2021

A-to-I RNA Editing in Human Cells

Tony Sun

A-to-I RNA editing in human cells

A Thesis Presented to the Faculty of

The Rockefeller University

in Partial Fulfillment of the Requirements for

the degree of Doctor of Philosophy

by

Tony Sun

June 2021

A-to-I RNA editing in human cells

Tony Sun, Ph.D.

The Rockefeller University 2021

RNA editing is a means of diversifying the transcriptome and regulating innate immunity. Among the different classes of enzymes that modify RNA, adenosine deaminase acting on RNA (ADAR) is a type that catalyzes adenosine-to-inosine editing on double-stranded RNA molecules to regulate cellular responses to endogenous and exogenous RNA. Of the three ADAR homologs in humans, dysregulation of ADAR1 editing due to inherited mutations leads to disorders such as Aicardi-Goutieres syndrome, an inflammatory disease that manifests in the brain and skin, and dyschromatosis symmetrica hereditaria, a skin pigmentation disorder. ADAR1 is the primary A-to-I editor of RNA in humans, and the majority of edit sites are found in a class of repetitive elements called Alu, many of which are located in introns and 3′ untranslated regions of RNA. The functional consequences of A-to-I editing are varied, although a complete lack of functional ADAR1 is usually not tolerated, as revealed by the MDA5-mediated embryonic lethality in mice lacking functional ADAR1. In human neural progenitor cells, loss of ADAR1 causes spontaneous upregulation of interferon and cell death, although the RNA triggers remain unknown. Given the importance of ADAR1-editing in maintaining homeostasis in various contexts, there is a need to understand in more detail how ADAR1 isoforms are regulated and how they individually contribute to the A-to-I RNA editome.

Two ADAR1 protein isoforms, p110 (110 kDa) and p150 (150 kDa), are expressed constitutively and in response to interferon, respectively, but the contribution of each isoform to the editing landscape remains incompletely characterized, largely because of the challenges in expressing p150 without p110. We revealed that the p110 isoform can be expressed from the canonical p150-encoding mRNA due to leaky ribosome scanning downstream of the p150 start codon. Synonymous mutations introduced in the region between the p150 and p110 start codons reduce leaky scanning and usage of the p110 start codon, and cells expressing p150 constructs with these mutations produce significantly reduced levels of p110.

With the ability to express p150 with significantly reduced levels of p110, the A-to-I editome can be classified in terms of p150-selective and p110-selective sites, allowing evaluation of the relative contributions of either isoform to global editing

levels. Our editing analysis revealed that the majority of ADAR1-edit sites are p150-selective, although a significant proportion of ADAR1-edit sites are also shared between p150 and p110, being not dependent on presence of either isoform for editing to occur. Of the sites that are putatively p110-selective, the majority are located in introns.

Finally, the ability of p150 mRNA to give rise to p110 means that p110 is also an interferon-inducible protein alongside the canonical interferon-stimulated ADAR1 isoform: p150. During the interferon response, the transcriptome changes, and many new mRNA structures, perhaps some immunogenic ones, will enter the nucleus and cytoplasm. The distribution of ADAR1 isoforms is such that p110 is mostly present in the nucleus, and p150 mostly in the cytoplasm. We propose that optimal editing in the nucleus and cytoplasm during the interferon response is achieved by the inducibility of p110 and p150, both of which share a large number of target sites.

I dedicate this thesis to Charles Rice and Hachung Chung.

ACKNOWLEDGMENTS

I acknowledge Charles Rice and Hachung Chung for their training and encouragement during my PhD years. The content in this thesis represents how I have learned to think as a scientist as a result of their guidance. Of course, nearly every member of the Rice lab, past and present, have also played some role in helping me reach this stage of my career, and I acknowledge their help over the years. I also thank members of the Rockefeller graduate school office and MD-PhD office for advice and help with organization and deadlines during my research years. I thank my committee members, Luciano Marraffini, Paul Bieniasz, and Charles Samuel, for their encouragement and advice over the years. I thank Abby Walczak for inspiring the phylogenetic work presented in section 2.5 of the thesis. Finally, I thank family and friends for general support, and in particular, my father, James Sun, for teaching me basic principles of using command line to access programs for analysis of sequencing data.

TABLE OF CONTENTS

ACKNOWLEDGMENTS iv
TABLE OF CONTENTS v
LIST OF FIGURES vi
LIST OF SUPPLEMENTARY MATERIALS viii
LIST OF ABBREVIATIONS ix
CHAPTER 1. Introduction to RNA editing 1
1.1 RNA can be modified after transcription 1
1.2 Discovery of adenosine deaminase acting on RNA 3
1.3 Evolution of the ADAR field 16
CHAPTER 2. Regulation and function of ADAR1..... 30
2.1 ADAR1 DNA regulation 30
2.2 ADAR1 splice variants 45
2.3 Expression of ADAR1 p110 and p150 54
2.4 Function of ADAR1 in neural progenitor cells 62
2.5 Evolution of ADAR1 isoforms 74
CHAPTER 3. Translational regulation of ADAR1 p110..... 79
3.1 ATG codon organization in ADAR1 79
3.2 Regulation of p150 and p110 translation 84
3.3 ADAR1 p150 and p110 isoform-specific cell lines 94
CHAPTER 4. Taxonomy of ADAR1 A-to-I edit sites..... 101
4.1 Clones for A-to-I editing analysis 101
4.2 Editing analysis workflow and taxonomy of sites ... 103
4.3 Implications of an A-to-I editing taxonomy 115
CHAPTER 5. Future directions and concluding remarks. 122
5.1 Endogenous expression of p110 and p150 122
5.2 Amplicon sequencing with unique molecular tags 125
SUPPLEMENTARY MATERIALS. 132
WORKS CITED. 136

LIST OF FIGURES

Figure 1.1.1 RNA editing of coxII in protists 2
Figure 1.1.2 RNA editing of ApoB in humans 3
Figure 1.2.1 RNA duplex unwinding in frog eggs 4
Figure 1.2.2 RNA duplex is degraded by RNase A 5
Figure 1.2.3 RNA duplex is structurally altered 6
Figure 1.2.4 The modified A migrates with G 7
Figure 1.2.5 The modified A elutes with inosine 8
Figure 1.2.6 Double-stranded RNA substrates are edited 9
Figure 1.2.7 Fractionation of frog egg extracts 10
Figure 1.2.8 Electrophoresis of deaminase-active regions . 11
Figure 1.2.9 Cloning of ADAR cDNA 12
Figure 1.2.10 Mapping of ADAR DNA sequence 14
Figure 1.2.11 Two ADAR protein isoforms 15
Figure 1.2.12 Two ADAR protein isoforms in situ 16
Figure 1.2.13 ADAR genomic organization 18
Figure 1.2.14 ADAR exon 1 upstream sequences 19
Figure 1.2.15 ADAR exon 1 is interferon responsive 20
Figure 1.2.16 Annotated sequences upstream of exon 1B 21
Figure 1.2.17 ADAR exon 1A promoter assays 22
Figure 1.2.18 ADAR exon 1B promoter assays 24
Figure 1.2.19 Relative expression of exons 1A and 1B 25
Figure 1.2.20 ADAR exons 6 and 7 splice isoforms 26
Figure 1.2.21 Expression of ADAR exon 6 and 7 variants ... 28
Figure 2.1.1 Plasmids for CRISPR 31
Figure 2.1.2 Transfection of 293T cells for KI 35
Figure 2.1.3 Flow analysis of bulk GFP-KI clones 36
Figure 2.1.4 Immunoblot analysis of bulk GFP-KI clones ... 37
Figure 2.1.5 NCBI RefSeq curated annotations for ADAR1 ... 39
Figure 2.1.6 Expression levels of ADAR1 RNA isoforms 40
Figure 2.1.7 ADAR1 exon 1C is an Alu element 43
Figure 2.2.1 Splice junctions of ADAR1 RNA isoforms 45
Figure 2.2.2 Splice donors downstream of exon 1C 47
Figure 2.2.3 Next generation sequencing bias 49
Figure 2.2.4 Novel 1A-1C splice junction 50
Figure 2.2.5 Nanopore dataset: ADAR1 RNA isoforms 53
Figure 2.3.1 Exon 1B knockout cells 56
Figure 2.3.2 Exon 1B/1C double-knockout cells 58
Figure 2.3.3 Exon 1B/1C/1A triple-knockout cells 60
Figure 2.4.1 Differentiation of hESCs into NPCs 64
Figure 2.4.2 NPC immunocytochemistry 65
Figure 2.4.3 Upregulation of IFN in ADAR1 KO NPCs 67
Figure 2.4.4 Cell death in ADAR1 KO NPCs 69
Figure 2.4.5 ISG levels in ADAR1 KO hepatocytes 71
Figure 2.4.6 ADAR1 KO hepatocytes respond to interferon .. 72
Figure 2.4.7 PKR activation in ADAR1 KO NPCs 73
Figure 2.5.1 ADAR1 Vertebrate Multiz Alignment 75
Figure 2.5.2 ADAR1 p150 evolution in vertebrates 76

Figure 2.5.3 ADAR1 locus in chimpanzees 78
Figure 3.1.1 ATG codons between p150 and p110 80
Figure 3.1.2 Distance between first two ATG codons 81
Figure 3.1.3 First two ATG codons in ISGs 83
Figure 3.2.1 Expression of p110 from p150 mRNA 85
Figure 3.2.2 Internal ribosome entry site analysis 86
Figure 3.2.3 Mutations to reduce p110 translation 88
Figure 3.2.4 Analysis of Kozak consensus sequences 89
Figure 3.2.5 Synonymous mutations to reduce p110 93
Figure 3.3.1 Standard curve to quantify ADAR1 levels 95
Figure 3.3.2 Generation of clones for editing analysis ... 97
Figure 3.3.3 Interferon response in 293T cells 99
Figure 4.1.1 Overexpression system in ADAR1 KO cells 102
Figure 4.2.1 Alignment and mismatch calling 104
Figure 4.2.2 Identification of ADAR1-edit sites 106
Figure 4.2.3 Additional filtering criteria 108
Figure 4.2.4 Classification of ADAR1-edit sites 110
Figure 4.2.5 Global trends in editing 112
Figure 4.2.6 Annotation of global ADAR1-edit sites 114
Figure 4.3.1 Annotation of grouped ADAR1-edit sites 116
Figure 4.3.2 Immunoprecipitation of ADAR1p150 118
Figure 4.3.3 Model of ADAR1 p110 and p150 editing 120
Figure 5.1.1 ADAR1 p150r knock-in 124
Figure 5.2.1 Amplicon sequencing with barcodes 126
Figure 5.2.2 Amplicon sequencing gives high coverage 128
Figure 5.2.3 Depth requirements in bulk analysis 130

LIST OF SUPPLEMENTARY MATERIALS

Appendix 1. qPCR primer pair sequences 132
Appendix 2. Guide RNA sequences 133
Appendix 3. NPC differentiation protocol 134

LIST OF ABBREVIATIONS

α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA)
Adenosine deaminase acting on RNA (ADAR)
Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC)
Adenosine-to-inosine (A-to-I)
Clustered regularly interspaced short palindromic repeats (CRISPR)
Cytomegalovirus (CMV)
4',6-diamidino-2-phenylindole (DAPI)
Diethylaminoethyl (DEAE)
Double/single-stranded (ds/ss)
Fluorescence-activated cell sorting (FACS)
Fluorescein isothiocyanate (FITC)
Glutamate ionotropic receptor (GluR)
Green fluorescent protein (GFP)
Human embryonic kidney (HEK)
Kilodalton (kDa)
Knockout (KO)
Lysogeny broth (LB)
Oligonucleotides (oligos)
Open reading frame (ORF)
Polymerase chain reaction (PCR)
Protospacer adjacent motif (PAM)
Rapid amplification of cDNA ends (RACE)
Single nucleotide polymorphism (SNP)
Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE)
Untranslated region (UTR)
Wild type (WT)

CHAPTER 1. Introduction to RNA editing

The first chapter of this thesis aims to: 1. introduce initial discoveries that suggest RNA can be modified following transcription; 2. discuss studies that led to the discovery of a major editor of mRNA, adenosine deaminase acting on RNA (ADAR); and 3. discuss how the ADAR field has evolved and present what is known about regulation, which relates directly to experiments that will be presented in the following chapter.

1.1 RNA can be modified after transcription

Biological information, as it gets transferred from DNA to RNA, can be diversified during and after transcription. Diversification here refers to changing the molecular structure of individual nucleosides, or adding and removing nucleosides, and is not limited to the four canonical bases, as numerous metabolic derivatives of nucleobases are found in cells, such as hypoxanthine, which more commonly in biological contexts is called by its name when it attaches to a ribose ring: inosine.

As transcription proceeds, the number of unique sequences beyond what would be specified directly by the DNA locus already begins to increase due to the error rate of RNA polymerase II: 10-100 errors each second in human cells (1). As the mRNA polymer grows in the nucleus, one can imagine that nucleoside-modifying enzymes present in the nucleus can catalyze reactions that further expand the number of unique sequences. After transcription and eukaryotic mRNA processing events such as polyadenylation and capping, the same possibilities for molecular modifications to nucleosides in the nucleus apply when mRNAs are exported to the cytoplasm.

Although modified nucleobases, such as 5-methylcytosine, have been known to exist on RNA in cells for more than one hundred years, use of the phrase “RNA editing” has been rather belated (2, 3). In 1986, nucleoside addition was discovered in coxII mRNA from two species of protists, although the mechanism of addition was not described in detail until about a decade later (4, 5). The insertion of uridines, which involves the breaking and reforming of phosphodiester bonds, is needed to ensure translation of a functional cytochrome c oxidase 2 enzyme (6). This editing of coxII mRNA occurs in the mitochondria of the organisms, where four additional uridines were discovered in the coxII mRNA that were not present in the cox2 coding sequence in the mitochondrial DNA (Figure 1.1.1).

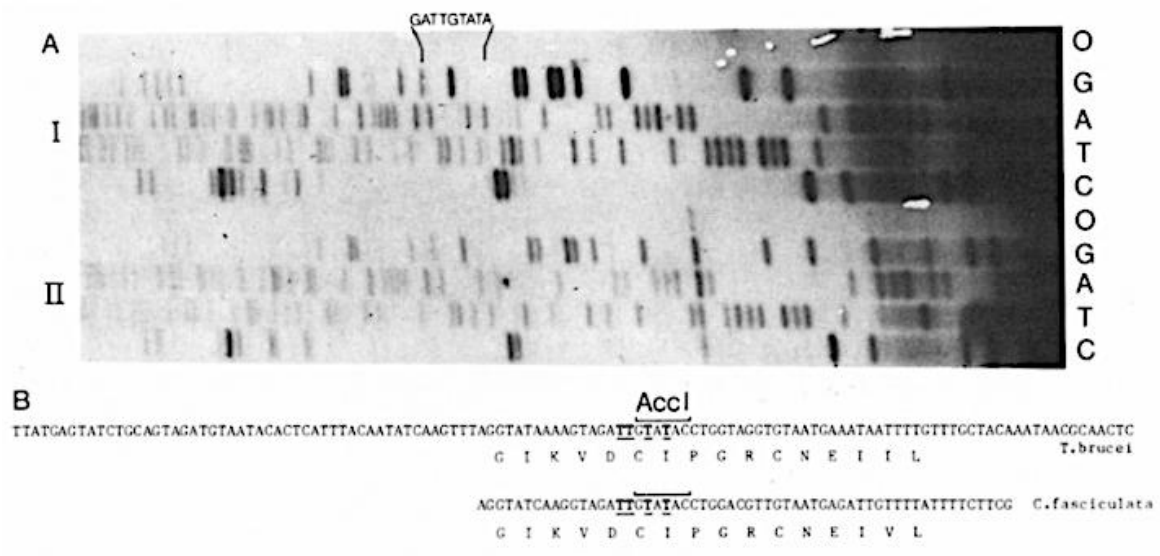


Figure 1.1.1 RNA editing of coxII in protists (4)
A. Trypanosoma brucei (I) and Crithidia fasciculata (II) total RNA was used as the template for reverse-transcription using a coxII-specific primer. The cDNA was sequenced using chain-terminating dideoxynucleotides followed by electrophoresis on a polyacrylamide gel. Electrophoresis is from left to right in the gel shown, and the additional thymidines seen in the cDNA are highlighted.
B. The addition of thymidines is shown in context with parts of the coding and amino acid sequences of the enzyme for both protist species. The addition restores the reading frame for translation of a functional protein. AccI indicates creation of a new restriction digest site by the nucleotide additions.

One year later, the field of RNA editing expanded to include mammals, with the discovery of cytidine-to-uridine (C-to-U) editing in the ApoB mRNA from human and rabbit small intestine (7–9). The editing creates an early stop codon that results in translation of a smaller apo-B48 isoform from the edited mRNAs molecules (Figure 1.1.2). The enzyme that catalyzes this editing reaction is part of a family of cytidine deaminases called APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like). Editing of ApoB mRNA occurs along with polyadenylation and splicing in the nucleus (10).

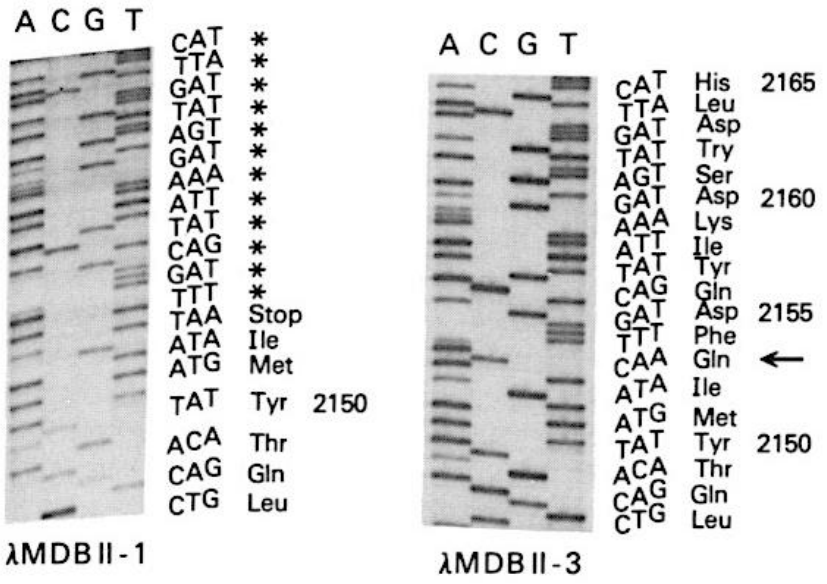


Figure 1.1.2 RNA editing of ApoB in humans (8)
Total mRNA from human intestinal tissue was extracted, reverse-transcribed into double-stranded cDNA, digested with restriction enzymes, and cloned into lambda phage vectors to generate a human intestinal cDNA library as a mixture of phage clones. Following infection of E. coli with the mix of phage clones, formation of individual plaques on the plate of E. coli, and transfer of DNA to nitrocellulose, screening of the clones was done by hybridization of two oligos corresponding to part of the ApoB cDNA, with one oligo containing the CAA-Gln and the other containing the edited TAA-Stop. Following hybridization, both oligos showed positive signals on the membrane, and DNA from those regions was extracted and sequenced using chain-terminating dideoxynucleotides followed by electrophoresis on a polyacrylamide gel. Shown left is an example of part of the sequence from a phage clone that hybridized with the TAA-Stop oligo, and right an example of part of the sequence from a phage clone that hybridized with the CAA-Gln oligo.

1.2 Discovery of adenosine deaminase acting on RNA

In the same year that C-to-U editing was discovered, another report was published that would set the stage for another type of RNA editing to be revealed. This would grow out of the field of developmental biology, and specifically the use of antisense RNA to inhibit mRNA expression in frog eggs (11, 12). Initially, antisense DNA oligos were used, but the abundance of single-stranded DNA endonucleases in frog eggs meant that antisense DNA oligos injected into eggs were quickly degraded and thus largely ineffective in getting around to forming the DNA/mRNA hybrids needed for inhibition of mRNA expression through the RNase H degradation pathway (13, 14). Use

of antisense RNA was therefore proposed for studying gene functions in frog embryo development, with the idea that having a long antisense RNA molecule anneal and hang off 5′ ends of target mRNAs would inhibit gene expression by blocking 40S ribosomal subunit scanning and assembly of translation initiation complexes. However, injection of antisense RNA did not produce the intended outcome, but rather, led to a surprising finding: disruption of the antisense/sense RNA duplex structure (Figure 1.2.1).

After attempting to reanneal the sense and antisense RNA molecules following incubation, a partially reannealed RNA was formed that was determined to be unprotected from degradation by RNase A, which has specificity to single-stranded RNA (Figure 1.2.2). Additionally, the partially reannealed RNA was found to have different migration properties on a native gel after the incubation with embryo extracts, but it retained the same migration properties on a denaturing gel, suggesting the mass is largely unchanged but the structure has been altered (Figure 1.2.3).

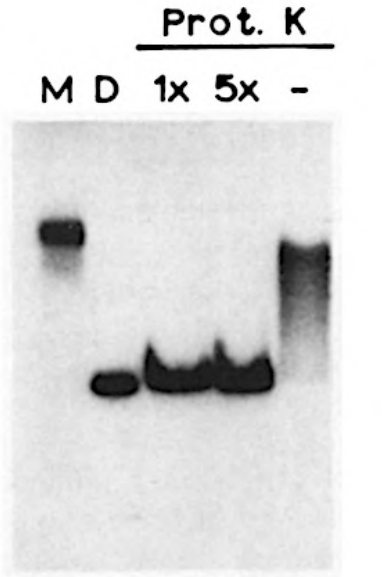


Figure 1.2.1 RNA duplex unwinding in frog eggs (15)
In this experiment, phosphorus-32 (ATP, CTP, GTP, UTP) was used to label sense RNA, and hydrogen-3 (UTP) was used to label antisense RNA during in vitro transcription. The annealed duplex RNA was incubated with Xenopus laevis oocyte extracts (with or without proteinase K pretreatment) for 2 hours at 25°C followed by electrophoresis on a native polyacrylamide gel. The "M" and "D" lanes show where the monomer and duplex RNA, respectively, would migrate prior to incubation with egg extracts. Without the proteinase K pretreatment, the extracts were able to disrupt the RNA duplex structure, as shown in the "-" lane. Pretreatment of extracts with different concentrations of proteinase K removed their ability to disrupt the RNA duplex structure, as shown in the "1x" and "5x" lanes, suggesting the agent responsible is a protein found in the extracts.

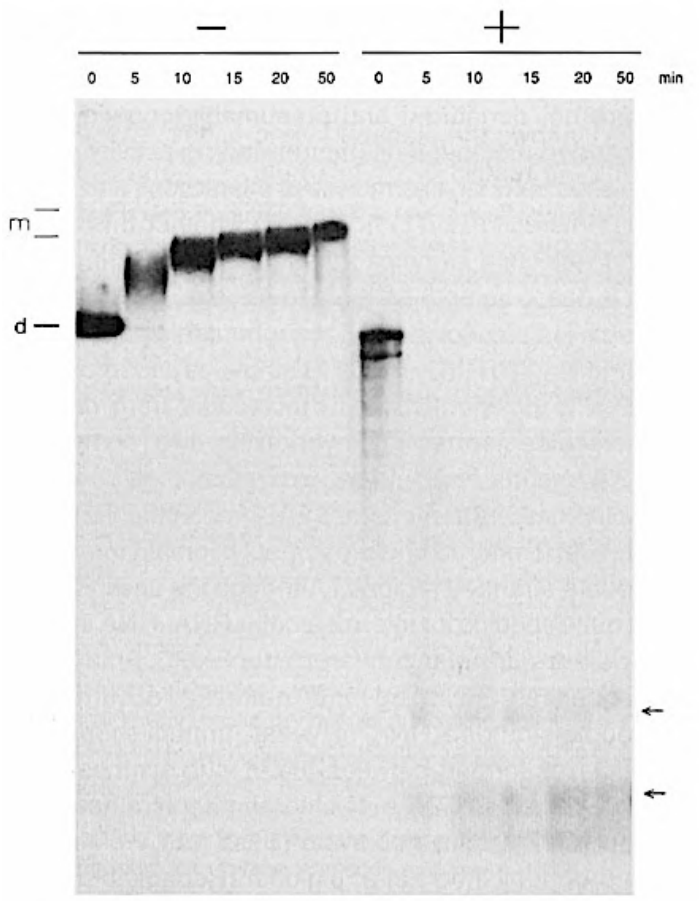


Figure 1.2.2 RNA duplex is degraded by RNase A (16)
This film image shows radioactivity signal produced by RNA labeled with phosphorus-32 (ATP, CTP, GTP, UTP) and hydrogen-3 (UTP) on its sense and antisense strands, respectively. The duplex RNA was incubated with Xenopus embryo nuclear extracts for the indicated times in minutes. The RNA was then extracted, depleted of proteins, and incubated at 25°C for 15 minutes in a salt solution with or without RNase A. Afterward, electrophoresis and transfer to a membrane revealed, for the RNase A (-) condition, radioactivity signals that start out where the duplex (d) is expected to migrate but then move upward toward where the monomer (m) is expected to migrate as the incubation times with embryo extract increased. For the RNase A (+) condition, radioactivity signals also started out where the duplex (d) is expected to migrate, but then disappear in the monomer-duplex gel mobility area. Faint signals are detected further down in the film, indicated by the arrows.

