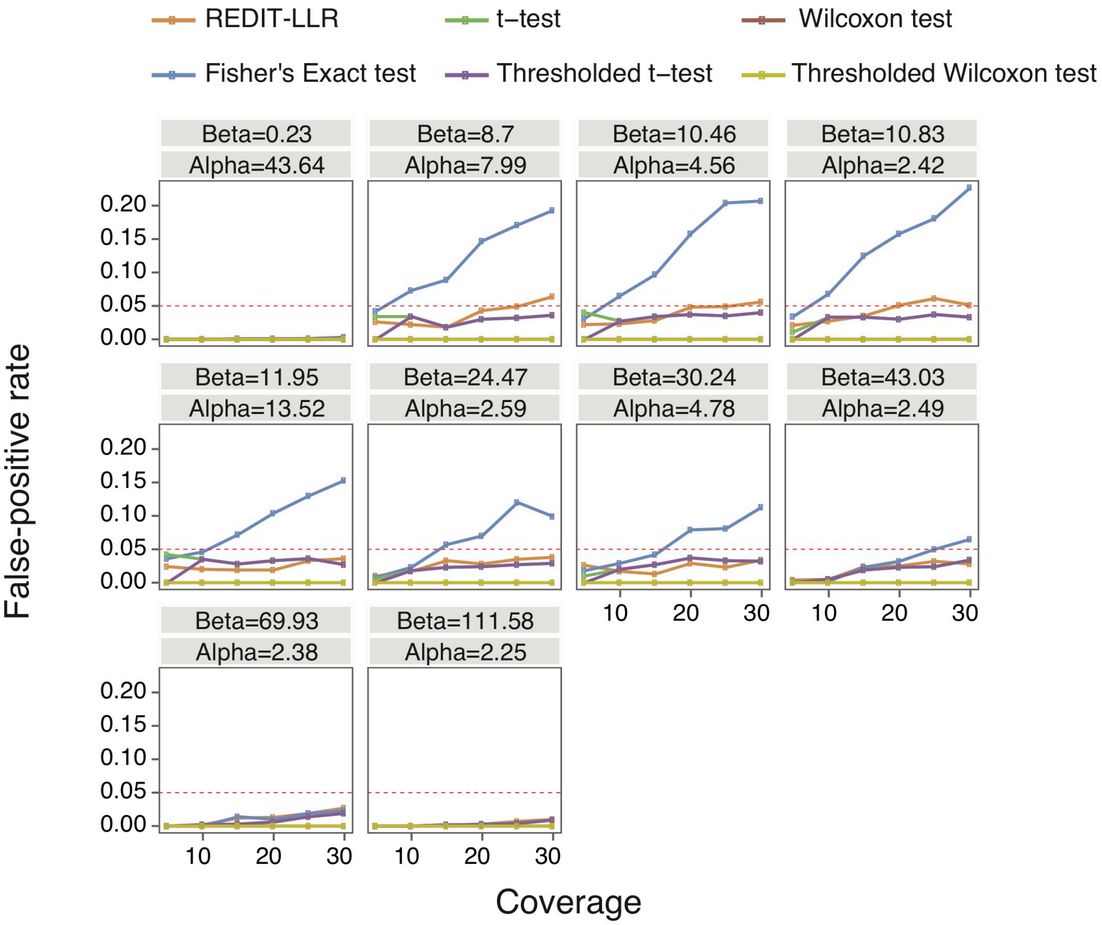
Supplemental Figure 4.2 False-positive rate and sensitivity of REDIT-LLR using simulated data

Expanding Figure 2a-b to more simulation parameters and sample sizes. Top row and rightmost column show underlying beta distributions characterizing editing levels from group1 and group2 respectively. The exact alpha and beta parameters are listed in next two rows and rightmost columns. Each bargraph shows proportion of editing sites called with p < 0.05 out of 1000 simulated editing sites. Thresholded t-test and Thresholded Wilcoxon test : t-test and Wilcoxon rank-sum test only on samples that have minimum of 10 read coverage. On-diagonal entries evaluate false-positive rates, where group1 and group2 samples are characterized by the same editing level distributions. Red lines denote the 5% false-positive threshold. Off-diagonal entries evaluate sensitivity where group1 and group2 samples are characterized by different editing distributions. Error bars show 25% and 75% quantiles from 100 independent simulations. a) Group1 and group2 have a sample size of 3. b) Sample size of 2. c) Sample size of 5. REDIT-LLR maintains highest sensitivity out of all methods that have false-positive rates at or below 5%.


b
REDIT-LLR
Fisher's Exact test
t−test
Thresholded t−test
Wilcoxon test
Thresholded Wilcoxon test
Editing level
Probability
Beta
Alapha
0.23
43.64
8.7
7.99
10.46
4.56
10.83
2.42
11.95
13.52
24.47
2.59
30.24
4.78
43.03
2.49
69.93
2.38
111.58
2.25
0 .25 .5 .75 1
1.00
0.75
0.50
0.25
0.00
Proportion of sites called p < 0.05
(out of 1000 simulated sites)