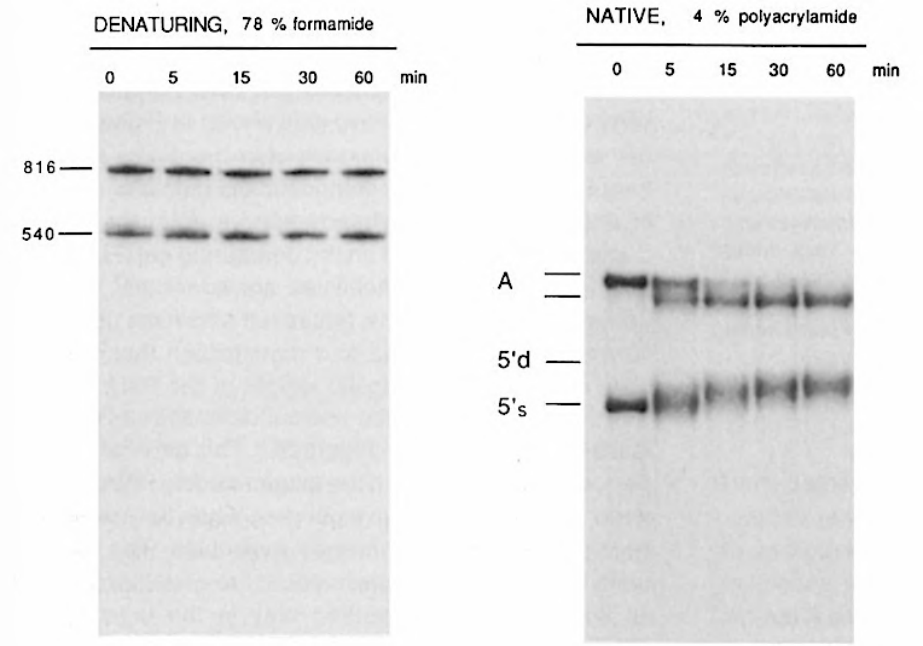


Figure 1.2.3 RNA duplex is structurally altered (16)
An 816-base antisense RNA (A) was annealed to a 540-base sense RNA (5's), and both strands consisted of nucleotides labeled with phosphorus-32 (ATP, CTP, GTP, UTP) incorporated during in vitro transcription. The annealed duplex (5'd) was incubated with embryo nuclear extracts for the indicated times in minutes. The film on the left shows radioactivity signal produced by the incubated duplex run out on a formamide-polyacrylamide denaturing gel after 2 minutes of 95°C heating. The numbers "816" and "540" refer to markers indicating where the antisense and sense RNA molecules, respectively, are expected to migrate. Incubation with the nuclear extracts over time did not change the pattern of migration for either molecule on a denaturing gel. The film on the right shows the same experiment except the heated RNA duplex was run on a native polyacrylamide gel. In contrast with the denaturing gel, the native gel reveals changes in the patterns of migration for both RNA molecules.

The reason for the structural change was revealed to be due to differences in the base-pairing properties of the RNA duplex before and after incubation with the extracts. These differences were determined to be a direct result of covalent modifications to adenosines that converted some of them into inosines. Some clues that pointed the scientists toward inosine as the modified base include the fact that a single and simple reaction, deamination, could convert adenosine into inosine, and also the fact that the modified adenosines migrated in a manner similar to that of guanosine, which is structurally similar to inosine; guanosine has an amine side chain on the 2-carbon that inosine lacks, but they are otherwise identical (Figure 1.2.4).

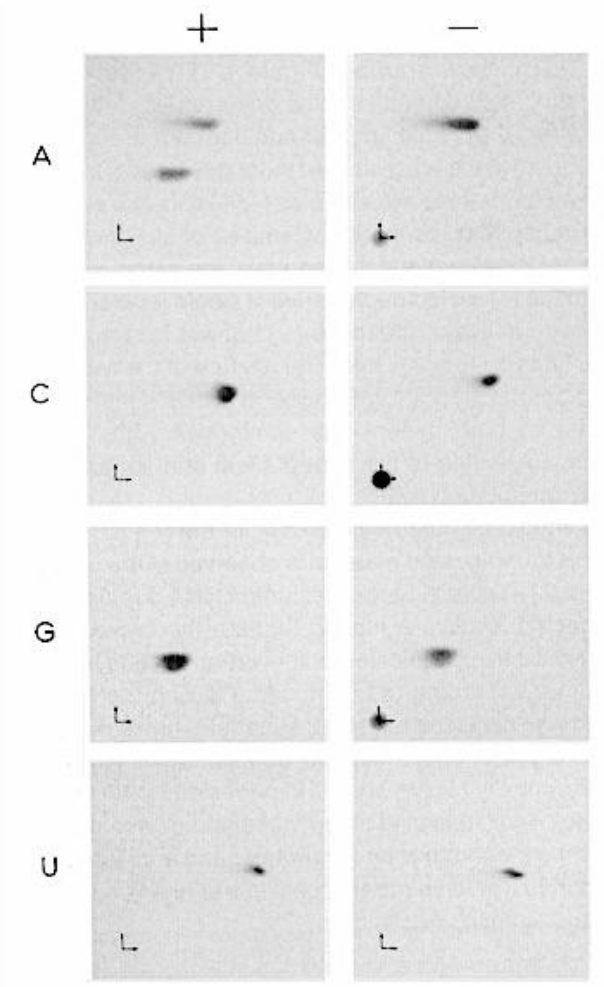


Figure 1.2.4 The modified adenosine migrates with guanosine (16) *Four RNA duplexes were labeled by in vitro transcription with radioactive (phosphorus-32) ATP, CTP, GTP, or UTP, and incubated for 1 hour with (+) or without (-) embryo nuclear extracts. The incubated RNA duplexes were then purified, depleted of proteins, and digested into mononucleosides using nuclease P1. 2D thin-layer chromatography was then used to analyze migration patterns of the digested mononucleosides. The labeled ATP was the only nucleoside, as expected, that had two discrete radioactive signals; the unmodified adenosine moved a bit farther to the right horizontally compared to the modified adenosine, which migrated to essentially the same location as the labeled GTP.*

Frog extracts were not the only candidates tested. Thin-layer chromatography performed on mononucleosides from digestion of RNA duplexes incubated with human lymphoma cell extracts revealed that radiolabeled adenosines appeared in two distinct regions, one at the expected region for adenosine and another near where guanosine would appear, consistent with what had been observed with prior Xenopus extract experiments. Fractionation

of labeled adenosines with unlabeled mononucleosides as standards, including unlabeled inosine, revealed that the radioactive eluates contained adenosine, as expected, and also inosine (Figure 1.2.5).

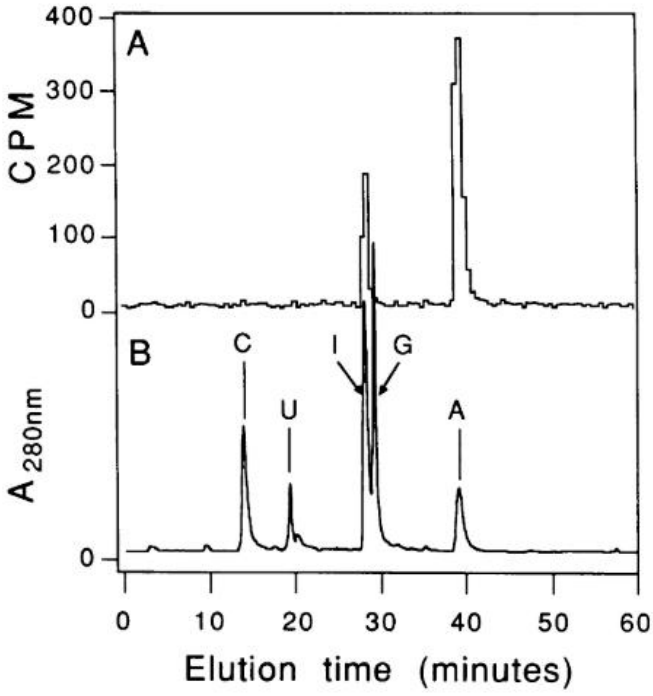


Figure 1.2.5 The modified adenosine elutes with inosine (17) *In this experiment, human beta-globin sense and antisense sequences were labeled with carbon-14 ATP during in vitro transcription and annealed to create a duplex. Following incubation with GM1500 (a human lymphoma cell line) extracts, the duplex was digested with nuclease P1 and analyzed on a reverse-phase liquid chromatography column along with unlabeled inosine as a standard. The elution fractions for the different nucleosides are shown by elution times on the horizontal axis, with readouts of both radioactivity (CPM) and UV-280nm absorbance throughout the fractionation. The peaks in radioactivity correspond to the inosine and adenosine absorbance peaks.*

Next, efforts were made to look into the mechanism of adenosine-to-inosine (A-to-I) modifications. By using fully carbon-13 labeled adenosines and oxygen-18 labeled water, the resulting modified adenosines were found to contain only carbon-13 and oxygen-18, suggesting there was no loss and gain of carbons during the reaction and that addition of water (hydrolytic deamination) is the mechanism of modification (18).

Naked adenosines on single-stranded RNA molecules do not appear to be efficient substrates for deamination, as having double-stranded structures, whether within a single molecule of

RNA or between two molecules, is essential for the reaction to occur (Figure 1.2.6).

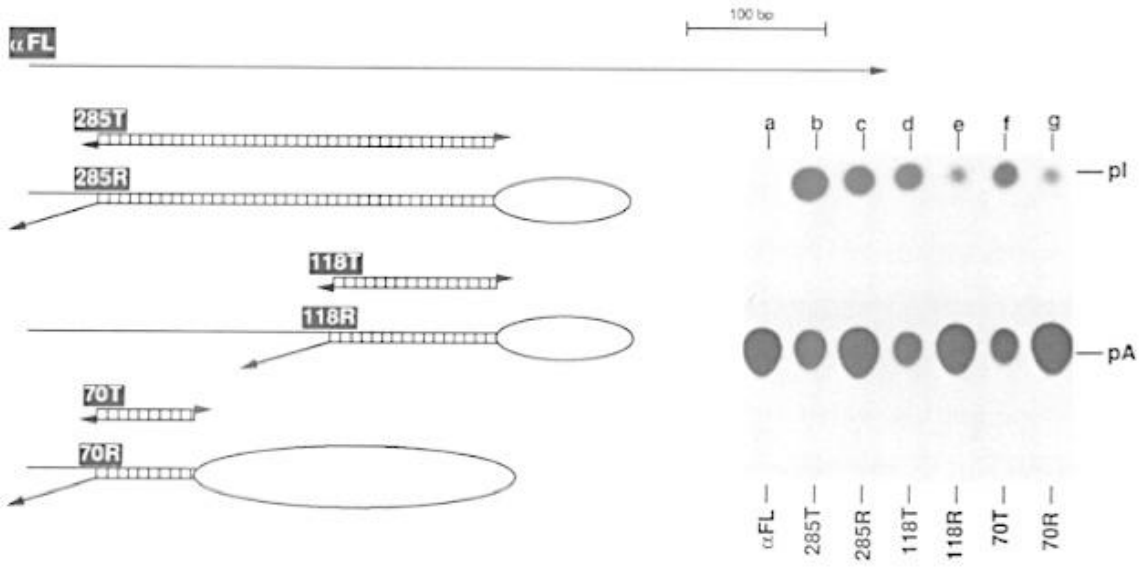


Figure 1.2.6 Double-stranded RNA substrates are edited (19) *Variations of alpha2-globin mRNA sequence were synthesized using SP6 in vitro transcription from cDNA cloned into expression plasmids. Alpha-phosphorus-32 ATP was used during the synthesis process for labeling. Several substrates for editing analysis were created from the single-stranded transcribed molecules: αFL, a linear RNA molecule; 285T and 285R, intermolecular and intramolecular substrates, respectively, created from 285-base sense and antisense single-stranded RNA molecules corresponding to part of the alpha2-globin mRNA; 118T/118R and 70T/70R, the same as 285T and 285R, except the sense and antisense RNA are shorter. The seven different substrates were incubated with HeLa cell extracts, then purified, depleted of proteins, and digested into mononucleosides using nuclease P1. The products were fractionated using 1D thin-layer chromatography, shown in lanes a-g. Radioactivity was then quantified. The non-duplex substrate in lane-a did not show signal at the inosine location. Inosine/adenosine signal intensity ratios for the intermolecular substrates, 285T, 118T and 70T, were about 40%, 40%, and 30%, respectively. Ratios for the intramolecular substrates, 285R, 118R and 70R, were about 17%, 17%, and 15%, respectively.*

The question that came next was to find the specific proteinase K sensitive agent or agent(s) responsible for this unwinding and deamination modification. Outside of test-tube conditions, in live cells, the deamination reaction was initially described to occur in both frog egg nuclear extracts and cytoplasm following breakdown of the nuclear membrane (15,

16). Thus, whole frog eggs were deemed to be excellent starting materials for purification of the protein in question. Ion-exchange chromatography was used to fractionate whole frog egg extracts, and the fractions were tested for deamination activity (Figure 1.2.7).

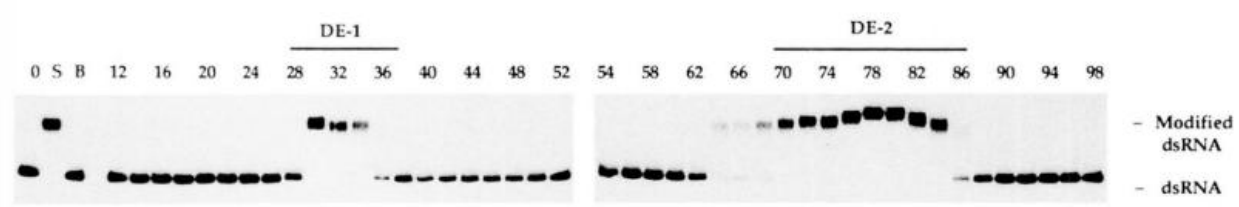


Figure 1.2.7 Fractionation of frog egg extracts (20)
Whole frog eggs extracts were fractionated using diethylaminoethyl (DEAE) groups exchange resin in a column. Fraction numbers 12-98 were collected and mixed with a 794-base duplex RNA as the substrate. The RNA was internally labeled with alpha-phosphorous-32 ATG. After incubation, the RNA was purified and resolved on a native polyacrylamide gel, taking advantage of the fact that the altered duplex RNA migrated more slowly than the unmodified RNA in proportion to the number of inosine-uridine mismatches. Two sets of fractions, labeled DE-1 and DE-2, corresponding to fractions 28-36 and 70-86, were found to slow down the duplex RNA on the polyacrylamide gel. The "0" column includes only the unmodified RNA, and the "B" column is the RNA with the DEAE exchange groups. The "S" column is the duplex RNA incubated with unfractionated frog egg extracts.

The fractions found to have deamination activity were purified using affinity chromatography, in which the ligand used was duplex poly-guanosine/poly-cytidine, which would be bound by the deaminase protein but presumably not be edited or destabilized by the protein during the purification process. The bound portions of the deaminase-active fractions from the ion-exchange chromatography were then analyzed by electrophoresis, revealing a band around the 120-kilodalton area (Figure 1.2.8).

Around the same time the protein was purified from whole frog eggs, homologs of this dsRNA adenosine deaminase protein was also purified for the first in mammals, from bovine liver and calf thymus, and revealed to have similar molecular weights of about 100 and 116 kilodaltons, respectively (21, 22).

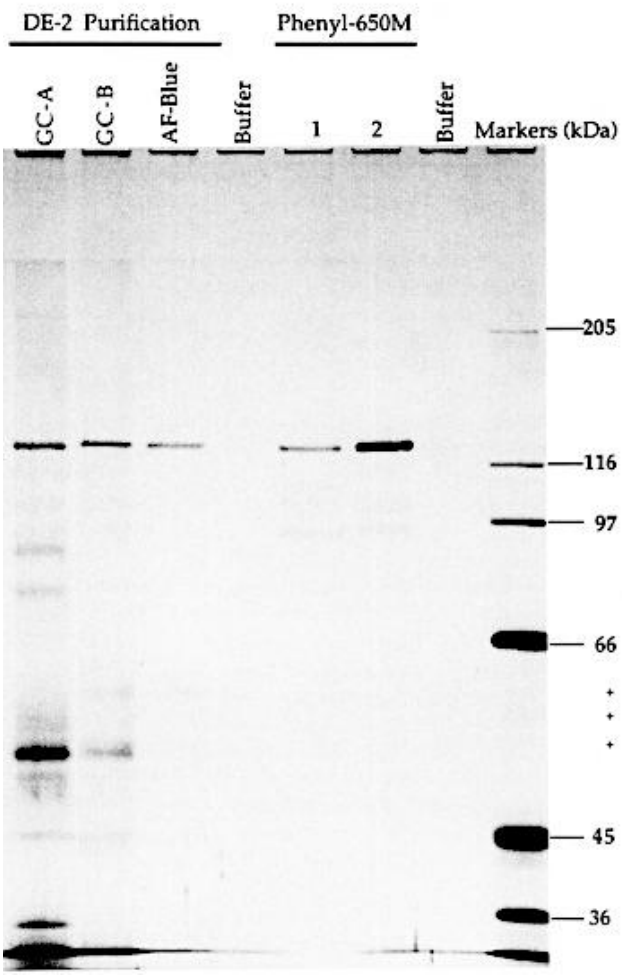


Figure 1.2.8 Electrophoresis of deaminase-active fractions (20) *The second deaminase-active set of whole frog egg fractions (fractions 70-86) was split into halves (GC-A and GC-B) and affinity purified using separation-pharmacia-agarose (Sepharose) attached to poly-guanosine and poly-cytidine RNA. The material in the fractions that was bound to the poly(G:C)-Sepharose was then analyzed using sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) and stained with silver for visualization. Further purification was done by utilizing different ligands in the column resins that have high affinity for proteins: a triazine dye (AF-Blue), and hydrophobic functional groups (Phenyl-650M). A clear band that corresponds to a molecular weight of about 120 kilodaltons appeared with the silver stain.*

Finally, editing assays, again with radioactively labeled duplex RNA as the substrate, confirmed the A-to-I editing activity of the proteins purified from Xenopus and calf thymus. With the dsRNA adenosine deaminase now purified, the task remained to sequence the protein, reverse infer possible mRNA

sequences, and test the inferences by screening cDNA libraries (23–25).

Of note, there was clear motivation to clone the gene encoding this enzyme, particularly after a study reported a significant functional consequence of A-to-I editing: the now well-known editing activity observed in mRNA encoding AMPA receptor subunit GluR-B. The editing occurs within a putative intron-exon duplex structure and creates a nonsynonymous substitution leading to a arginine being incorporated in the polypeptide rather than a glutamine, with resulting changes in ion flow through the ionotropic AMPA receptor (26, 27).

Cloning of the dsRNA-specific deaminase was achieved by screening human cDNA libraries with sets of oligos that correspond to the codons of part of the sequenced deaminase protein, taking into account the fact that different codons can specify the same amino acid, especially due to diversity at the third (wobble) position of codons (Figure 1.2.9).

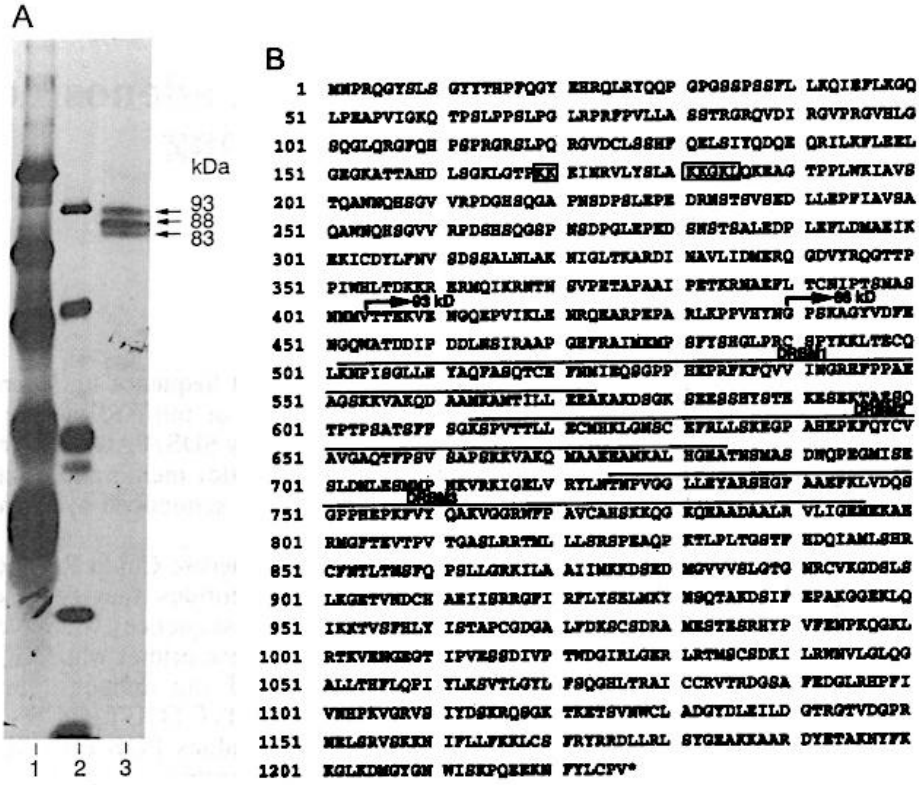


Figure 1.2.9 Cloning of ADAR cDNA (23)
A. Bovine liver extracts were fractionated to search for portions with deaminase activity. The fractions that showed activity were purified using affinity chromatography and then analyzed by SDS-PAGE, revealing several polypeptides of 93, 88, and 83 kDa, show in lane 3. Lanes 1 and 2 show the molecular mass standards. The 93 and 88 kDa bands were cut and sequenced from their N-terminal ends by Edman degradation. The mRNA sequence possibilities were inferred from the amino acid

sequences toward the N-terminal end of the protein, taking into account the degeneracy of the genetic code. For both the 93 and 88 kDa polypeptides, three sets of oligos were designed: antisense primers for first-strand cDNA synthesis, sense primers for first-round PCR, and internal probes corresponding to the deduced mRNA sequences between the sense and antisense primer binding sites. The first-strand cDNA synthesis and subsequent first-round PCR were carried out using total RNA from bovine endothelial cells as the starting material. Analysis of the first-round PCR products by Southern blot revealed PCR products that hybridized to the internal probes, and these products were isolated and inserted into plasmids for storage using the restriction sites that were part of the 5′ ends of the sense and antisense primers. These inserts were then used to screen a human natural killer cell cDNA library, which revealed positive cDNA clones. The inserts of the positive cDNA clones were then used as probes themselves for subsequent rounds of screening to identify the other inserts that allowed piecing together the full-length sequence of the cDNA.
B. The open-reading frame of the human cDNA was determined to encode 1226-amino acids, shown in the schematic with the N-terminal ends of the bovine 93 and 88 kDa polypeptides labeled with arrows. The full protein includes three regions that share homology with dsRNA binding motifs (DRBM). Finally, the nuclear localization signals are shown in boxes.

The sequence of the full-length open-reading frame for human ADAR was deduced from the overlapping inserts of various cDNA clones that showed up as positive during several rounds of screening. Consistent with findings that ADAR binds to dsRNA and that A-to-I activity is prominently observed in nuclear extracts, the protein also has three dsRNA binding motifs and a nuclear localization signal (24, 28).

With the protein sequenced and the mRNA sequence assembled, scientists were interested to know more about the DNA locus that contains the information for ADAR expression. Of interest, the human genome project had not started until September of 1999, so it is perhaps no surprise that the ADAR field progressed in a direction opposite to that of the central dogma of information transfer in biology (29). It was revealed by fluorescence in situ hybridization that the ADAR sequence is located on chromosome 1 (30, 31). Inserts from selected lambda phage vectors corresponding to fragments of human placental DNA were used in this in situ hybridization experiment (Figure 1.2.10).

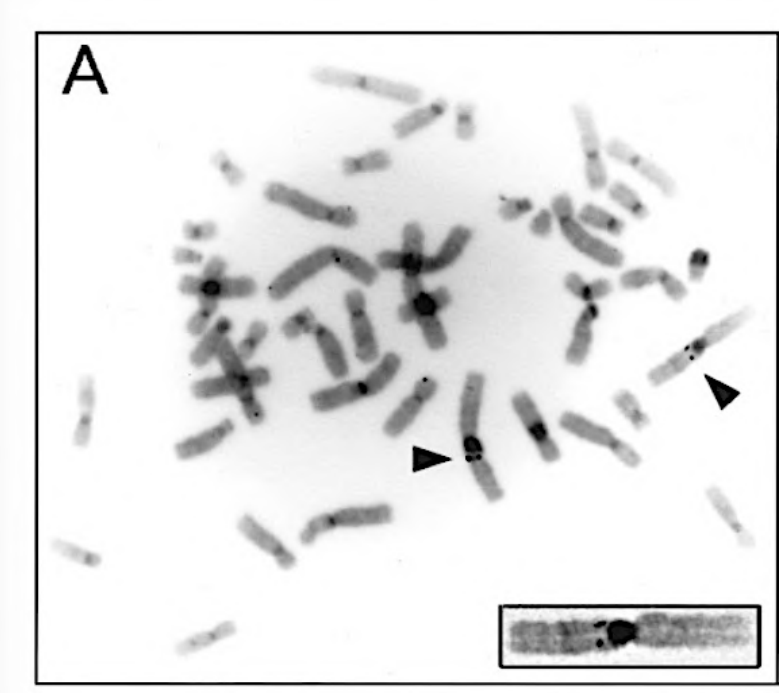


Figure 1.2.10 Mapping of ADAR DNA sequence (30)
A. Metaphase chromosomes from human lymphocytes are shown here fixed and hybridized with digoxigenin-labeled probes specific for the ADAR genomic sequence. The probes are inserts from lambda phage clones prepared from human placental DNA that tested positive in a screen with ADAR cDNA probes, which had been used previously in a screen for interferon-stimulated genes, including the ADAR gene. Here, DAPI was added for visualization of chromosomes, and anti-digoxigenin antibodies labeled with the rhodamine dye were used for visualization of the digoxigenin-labeled probes. The smaller box shows a homologous metaphase chromosome 1 zoomed in to show more details, and the darker signals near the centromere correspond to the location of the ADAR1 locus, at the long arm, region 2, bands 1.1-1.2 (q21.1–21.2).

Around the same time the ADAR genomic organization was being mapped, two protein isoforms of ADAR were also reported, one inducible by several interferon isoforms, and the other constitutively expressed. The inducible isoform is about 150 kDa and can be found in both the nucleus and cytoplasm, whereas the other isoform is smaller, about 110 kDa, and is found in the nucleus (32). Antibodies specific to different portions of the full-length ADAR were used to probe nuclear and cytoplasmic extracts prepared from human neuroblastoma cells (Figure 1.2.11). Furthermore, antibodies specific to the larger isoform and selective for the smaller isoform in cells untreated with interferon were used in immunofluorescence microscopy to visualize the two isoforms in situ (Figure 1.2.12).