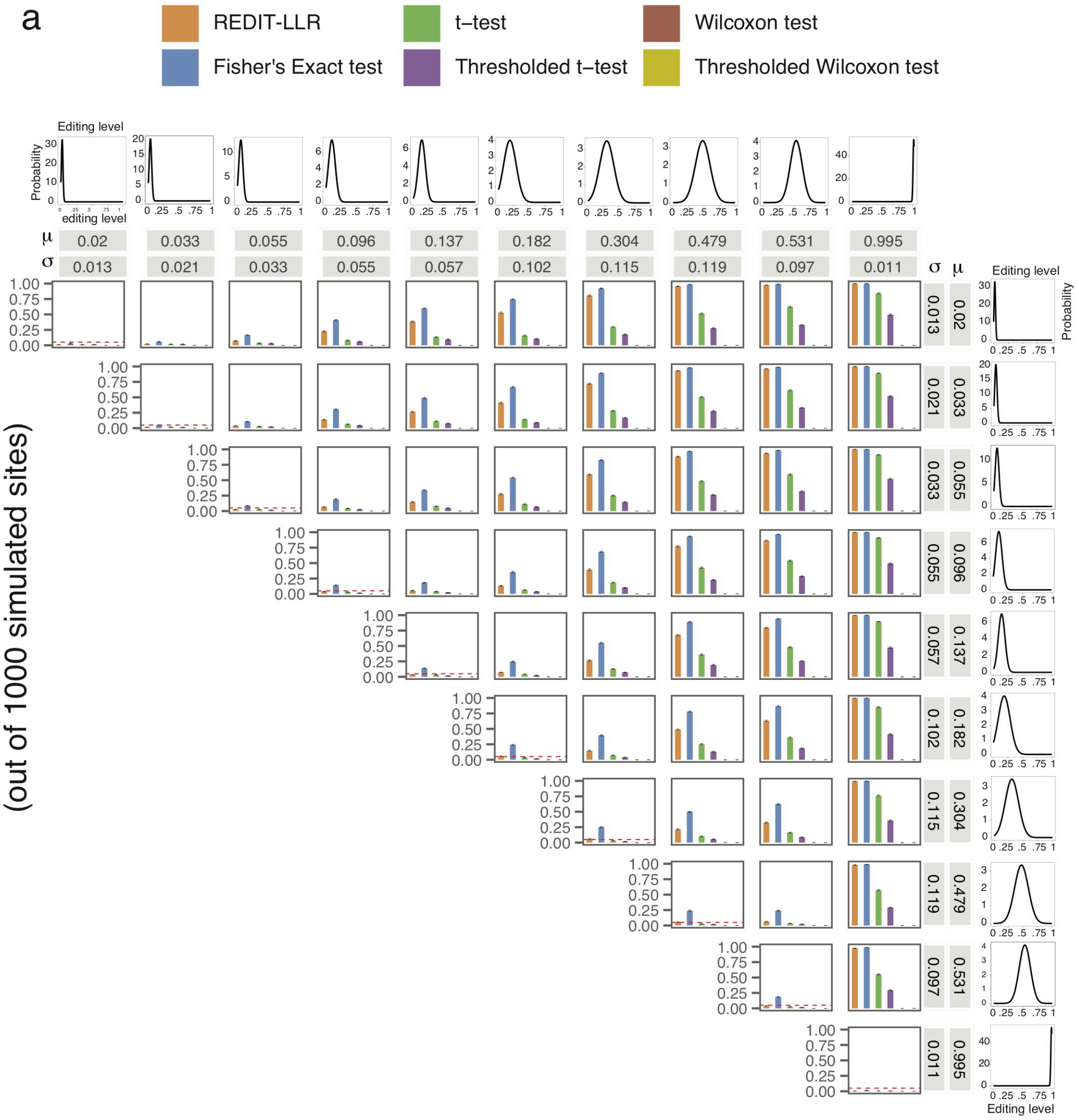
c


REDIT-LLR
t−test
Wilcoxon test
Fisher's Exact test
Thresholded t−test
Thresholded Wilcoxon test
Editing level
Probability
Beta
Alapha
0.23
8.7
10.46
10.83
11.95
24.47
30.24
43.03
69.93
111.58
43.64
7.99
4.56
2.42
13.52
2.59
4.78
2.49
2.38
2.25
1.00
0.75
0.50
0.25
0.00
0 .25 .5 .75 1
Proportion of sites called p < 0.05
(out of 1000 simulated sites)



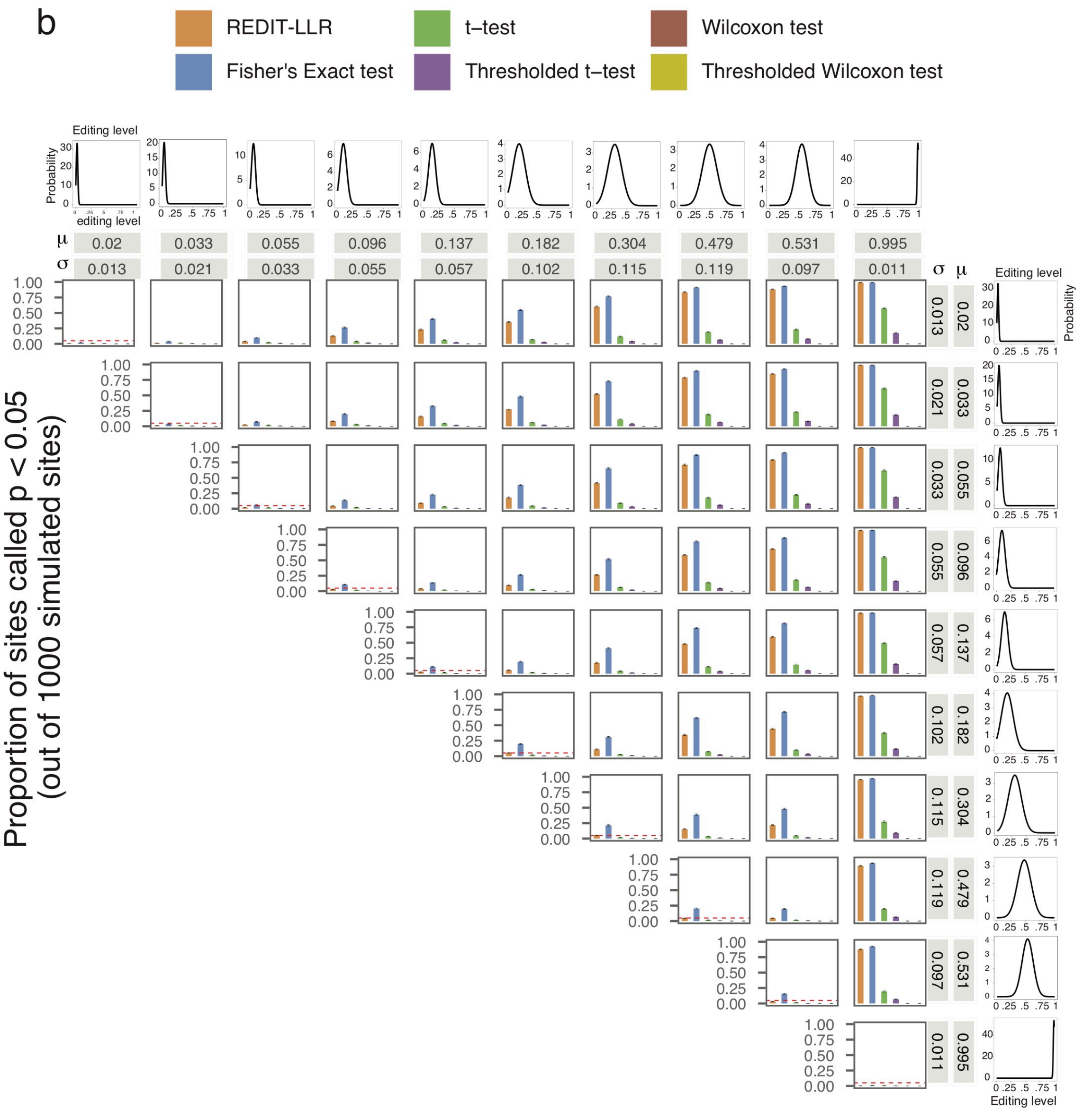
Supplemental Figure 4.3 Effect of increasing coverage on false-positive rate

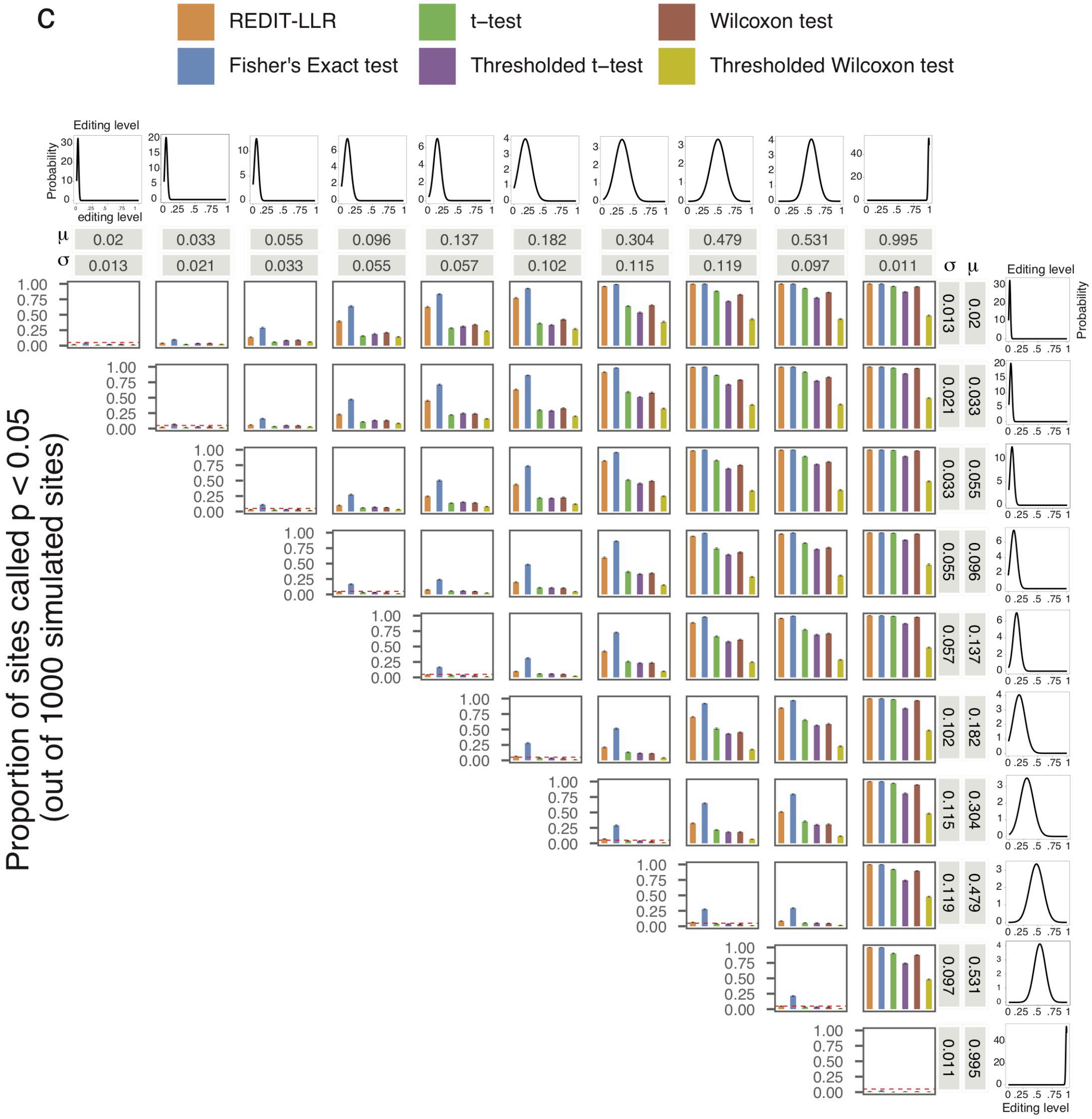
Linegraphs show the effect of increasing covering on false-positive rate. Coverage : number of reads covering the editing site. Each graph shows false-positive rate (number of sites called with $p < 0.05$ out of 1000 simulated sites) where group1 and group2 editing levels are characterized by identical beta distributions. Exact alpha and beta parameters are annotated. For all simulations both groups had a sample size of 3. Thresholded t-test and Thresholded Wilcoxon test : t-test and Wilcoxon rank-sum test only on samples that had minimum of 10 read coverage. The false-positive rate of Fisher’s Exact test inflates with increasing coverage.

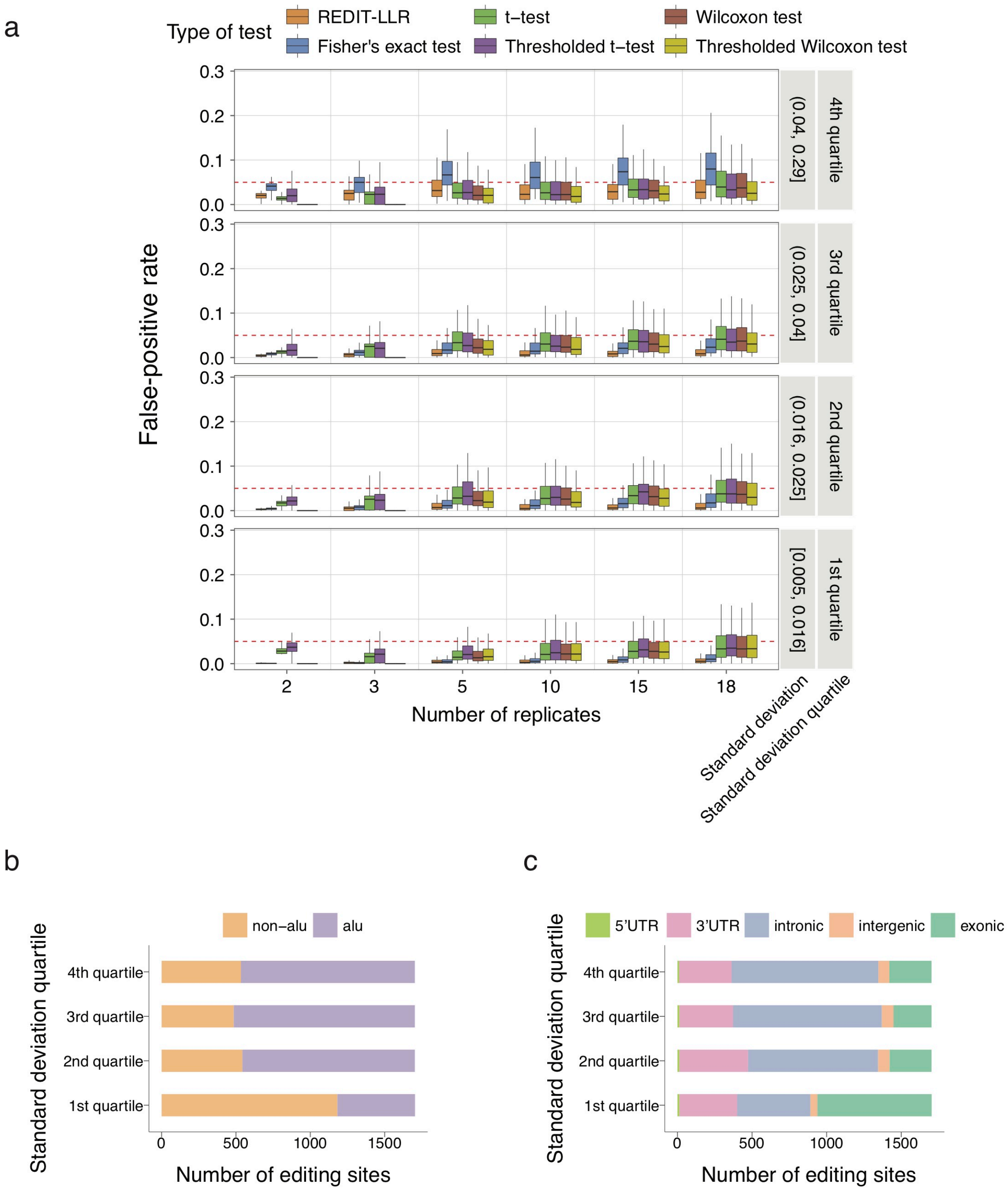


Supplemental Figure 4.4 Evaluating robustness of REDIT-LLR relative to underlying editing distribution

Similar as Supplemental Figure 2 but using truncated normal distributions instead of beta distributions to characterize editing level distributions. Top three rows and rightmost three columns depict truncated normal distributions and respective mean (μ) and standard deviation (σ) parameters. Simulations were performed for varying sample sizes. a) Group1 and group2 both had sample size of 3. b) Sample size of 2. c) sample size of 5. Even with the change in underlying distribution, REDIT-LLR retains highest sensitivity out of all methods that mitigate false-positive rates at or below 5%.