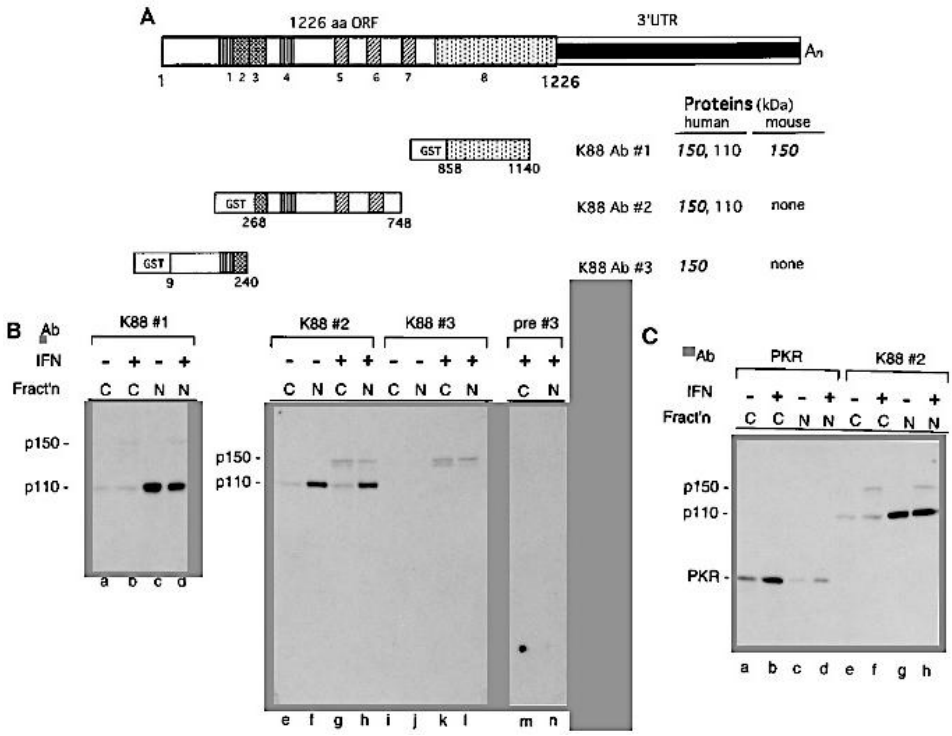


Figure 1.2.11 Two ADAR protein isoforms (32)
A. A schematic is shown of the full-length ADAR open-reading frame of 1,226 amino acids, corresponding to the assembled sequence of a set of cDNA clones, called K88 (the set is made up of several clones obtained through a few rounds of screening), previously isolated in a screen for interferon-regulated genes. Numbers 1-4 represent tandem (adjacent) repeat sequences; numbers 1 and 4 specifically share homology with the vaccinia virus E3L protein, which has the ability to bind Z-DNA. Numbers 5-7 correspond to the dsRNA binding domains, and number 8 is the deaminase catalytic domain. Three immunogens were prepared using a plasmid expression system in E. coli; the expression plasmids encode glutathione S-transferase (GST) fusion proteins, K88#1, K88#2, and K88#3, corresponding to different regions of the ADAR protein: residues 858-1140, 268-748, and 9-240, respectively. Rabbits were immunized with these polypeptides, and antibodies specific to the different regions of the ADAR protein were used to probe nuclear (N) and cytoplasmic (C) extracts from human SH-SY5Y neuroblastoma cells treated or untreated with Sendai virus-induced leukocyte alpha-interferon.
B. Antibody-antigen signal was detected on film by radioactivity emitted from iodine-125 attached to protein A, which binds antibodies. The C-terminal antibody, K88#1, detected antigen of sizes 150 kDa and 110 kDa in both the cytoplasmic and nuclear fractions, but only in the presence of interferon; without interferon, only the 110 kDa protein was detected, and in much higher amounts in the nucleus compared to cytoplasm. The K88#2

antibody, which recognizes the middle regions of ADAR, produced the same pattern as the C-terminal antibody. Finally, the N-terminal antibody, K88#3, produced bands that correspond only to the 150 kDa isoform, present in both nucleus and cytoplasm. The pre-immunized serum (pre #3) did not produce signal.
C. Although both PKR and ADAR have conserved arginine motifs in the dsRNA binding domains, no cross-reactivity was detected in the antibodies used to detect RNA-dependent protein kinase (PKR) and the middle portion of ADAR.

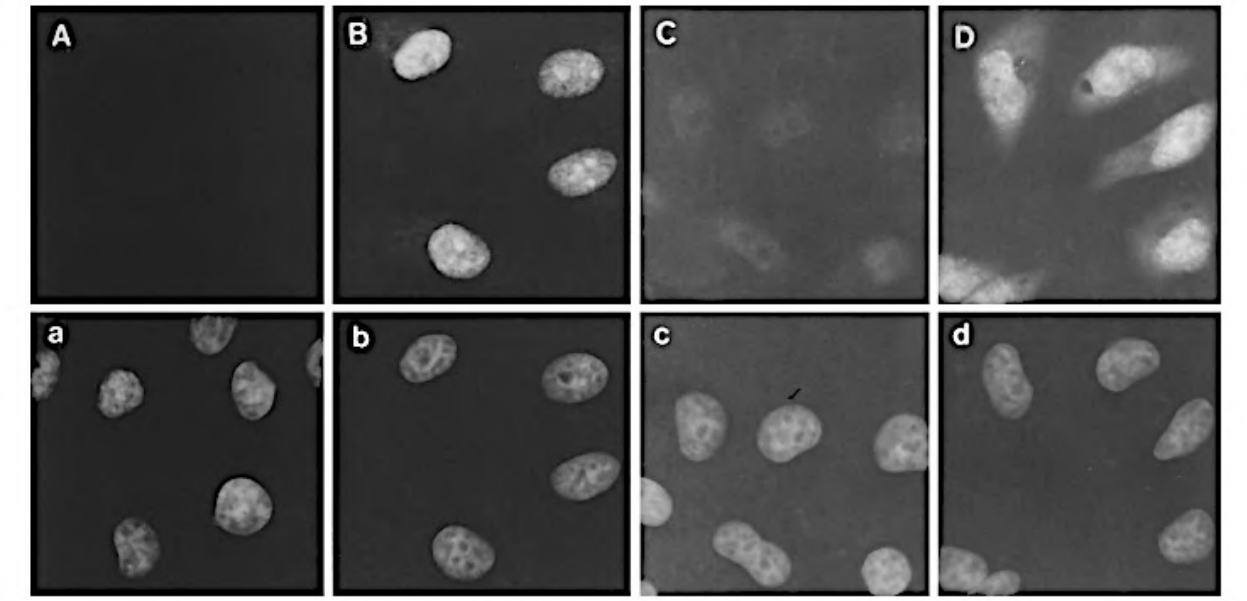


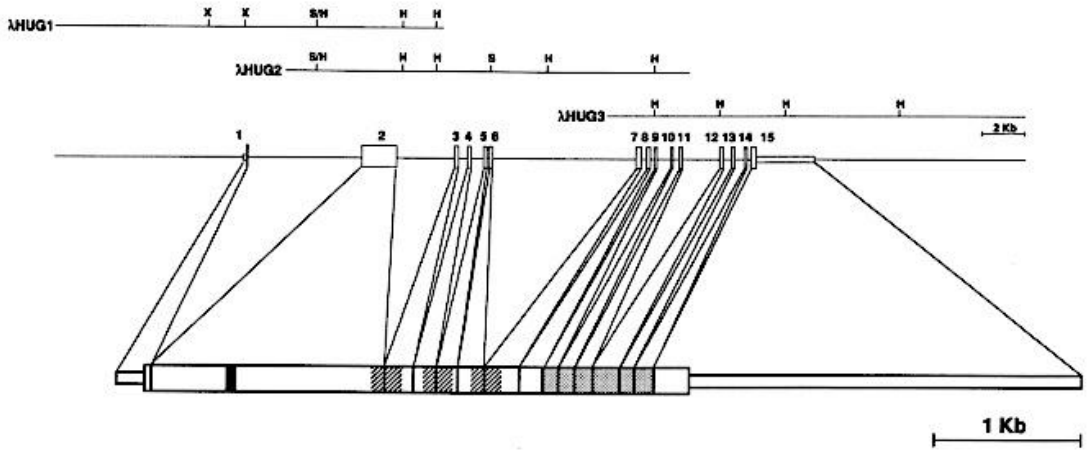
Figure 1.2.12 Two ADAR protein isoforms in situ (32)
Human amnion U cells were examined by immunocytochemistry, with antibodies that bind to epitopes in the middle (K88#2) and N-terminal ends (K88#3) of full-length ADAR. Secondary antibodies conjugated to FITC were used to amplify signal from the primary rabbit antibodies. Panel A shows cells treated with interferon and probed with pre-immune serum. Panel B shows untreated cells probed with K88#2, which reacts with both isoforms but is selective for the smaller isoform (p110) in the absence of interferon. The selectivity is largely due to low expression levels of the larger isoform (p150) without interferon treatment. Panel C shows untreated cells probed with K88#3, and panel D shows interferon-treated cells also probed with K88#3. Lowercase a-d shows the same microscope fields as the top panels, but stained with DAPI.

1.3 Evolution of the ADAR field

After the discovery and cloning of ADAR, the field has evolved in three main directions: 1. regulation of ADAR expression; 2. ADAR editome analysis; and 3. role of ADAR proteins in regulating immunity, among other aspects of biology.

The third category, specifically ADAR1 regulation of innate immunity, will be touched on briefly in the second half of chapter two. The second category, and specifically ADAR1 isoform-specific editome analysis, follows directly from work that will be presented in chapters two and three related to understanding isoform-specific ADAR1 expression, and will be presented in chapter four. Studies related to the ADAR1 editome and the role of ADAR1 in immunity have both benefited immensely from advances in sequencing technologies, stem cell techniques, CRISPR, and other methods. There has been less extensive work, taking advantage of advances in lab techniques, done on understanding regulation of isoform-specific ADAR1 expression in human cells.

The early studies on ADAR regulation highlight the complexity of this system and will be reviewed briefly. The ADAR genomic locus was originally reported to have 15 exons, with the first exon being the one that includes the methionine codon for the p150 full-length open-reading frame (Figure 1.1.13). The published start of the exon 1 sequence is actually that of exon 1A, one of three exons that have promoters upstream and that all attach to exon 2 during splicing of the pre-mRNAs.



(a)

EXONS		INTRONS	
GGGCCCGGCG...EXON1	(179 bp)...TCCGCGGCAG/	gtaagccggg...INTRON1	(5.4 Kb)...tattctgcag/
GGGTATTCCC...EXON2	(1586 bp)...ATGAACCTCG/	gtaagagacc...INTRON2	(2.5 Kb)...ttccgtcaag/
ATTTAAATTC...EXON3	(184 bp)...ATCAGAGAAG/	gtaggtgtcc...INTRON3	(0.4 Kb)...ttttctctag/
ACTGCAGAGT...EXON4	(149 bp)...ATGAACCCAA/	gtatgtccta...INTRON4	(571 bp)...ctcctgtcag/
GTTCCAATAC...EXON5	(145 bp)...TGATAACCAG/	gtagggcgtt...INTRON5	(127 bp)...tctcctttag/
CCTGAAGGTA...EXON6	(191 bp)...ACGAGCCCAA/	gtgagtgtcc...INTRON6	(6.5 Kb)...catcccaaag/
GTTCGTTTAC...EXON7	(226 bp)...GCCAAAGACA/	gttaagacgt...INTRON7	(255 bp)...ttccccacag/
CTCCCTCTCA...EXON8	(172 bp)...TTGGGAACAG/	gtgagtgagg...INTRON8	(235 bp)...acctccctag/
GGAATCGCTG...EXON9	(94 bp)...GCTTCATCAG/	gtgagcgagg...INTRON9	(0.6 Kb)...ctttttgtag/
GTTTCTCTAC...EXON10	(123 bp)...TGTATATCAG/	gtctgtacag...INTRON10	(292 bp)...tgtttttcag/
CACTGCTCCG...EXON11	(134 bp)...GTGGAGAACG/	gtgagtgata...INTRON11	(1.6 Kb)...tctcacacag/
GAGAAGGCAC...EXON12	(183 bp)...GTCACATTGG/	gtaaggggcc...INTRON12	(315 bp)...ttgtactcag/
GTTACCTTTT...EXON13	(113 bp)...CCACCCCAAG/	gtgctataac...INTRON13	(408 bp)...ggattcctag/
GTTGGCAGAG...EXON14	(128 bp)...CTGTGGATGG/	gtaaggaaac...INTRON14	(173 bp)...gtttctctag/
GCCACGGAAT...EXON15	(2975 bp)...CAAGAATCTG/		

(b)

Figure 1.2.13 ADAR genomic organization (31)
A. Clones from a lambda phage genomic DNA library made from human lymphoid cells were screened using an insert from a plasmid that corresponds to part of the human ADAR cDNA sequence. The three overlapping clones that were selected from the screen, called lambda HUG1, lambda HUG2, and lambda HUG3, together encompass the 15 exons (boxes) and 14 introns (lines) that make up the human ADAR gene locus on chromosome 1. The letters shown on the schematics of the lambda clones refer to restriction sites: HindIII (H), XhoI (X), and SalI (S). Below the schematics of the lambda phage clones and the ADAR DNA locus is a schematic of the full-length open-reading frame, with smaller boxes denoting untranslated regions and larger boxes denoting coding regions.
B. The starting and ending sequences of the 15 exons and 14 introns are shown. The exon-intron junctions are denoted by slashes, and splice donor (gt) and acceptor (ag) sites in the introns are underlined.

With the exons and introns mapped and sequenced, scientists now turned to understanding how expression from the promoter(s) is regulated. First, Northern blots done with probes specific to ADAR mRNA suggested the size of the mRNA could be larger than what was predicted from assembling sequences directly from overlapping cDNA clones isolated from rounds of library screenings (23, 31). To examine the ends of ADAR mRNAs, 5′ rapid amplification of cDNA ends (RACE) was performed on total RNA using antisense ADAR specific primers for first-strand cDNA synthesis (33). These RACE studies revealed that transcription could potentially start at multiple sites upstream of the translational start site for p150 (Figure 1.2.14). One study done in HeLa and Raji cells (a cell of hematopoietic origin isolated from a patient with Burkitt's lymphoma) placed the transcriptional start sites at between 120 and 160 bases upstream of the translational start site; another study done using cells isolated from human placental tissue placed the transcriptional start site at a maximum distance of 186 bases upstream of the translational start site (31, 34, 35). Taking into account biological variability, there is perhaps no need to assume that transcription starts at the exact same nucleotide position each time. Maybe this variability is part of the reason for the evolution of 5′ untranslated regions (UTRs) in translatable mRNAs, that is, to ensure that the first methionine codon is not missed or broken apart because of variation in where transcription starts. Of course, structures of 5′ UTRs are also intimately associated with 40S ribosomal subunit scanning and regulation of translation initiation.

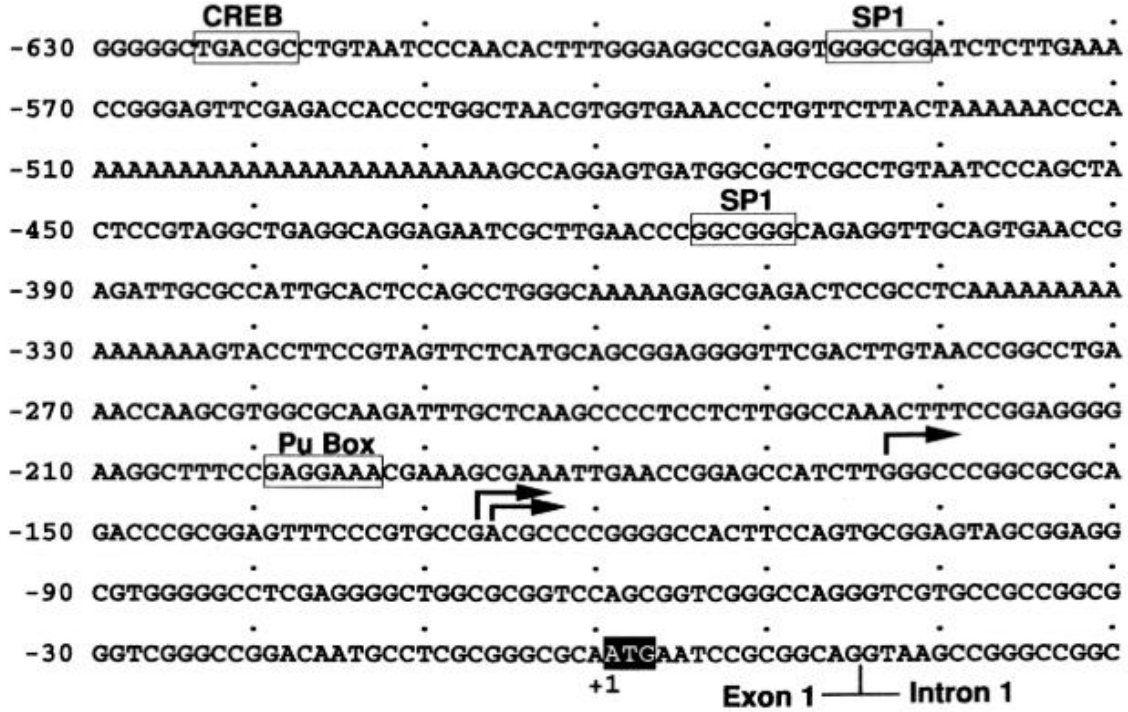


Figure 1.2.14 ADAR exon 1 upstream sequences (31)
Poly(A)-selected RNA from HeLa cells was reverse-transcribed using an ADAR specific antisense primer. Second-strand cDNA synthesis was achieved by adding an enzyme cocktail of RNase H, DNA Polymerase I, and DNA ligase. Blunt ligation of 5′ RACE adapter sequences to the cDNA ends was done using T4 DNA ligase, and this was followed by PCR amplification of ligated fragments. Products were cloned into vectors (using restriction sites present in the RACE adapter sequences) for DNA sequencing. In the genomic sequence shown above, the arrows show variations seen in the 5′ ends of the sequenced inserts, corresponding potentially to differences in transcriptional start sites. The +1 label and highlighted ATG show the first methionine codon of the full-length ADAR p150. A purine rich (Pu Box) region is upstream of the putative transcriptional start sites, and is highlighted by a box. Further upstream, boxed, are two SP1 transcription factor binding sites. Even further upstream is a cAMP responsive element binding (CREB) site. CREB is a transcription factor that is part of the cAMP pathway (36).

Of curiosity, unlike many other eukaryotic genes, ADAR lacks a clear canonical TATA-type promoter signal upstream of the putative transcriptional start sites. A few years later, the sequences upstream of this exon were determined to be responsive to interferon, and an interferon-stimulated response element (ISRE) and kinase conserved sequence (KCS) were also identified (Figure 1.2.15). KCS was found to be required for efficient

transcription of the PKR gene both in the presence and absence of interferon (37).

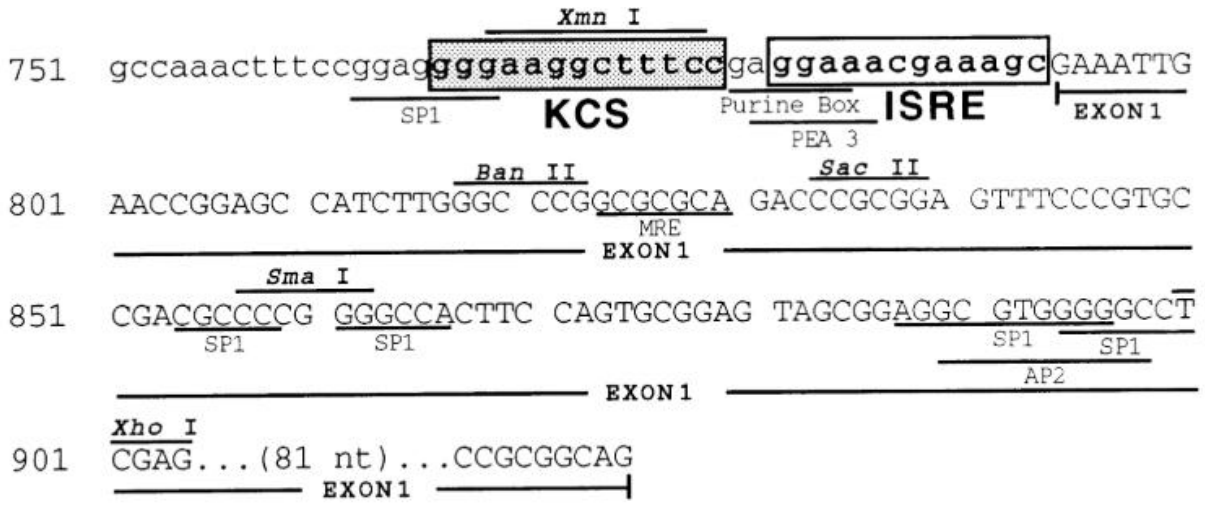


Figure 1.2.15 ADAR exon 1 interferon responsive sequences (31) *In this experiment, total human placental RNA was used as the starting material, along with random hexamer priming for reverse transcription and completion of cDNA synthesis using the traditional cocktail of RNase H, DNA Polymerase I, and DNA ligase. After ligation of RACE adapters, several rounds of PCR, including nested PCRs, were performed on the cDNA using sense adapter-specific primers and antisense ADAR-specific primers. The PCR products were then cloned, sequenced, and mapped to the genomic locus based on Southern blot hybridization of cDNA probes (available from earlier experiments) to restriction-digested lambda or P1 phage library DNA. The digested genomic DNA fragments that gave signals on the Southern blot could then be subcloned from the original giant phage vectors into smaller plasmids, sequenced, and aligned to the RACE cDNA sequences to reveal the exact sequences upstream of transcriptional start sites. Shown above, upstream of the start of exon 1, in boxes, are the ISRE and KCS, identified based on sequence homology (38). Of interest, the KCS sequence has only been found in the PKR promoter, and now, the ADAR promoter (39). Part of the ISRE had been referenced in an earlier study as a purine-rich transcriptional enhancer. Also labeled are consensus sequences for proximal specificity protein 1 (SP1), activating protein 2 (AP2)(40), and metal response element (MRE) binding factors, along with restriction digest sites.*

The RACE studies proved to be fruitful not only in defining potential ADAR exon 1A transcriptional start sites, but also identification of another exon 1B further upstream that shares the same linkage site with exon 2 (Figure 1.2.16). Unlike exon 1A, the newly identified exon 1B lacks a methionine codon in the

ADAR open-reading frame when spliced to exon 2 and thus serves as an extended UTR during translation. The first start codon in these 1B splice variants (in the correct reading frame) is located at the 296th position in exon 2 and encodes p110.

Exon 1B is shorter than 1A, with the longest RACE cDNA clone showing a 107-base sequence starting from the splice junction with exon 2, in contrast to exon 1A, which extends up to 201 bases from the junction. Of interest, the modern human genome build, hg38, and the associated curated NCBI Reference Sequence (RefSeq) database, annotates exons 1A and 1B with a maximum length of 257 and 126 bases, respectively.

The curated annotations, according to the NCBI, are determined based on a combination of: 1. analysis of publicly-available sequencing data, including data uploaded during publication of both related and unrelated studies; and 2. a review of relevant literature before the current trend of uploading sequencing data during the publication process. The increased maximum lengths of exons 1A and 1B as seen in the curated NCBI RefSeq database compared to the earlier RACE studies probably reflect the increased sensitivity of deep-sequencing methods, ranging from the popular fragmentation and sequencing-by-synthesis methods to nanopore-based single-molecule sequencing.

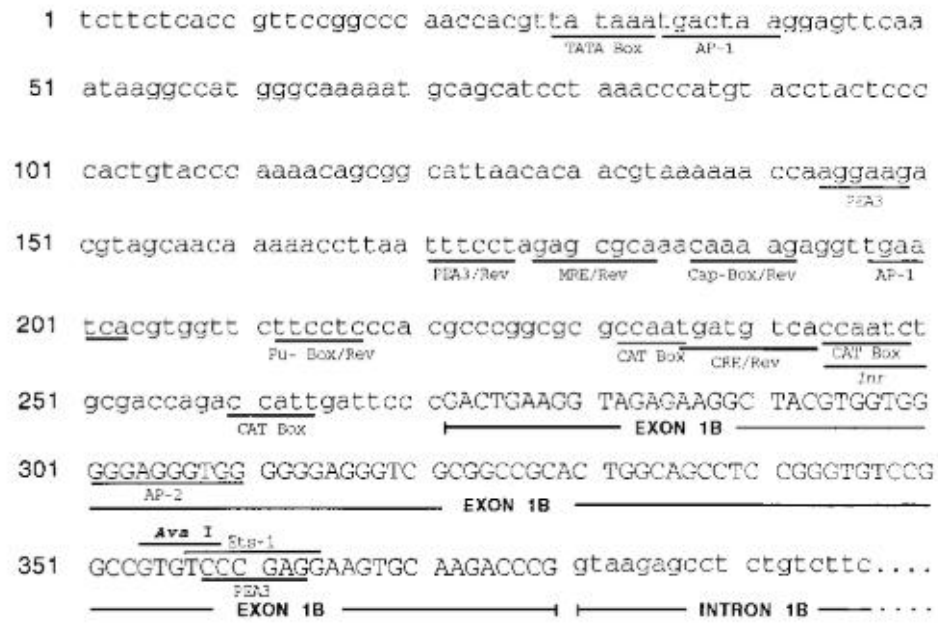
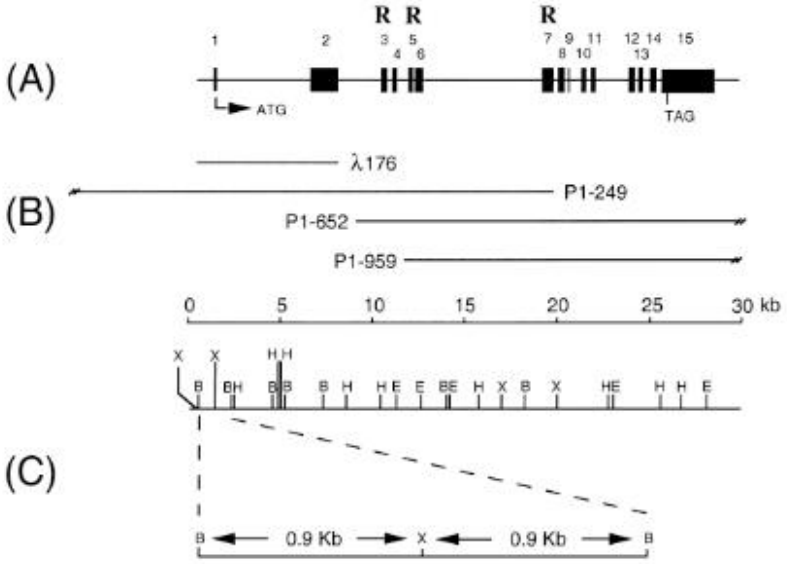


Figure 1.2.16 Annotated sequences upstream of exon 1B (41)
Here, the ADAR exon 1B sequence is shown, based on the largest 5'RACE product, along with upstream genomic sequences and downstream intron sequences. Unlike exon 1A, this exon has a canonical eukaryotic promoter, labeled as TATA Box. Putative transcription factor binding sites are labeled for activator

protein 1 (AP-1)(42), polyoma enhancer activator 3 (PEA3), MRE binding factors, catabolite activator protein (Cap-Box) (43), purine-rich binding factors (Pu-Box), CCAAT box binding factors (CAT Box), cAMP response element factors (CRE), transcription factor IID (Inr)(44), AP-2, and erythroblast transformation specific variant 1 (Ets-1)(45). The AvaI restriction site shown is one of two restriction sites in the genomic DNA insert of a large P1-phage clone that allowed for digestion of the fragment containing ADAR exon 1B.