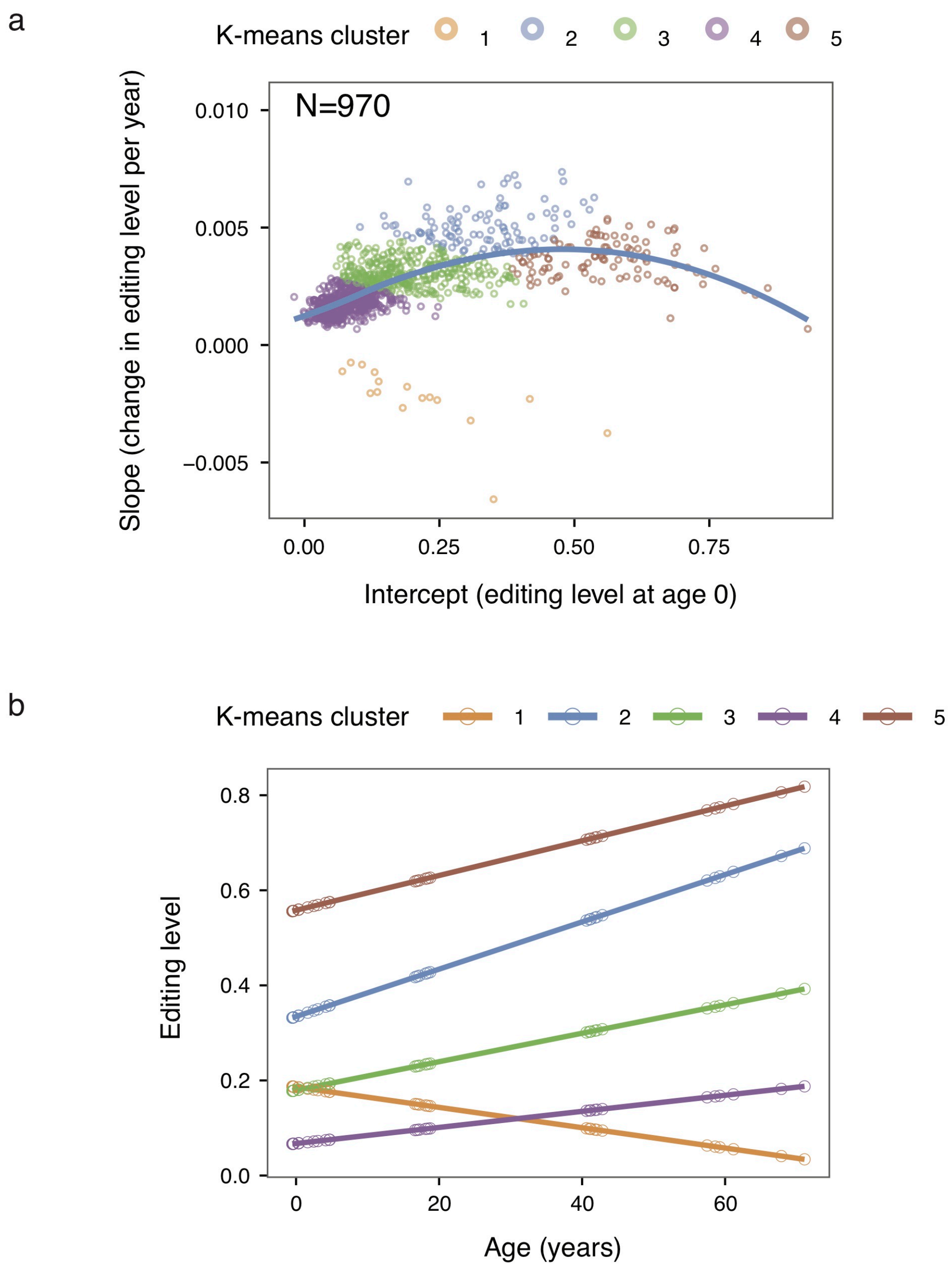



c
REDIT-LLR
Fisher's Exact test
t–test
Thresholded t–test
Wilcoxon test
Thresholded Wilcoxon test
Editing level
Probability
editing level
0 .25 .5 .75 1
μ
σ
0.02 0.033 0.055 0.096 0.137 0.182 0.304 0.479 0.531 0.995
0.013 0.021 0.033 0.055 0.057 0.102 0.115 0.119 0.097 0.011
1.00 0.75 0.50 0.25 0.00
Proportion of sites called p < 0.05
(out of 1000 simulated sites)


a
Type of test
REDIT-LLR
Fisher's exact test
t−test
Thresholded t−test
Wilcoxon test
Thresholded Wilcoxon test
False-positive rate
0.0
0.1
0.2
0.3
4th quartile
(0.04, 0.29]
3rd quartile
(0.025, 0.04]
2nd quartile
(0.016, 0.025]
1st quartile
[0.005, 0.016]
2
3
5
10
15
18
Number of replicates
Standard deviation
Standard deviation quartile
b
non−alu
alu
Standard deviation quartile
4th quartile
3rd quartile
2nd quartile
1st quartile
0
500
1000
1500
Number of editing sites
c
5'UTR
3'UTR
intronic
intergenic
exonic
Standard deviation quartile
4th quartile
3rd quartile
2nd quartile
1st quartile
0
500
1000
1500
Number of editing sites

Supplemental Figure 4.5 False-positive evaluation of the REDIT-LLR on real data

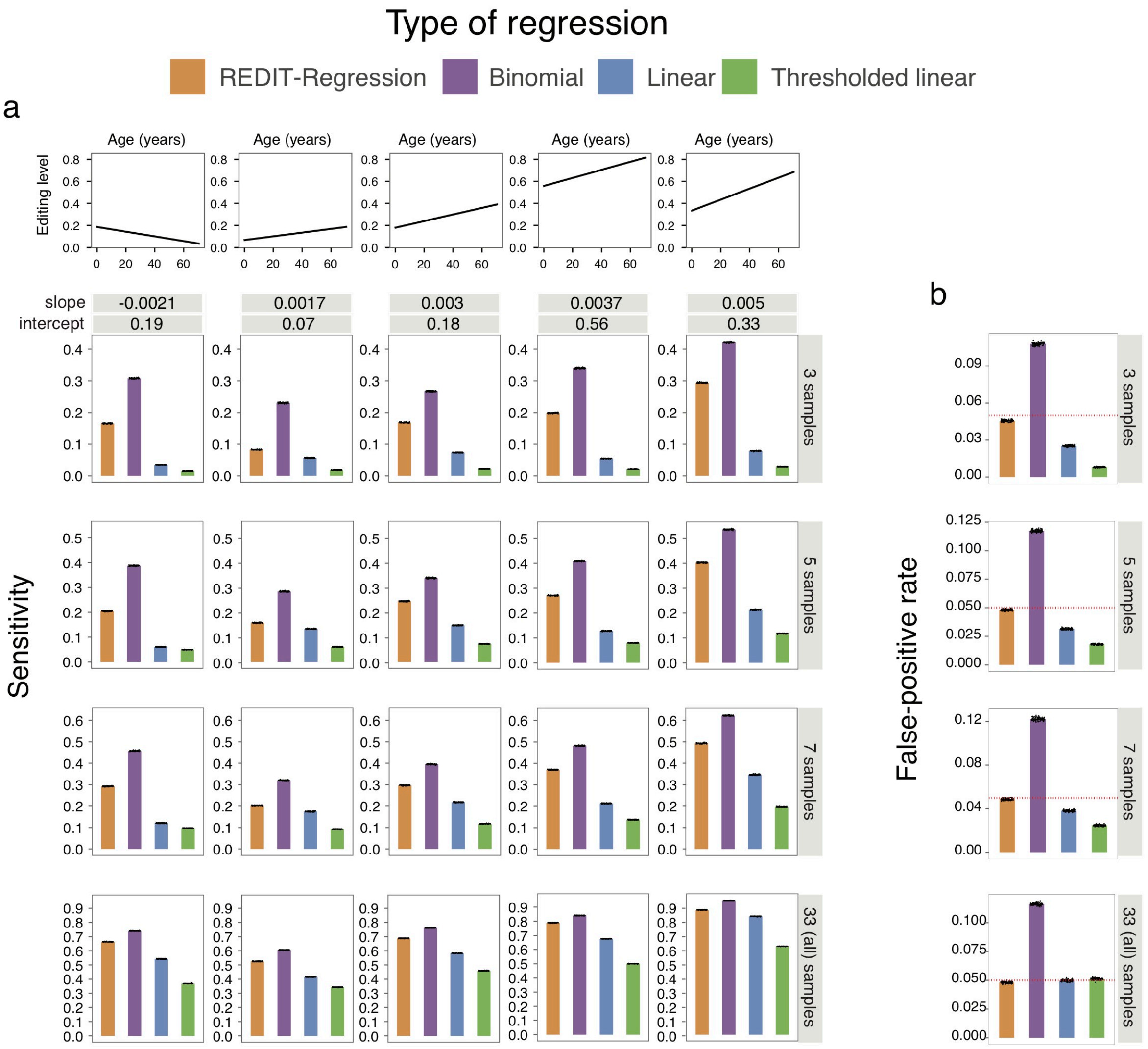
a) False-positive rate (proportion of sites called with $p < 0.05$) of various statistical tests on real data of 6814 editing sites. For each given sample size (x axis), 100 random subsamplings of control samples were compared against each other. Editing sites tested are partitioned into four quartiles according to standard deviation calculated using all 18 control samples. Red dotted line denotes the 5% false-positive threshold. Thresholded t-test = t-test on samples with at least 10 reads, Thresholded Wilcoxon test = Wilcoxon test on samples with at least 10 reads. b) Alu compsition of editing sites within each standard deviation quartile. c) Similar to (b) but partitioned according to genomic locations of the editing sites.



Supplemental Figure 4.6 Simulations for REDIT-Regression

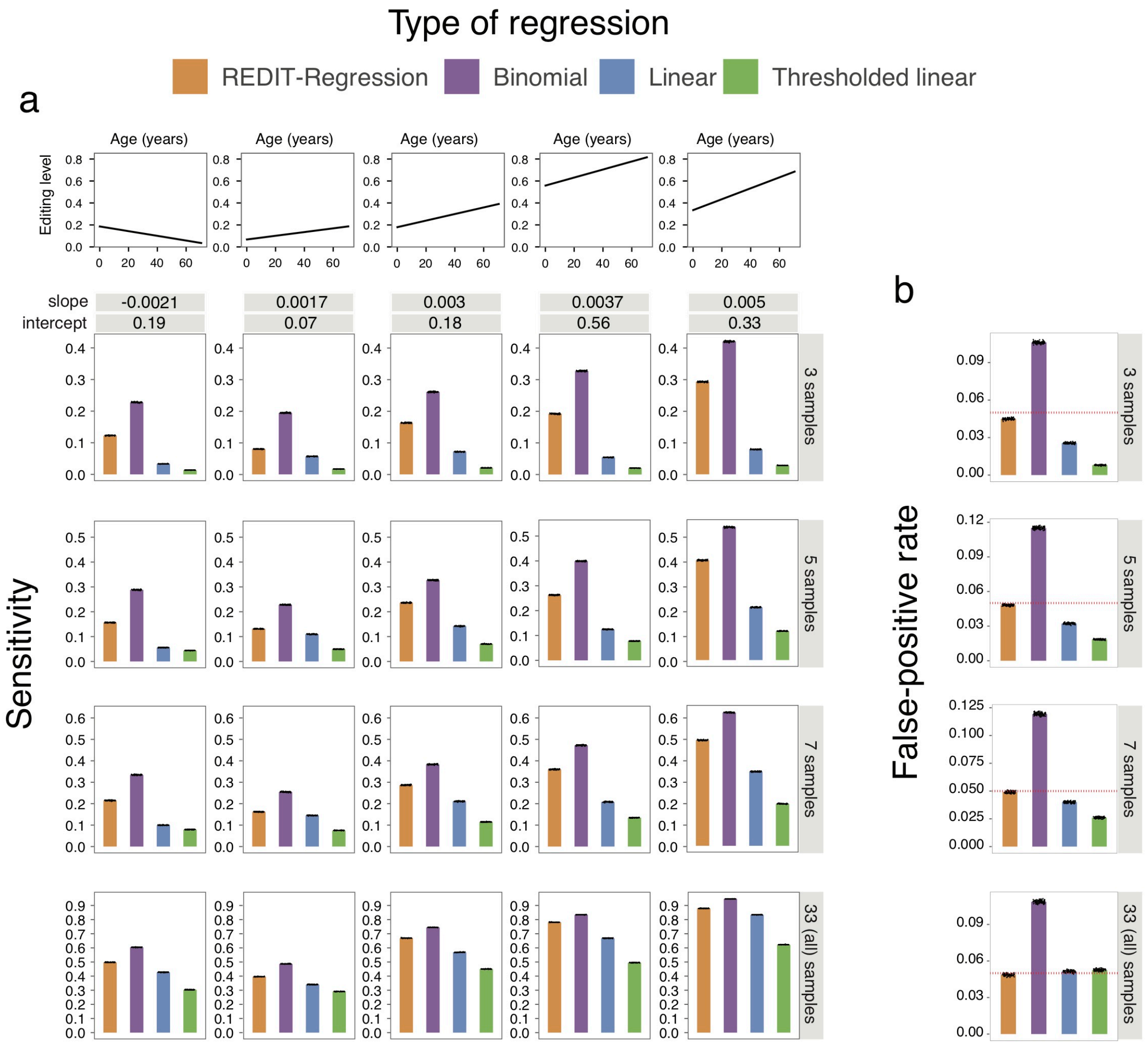
REDIT-Regression was tested on data simulating effects of age on RNA editing.