The exon 1A and 1B ADAR splice variants identified from the RACE sequences have distinct promoter activities. The genomic sequences upstream of exons 1A and 1B were formally tested for level of constitutive activity and responsiveness to interferon by chloramphenicol acetyltransferase (CAT) promoter assays (Figure 1.2.17-1.2.18).



(D)

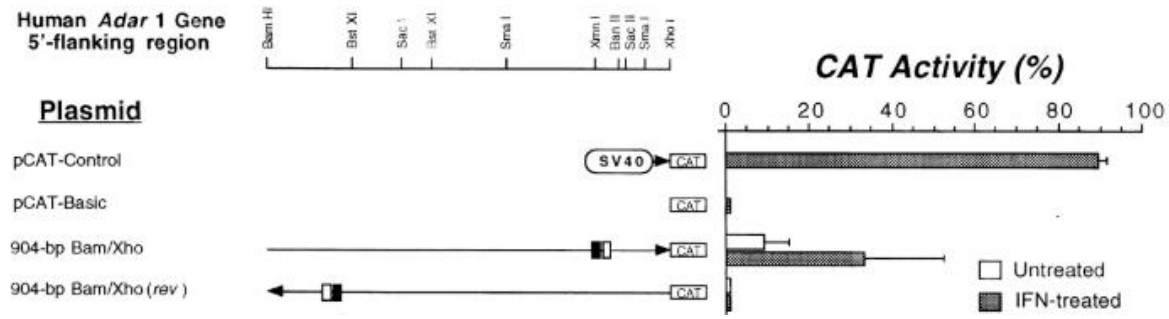


Figure 1.2.17 ADAR exon 1A promoter assays (35)
A. Shown here is the genomic organization of the ADAR locus, with exon 1A denoted by "1" and the other 14 exons denoted by 2-

15. The letter "R" appears above the exons that encode the three dsRNA-binding domains.
B. Shown here are overlapping insert sequences from the lambda (human placental DNA) and P1 (human fibroblast DNA) phages isolated from screens using fragments of ADAR cDNA as probes.
C. Shown here are restriction sites present in the assembled genomic DNA sequence, and in particular, the 904-base BamHI/XhoI fragment that contained the sequences upstream of exon 1A is magnified.
D. Shown here are results from an experiment in which the 904-base sequence was tested using the CAT-reporter plasmid. The reporter contains the chloramphenicol acetyltransferase gene, which, when expressed, catalyzes addition of an acetyl functional group to chloramphenicol; this process originally evolved in E. coli to prevent chloramphenicol from binding bacterial 23S rRNA and inhibiting translation (46). In this experiment, the BamHI/XhoI fragment was first blunted by Klenow fragment, a protein fragment created from cleavage of DNA Polymerase I with the subtilisin protease (47). The blunted DNA fragment was inserted into a single blunted cut introduced into the reporter plasmid and the orientation was confirmed by sequencing. The pCAT-Control plasmid, which contained the simian virus 40 (SV40) promoter/enhancer, was used as a positive control for signal, and the pCAT-Basic plasmid, which lacked a promoter, was used as a negative control. The four plasmids were transfected into human amnion U cells using a traditional DEAE-dextran transfection method, and cells were either treated or untreated with alpha-interferon 24 hours after transfection. Cells were harvested up to 72 hours after transfection, and extracts from the cells were tested on chloramphenicol substrates labeled with carbon-14. The reaction products were analyzed by thin-layer chromatography and measurement of carbon-14 radioactivity levels, taking advantage of the fact that acetylated chloramphenicol migrates differently than the original, unmodified molecule. The protein-normalized enzymatic activities of CAT are shown for the four transfection conditions with and without interferon. The BamHI/XhoI fragment drives expression of the CAT gene only in sense orientation and in the presence of interferon. The black and white rectangles show the locations of the KCS and ISRE elements, respectively, as referenced earlier in Figure 1.2.15.

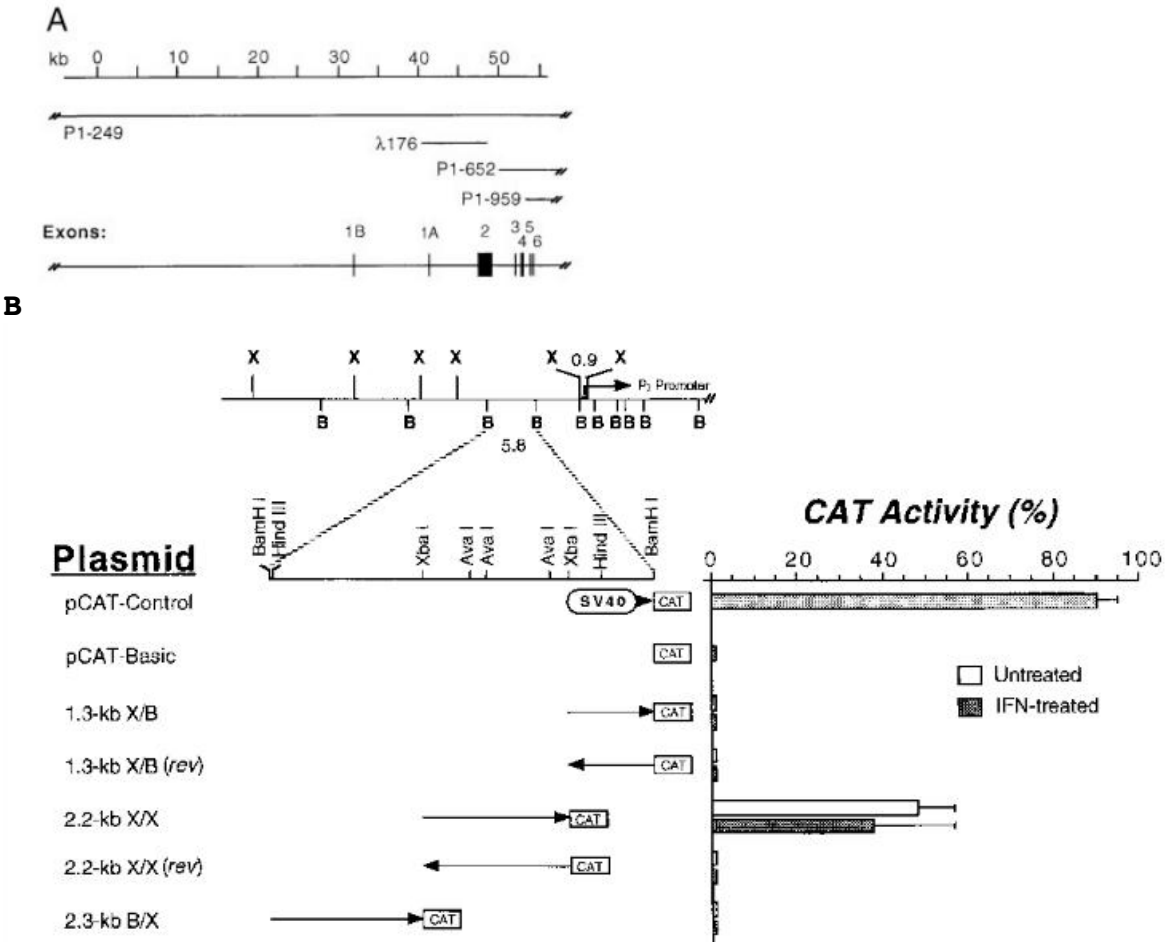
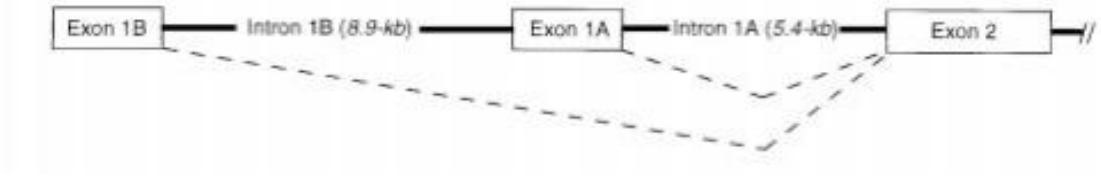


Figure 1.2.18 ADAR exon 1B promoter assays (48)
A. The same four lambda (human placental DNA) and P1 (human fibroblast DNA) phages as the ones described in Figure 1.2.17 are shown here with an extension further upstream of the ADAR exons. Exon 1B is indicated in the schematic along with several downstream exons, including exon 1A and its interferon-inducible promoter (PI Promoter). Like with the previously mapped exon 1A, the exon 1B sequence from 5' RACE sequencing was used to design 1B-specific probes, which were found to hybridize only to DNA in the phage P1-249 clone but not the other two P1 clones or the lambda phage clone.
B. The 5.8-kb BamHI/BamHI fragment digest from the P1-249 insert is shown, and subsequent restriction digests within this fragment were tested using the CAT-reporter system as described in Figure 1.2.17. Only the sense-directional 2.2-kb XhoI/XhoI had promoter activity, which turned out to be constitutive.

With the promoter and splicing activities for exons 1A and 1B formally tested, scientists now had a better understanding of the expression of ADAR in human cells. The promoter activities

and expression levels for these two splice variants were tested in vivo in amnion U cells (Figure 1.2.19).

A



B

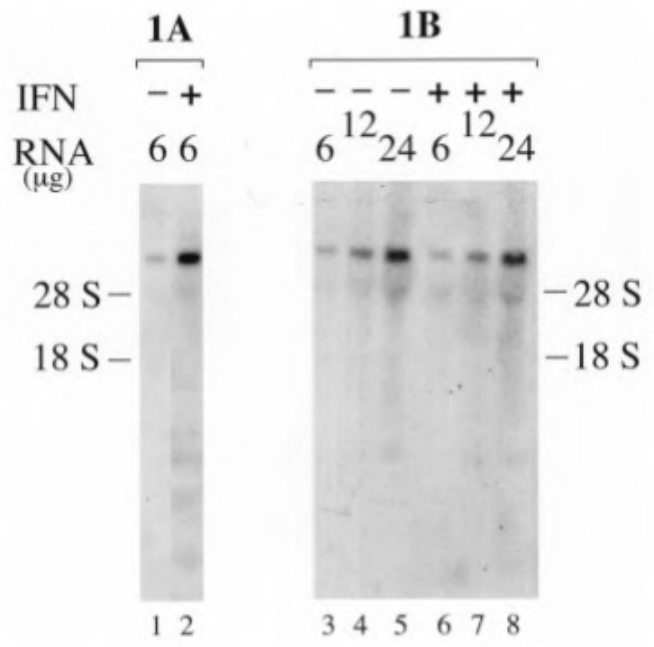


Figure 1.2.19 Relative expression levels of exons 1A and 1B (41)
A. Shown here is a schematic of the first three exons and two introns of the ADAR locus, with splice linkages indicated by dotted lines.
B. An experiment is shown here in which human amnion U cells were treated with 1,000 units/ml of alpha-interferon (from leukocytes stimulated by Sendai virus) for 24 hours. Total RNA was isolated from the cells and probed with phosphorus-32 end-labeled probes specific for the exon 1A/exon 2 junction and the exon 1B sequence. The 28S and 18S eukaryotic rRNA positions are indicated next to the Northern blots. The increase in band intensity seen with the exon 1A/exon 2 probe corresponds with interferon treatment. By contrast, interferon treatment is not associated with an increase in band intensity when probed with the exon 1B probe, for the same amount of starting RNA material.

Of note, there are additional possibilities for both transcription and splicing upstream of exon 2, in addition to exons 1B and 1A, as will be shown in the next chapter. In terms

of alternative splicing downstream of exon 2, truncations were observed in ADAR exons 6 and 7 that led to variations in the second and third dsRNA binding domains for the full-length ADAR isoform (Figure 1.2.20). Although all variants had comparable A-to-I editing activity, mutation of the conserved lysine residues (49, 50) responsible for dsRNA binding in either the first, second, or third dsRNA binding domains of the variants affected their editing efficiencies differently. For example, abolishing the dsRNA binding activity of the second domain caused reduced editing for the binding-abolished non-truncated version relative to the binding-competent non-truncated version. However, this same lysine mutation in the second domain increased the efficiency of editing in the binding-abolished truncated versions, compared to the corresponding binding-competent truncated versions, at least in terms of in vitro editing levels (51).

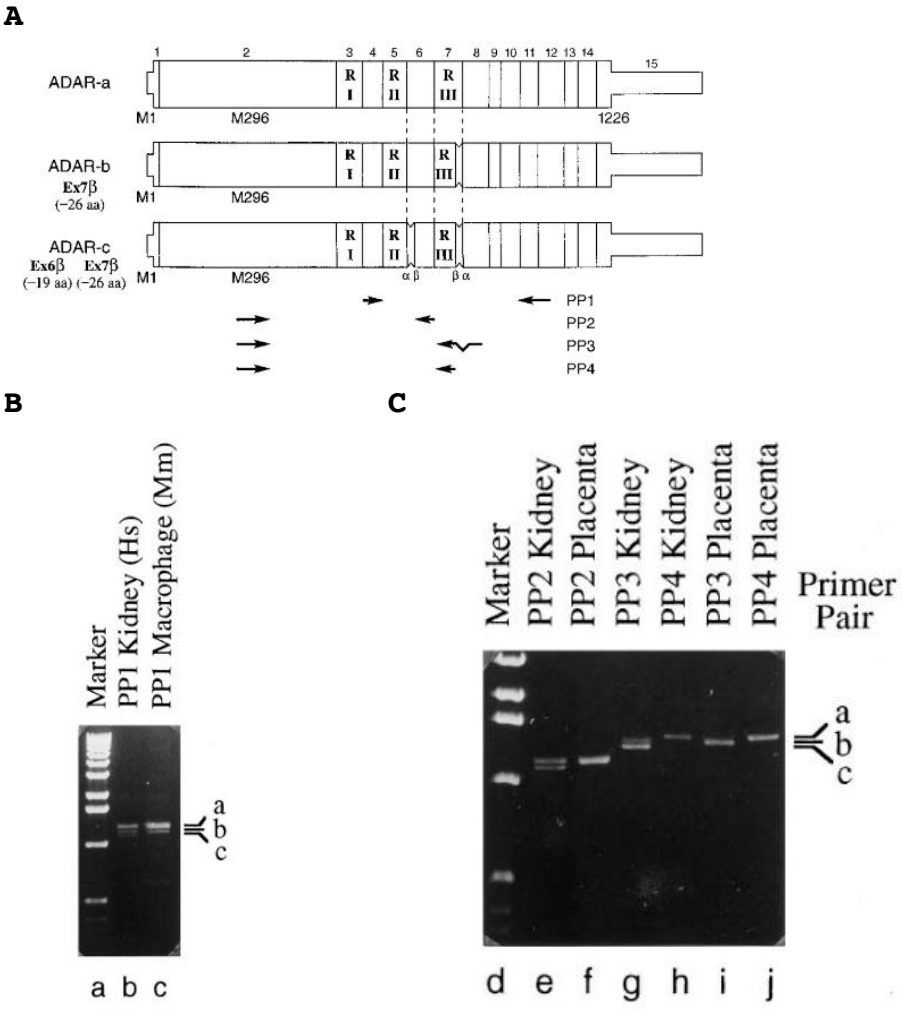


Figure 1.2.20 ADAR exons 6 and 7 splice isoforms (51)
A. The schematics here show the open-reading frames of three full-length ADAR variants with changes in the splice junctions between exons 5-8. In the non-truncated version, ADAR-a, the

open-reading frame encodes a 1,226 amino acid polypeptide, the theoretically longest isoform. ADAR-b is made up of 1,200 amino acids and results from a truncated exon 7; ADAR-c is 1,181 amino acids in length and results from truncations in exons 6 (exon 6-beta) and 7 (exon 7-beta), all due to alternative splicing. The four primer pairs (PP1, PP2, PP3, and PP4) were used to make PCR products from cDNA libraries prepared from human kidney, human placenta, and mouse macrophages. PP4 was designed to screen for a fourth theoretically possible cDNA variant (truncation in exon 6 but not exon 7) given the observed alternative splice sites in exons 6 and 7.
B. Shown here is 1% agarose gel electrophoresis of cDNA products, stained with ethidium bromide. Lanes b and c show the presence of cDNA encoding the ADAR-a, ADAR-b, and ADAR-c (faint smallest band) variants, for human kidney, and only ADAR-a and ADAR-b variants for mouse macrophages.
C. Testing with PP2 in human kidney resulted in two bands and confirmed the presence of alternative splicing leading to a truncated exon 6, as shown in lane e. In human placenta, this truncation in exon 6 was not observed, as PP2 only gave one PCR product, as shown in lane f. Therefore, human kidney is able to express both truncated variants, while placenta only expresses the ADAR-b truncated variant. PP3 is used to further test for the presence and absence of ADAR-b and ADAR-c variants; it includes a reverse primer that binds at the splice junction between a truncated exon 7 and exon 8. PCR with PP3 gave two bands, as expected, for kidney cDNA, and only one band for placental cDNA, as shown in lanes g and i, respectively. Finally, PP4 was designed to test for the presence of the fourth theoretical isoform, with a full-length exon 7 (7-alpha) and a truncated exon 6 (6-beta). Although the reverse primer in PP4 is drawn to appear as if the binding site is solely within truncated exon 7 (7-beta), the agarose gel results are more consistent with usage of a primer that spans the splice junction between full-length exon 7 (7-alpha) and exon 8. The cDNA from kidney and placenta yielded only one product, as shown in lanes h and j, respectively, suggesting this fourth theoretical splice variant is not expressed in these tissues. Markers for ADAR-a, ADAR-b, and ADAR-c shown to the right of the gel apply only for the PCR products shown in lanes g-j.

The two exon 7 truncated ADAR isoforms either lacked 26 amino acids in third dsRNA binding domain, or that plus lacking another 19 amino acids in the region between the second and third dsRNA binding domains; these truncations did not seem to affect the conserved lysine residues important for dsRNA binding activity (51). The cDNA encoding these three variants were cloned into expression vectors without 5′ UTR sequences, which

seemed to inhibit expression of the proteins (51, 52). This will be relevant in experiments presented in later chapters that also utilize expression vectors with ADAR cDNA variants lacking their 5′ UTR sequences. Truncation constructs were made in both the p150 and the p110 forms to test how well these cDNA variants could be translated in vivo (Figure 1.2.21).

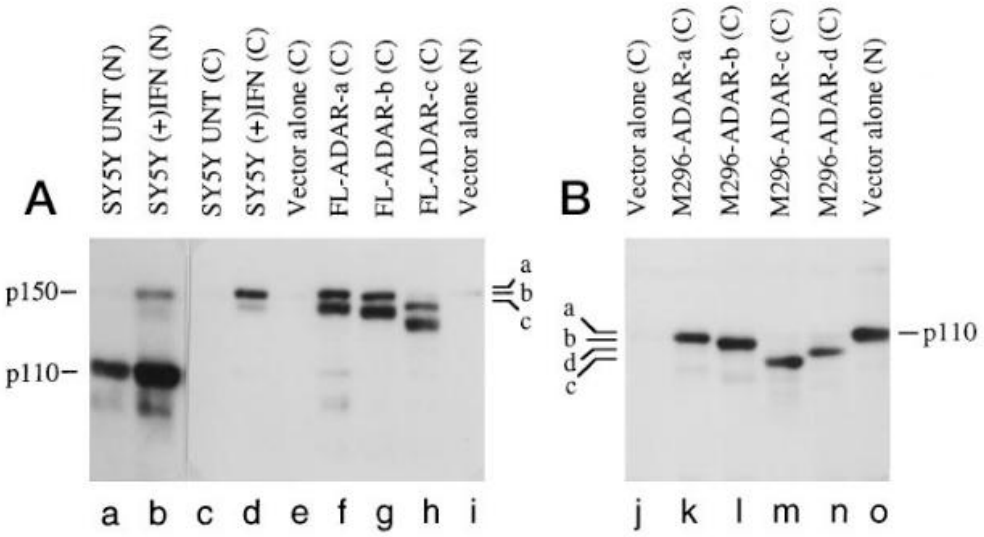


Figure 1.2.21 Expression of ADAR exon 6 and 7 variants (51)
A. The four theoretically possible ADAR exon 6 and 7 variants (a, b, c, and the theoretical d) were cloned into expression vectors and transfected into monkey kidney COS-1 cells using traditional DEAE-dextran transfection methods. As benchmarks, nuclear (N) and cytoplasmic (C) extracts from human SY5Y neuroblastoma cells with and without interferon treatment were included in the immunoblot. Lanes a and b show extracts probed with an anti-ADAR antibody recognizing the middle of the protein, with the bands smaller than p150 and p110 corresponding to the putative truncations (a, b, and c) of the full-length (methionine-1) and methionine-296 ADAR versions. Extracts in lanes c-i were probed with a p150-specific antibody. Of note, expression of the full-length ADAR, without truncations (lane f), appeared to produce a smaller isoform around the size of p110, but because the antibody used was a p150 N-terminal specific antibody, this band and the other bands smaller than the largest bands were regarded as degradation products.
B. The two alternatively spliced exons were introduced into the p110 open-reading frame; these constructs also showed efficient translation in COS-1 cells. One curiosity is that the band in lane letter l (ADAR-b) appears to be higher than the band in lane n (the theoretical ADAR-d variant), even though the ADAR-b variant, which is truncated by 26 amino acids, should be smaller than ADAR-d, which is truncated by 19 amino acids, both relative to the 931 amino acid full-length p110 version of ADAR-a.

In conclusion to this introductory chapter, it is worth mentioning that the human ADAR family of proteins include ADAR1; ADAR2, discovered and cloned in 1997 (53); and ADAR3, discovered and cloned in 2000 (54). The years here refer to when the human versions were isolated; the rat homologs, called RED1 and RED2, were identified earlier (55, 56). ADAR1 maps to chromosome 1, while ADAR2 and ADAR3 map to chromosomes 21 and 10, respectively. ADAR2 and ADAR3 lack the Z-DNA/Z-RNA binding domains found in ADAR1 and have two rather than three dsRNA domains (57). Furthermore, while the catalytic domains in ADAR1 and ADAR2 are functional, this domain has not been found to be active in wild-type ADAR3 (58). The experiments reviewed above and ones that will be presented in later chapters focus on ADAR1.

CHAPTER 2. Regulation and function of ADAR1

As reviewed in the first chapter, the expression and regulation of ADAR1 is complex and is worth re-examining using advances made in sequencing and other molecular biology techniques during the last two decades. The second chapter of this thesis aims to: 1. describe ADAR1 regulation at the DNA level with an application to screening of CRISPR knock-in clones; 2. describe the types of ADAR1 splice variants; 3. describe how splice variants contribute to ADAR1 p110 and p150 protein levels; 4. describe the role of ADAR in maintaining immune homeostasis in human neuronal cells; and 5. discuss the belated evolution of longer ADAR isoforms compared to shorter ones in the animal kingdom.

2.1 ADAR1 DNA regulation

Human embryonic kidney (HEK) 293T cells were selected for nearly all experiments, largely because of convenience with regard to cell culture maintenance, and efficiency in terms of genetic manipulations. The number 293 refers to the experiment number from the lab in which this cell line was originally made, by transformation of primary human embryonic kidney cells with adenovirus type 5 DNA, which has regions that stimulate primary cell proliferation (59, 60).

For purposes of studying ADAR1, 293T cells are easily responsive to human interferon treatment and express both p110 and p150 isoforms, along with extensive global editing of RNA (61). Although ADAR2 is also expressed in 293T cells, significant ADAR2-associated A-to-I editing has not been detected in 293T cells, at least by comparing the editomes of ADAR2-KO and WT 293T cells (61). As the focus of this thesis is on ADAR1 and ADAR1-associated A-to-I editing, the use of 293T cells makes sense given that ADAR1 is the primary A-to-I editor in these cells. Previous work has suggested that ADAR1 is the primary editor of repetitive, non-coding RNA sequences in nearly all human tissue types, whereas ADAR2 targets more non-repetitive, protein-coding RNA sequences (62).

As had been reported previously, ADAR exon 1A has an interferon-stimulated response element (ISRE) upstream of the transcriptional start site (35). We took advantage of this element both in terms of validating the ability of ADAR to be induced by interferon-beta and determining the effectiveness of a CRISPR knock-in method, in which the goal here was to knock in the open-reading frame of GFP (minus the first ATG codon of GFP) directly after the p150-ATG codon. By stimulating the ISRE upstream of exon 1A with interferon, induction of GFP-tagged ADAR would enable quantification of GFP-positive cells to compare efficiencies of guide RNAs, among other factors.