Simulated parameters of effect of age (editing level= β_{age}*age+ β_0) on editing level was calculated using a dataset of 33 postmortem brains spanning human development (methods). a) K-means clustering of slope (β_{age}) and intercept (β_0) parameters that were estimated from linear regression on 970 editing sites that all had high coverage (methods). Blue line: loess curve depicting density of the points b) Graphical depiction of the medians of the five k-means clusters. Open circles show the ages of the 33 samples.



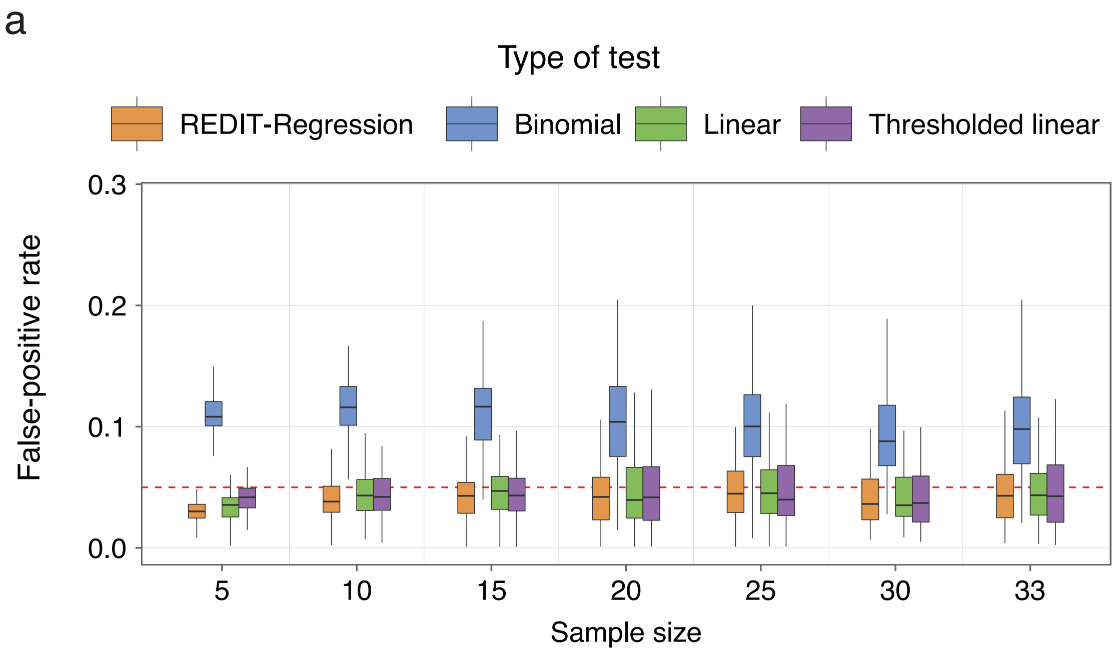
Supplemental Figure 4.7 False-positive rate and sensitivity of REDIT-Regression on simulated data

Expanding Figure 2c,d to more simulation parameters and sample sizes. a) Sensitivity (true-positive rate) evaluated using simulated data (Methods). Editing levels were characterized as beta distributions with mean editing level having linear dependence on age. The linear dependence is illustrated and annotated in the top 3 rows. Varying sample sizes are shown in columns. b) Similar to (a) but evaluating false-positive rate where editing has no dependence on age. Red dotted line indicates the 5% false positive threshold. Individual points show 100 independent replications of the simulations. Binomial : binomial regression; linear : linear regression; Thresholded linear : linear regression on samples that have minimal 10 read coverage per editing site. Out of methods that mitigate false-positives to 5%, REDIT-Regression has the highest sensitivity.


Type of regression
🟧 REDIT-Regression
🟪 Binomial
🟦 Linear
🟩 Thresholded linear
a
Age (years)
Editing level
0.0
0.2
0.4
0.6
0.8
0
20
40
60
slope
intercept
-0.0021
0.19
0.0017
0.07
0.003
0.18
0.0037
0.56
0.005
0.33
Sensitivity
3 samples
5 samples
7 samples
33 (all) samples
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
b
False-positive rate
0.00
0.03
0.06
0.09
0.12
0.000
0.025
0.050
0.075
0.100
0.125

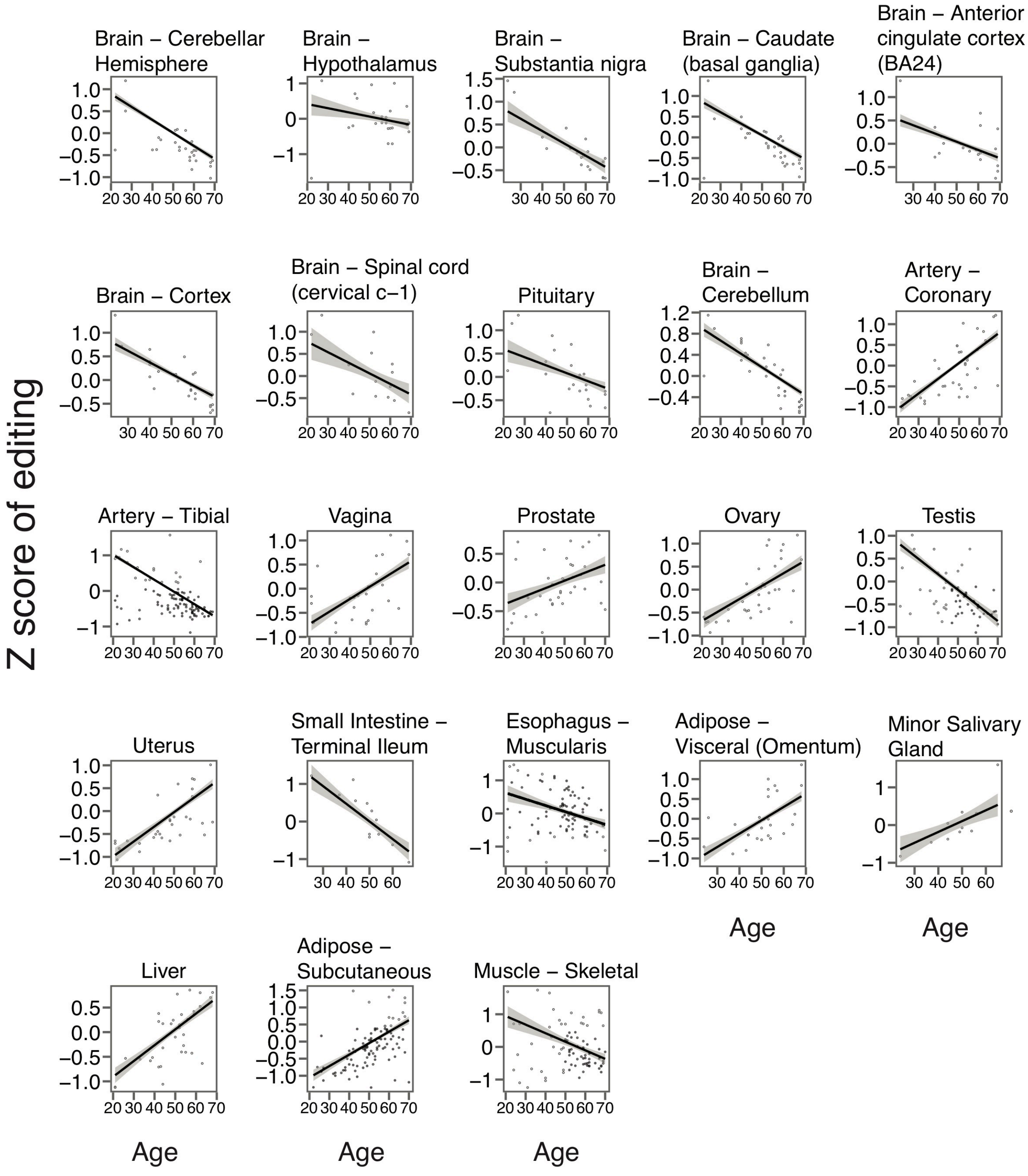
Supplemental Figure 4.8 Robustness of REDIT-Regression relative to underlying distribution of editing

Similar to Supplemental Figure 7 but characterizing editing levels using a truncated gaussian distribution rather than beta distribution a) Evaluations of sensitivity. b) Evaluations of false-positive rate.



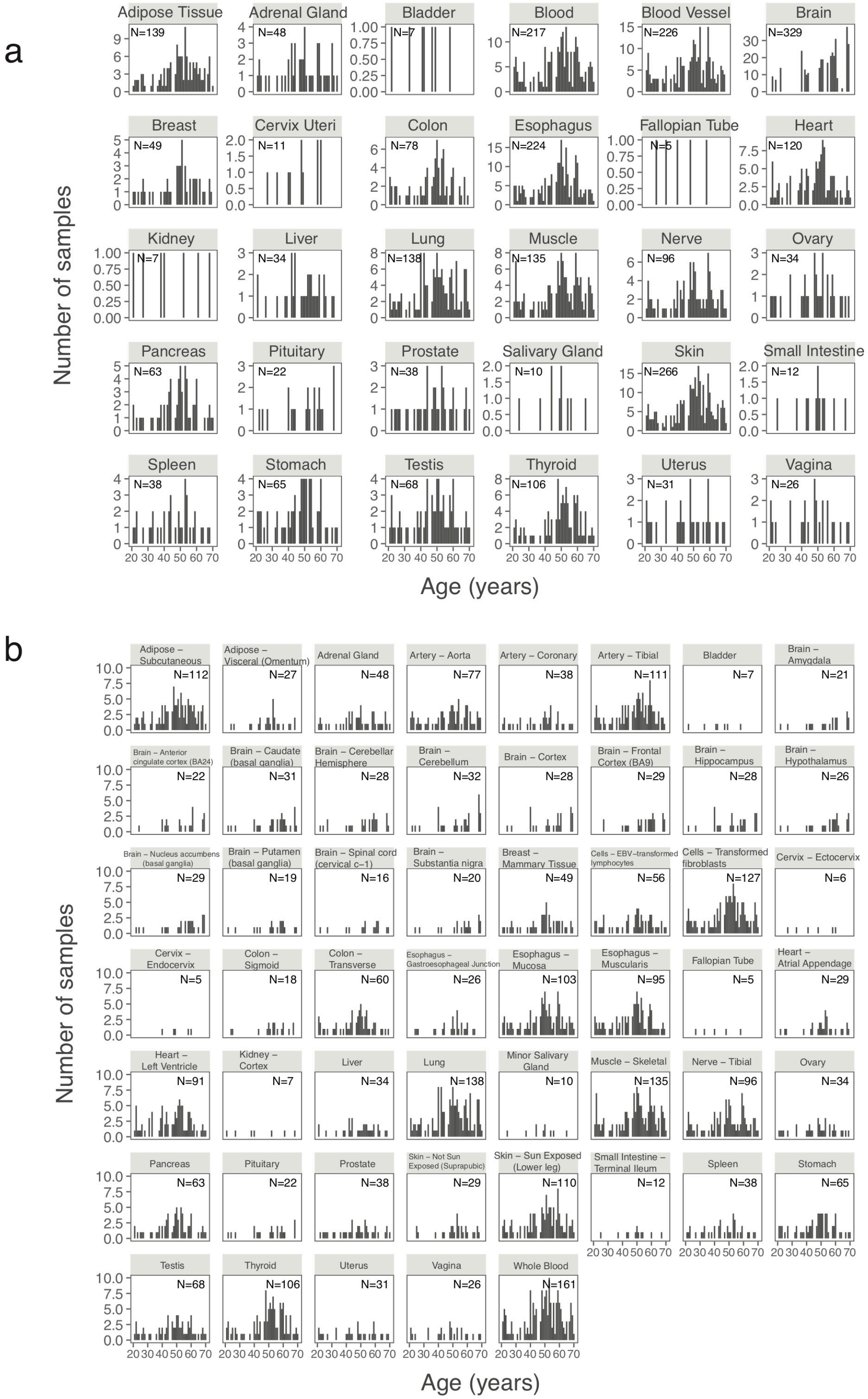
Supplemental Figure 4.9 False-positive evaluation of REDIT-Regression using real data

a) False-positive rate (proportion of sites called with p < 0.05) of various statistical tests on real data of 267,766 editing sites from 33 postmortem brains. For each given sample size (x axis), 100 random subsamplings of samples were selected and assigned a randomized covariate value (Methods). binomial : binomial regression, linear : linear regression, thresholded linear : linear regression using samples with at least 10 reads. Red dotted line shows the 5% false-positive threshold.


Z score of editing
Brain – Cerebellar Hemisphere
Brain – Hypothalamus
Brain – Substantia nigra
Brain – Caudate (basal ganglia)
Brain – Anterior cingulate cortex (BA24)
Brain – Cortex
Brain – Spinal cord (cervical c–1)
Pituitary
Brain – Cerebellum
Artery – Coronary
Artery – Tibial
Vagina
Prostate
Ovary
Testis
Uterus
Small Intestine – Terminal Ileum
Esophagus – Muscularis
Adipose – Visceral (Omentum)
Minor Salivary Gland
Liver
Adipose – Subcutaneous
Muscle – Skeletal
Age
Age
Age
Age
Age