Furthermore, fluorescence-activated cell sorting (FACS) by GFP-signal would allow isolation of single-cell clones to determine knock-in efficiency in terms of homozygosity and heterozygosity. Here, use of the terms homozygosity and heterozygosity is not perfectly precise, given that these terms are typically used when discussing diploid cells. 293T cells are not exactly diploid, but yet also not fully triploid; some cells in culture may have more than two copies of chromosome 1, where the ADAR1 gene locus is located, and some may have exactly two copies. The total number of chromosomes in 293T cells varies from 56 to 78, and about a third of cells in culture was found to carry 64 chromosomes (63–65). For all subsequent experiments, unless otherwise stated, 293T cells were maintained at 37°C and 10% CO_2, and cultured using Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum and MEM non-essential amino acids (NEAA) solution.

The CRISPR knock-in method was adapted from two reports, one describing a plasmid-based Cas9 expression system with built-in guide RNA cloning sites, and another describing a plasmid-based donor DNA strategy for knock-in and replacement of target sequences by homologous recombination, after double-stranded breaks are introduced in both the donor DNA plasmid and chromosomal DNA (66, 67). Our experiment was done in WT 293T cells transfected with a Cas9/sgRNA expression plasmid and donor DNA plasmid carrying the insert containing the GFP open-reading frame (Figure 2.1.1).

A **B**

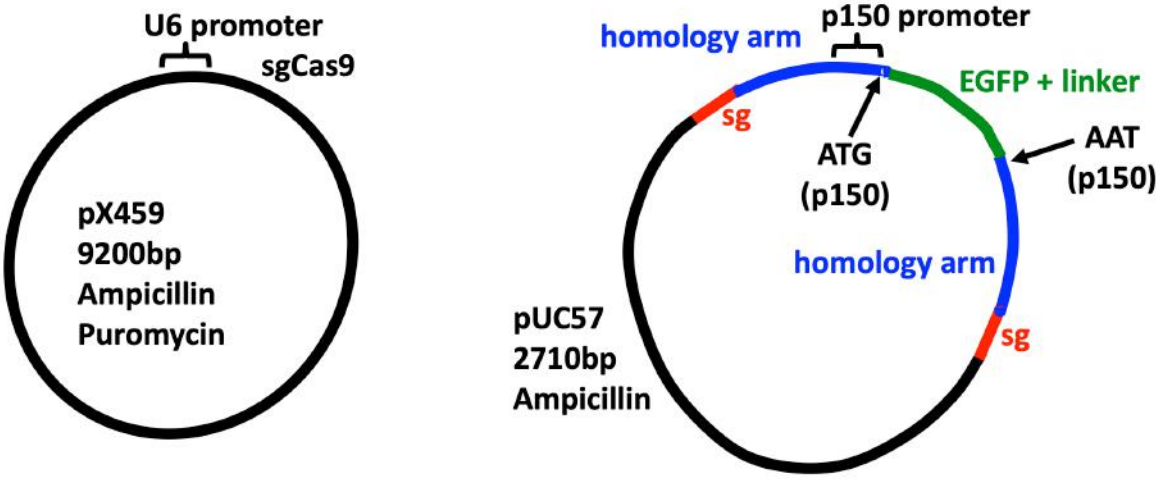


Figure 2.1.1 Plasmids for CRISPR
A. The plasmid shown on the left, PX459, is available from Addgene and has a U6 promoter that promotes expression of the single guide RNA (sgRNA), which has both a target-sequence specific CRISPR RNA and a trans-activating CRISPR RNA fused into a single molecule. A CMV enhancer element is upstream of the

open-reading frame encoding Cas9 connected to puromycin-resistant protein (not applicable for this particular experiment) with a T2A self-cleaving peptide in between. The T2A works by causing the translating ribosome to skip forming a peptide bond at the C-terminal end of the T2A peptide (68). Finally, the ampicillin-resistance gene in the plasmid has its own promoter and allows propagation of the plasmid in transformed DH5-alpha E. coli cells cultured on Lysogeny broth (LB) agar plates with carbenicillin for selection.
B. The donor DNA plasmid is a pUC57-based plasmid that encodes resistance to ampicillin. The insert contains the DNA sequence corresponding to ADAR1 exon 1A and the flanking sequences extracted from the most recent build of the human genome (hg38). The methionine codon in exon 1A now serves as the beginning of the open reading frame for GFP, and the GFP stop codon is removed and replaced with a linker peptide sequence, followed by the rest of exon 1A, which correspond to amino acids 2-5 of the p150 protein. The ends of the sequences flanking exon 1A (with the inserted enhanced GFP and linker proteins) includes the sgRNA sequences to be tested. The concept here is to have the guide RNA bring Cas9 to both the genomic locus and the dsDNA plasmid, thus introducing double-stranded breaks in both substrates. During the process of homology-directed DNA repair, the hope is that some molecules of liberated donor DNA, with the extended homology arms, will recombine with the chromosomal DNA, completing the knock-in.

The donor DNA plasmid, a customized pUC57 plasmid, was synthesized by the Gene Universal company. This small plasmid encodes an ampicillin-resistance gene and was derived from the original pUC (University of California) series of plasmids with multiple cloning sites (69, 70). After delivery, the plasmid was transformed into DH5-alpha E. coli cells using heat-shock for 45 seconds followed by plating on carbenicillin-selective LB agar plates and incubation at 37°C overnight. Single colonies were selected the next morning for single-tube cultures in LB broth with carbenicillin to maintain selective expansion of the transformed bacterial cells. After reaching the optical density threshold, the bacteria in suspension were pelleted, and plasmid DNA was extracted using the Qiagen miniprep kit, following manufacturer guidelines. 500ng of individual minipreps were mixed with a primer that binds to the insert and sent to Macrogen for sequence verification. Shown below is the insert sequence contained in the pUC57 plasmid:

5′-**GGGCGCAATGAATCCGCGGC(AGG)**TGCGCTCCCGCCCATCCCCTCCCCCCCCCCTCCACATTGGAGACGCGGC
CACCACCGCGCTGGCGCGGAGAGAGGGAGGACCGGGCGTCATGCTGTTTCTGGCCTGAGGTTTTGTGTGCCTTTGTT
TTCCTTTTGCTCTATTCGTGTATTCCTGCCTACGGCCTGTGCGGGGAATTAGGAGCTCAGTACTGAAACGGCGGTTT
TCCTAAACAGTACCGGACGGGCGCGGGGGCTGACGCCTGTAATCCCAACACTTTGGGAGGCCGAGGTGGGCGGATCT

CTTGAAGCCGGGAGTTCGAGACCACCCTGGCTAACGTGGTGAAACCCTGTTCTTACTAAAAATACAAAAAAAAAAAA
AAAAAAAAAGCCAGGAGTGATGGCGCTCGCCTGTAATCCCAGCTACTCCGTAGGCTGAGGCAGGAGAATCGCTTGAA
CCCGGGGGGCAGAGGTTGCAGTGAGCCGAGATTGCGCCATTGCACTCCAGCCTGGGCAAAAAGAGCGAGACTCCGCC
TCAAAAAAAAAAAAAAAAGTACCTTCCGTAGTTCTCATGCAGCGGAGGGGTTCGACTTGTAACCGGCCTGAAACCAA
GCGTGGCGCAAGATTTGCTCAA__GCCCCTCCTCTTGGCCAAACTTTCCGGAGGGGAAGGCTTTCCGAGGAAACGAA
AGCGAAATTGAACCGGAGCCATCTTGGGCCCGGCGCGCAGACCCGCGGAGTTTCCCGTGCCGACGCCCCGGGGCCAC
TTCCAGTGCGGAGTAGCGGAGGCGTGGGGGCCTCGAGGGGCTGGCGCGGCCCAGCGGTCGGGCCAGGGTCGTGCCGC
CGGCGGGTCGGGCCGGGCAATGCCTCGCGGGCGCAATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCC
ATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTA
CGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGA
CCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAA
GGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGG
CGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGG
AGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATC
CGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGT
GCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGG
TCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGGGTGGAGGCGGTTCAGGC
GGAGGTGGCTCTGGCGGTGGCGGATCGCTCGAGAATCCGCGGCAG__GTAAGCCGGGCCGGCCTTGGACCTTCGCCG
CCGTCTGGGTTCGTTTACAACCTCACAGGCTTTGTGTTGCAGTGCGTAGCGTGTGCGTCTTGTGAGTGTTAGAGTGT
GTGTGTGTGTGTCGTCTTGCCAAGCAGCATTGCTGGTTTAGGAATTTGTGCGTCTTGTGAGTGTGTGTGTGTGGGTG
TGTGTCGTCTTGCCAAGCAGCATTGCTGGTTTAGGAATTTGTGCGTCTTGTGAGAGTGTGTGTGTGTGTGTGTGCGT
GTGTGTGTAGTCTTGCCAAGCAGCATTGCTGGTTTAGGAATTTGTGAATTTGTATCCTGCTCATTAATTCTGCAGAA
TGGAGCAGTGCGTGAAGAGGGCTTGGGGGAAAATGCGCCCCCGTCTGAGTAGGAAGGCCTGAGCCCATGTCAAGGCA
GACACATCGTCTCCCTTTCTGCTAGGGCCCCTTGTGGAACCCCCTACCCCCGCTTTAGCCCCACTTGAACAACGTTC
GGACTTTGAGCAGCGCACACTATCCTCAGCTCACCTTATCCACCTCCTGAAGGCCTTCTGGGAGTTAAAAATGGCAC
TTAAGCTGTAGGAGAAAGCTTGTTAACCACTTTATAG**(CCT)GCCGCGGATTCATTGCGCCC**-3′

The hg38 annotated ends (upstream genomic DNA/exon 1A junction and exon 1A/intron junction) of exon 1A are shown above with the "__" spaces. The eGFP sequence is shown in green, and the linker peptide sequence, comprised largely of glycine codons, is shown in blue. Note the linker peptide connects the final amino acid of eGFP (lysine-AAG) with the second amino acid of ADAR1 p150 (asparagine-AAT).

The pUC57 plasmid containing the sequence above was transfected along with the PX459 plasmid that contains the GFP-KI guide#1 oligos (shown on the next page). The pUC57 plasmid containing the same sequence above, but with the bolded target sequences at the ends replaced with the GFP-KI guide#2 target sequence, was transfected along with the PX459 plasmid that contains the GFP-KI guide#2 oligos.

To knock-in the GFP open-reading frame within the ADAR1 exon 1A locus, a double-stranded break in the chromosome would need to be introduced by Cas9, targeted by sgRNA. The double-stranded break in the target sequence is expected to be between the third and four bases upstream of the first base of the protospacer adjacent motif (PAM) site (71). This is relevant for design of the homology arms in the donor DNA plasmid. Given that the corresponding guide RNA sequences are attached to the ends of the homology arms, and oriented such that the PAM sequence is proximal to the ends of the homology arms, the break introduced by Cas9 would result in a dsDNA fragment that contains the 3-base PAM along with an additional 3 bases corresponding to part

of the guide RNA. Therefore, the homology arms should be designed to include only intron sequences to avoid disturbing either UTRs or coding regions of exons. Of course, not all side effects can be predicted, and there is always the chance that altering genomic intron sequences could change DNA regulatory elements or splicing elements. The following guide RNA sequences shown below were chosen based on availability of PAM sites around the p150 methionine. The PAM sequences are shown in parentheses.

GFP-KI guide#1:
Target sequence (sense strand) with PAM:
5′-GGGCGCAATGAATCCGCGGC(AGG)-3′
ssDNA oligos for cloning into PX459:
5′-***CACC***GGGCGCAATGAATCCGCGGC-3′
    3′-CCCGCGTTACTTAGGCGCCG***CAAA***-5′
GFP-KI guide#2:
Target sequence (antisense strand) with PAM:
5′-TCGCGGGCGCAATGAATCCG(CGG)-3′
ssDNA oligos for cloning into PX459:
5′-***CACC***GCGCGGGCGCAATGAATCCG-3′
    3′-CGCGCCCGCGTTACTTAGGC***CAAA***-5′

The G/C highlighted in red for GFP-KI guide#2 is an extra base added to create a preferred transcriptional start site (G) for RNA polymerase III, and does not seem to affect targeting efficiency (66, 72). The oligos shown above were annealed at 95°C for 5min and then cooled to 4°C at a rate of 0.1°C/s. This produced the inserts with overhangs that were then ligated into recipient Cas9/sgRNA expression plasmids (PX459) following digestion with BbsI. Restriction digests and ligations with T4 DNA ligase were done according to manufacturer (NEB) guidelines. Ligated produced were transformed into DH5-alpha bacteria, and carbenicillin-selected clones were submitted for sequence verification of the guide RNA inserts.

With the sets of donor DNA and Cas9/sgRNA plasmids ready, an experiment was set up to transfect them into WT 293T cells in the presence and absence of nocodazole, a small molecule that has been shown to increase the efficiency of homology-directed repair (HDR) and inhibit non-homologous end joining (NHEJ) (67, 73–75). After the double-stranded break, HDR is preferred over NHEJ because the donor DNA fragment needs to be introduced into the genomic locus during the DNA repair process (76). NHEJ is overall more efficient than HDR, and also active in more phases of the cell cycle; NHEJ is ideal when a sequence needs to be removed from the genomic DNA, as will be the case for the exon-deletion experiments shown later in this chapter. Nocodazole prevents microtubule polymerization and arrests the process of

mitotic spindle formation, synchronizing cells in the mitotic phase (77). Because NHEJ is largely inactive during mitosis, synchronization of cells in this stage is thought to inhibit NHEJ and promote HDR activity (78, 79).

WT 293T cells were plated in regular medium with nocodazole for 12 hours before transfection, followed by expansion of cell cultures for interferon treatment, single-cell sorting, and downstream analysis (Figures 2.1.2, 2.1.3, and 2.1.4).

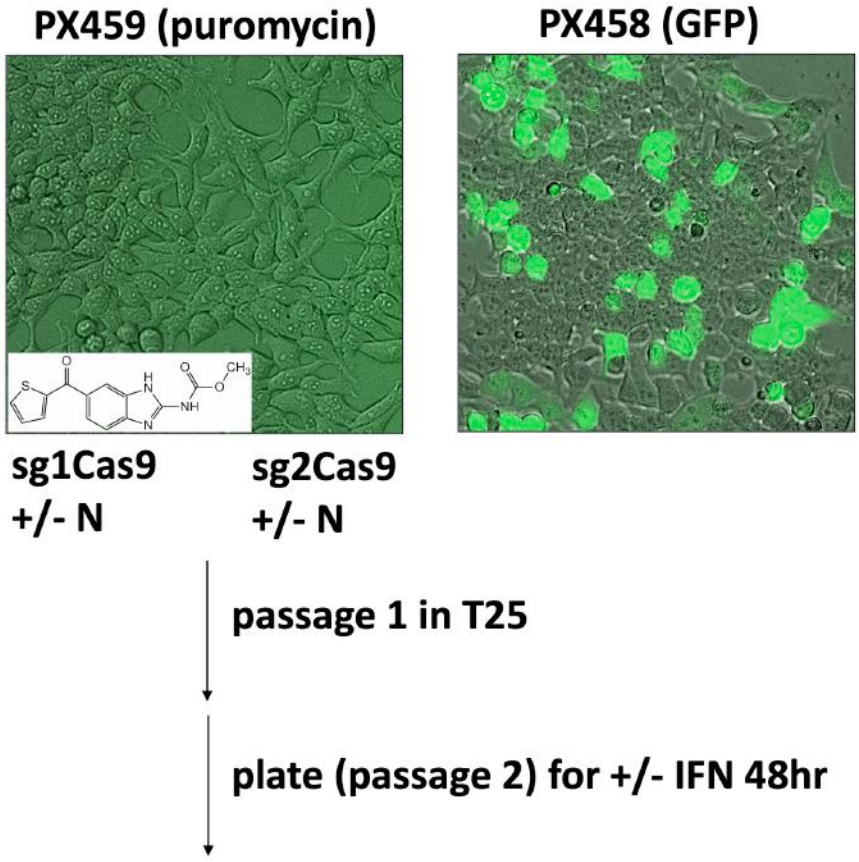


Figure 2.1.2 Transfection of 293T cells for KI
WT 293T cells were plated into wells (30,000 cells/cm^2) coated with polymerized L-enantiomer lysine. Regular 293T medium was added with and without nocodazole at a concentration of 100ng/µl for 12 hours before transfection. Lipofectamine 3000 lipid-based transfection was performed following manufacturer guidelines. The +/- nocodazole groups were transfected either with the PX459sg1/pUC57sg1 or PX459sg2/pUC57sg2 combinations for transient expression of Cas9 and sgRNAs. As a control for transfection efficiency, cells were also transfected with PX458, which is the same as PX459 except the puromycin-resistance gene is substituted by a GFP gene. Transfected cells were cultured under standard conditions for 24 hours and imaged, as shown above, with a mix between the phase contrast and GFP channels. 24 hours after transfection, the media was changed to remove

nocodazole and lipid-DNA complexes. After growing to about 80% confluency, cells were passaged once and expanded (T25 refers to the 25cm^2 flask used for the expansion) before plating (30,000 cells/cm^2) for further experiments.

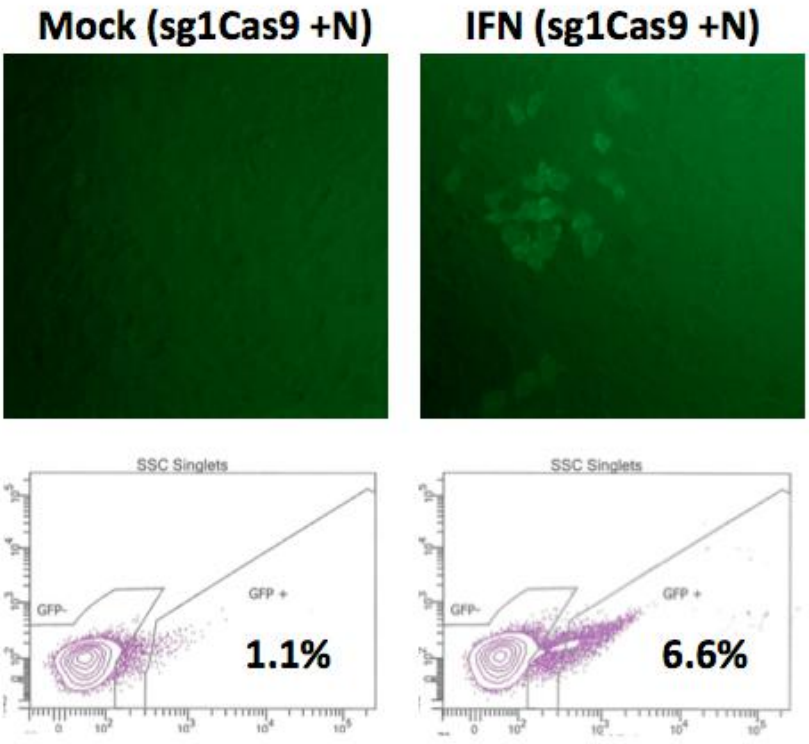
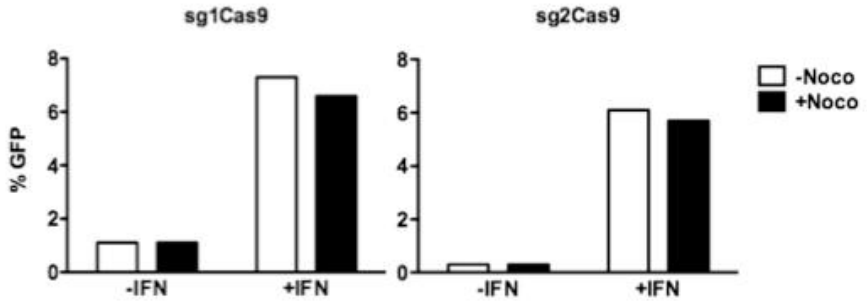


Figure 2.1.3 Flow cytometry analysis of bulk GFP-KI clones
Bulk transfected cells were plated (30,000 cells/cm^2) and treated with human interferon beta 1a by addition of recombinant protein (PBL Assay Science) to regular culture media at 1nM concentration for 48 hours. After 48 hours, cells were imaged using the GFP channel of a Zeiss microscope, and colonies of green cells were observed in the interferon-treated groups. GFP-positive cells were sorted into 96-well (0.32cm^2) flat-bottom plates using the BD FACSAria II flow cytometer at The Rockefeller University Flow Cytometry Resource Center. For single-cell culturing, media in the 96-well plates consisted of 50% regular medium with 50% conditioned medium, which is supernatant harvested and filtered from other 293T cell cultures. After 2 weeks, single-cell clones were observed in, on average, about 20% of the wells. Clones were then lifted from the wells with Accutase gentle cell dissociation solution and transferred to 6-well (9.6cm^2) plates for expansion.

A



B

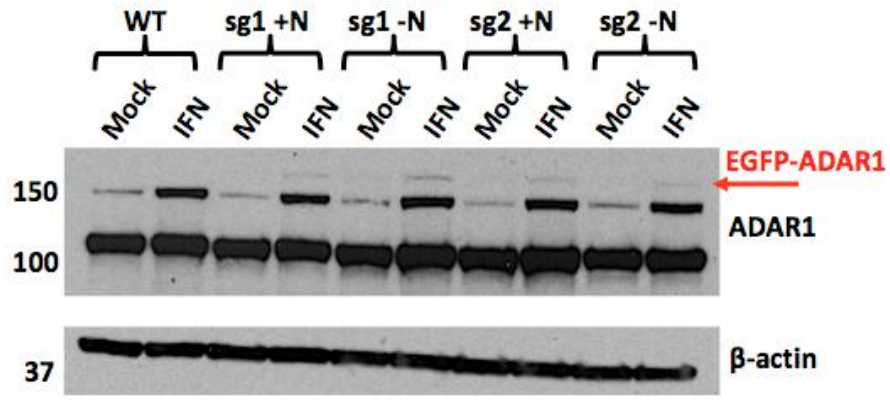


Figure 2.1.4 Immunoblot analysis of bulk GFP-KI clones
A. Originally, after 48 hours of interferon treatment, some wells were submitted for single-cell sorting by GFP signal and other wells were prepared for immunoblot. The graph here shows percent GFP-positive cells in the bulk cell population as measured by flow cytometry. Pre-treatment with nocodazole did not appear to significantly change the knock-in efficiency for this experiment. GFP-KI guide#1 may have a slightly different effect compared to guide#2 with regard to KI efficiency.
B. For immunoblot, cells were lysed using 2X SDS-PAGE Sample Buffer with 400mM dithiothreitol (DTT) as a reducing agent, passed through a 26G needle 10 times to shear DNA, boiled (100°C) for 10min, and centrifuged at 10,000G for 10min. Cell lysate supernatants were loaded into 4-12% Bis-Tris gels and run at 130V in 1X 3-Morpholinopropane-1-sulfonic acid (MOPS) buffer for 2 hours at room temperature. Next, proteins were transferred from gels onto nitrocellulose membranes at 300mA for 2 hours at 4°C. 10X transfer buffer was made using Tris-glycine and mixed with water and methanol to create 1X transfer buffer with 20% methanol. Following transfer, membranes were blocked for 1 hour at room temperature with 5% milk in TBS with 0.1% Tween20 detergent. Following blocking, membranes were incubated with anti-ADAR1 antibodies (Santa Cruz D-8 clone, which recognizes an epitope within amino acids 1051-1226 in the C-terminal end of

ADAR1) and anti-beta-actin antibodies overnight at 4°C, followed the next day by three 15-minute washes, incubation with secondary antibodies conjugated to horseradish peroxidase (HRP) for 1 hour at room temperature, and another set of three 15-minute washes with TBS/0.1% Tween-20. For chemiluminescence, membranes were incubated for 5min using Pico chemiluminescent HRP substrate. The immunoblot revealed presence of a larger band above the interferon-treated groups, corresponding to the GFP-tagged ADAR1p150. WT 293T cell lysates are also shown in the first two lanes to serve as benchmarks.

The experiment described above is an example of how ADAR1 expression is regulated at the DNA level by interferon-beta. The interferon-inducible promoter of exon 1A allowed for sorting of cells that grew from an original clone in which HDR successfully repaired the Cas9-induced double-strand break within exon 1A using the insert from the donor plasmid. The rate of heterozygous to homozygous clones is about 20:1, as determined by single-cell clone immunoblot screening. Another way to screen the clones is genomic DNA PCR using primers that extend into genomic DNA beyond the ends of the homology arms, to avoid amplifying residual plasmid DNA, which may be present in the cell or potentially integrated randomly into parts of the genome.

Of note, one concern with the transfection of plasmids as compared to delivery of Cas9/sgRNA ribonucleoprotein complexes (RNPs) with single-stranded oligo DNA nucleotides (ssODNs) is random integration of expression plasmids, which can potentially generate low-level amounts of Cas9/sgRNA that could continue disrupting the gene locus, especially if the original guide RNA binding sites in the genomic DNA remain unmodified by the knock-in, or if the PAM site is not altered in the donor DNA. In this experiment, successful knock-in would split the guide RNA binding sites, reducing the chance of repeated targeting of the locus by Cas9. Another method to reduce repeated targeting by lingering Cas9/sgRNA complexes would be to alter the PAM, and for guide#1, changing the (C)AGG to (C)AAG would be a way to preserve both the coding sequence of the protein at the fifth amino acid position (CAG/CAA: glutamine) and the splice donor (GT) site of intron 1A beginning with the final guanosine of the AGG PAM site. For guide#2, disrupting the PAM would mean changing the 5′ UTR non-coding sequence of exon 1A.

To continue the story on regulation of ADAR1 expression at the DNA level, we examined the relative expression of different RNA isoforms in a few different cell types. The isoforms selected for testing by quantitative PCR are based on annotations curated by NCBI RefSeq (Figure 2.1.5).

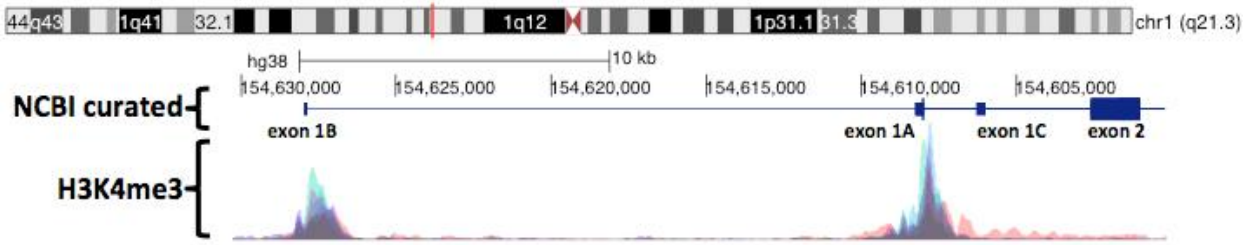


Figure 2.1.5 NCBI RefSeq curated annotations for ADAR1
Shown are selected tracks from UCSC Genome Browser, an interactive web-interface tool for browsing genomic DNA sequence builds from different species and viewing a wide range of genomic annotations. For the ADAR1 locus on chromosome 1, three transcriptional start sites corresponding to exons 1B, 1A, and 1C are annotated, all with linkages to exon 2. Shown below the exon-intron schematic is information extracted from the Encyclopedia of DNA Elements (ENCODE) project (80), showing chromatin immunoprecipitation (ChIP) sequencing data from seven different human cell lines, color-coded and overlayed: GM12878, a lymphoblastoid cell line; H1-hESC, a human embryonic stem cell line; HSMM, human skeletal muscle myoblasts; HUVEC, human umbilical vein endothelial cells; K562, a human leukemia cell line; NHEK, human epidermal keratinocytes; and NHLF, human lung fibroblasts. Here, chromatin was immunoprecipitated using an antibody specific to H3K4me3, which is histone H3 protein with addition of three methyl groups to the lysine 4 residue. DNA was then extracted from the antibody-bound portions for library preparation and sequencing (81). H3K4me3 is thought to occur near transcriptional start sites (82), and the peaks cluster around ADAR1 exons 1B, 1A, and 1C.

Using information from publicly available databases as a starting point, we aimed to see how ADAR1 RNA isoforms that begin with exons 1B, 1A, and 1C are expressed in 293T cells, human embryonic stem cells (hESCs), and induced hepatocyte-like cells (iHeps), in the presence and absence of interferon (Figure 2.1.6). Details for culturing of hESCs and iHeps will be described later in this chapter.

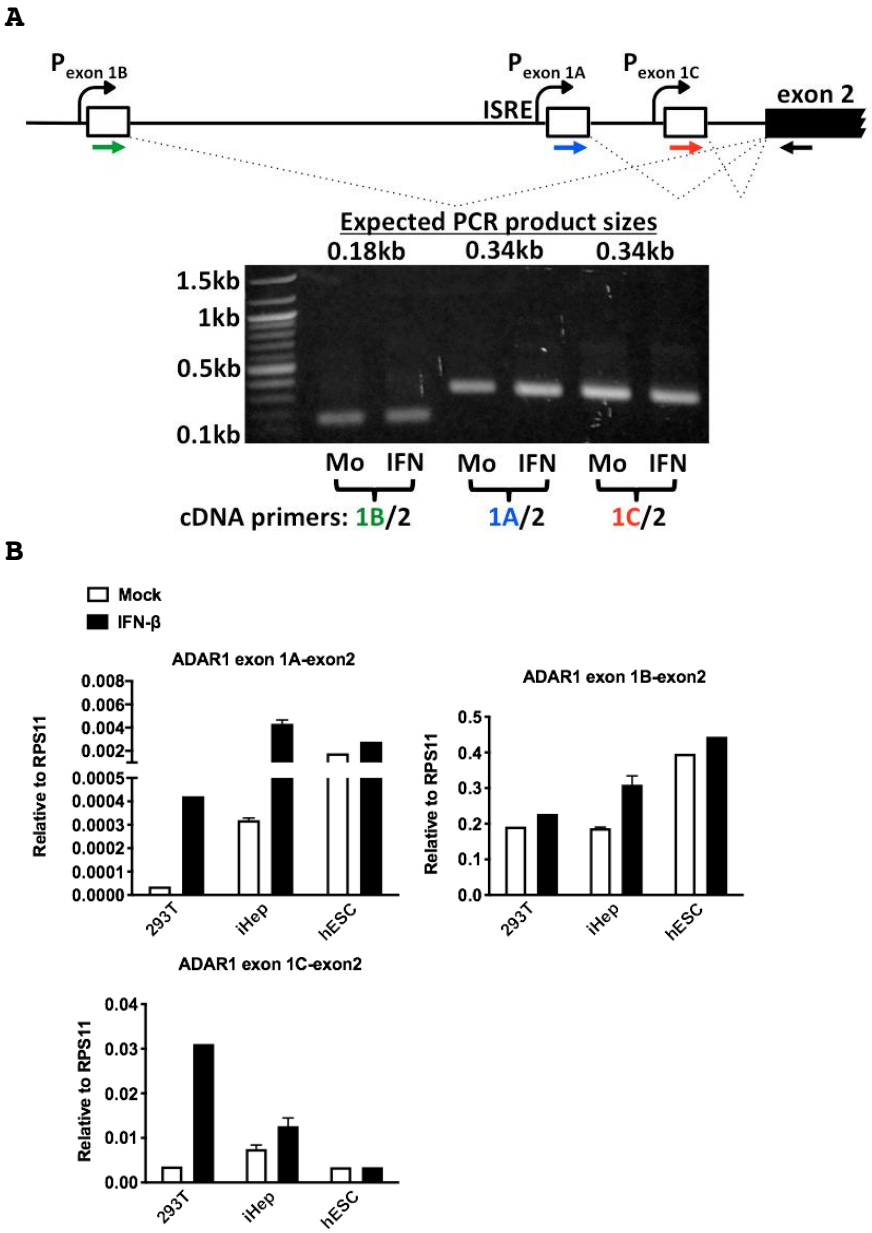


Figure 2.1.6 Expression levels of ADAR1 RNA isoforms

A. Total RNA was extracted from WT 293T cells, iHeps, and hESCs, reverse-transcribed using random hexamers and SuperScript III RT, followed by RNase H treatment, which resulted in single-stranded cDNA. Quantitative PCR (qPCR) reactions were set up in duplicates for each gene (color-coded primer pairs) and sample combination with SYBR master mix. This mix contains DNA polymerase, dNTPs, and SYBR Green dye, which binds to double-stranded DNA and forms a complex that absorbs blue light and emits green light, allowing relative quantification of DNA in each PCR cycle (83). Cycle threshold (Ct) values were defined for each reaction as the PCR cycle number at which the change (second derivative) in the slope (first derivative) for the

plotted fluorescence intensities goes from positive to negative. Ct values were averaged for each gene/sample combination in duplicate, and the differences in Ct values between the genes of interest and housekeeping gene (ribosomal protein S11) were calculated, and plotted as $(1/2)^{\Delta ct}$ to show expression relative to RPS11 (approximate copies per RPS11). A 1% agarose gel shows the qPCR products with their expected sizes.
B. The exon 1A variant showed increased expression following interferon treatment in the 293T cells and iHeps but less so for the hESCs, which maintained high expression of the exon 1A variant even at baseline, consistent with the finding that ISGs are constitutively expressed in pluripotent stem cells (84). The exon 1B variant was relatively consistent in expression with or without interferon treatment in 293T cells and hESCs, but slightly upregulated in iHeps with interferon treatment. The exon 1C variant was surprisingly upregulated with interferon treatment in the 293T cells and iHeps. The primer pairs used in this experiment are listed in appendix 1.

Given the availability of total RNA-sequencing data previously published for WT 293T cells, another method to analyze gene expression in addition to qPCR analysis of individual isoforms, was to look for reads that aligned to exons 1B, 1A, and 1C. Exon 1C posed a challenge, however, because it is a repetitive element of the AluSc subfamily within the class of short interspersed nuclear elements (SINEs), and the presence of repetitive motifs and homopolymers creates a challenge for alignment and base-calling fidelity in sequencing machines. In one 293T dataset, many reads aligned for exons 1B and 1A, but less aligned for exon 1C. In the same dataset, the number of aligned reads also decreased in areas with Alu repeats, such as in the 3′ UTR of a reference gene (Figure 2.1.7). Several reasons could account for this observation of decreased read depth: 1. repetitive elements in RNA form secondary structures that resist the processive activity of reverse transcriptase, thus resulting in decreased formation of cDNA corresponding to repetitive elements; 2. alignment programs have difficulty assigning reads with repetitive sequences to a specific locus; and 3. lowered accuracy of base-calling tends to happen downstream of repetitive sequences, and this decreases the chance of finding a concordant alignment for the read.

In the original experiment that produced the data shown in figure 2.1.7, total RNA was extracted from WT 293T cells and prepared using the Illumina stranded RNA-sequencing kit, following manufacture guidelines with two modifications to improve alignment of repetitive elements: 1. RNA fragmentation time was shortened to create larger fragment sizes; 2. 60% of the suggested volume of AMPure XP magnetic beads was used for

clean-up after PCR, to enrich for larger fragments of 400 bases (insert size around 300 bases) or larger. Of note, stranded library kits allow creating libraries that retain the sense/antisense information contained in each read. This is achieved by the use of dUTP in second-strand cDNA synthesis prior to ligation of adapters. Adapters pairs are oriented in a way to ensure the P5 (read 1) and P7 (read 2 in paired-end sequencing) primed reads correspond to sense and antisense sequences consistently for each fragment. Future rounds of PCR result in amplification of the first-strand antisense cDNA, which has a consistent P5/P7 orientation; second-strand cDNA, which has the opposite P5/P7 orientation, remains unamplified because it contains PCR-quenching uridine.

The consistent orientation of P5 and P7 adapters around PCR-amplified fragments allow retention of strand information. Generally, read 1 contains the exact sense strand of the original mRNA, so the P5 sequence would typically be ligated on the 3′ end of the first-strand antisense cDNA and also the 3′ end of the uridine-quenched second-strand sense cDNA. The P7 sequence would be ligated onto the 5′ ends of the first and second strand cDNA, and because the second strand will not amplify, the P7 primed read 2 should contain the sequence of the antisense strand.

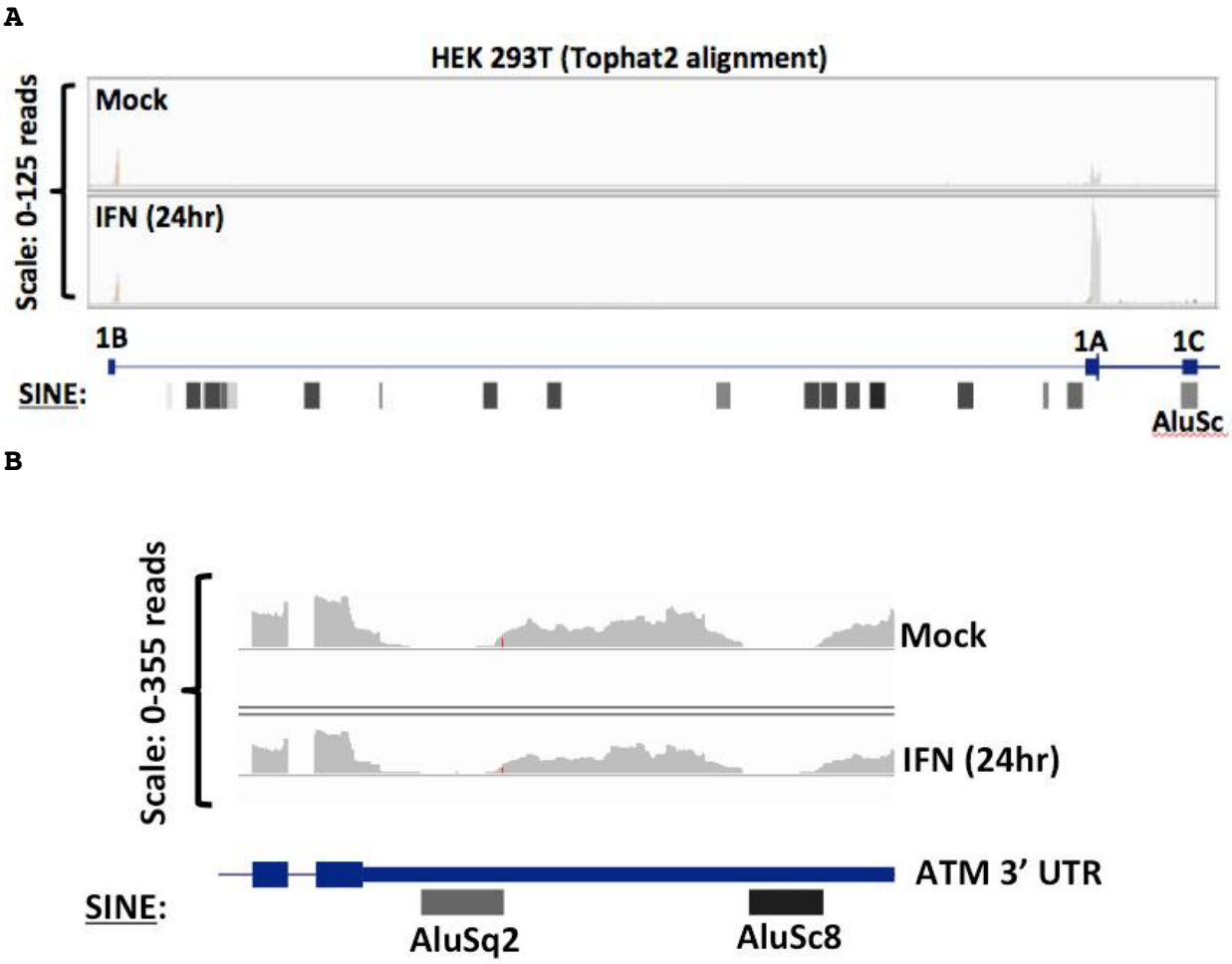


Figure 2.1.7 ADAR1 exon 1C is an Alu element
A. Aligned BAM files created with TopHat2 were indexed using SAMtools for compatibility with Integrative Genomics Viewer (IGV), which was used to generate the image tracks. Peaks corresponding to RNA-sequencing reads can be observed for the regions where exons 1B and 1A are located. Of note, because the scale is 0-125 reads, the reads that aligned to introns, although present (because total RNA was used to make the libraries), are present in low abundance relative to reads that mapped to exons, probably because spliced mRNAs are present in larger quantities than pre-mRNAs, and removed introns are degraded by various cellular mechanisms (85, 86). Without quantitative analysis of ADAR1 RNA isoforms, such as normalization of fragments per kilobase million (FPKM) or transcripts per kilobase million (TKM), it can be difficult to compare directly across the different isoforms and samples, but the trend suggests that 1B has higher basal expression levels than 1A, and that 1A is induced following interferon treatment. This is consistent with both qPCR data shown in figure 2.1.6 and

previous studies showing Northern blots and functional analysis of sequences upstream of these exons (35, 48).
B. Few reads aligned to the genomic locus corresponding to exon 1C, even with interferon treatment. Since exon 1C shares homology with Alu elements, it can be difficult to draw conclusions about 1C expression levels from this dataset. The color densities of boxes in the SINE lanes, generated using RepeatMasker, indicate the relative percentage of homology to a consensus sequence, with darker boxes corresponding to higher homology, meaning less base mismatches and insertions/deletions when aligned to the consensus sequence. The alignment results of other genomic regions with Alu repeats were also examined, with an example shown. The 3' UTR of the Ataxia-telangiectasia mutated (ATM) kinase gene has several repetitive elements, including Alu elements, and aligned read counts are decreased in those regions compared to neighboring regions.

Collectively, the qPCR and total RNA dataset provide evidence for expression of ADAR1 exon 1B, 1A, and 1C variants in three different cell types, with the 1C variant being difficult to see in the TopHat2 alignment due to its repetitive sequences. The qPCR and RNA-sequencing results are consistent with presence of an interferon-inducible promoter upstream of the exon 1A sequence. Unexpectedly, we also observed an increase in exon 1C-containing isoforms upon interferon treatment. This observation prompted further experiments to define the ADAR1 splice junctions and look for additional splicing events not previously described.

2.2 ADAR1 splice variants

There is expression of exon 1B, 1C, and 1A variants in 293T cells as assayed by SYBR Green I. Next, the qPCR products were resolved on agarose gel, extracted, and submitted for Sanger sequencing to define the splice junctions (Figure 2.2.1).

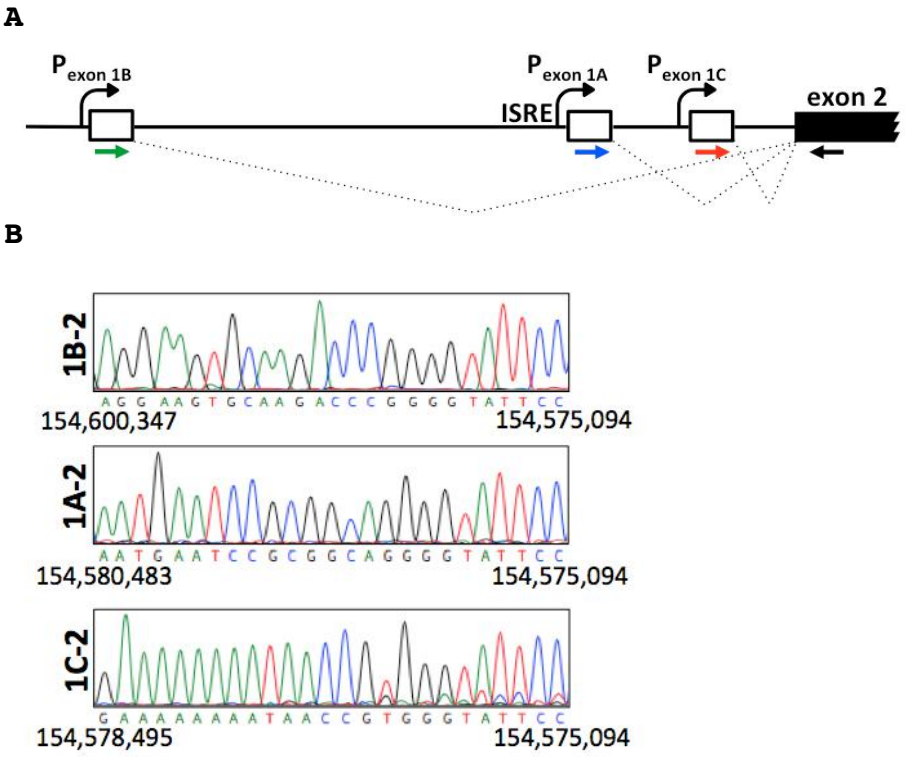


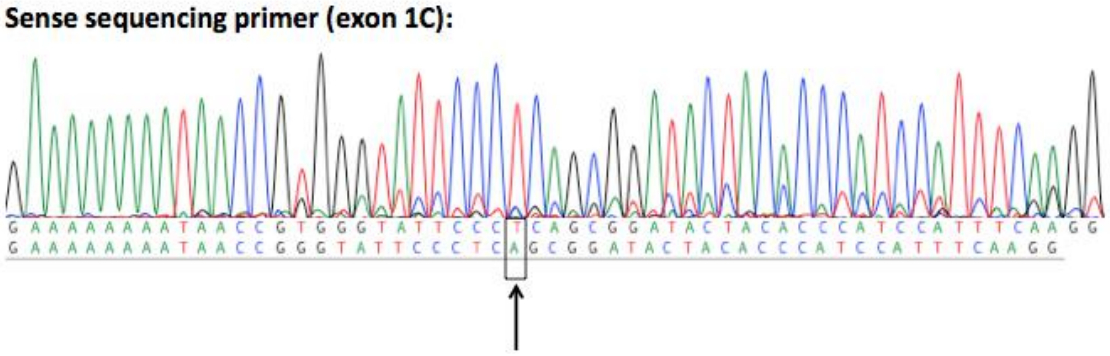
Figure 2.2.1 Splice junctions of ADAR1 RNA isoforms
A. The PCR products shown in the agarose gel in figure 2.1.6 were extracted using the Qiagen gel extraction kit, following manufacturer guidelines. Forward primers used to generate the PCR products (shown in the schematic of the ADAR1 gene locus as color-coded arrows) were mixed with the corresponding gel-extracted products and submitted for Sanger sequencing.
B. Sequence chromatograms show the splice junctions between exons 1B, 1A, and 1C with exon 2 and are consistent with the annotated exon-intron boundaries. During the chain-termination sequencing reactions using ddNTPs labeled with fluorescent tags, fragments of different sizes are generated that are then separated by micro-electrophoresis. Base calls are made in the order of smallest to largest fragments, and for each position, if there are multiple bases, the area under the curves for the multiple bases correspond approximately to the ratio of the fragments possessing the different bases. The base called by the sequencer is based on the nucleotide peak with the highest area under the curve, but manual examination of individual positions

can reveal sites with multiple peaks, as is seen for the exon 1C-exon 2 chromatogram.

Sanger DNA sequencing chromatograms reveal splice junctions that line up exactly with the RefSeq annotated splice junctions for the exon 1B-exon 2 and exon 1A-exon2 spliced isoforms (exon genomic sequences shown in uppercase and introns in lowercase):
RefSeq exon 1B/intron 1B: …AAGTGCAAGACCCGgtaagagcc…
RefSeq exon 1A/intron 1A: …TGAATCCGCGGCAGgtaagccgg…
Refseq intron/exon 2: …ctgcagGGGTATTCCCTCAGCGGATAC…
RefSeq exon 1C/intron 1C: …GAAAAAAAATAACCgtgtgagta…
The canonical splice donor site is GT (GU in mRNA) at the beginning of the intron start site, often also with the exon preceding the splice donor site ending in a guanosine, and the splice acceptor site is AG at the end of the intron just before the junction with the downstream exon (87–90). Of curiosity is the exon 1C-exon 2 junction as seen in the Sanger chromatograms, in which there is an extra two bases (GT) between the RefSeq annotated end of exon 1C and start of exon 2. In addition, there appears to be a mix of nucleotides for each position starting at the second base of the extra GT sequence, which also corresponds to the first potential splice donor site (gt) in the intron immediately following exon 1C. Notably, there is a second putative splice donor site right after the first one (gtgt), and one can imagine this could lead to mixed splice donor activity (Figure 2.2.2). If both splice donor sites are active and join with the common splice acceptor site upstream of exon 2 (ag), then the possible splice isoforms based on usage of the second and first splice donor (gt) sites are, respectively (exon 2 sequence is underlined):

1. GAAAAAAAATAACCGTGGGTATTCCCTCAGCGGATACTACACCCATCCATTTCAAGG
2. GAAAAAAAATAACCGGGTATTCCCTCAGCGGATACTACACCCATCCATTTCAAGG

A



B

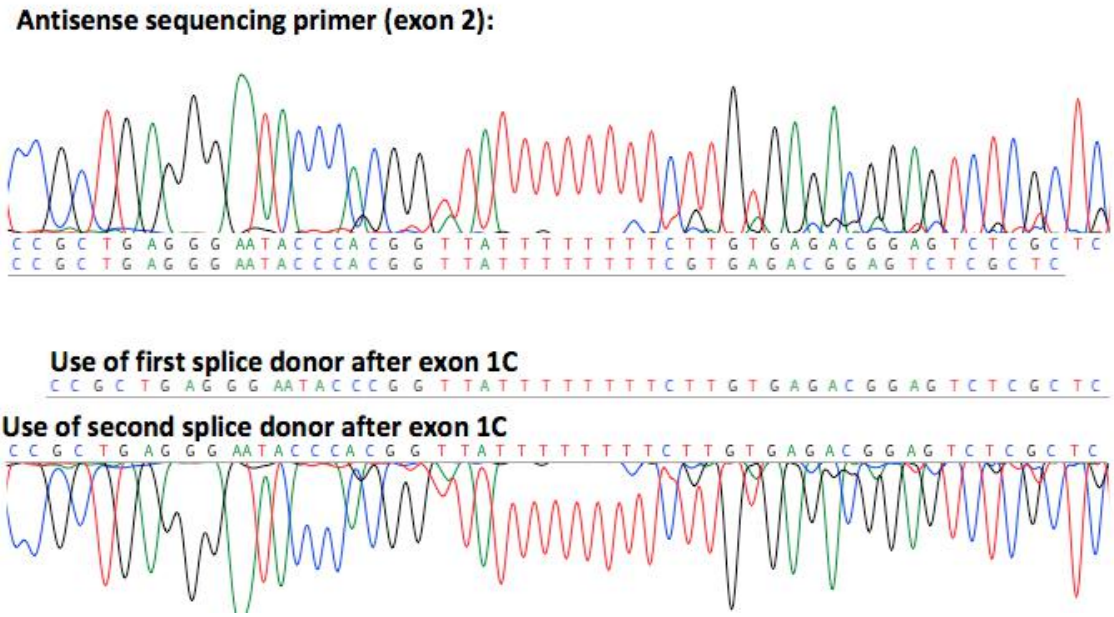


Figure 2.2.2 Splice donors downstream of exon 1C

A. The chromatogram shown in Figure 2.2.1 for the exon 1C-exon 2 junction is shown here extended and with nucleotide annotations at each position for the two possible splice variants depending on which splice donor site downstream of exon 1C is used. Usage of the second splice donor site would result in the sequence shown closest to the chromatogram, whereas usage of the first splice donor site would result in the shorter sequence shown underneath. Following the chromatogram traces for each position, and taking into account both the highest and lower peaks, both sequences appear to be reasonable interpretations based on the chromatogram shown, except for the single boxed position that shows a smaller peak (C) inconsistent with the second sequence shown below at that position (A). Initially, we considered the possibility that this mismatch could result from a variant at that position in the cDNA either due to a PCR error (rare event, because KOD high-fidelity DNA polymerase was used in all PCR reactions) or single-nucleotide polymorphism (SNP) in the

genomic DNA (A-to-C), but this proposed A-to-C SNP was not observed at that position in the exon 1B and exon 1A containing mRNA variants. Even if the SNP was heterozygous, and the WT chromosome(s) or part of it happened to be silenced, this should still be uniform across the three qPCR products, which were produced from cDNA made from bulk RNA coming from the same cell population. Because of this inconsistency, we looked at other possibilities, starting with simply examining how the chromatogram would look by sequencing with the reverse primer, which would avoid the exon 1C homopolymer tail with regard to sequencing of exon 2.
B. There is no SNP (T-to-G) in question seen in the antisense-primed sequence, and the presence of the polyT stretch also created an alternate sequence with a 2-base displacement (deletion) starting downstream of the homopolymer, much like the polyA stretch, in which case the displacement seemingly corresponded to usage of the first splice donor site. Taking into consideration chromatograms generated from both sense and antisense sequencing primers, we conclude that this displacement is an artifact of the sequencing, and that the actual splice site between exon 1C and exon 2 is two bases downstream of what is annotated in RefSeq. In other words, the second splice donor site downstream of exon 1C is what appears to be involved in the splicing of exon 1C with exon 2, at least in 293T cells.