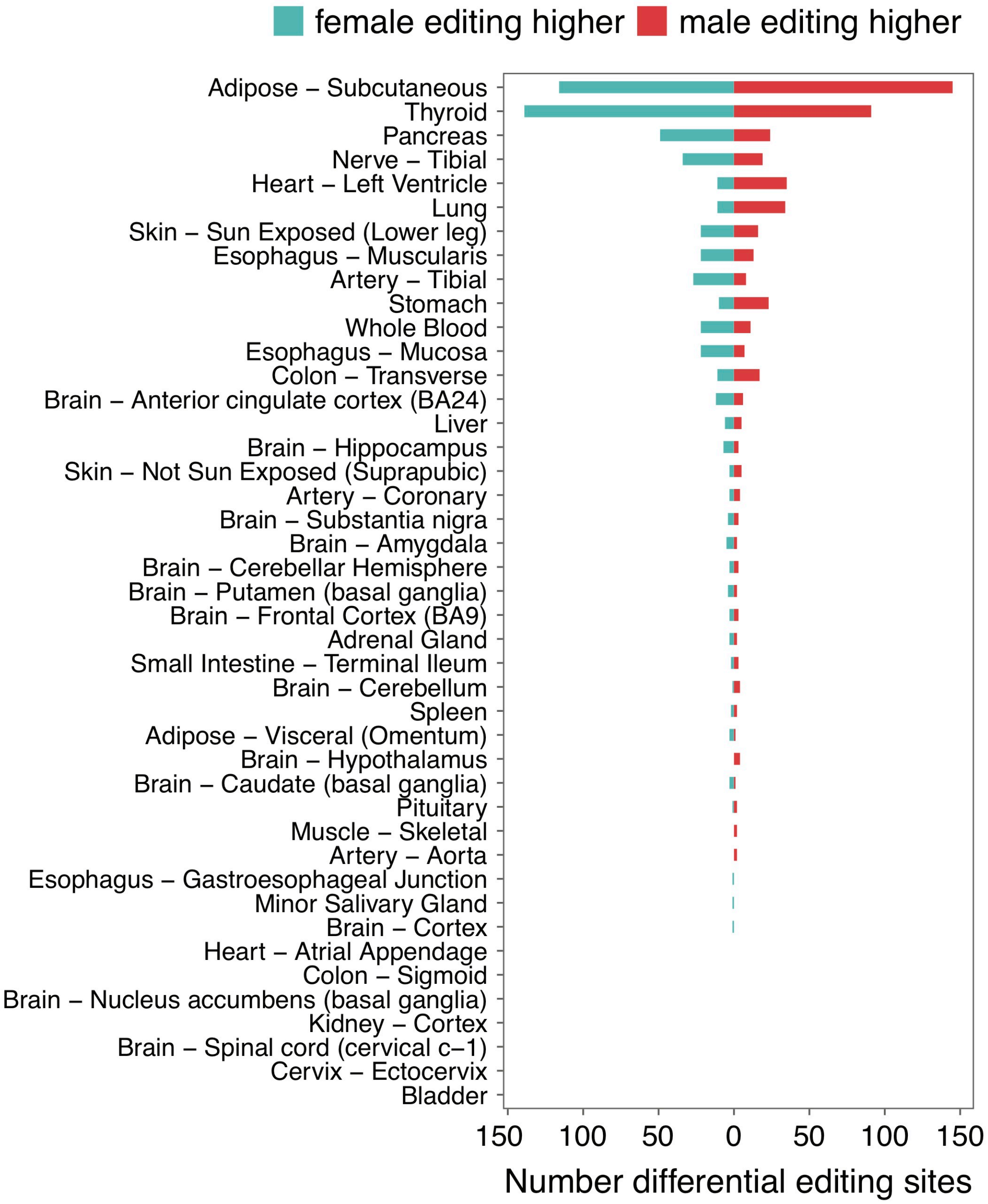
Supplemental Figure 4.10 Overall projections of RNA editing over age partitioned by body sites

Same as Figure 3 but partitioning samples by body site. Only body sites exhibiting homogenous trajectories of editing over age are plotted (Supplemental Table 2b).


a
Number of samples
Age (years)
Adipose Tissue N=139
Adrenal Gland N=48
Bladder N=7
Blood N=217
Blood Vessel N=226
Brain N=329
Breast N=49
Cervix Uteri N=11
Colon N=78
Esophagus N=224
Fallopian Tube N=5
Heart N=120
Kidney N=7
Liver N=34
Lung N=138
Muscle N=135
Nerve N=96
Ovary N=34
Pancreas N=63
Pituitary N=22
Prostate N=38
Salivary Gland N=10
Skin N=266
Small Intestine N=12
Spleen N=38
Stomach N=65
Testis N=68
Thyroid N=106
Uterus N=31
Vagina N=26
20 30 40 50 60 70
b
Number of samples
Age (years)
Adipose – Subcutaneous N=112
Adipose – Visceral (Omentum) N=27
Adrenal Gland N=48
Artery – Aorta N=77
Artery – Coronary N=38
Artery – Tibial N=111
Bladder N=7
Brain – Amygdala N=21
Brain – Anterior cingulate cortex (BA24) N=22
Brain – Caudate (basal ganglia) N=31
Brain – Cerebellar Hemisphere N=28
Brain – Cerebellum N=32
Brain – Cortex N=28
Brain – Frontal Cortex (BA9) N=29
Brain – Hippocampus N=28
Brain – Hypothalamus N=26
Brain – Nucleus accumbens (basal ganglia) N=29
Brain – Putamen (basal ganglia) N=19
Brain – Spinal cord (cervical c–1) N=16
Brain – Substantia nigra N=20
Breast – Mammary Tissue N=49
Cells – EBV–transformed lymphocytes N=56
Cells – Transformed fibroblasts N=127
Cervix – Ectocervix N=6
Cervix – Endocervix N=5
Colon – Sigmoid N=18
Colon – Transverse N=60
Esophagus – Gastroesophageal Junction N=26
Esophagus – Mucosa N=103
Esophagus – Muscularis N=95
Fallopian Tube N=5
Heart – Atrial Appendage N=29
Heart – Left Ventricle N=91
Kidney – Cortex N=7
Liver N=34
Lung N=138
Minor Salivary Gland N=10
Muscle – Skeletal N=135
Nerve – Tibial N=96
Ovary N=34
Pancreas N=63
Pituitary N=22
Prostate N=38
Skin – Not Sun Exposed (Suprapubic) N=29
Skin – Sun Exposed (Lower leg) N=110
Small Intestine – Terminal Ileum N=12
Spleen N=38
Stomach N=65
Testis N=68
Thyroid N=106
Uterus N=31
Vagina N=26
Whole Blood N=161
0.0 2.5 5.0 7.5 10.0
20 30 40 50 60 70

Supplemental Figure 4.11 Age ranges of GTEx samples

Histograms of the ages of samples from GTEx in REDIportal. a) Samples partitioned by histological type. b) Samples partitioned by body site. N = number of total samples.


female editing higher
male editing higher
Adipose – Subcutaneous
Thyroid
Pancreas
Nerve – Tibial
Heart – Left Ventricle
Lung
Skin – Sun Exposed (Lower leg)
Esophagus – Muscularis
Artery – Tibial
Stomach
Whole Blood
Esophagus – Mucosa
Colon – Transverse
Brain – Anterior cingulate cortex (BA24)
Liver
Brain – Hippocampus
Skin – Not Sun Exposed (Suprapubic)
Artery – Coronary
Brain – Substantia nigra
Brain – Amygdala
Brain – Cerebellar Hemisphere
Brain – Putamen (basal ganglia)
Brain – Frontal Cortex (BA9)
Adrenal Gland
Small Intestine – Terminal Ileum
Brain – Cerebellum
Spleen
Adipose – Visceral (Omentum)
Brain – Hypothalamus
Brain – Caudate (basal ganglia)
Pituitary
Muscle – Skeletal
Artery – Aorta
Esophagus – Gastroesophageal Junction
Minor Salivary Gland
Brain – Cortex
Heart – Atrial Appendage
Colon – Sigmoid
Brain – Nucleus accumbens (basal ganglia)
Kidney – Cortex
Brain – Spinal cord (cervical c−1)
Cervix – Ectocervix
Bladder
150
100
50
0
50
100
150
Number differential editing sites

Supplemental Figure 4.12 Gender-biased RNA editing in various tissues

Similar to Figure 4, but partitioning GTEx samples according to body site.