The presence of an 8-base adenosine homopolymer upstream of the splice junction between exon 1C and exon 2 seems to inhibit Sanger base-calling fidelity downstream, regardless of whether the sequencing was done in a sense or antisense direction. In correspondence with technical support at Macrogen, we learned that presence of homopolymer stretches, and particularly ones of the polyA/polyT type found so often at the tail ends of Alu repeat elements (91), results in the reporting of multiple possible DNA traces downstream, and often these traces are displaced by a few nucleotides. One study looking at the types of bias found in next-generation sequencing methods reported increased rates of indels in homopolymer regions as a function of polymer length, as compared to individual base mismatch rates in the same regions (Figure 2.2.3). Other informatics studies have also reported the need for error-correction protocols when analyzing next-generation sequencing data from regions with repeat elements and homopolymer motifs (92, 93). Although Sanger sequencing is different from next-generation sequencing both in terms of base-calling chemistry and downstream analysis methods, the bias toward indels seen in a few NGS platforms seems to apply also to Sanger sequencing, but in this case more downstream of polyA/polyT regions rather than within the homopolymers themselves.

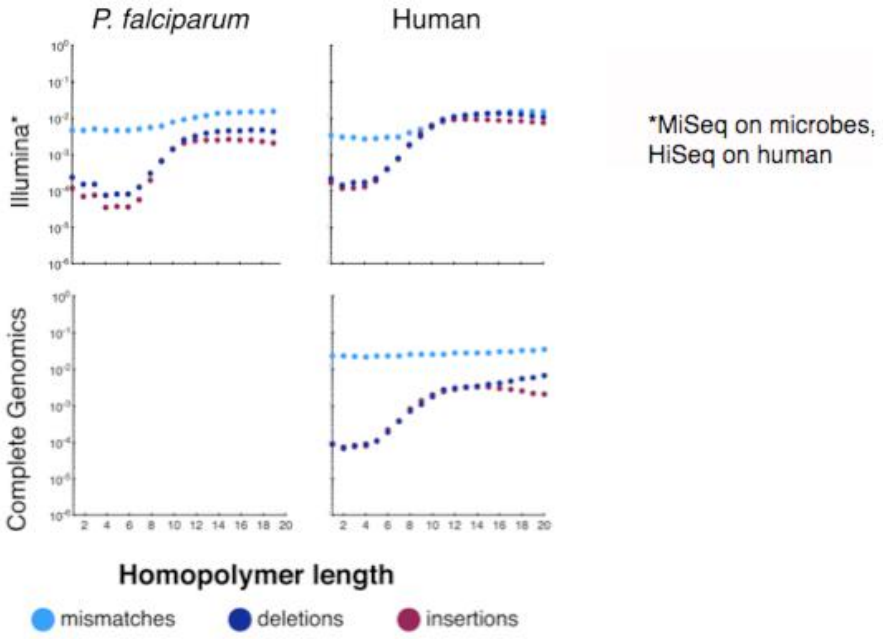
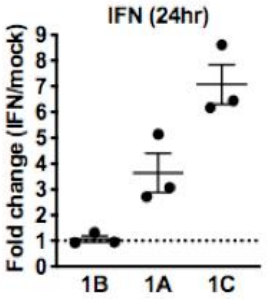


Figure 2.2.3 Next generation sequencing bias (94)
In this study, data from several NGS platforms: Illumina sequencing-by-synthesis, ion-torrent sequencing by release of hydrogen ions (not shown), Pacific Biosciences single-molecule sequencing (not shown), and Complete Genomics sequencing by circular-replication, are shown in terms of the rate of mismatches, deletions, and insertions (compared to reference genome builds) within homopolymers as a function of homopolymer length. For human sequencing data from HiSeq, the rate of mismatches largely remains the same, and increases slightly as the length of homopolymers increases past 8 bases. By contrast, the rate of indels, whether "real" or artifactual, begins to increase as the length of homopolymers increases past 4 bases, and the increase is up to two log-fold.

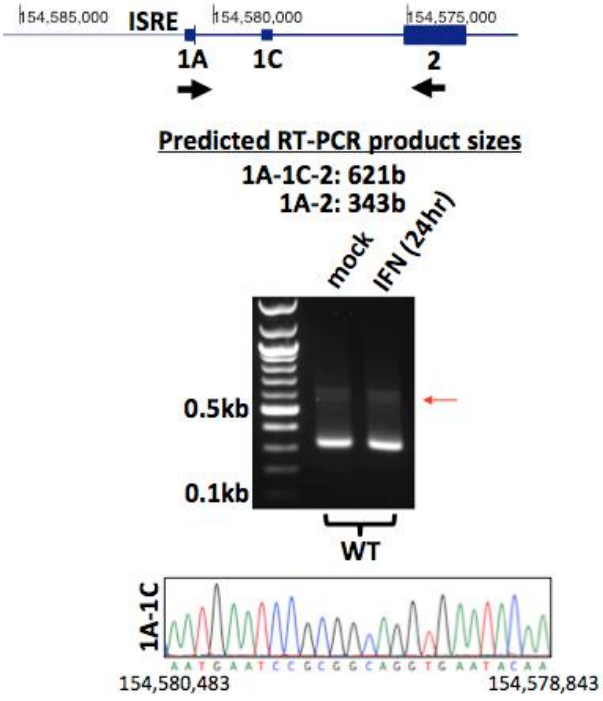
Having formally examined the splice junctions for ADAR1 mRNA containing exons 1B, 1A, and 1C, there is one more curiosity from figure 2.1.6 that is worth mentioning: the observation that exon 1C-containing variants are also increased following interferon treatment. The experiment was repeated in 293T cells, comparing the fold change across the different splice isoforms, with a focus on the band(s) generated by 1A-specific forward and exon 2-specific reverse primers. We examined the hypothesis that there is alternative splicing of 1A to 1C within 1A pre-mRNAs. Specifically, we are checking if exon 1C can function as both a retained exon and excised intron within the pre-mRNA that is formed from transcriptional

initiation upstream of exon 1A (Figure 2.2.4). Of note, examination of the sequence upstream of exon 1C does not reveal a KCS-like or ISRE-like element (35, 39), although this by itself does not rule out the possibility that the genomic DNA there could be responsive to interferon treatment.

A



B



C

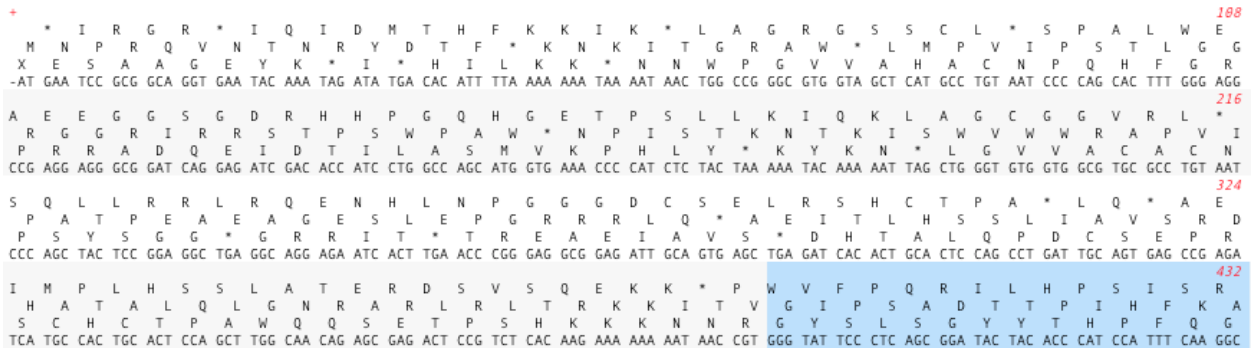


Figure 2.2.4 Novel 1A-1C splice junction

A. Repeating the qPCR experiment in figure 2.1.6 using the same primers yielded the fold change results shown above in panel A, with 24 hours of interferon treatment. Although the fold-change in transcripts containing exon 1C is even more than that for transcripts containing exon 1A, this could be because the basal levels of 1C are lower than 1A. Without interferon treatment, the primer pair that binds to 1A-containing transcripts reach the inflection point for change in SYBR Green fluorescence earlier than the pair for 1C-containing transcripts, but differences in primer binding efficiencies make comparisons between primer pairs difficult.

B. Examination of the PCR products resulting from the primer pair shown in the schematic (1A-2) reveals two bands upon close inspection: the expected 343-base band for the exon 1A-exon 2 splice isoform, and another band a bit above the 600-base benchmark, highlighted by the red arrow. This band is thought to be the product of exon 1A spliced to exon 1C and then to exon 2. Using the RefSeq annotated splice junctions, a band of 619-bases is expected for a 1A-1C-2 splice isoform, but based on figure 2.2.2, we can modify the expected size to 621-bases. The larger band was gel extracted, and a nested PCR was performed using a forward primer that binds within exon 1A (5'-AAGCGAAATTGAACCGGAGC-3') and a reverse primer that binds within exon 1C (5'-CAGGATGGTGTCGATCTCCTG-3'). The nested PCR product was sequenced using both forward and reverse primers, revealing the 1A-1C splice junction as shown. Of note, the junction is 77 bases upstream of the RefSeq annotated start of exon 1C. The consensus splice acceptor site is present in the intron sequence immediately upstream of the extended exon 1C sequence, and the consensus splice donor site immediately downstream of exon 1A has already been referenced earlier (exon sequences capitalized and introns in lowercase):
…AATGAATCCGCGGCAGgtaagccgg…tatccccagGTGAATACAA…

C. Naturally, one might wonder if this splice isoform could produce an ADAR1 isoform larger than p150, but examination of the open-reading frame for the 1A-methionine in this 1A-1C-2 isoform shows only the potential to translate small peptides upstream of the p110-methionine in exon 2, because of premature stop codons. Shown is the assembled 1A-1C-2 splice isoform with sense-strand translations for the three possible reading frames. The highlighted portion of the sequence shows the start of exon 2, and the amino acid sequence on the bottom closest to the nucleotide sequence shows the p150 open-reading frame (GYSLSGYYTHPFQG…). In this frame, there is no methionine between the most distal stop codon before exon 2 and the glycine codon (GGG) that begins exon 2. Here, we are considering only the canonical translational start codon (ATG), although there is the

formal possibility that translation initiation can start somewhere other than at AUG codons (95).

With the identification of this 1A-1C splice isoform, the observation that exon 1C containing mRNA isoforms are upregulated following interferon treatment can be attributed to increased transcription at the exon 1A start site coupled to alternative splicing. The 1A pre-mRNA molecules that contain exon 1C will be increased in abundance with interferon treatment, and the alternative splicing of 1A to 1C will ensure that not all exon 1C sequences are excised out and degraded as introns. This retention of exon 1C in between exon 1A and exon 2 is one explanation for why the 1C-2 primer pair gives products that reach amplification threshold at an earlier PCR cycle for the interferon treated samples compared to untreated samples.

Although this 1A-1C-2 splice isoform does not encode p150 or an isoform larger than p150, it does have the potential to give rise to p110, suggesting that p110 is also interferon-inducible by means of alternatively spliced pre-mRNAs that originate from an interferon-responsive promoter. More data suggesting that p110 is induced following interferon treatment will be presented in the next section, in which exon 1B, 1A, and 1C deletions will be made using CRIPSR to understand the contributions of the various splice isoforms to ADAR1 protein levels.

To conclude this section, the splicing results presented above are corroborated with publicly available single-molecule sequencing data, done using a nanopore sequencer designed by Oxford Nanopore Technologies. The human GM12878 B-cell line was sequenced and assembled using the minimap2 aligner, taking GRCh38 as a reference genome (Figure 2.2.5). The advantage of this sequencing method compared to Illumina methods is the extended length of contiguous reads, which typically go over 100kb for DNA samples and 21kb for RNA samples, allowing increased confidence when determining presence of spliced isoforms (96, 97). Fragmentation-based methods rely more on read depth, overlapping reads, and informatics methods for assembling reads and inferring splice isoforms, although the assembly is complicated by the presence of repetitive sequences, many of which can be longer than the average read length at which sequencing-by-synthesis platforms can offer high base-calling confidence.

In single-molecule Nanopore RNA sequencing, the poly(A) fraction can be captured using a poly(T)-bead based system. Reverse transcription of the mRNA is not necessary because the first-strand cDNA is not sequenced, but according to guidelines from the company, generating a cDNA/mRNA hybrid strand seems to improve throughput during the sequencing reaction. The poly(T)

used to capture RNA for sequencing includes additional sequences that enable sequencing tethers, which are single-stranded overhangs attached to molecular tethers, to anneal onto the captured poly(A) RNA, which then passes through modified transmembrane pore proteins or synthetic alloy pores, both in the single nanometer range in terms of diameter. Both the nanopore and captured RNA are immersed in a solution with ions, allowing detection of current when electricity is applied. As the RNA molecule passes through the pore (which is about 10nm or more in length) in response to an electric gradient set up in the solution, the presence of different bases in the RNA will briefly alter the structure for each sequence-nanopore combination, and these changes in structure affect the electric current around the nanopore and thus provide information about what is passing through.

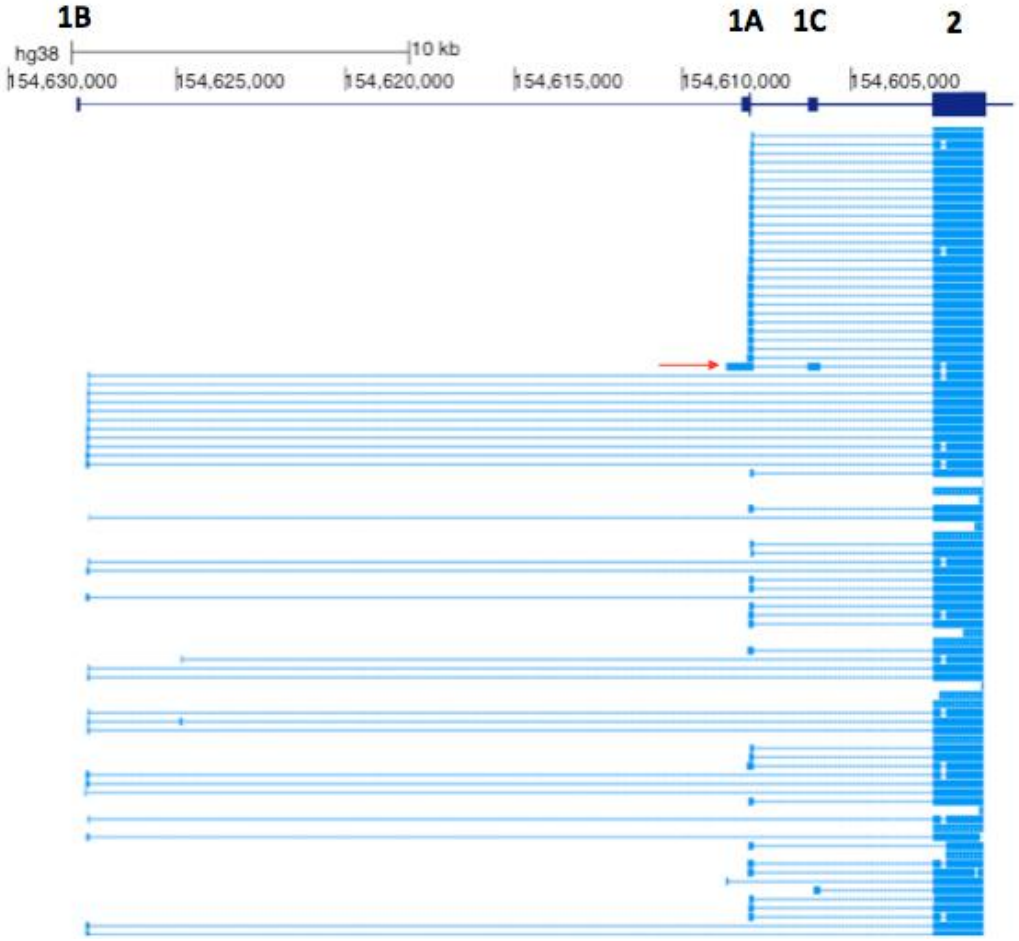


Figure 2.2.5 Nanopore dataset: ADAR1 RNA isoforms (97)
Total RNA was extracted from the human GM12878 lymphoblastoid cell line, and the poly(A) fraction was isolated for sequencing. Shown above is the nanopore sequencing data corresponding to ADAR1, as mapped to the hg38 human-genome build on the UCSC Genome Browser. The alignments show presence of 1B-2, 1A-2, 1C-

2, and 1A-1C-2 (red arrow) splice isoforms, along with other potential isoforms, including some that seem to start within exon 2. Of interest, the putative 5' ends of the RNA isoforms seem to vary, recapitulating what was observed in the 5' RACE studies (31, 35, 41). Whether these correspond to variations in the transcriptional start sites because of RNA polymerase II slippage, or whether they represent experimental variation during sample prep and sequencing is unclear, although both could be possibilities. Of note, the 1A-1C-2 splice isoform corresponds to an extended exon 1C as compared to the RefSeq annotation, and this is consistent with our Sanger sequencing data. Because the tracks here correspond to direct sequencing of RNA molecules captured during poly(A) selection, as compared to amplification of cDNA fragments, the numbers of tracks corresponding to each isoform can be taken as an approximate measure of the relative amounts of these isoforms in the original RNA sample. This figure shows only one snapshot of all the tracks that aligned to the ADAR1 locus, but even within this snapshot, it appears that exon 1C containing variants are present in much lower quantities than splice variants that contain exons 1B and 1A.

2.3 Expression of ADAR1 p110 and p150

With the formal examination of ADAR1 splice junctions (for 293T cells) in place, we next aimed to investigate the contributions of RNA variants containing exons 1B, 1A, and 1C, to ADAR1 protein levels. By adopting a CRISPR-based knockout approach that aims to delete exon sequences along with upstream promoter sequences, we were able to examine protein phenotypes of 293T cells when all three exons or combinations of the three exons and their promoters were removed from the genomic DNA.

The method used to generate the knockouts is transient transfection of Cas9/sgRNA expressing plasmids, much like the method used to make the GFP-KI. In this case, the plasmid used (PX458) encodes a GFP protein that correlates with expression of the Cas9, because the two proteins are linked via a T2A self-cleaving peptide. The approach is to use two guides, one that binds upstream of the exon and promoter sequences, if known, and one that binds downstream of the exon. The two break sites would result in removal of the insert containing the exon and promoter, and NHEJ repair mechanisms could then join the two break sites, resulting in the removed piece of chromosomal DNA being degraded.

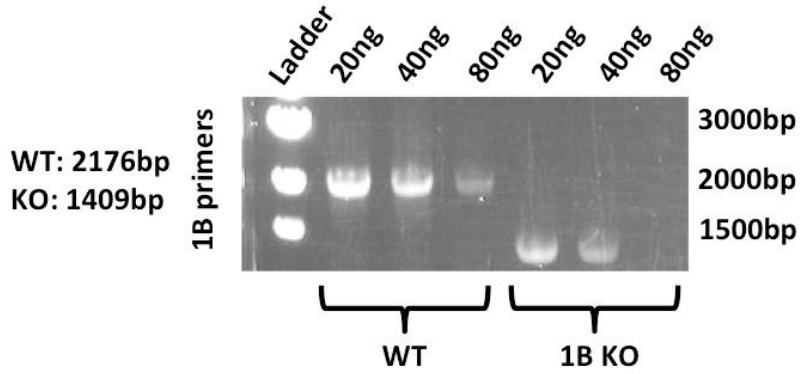
For guides designed to delete exon 1B and its promoter, the decision to make the upstream binding site about 600 bases upstream of the annotated transcriptional start site is based on prior studies examining promoter activity of genomic DNA sequences upstream of exon 1B (41). Single-cell cloning of GFP-

positive cells was then done on the PX458 bulk-transfected cells, and clones were screened by genomic DNA PCR. Finally, the protein phenotypes of selected clones were analyzed by immunoblot (Figure 2.3.1). Guide RNA sequences are shown in appendix 2.

To prepare for transfection of plasmids encoding Cas9 and exon 1B KO guide RNA, WT 293T cells were plated into PLL-coated wells at a density of 30,000 cells/cm^2. Lipofectamine 2000 lipid-based transfection was performed following manufacturer guidelines, including amount of plasmid DNA and volume of transfection reagents. 48 hours after transfection, with no need for media change, single-cell sorting was done in 96-well flat-bottomed plates containing 50% conditioned 293T media and 50% fresh 293T. To prevent evaporation from inner wells, PBS was added to the outside ring of wells, in which there will be no addition of single-cell clones during sorting. Three plates were prepared per pair of guides being tested.

Prior to sorting, cells were dissociated gently in FACS media: 0.5mL PBS/20mM HEPES/0.1% BSA, with DAPI added at 2ng per 1 million cells for live/dead cell gating. Following sorting, all wells were replenished with fresh 293T media at 24hr and 72hr time-points. After 2 weeks of growth as single-cell clones, the cells were transferred to 9.6cm^2 plates for expansion. Following growth to 80-90% confluency, each well, which corresponds to one original single-cell clone, was treated with Accutase gentle dissociation media, and the lifted cells were split in half, one half for extraction of genomic DNA and the other half for resuspension in cryopreservation media (40% DMEM, 40% FBS, and 10% DMSO). Freezing was done by incubating vials in isopropanol-infused Mr. Frosty containers, which are designed to achieve a temperature change of about -1°C per minute. After 24 hours of cooling, cell clones can be transferred to liquid nitrogen, particularly if screening of genomic DNA is delayed. This system of storing and screening single-cell clones allows direct correlation of genotype to phenotype.

A



B

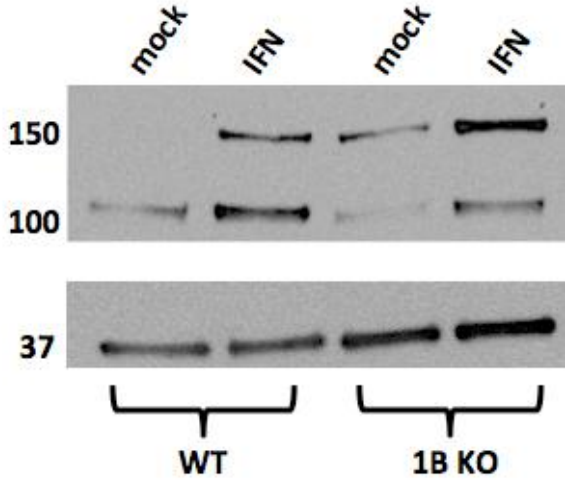


Figure 2.3.1 Exon 1B knockout cells

A. Genomic DNA was extracted from sorted single-cell clones using the Qiagen DNeasy Blood & Tissue kit, and PCR reactions were set up using primers that bind more than 500 bases away from the predicted double-stranded break site. The reason for this primer design is to minimize the chance of false negatives, which can be caused by deletion of the primer-binding site during NHEJ DNA repair but retention of some or all of the exon sequence, particularly if only one of the two guide RNA target sites are efficiently cut. 20 clones were screened, and selected clones were further screened using another set of primers, 5'-GGAAGTTTCCTTCTCTTTTCCCC-3' and 5'-CTCCGCTAATTGCATACTTGGG-3', and testing various input amounts of genomic DNA. The 1B-KO clone shown in panel A is the clone that was selected for the experiment shown in panel B, as the PCR pattern suggests a homozygous deletion of exon 1B.

B. The corresponding frozen cell vial was thawed by infusion in a 37°C water bath, and just prior to the disappearance of ice crystals, the contents of the vial were transferred to a T25 flask pre-filled with warmed 293T media. Cells were later plated at a density of 30,000 cells/cm^2 the day before interferon

treatment. Interferon-beta was added to regular 293T media at a concentration of 1nM, and this interferon-spiked media was added to cells for 24 hours. Cells were then prepared for ADAR1 immunoblot as described previously. Of note, the p150 band in the 1B-KO cells is present slightly even in the absence of interferon, and this band, along with the p110 band, both increased in intensity following interferon treatment.

Of note, deletion of exon 1B did not result in removal of p110, suggesting that 1B-containing mRNA isoforms, which have the p110-ATG as the first start codon in the ADAR1 reading frame, are not the only mRNA variants contributing to p110 protein levels. Furthermore, treating 1B KO cells with interferon resulted in upregulation of p110, suggesting the presence of mRNA with both p110-coding potential and the ability to be induced by interferon. The exon 1A-1C-2 splice variant is a possibility. Thus, we aimed to delete exon 1C, and to investigate how this deletion would impact p110 protein levels (Figure 2.3.2). As the promoter for exon 1C is not well annotated, the binding site for the upstream guide RNA was chosen to be located more than 1kb upstream of the annotated start of exon 1C. This distance was chosen to maximize deletion of upstream regulatory DNA elements that could promote transcriptional initiation without being so far upstream of exon 1C that there is a risk of disturbing exon 1A.

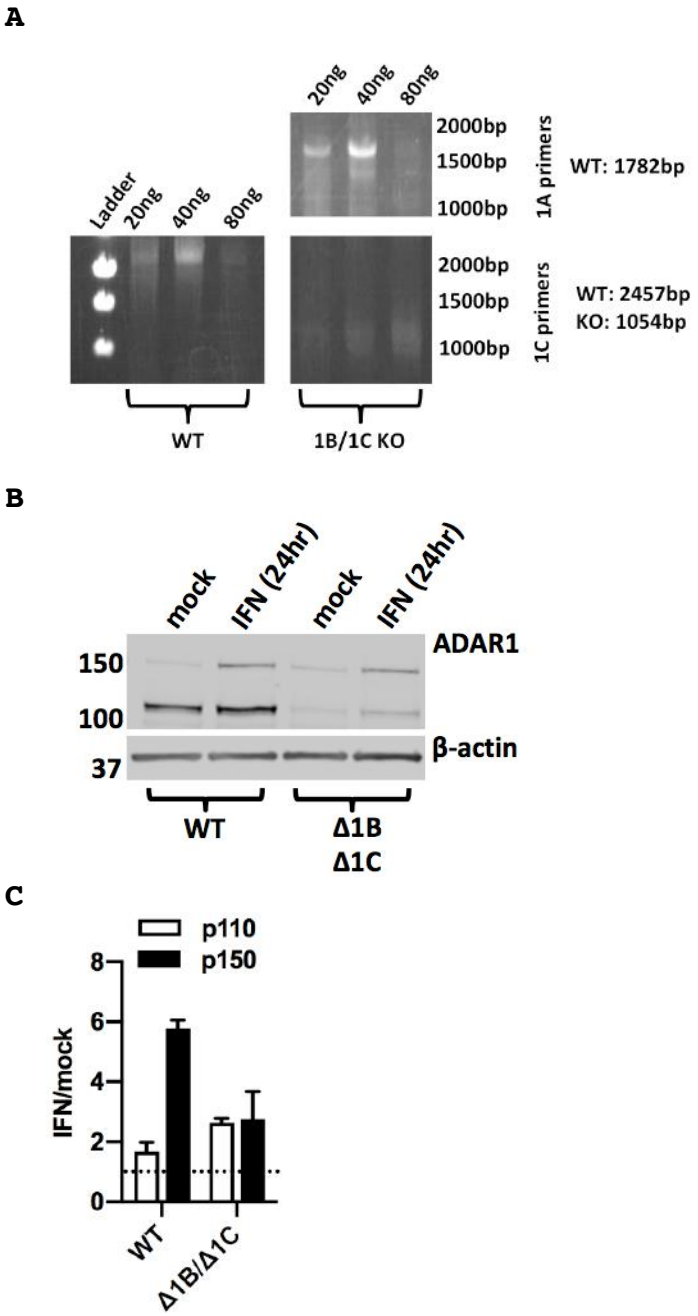


Figure 2.3.2 Exon 1B/1C double-knockout cells

A. The exon 1B knockout clone shown in figure 2.3.1 was transfected with pairs of guides to delete exon 1C. The method for designing guides, transfection, single-cell cloning, and genomic DNA screening is identical to that used to make the exon 1B KO. The 1B/1C-KO clones were screened by genomic DNA PCR using a set of primers, 5'-GGAATTTGTGCGTCTTGTGAGTGTG-3' and 5'-TCCCACCTTCTTCAGCCCTTGG-3', that bind more than 500 bases outside of the 1C-KO guide RNA binding sites, and PCR products corresponding to a selected clone is shown, generated by a second set of screening primers for confirmation,

5′-GTGTGTCGTCTTGCCAAGCAGC-3′ and 5′-CAGCCCTTGGGGTTTCCTCTCC-3′, and different input amounts of genomic DNA were tested, as was the case for exon 1B screening. Furthermore, the genomic DNA from this clone was screened using primers that bind outside of exon 1A to examine whether the exon 1A locus, which is about 2kb upstream of exon 1C, was affected by the DNA repair process. The selected clone had exon 1C removed and retained an intact exon 1A locus, and was used for downstream analysis.
B. Examination of the ADAR1 protein phenotype in the presence and absence of interferon revealed a pattern similar to that of exon 1B KO cells.
C. Quantification, using ImageJ, of the relative p150 and p110 band intensities between treated and untreated samples is shown here. Both WT and exon 1B and exon 1C double-deletion (Δ1B/Δ1C) cell clones showed positive (greater than 1) fold induction of p110 and p150 following interferon treatment.

Of note, removing exons 1B and 1C along with upstream DNA sequences did not abolish p110 protein levels, suggesting that additional mRNA variant(s) other than 1B and 1A-1C-2, potentially the 1A-2 isoform, or maybe a variant originating from an unidentified promoter, are contributing to p110 protein levels. Furthermore, quantification of the relative band intensities between interferon-treated and untreated 1B/1C-KO cells reveals upregulation of both p150 and p110 with exposure to interferon. Here, upregulation of p150 protein can be traced back to increased transcriptional activity at the promoter regions upstream of exon 1A, and upregulation of p110 protein suggests that the mRNA(s) in question with p110-coding potential is also upregulated following interferon treatment. Several possibilities were considered, including the hypothesis that 1A-containing mRNAs could also give rise to p110 through translational initiation at the downstream methionine at codon 296. To test if removing exon 1A would remove p110, the same CRISPR knockout system described above to remove exons 1B and 1C was used to remove exon 1A and its promoter, which has been formally examined in the literature (35). Deletion of exon 1A along with exons 1B and 1C resulted in clones that lacked detectable p150 and p110 protein levels as well as detectable A-to-I editing at known editing sites (Figure 2.3.3).

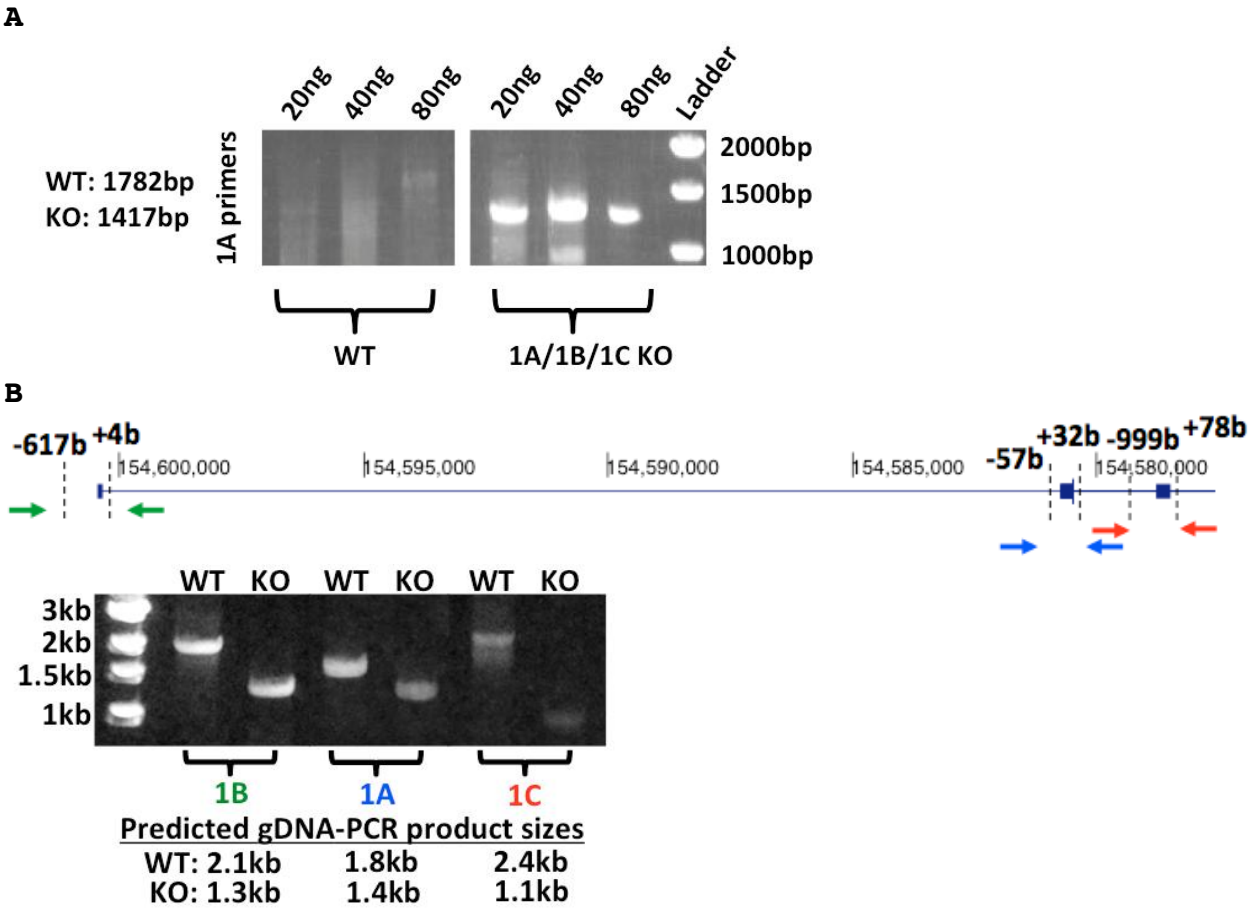


Figure 2.3.3 Exon 1B/1C/1A triple-knockout cells

A. The 1B/1C KO clone shown in figure 2.3.2 was used as a starting point to make a 1B/1C/1A triple KO. Following CRISPR KO single-cell cloning, genomic DNA from 20 clones was screened, and the selected clone shown above was screened again using a separate set of primers, 5'-TTACTAAAGCCTTCAGACCTG-3' and 5'-CTATAAAGTGGTTAACAAGCTTTC-3', and the PCR products are shown on an agarose gel.

B. Shown here is a schematic of part of the human ADAR1 locus on chromosome 1, namely the three alternative exon 1 structures (1B, 1A, and 1C from left to right) that each have distinct promoters, including the type I interferon inducible promoter of

exon 1A. Exon 2 is the common splice acceptor for the three exon 1 structures shown. Exon 1A has a start codon for the ADAR1 p150 isoform, and exon 2 has a start codon for the ADAR1 p110 isoform. The dotted vertical lines together show a summary of the predicted break sites in the DNA locus from the iterative sets of deletions introduced by the Cas/sgRNA combinations. Color-coded arrows show the approximate binding sites of primer pairs used to generate the PCR products shown in the gel. The substrate used for the PCR reactions is genomic DNA from the 1B/1C/1A knockout clone selected for downstream experiments; the bands on the gel represent a second-round confirmation that the clone has homozygous deletions in exons 1B, 1A, and 1C.
C. Shown here is an immunoblot revealing ADAR1 isoforms in WT and triple-deletion (Δ1B/Δ1C/Δ1A) cell clones. By chemiluminescence-level sensitivity, p150 and p110 levels are undetectable. Analysis of a known A-to-I edit site in both clones was done using Sanger sequencing of cDNA PCR products made from the ATM gene 3' UTR region, showing a double-peak (A and G) at that site only in the WT clone. Of note, inosine will appear as guanosine in sequencing reactions because of its base-pairing properties (98).

In summary, CRISPR-Cas9 technology was used to initially delete exons 1B and 1C in 293T cells along with their promoters, which initiate constitutive transcription of mRNA variants that have p110-coding potential. PCR was used to confirm homozygous deletion of the exons along with their promoters at the genomic DNA level. Immunoblotting for ADAR1 with a C-terminal specific antibody (amino acids 1051-1226) showed that while p110 levels were reduced following deletion of exons 1B and 1C, expression of p110 not entirely ablated. Furthermore, p110 was induced following treatment of exon 1B/1C knockout cells with interferon, raising the possibility that the interferon-inducible exon 1A variant could be contributing to p110 protein levels.

When exon 1A was deleted in addition to exons 1B and 1C, p110 was no longer detected at immunoblot sensitivity levels, suggesting that the exon 1A mRNA variant contributes significantly to the remaining p110 levels in Δ1B/Δ1C cells. As expected, p150 was also undetectable in Δ1B/Δ1C/Δ1A cells, and Sanger sequencing of reverse transcription (RT) PCR products from these cells confirmed a lack of A-to-I editing at known ADAR1-editing sites.

More details about p110 expression from exon 1A-containing mRNAs will be presented in Chapter 3. Here, as brief side notes, two separate stories will be presented, one about the function of ADAR1 in neuronal cell biology, and another about the belated evolution of p150 compared to p110.

2.4 Function of ADAR1 in neural progenitor cells

When discussing functions of proteins, one way to start is to see how knockouts affect biology, or how mutations in various domains of the protein affect function. In the case of ADAR1, complete lack of editing-capable protein is embryonic-lethal in mice, but this can be rescued by removal of the cytoplasmic dsRNA sensor MDA5 (melanoma differentiation-associated protein 5). MDA5 is part of the RIG-I (retinoic acid inducible gene I) pathway, one of several innate pattern-recognition pathways that is stimulated upon interaction with various nucleic acid ligands (99—101). Broadly speaking, the MDA5/RIG-I family of receptors, upon encountering dsRNA in the cytoplasm, engages another protein called MAVS (mitochondrial antiviral-signaling protein), and this complex then activates various transcription factors from the IRF (interferon regulatory factors) family and also the well-known signaling complex, NF-κB (nuclear factor kappa B). This leads to transcription of various interferon genes, and subsequent synthesis and release of interferon isoforms that have their own autocrine and paracrine effects.

For type I interferons, such as the alpha and beta types, binding to cognate receptors results in phosphorylation of various members of the Janus kinase family, which then further phosphorylate STAT (signal transducer and activator of transcription) proteins, leading to binding at various sequence motifs such as ISREs and GAS (Interferon-Gamma Activated Sequence) elements to initiate transcription of ISGs, including ADAR1. As a side note, while sensing of dsRNA is achieved in the cytoplasm by RIG-I and RIG-I-like receptors (RLRs), sensing of dsDNA in the cytoplasm is accomplished by cGAS (cyclic GMP-AMP synthase), among other proteins, which can interact with STING (endoplasmic reticulum resident transmembrane protein stimulator of IFN genes) to stimulate downstream IRF activity to initiate transcription of interferon genes (102—104).

These pathways and others that have evolved to recognize nucleic acid patterns from viral pathogens can also be activated by endogenous DNA and RNA structures that mimic viral nucleic acids. Up to half the human genome includes repetitive elements that are believed to have viral origins (105); one can imagine that transcription of some of these elements, often by RNA polymerase III (106), can result in formation of RNA structures that resemble viral RNA, thus leading to activation of innate response elements such as the RIG-I/MDA5 pathway. These viral structures may also be present in the genomic DNA, but they may remain hidden by the chromatin structure and thus not be extensively revealed to nuclear pattern recognition receptors.

Because ADAR1 has the ability to bind and alter the structure of dsRNA molecules (in fact, this was how ADAR1 was originally discovered in Xenopus, as was discussed in chapter 1)

by modifying adenosines and destabilizing base pairings, one can imagine that one of the functions of ADAR1 could be to target endogenous viral elements when they are transcribed. By binding to endogenous viral dsRNA, ADAR1 could compete on a binding-level with factors such as MDA5 and RIG-I, and/or ADAR1 could permanently modify these substrates, potentially rendering them less immunogenic from a structural perspective.

Although all cell types in an organism should have the same genomic DNA sequence, the transcriptomic profiles of cells vary, and can even be used to define a cell as being distinct from another cell. Repetitive DNA sequences, if expressed, are part of the transcriptomic profile, and of particular interest are the transcriptomes of cell types in the neuronal lineage, some of which have been shown to express Alu-repeat associated circular RNA molecules (107—110). These unique RNA elements are thought to play a role in development and differentiation down the neuronal lineage (111), but their association with transcription of Alu repeat elements raises the question of whether innate-sensing of dsRNA repeat elements in neuronal cells is regulated by ADAR1.

In a human genetics study, mutations in ADAR1 leading to a catalytically inactive protein was linked to Aicardi-Goutieres syndrome (AGS), which manifests as developmental inflammation in the brain and skin. To further examine the role of ADAR1 in neuronal cells, previously generated ADAR1-knockout human embryonic stem cells (61, 84) were differentiated down the neuronal lineage, and the expression of interferon-stimulated genes was measured, among other measurements, as a readout for the extent of innate immune activation in these cells during the differentiation process. Of note, previous quantitative RT-PCR examination of ISG expression in whole blood taken from 10 AGS patients with mutations in ADAR1 revealed upregulation of genes such as ISG15 and IFIT1 (112).

The process of maintaining and differentiating human embryonic stem cells (hESCs) down a neuronal lineage is different from culturing 293T cells in several ways. First, the basal media used is a trademark formula sold by STEMCELL Technologies called mTeSR, to which is added a supplement mix, also from the same company. This supplemented media must be changed daily to maintain pluripotency in the cultured cells. The only exception, during the neuronal differentiation pathway, is that once axon-like structures appear in culture, only half of the media is removed and replaced with fresh media, so as to not expose neuronal structures to air-drying conditions, even if briefly. This was found empirically to result in more successful culturing of neural progenitor cells (NPCs) and their progeny cell types, including neurons, astrocytes, and oligodendrocytes.

Of note, planning experiments that use hESCs require different considerations than when 293T cells are used. Freeze-thawing hESCs more than necessary should be avoided, so each batch of thawed ESCs should be used for about two months of experiments to maximize its potential. After more than two months of passage with daily media changes, the hESCs should be discarded. Another difference from 293T cell culture techniques is usage of the small-molecule inhibitor of ROCK (Rho-associated, coiled-coil containing protein kinase), called Y-27632, during cell passaging. This inhibitor, when added to dissociated cell suspensions, seems to improve cell viability during passaging (113). Finally, media used for passaging and daily replenishing must be heated evenly at 37°C before adding to cells.

The process of generating NPCs from hESCs begins with the coating of 12-well plates (3.5 cm^2) with Matrigel at 37°C for at least one hour. Matrigel is a trademark mixture of gelatinous protein secreted by mouse cancer cells, and helps in the expansion of hESCs and NPCs in culture (114). Once the Matrigel-coated plates are prepared, the neuronal differentiation path is set in motion by addition of Neural Induction Media (NIM), followed by Neural Progenitor Media (NPM), both from STEMCELL Technologies. The differentiation process can take up to a month, as summarized in figure 2.4.1, although our protocol has been modified slightly from the original, in particular, the earlier addition of NPM at day 12. Pellets can continue to be harvested after the time points shown or in between the time points if cell counts on passage days permit the plating of additional wells.

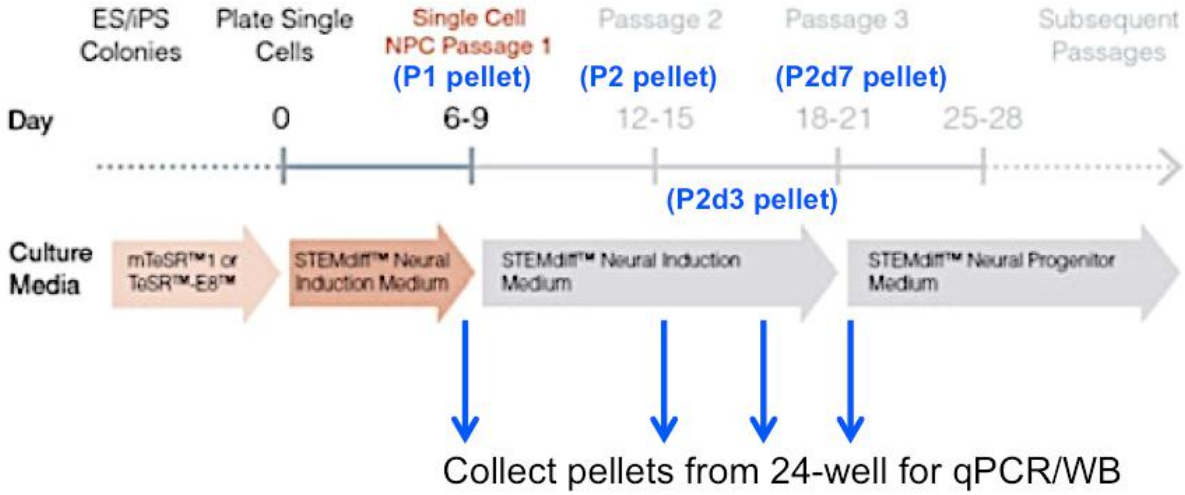


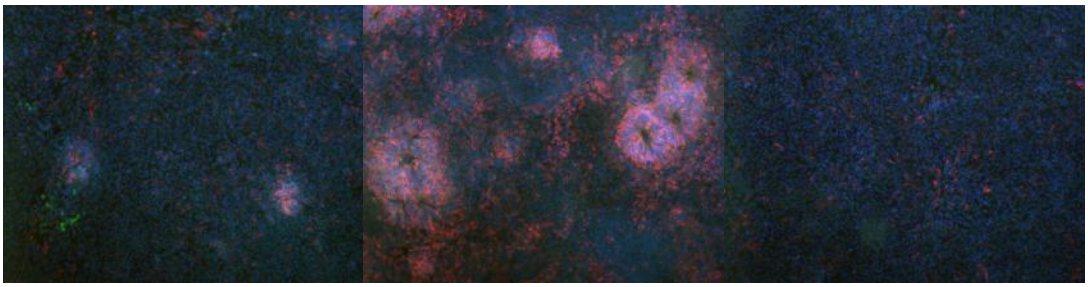
Figure 2.4.1 Differentiation of hESCs into NPCs
The original protocol from STEMCELL Technologies shown above has been modified slightly. Details are in appendix 3. In summary, the protocol from STEMCELL Technologies was shortened so as to have 12 days of culturing in NIM media before switching to NPM

media. Time points for monitoring ISG expression include passage 1, which is 6 days after culturing in NIM media, at which point the pluripotent hESCs have started to commit to a neural-like multipotent lineage; passage 2, which is after 12 days of exposure to NIM media, at which point most all cells can be considered multipotent neural stem cells, which have the ability to give rise to neurons, astroctyes, and oligodendrocytes; and day 3 NPM, which is 15 days after start of culture; and finally day 7 NPM, or 19 days after culture.

Examination of cell-type specific markers was done throughout the differentiation process (Figure 2.4.2). Analysis of gene expression revealed, for two different ADAR1 KO clones, upregulation of STAT1 (signal transducer and activator of transcription 1) and ISG15 (interferon-stimulated gene 15) as compared to WT starting at passage 2. The upregulation continued through the end of the time points. Furthermore, and consistent with the observed ISG upregulation, there was also spontaneous upregulation of interferon-beta starting from day 12, although the endogenous trigger(s) for this upregulation remains to be determined in these cells (Figure 2.4.3).

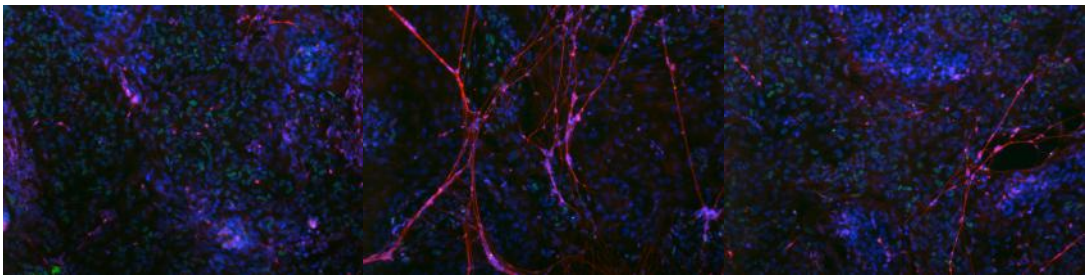
P1: Day 6 NIM (Nestin, Oct4):

N2 **N3** **N16**



P2d3: Day 3 NPM (Tuj1, Sox2):

N2 **N3** **N16**



P2d3: Day 3 NPM (phase contrast):

N2 **N3** **N16**

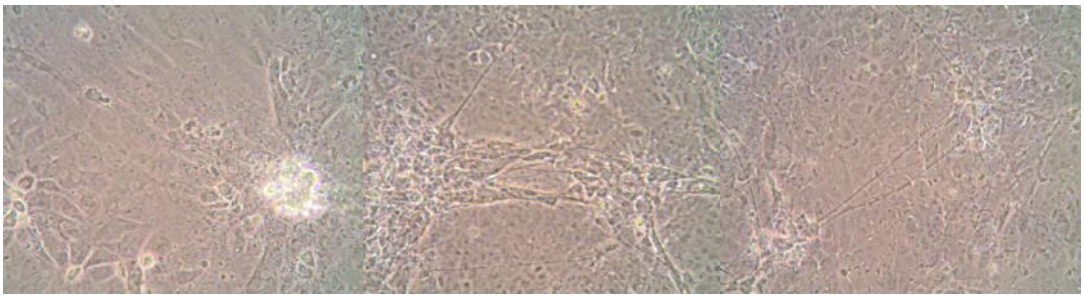


Figure 2.4.2 NPC immunocytochemistry

Cells on coverslips, after fixation, were incubated in PBS with 5% goat serum and 0.1% Triton X-100 to reduce non-specific binding of antibody Fc regions and to permeabilize cell membranes for access to intracellular antigens, respectively. Cells from P1 and P2d3 were then incubated for 1 hour at room temperature with mouse anti-Nestin/rabbit anti-Oct4 and mouse anti-Tuj1/rabbit anti-Sox2, respectively. Oct4 is a transcription factor used often as a marker for pluripotency (115), and Nestin is filament protein expressed in neural stem cells (116); Tuj1 is a tubulin protein expressed in both neurons and precursor cells committed to a neuronal lineage (117), and Sox2 is a transcription factor important for maintaining the biology of neural progenitor cells (118). After incubation with primary antibodies, three washes of 15 minutes each were done with blocking solution. Secondary staining using goat antibodies was then done for 1 hour at room temperature, followed by another set of washes. Finally, DAPI was added for 10 minutes to

stain the nucleus. The coverslips were then mounted onto microscope slides, sealed with nail polish, and imaged with a Zeiss microscope. The images correspond to cells plated in 24-well plates that were imaged at the indicated time points. N2 refers to a WT hESC clone, and N3 and N16 refer to two ADAR1 KO hESC clones. After 6 days of incubation with NIM, clusters of cells are beginning to express Nestin. After switching to NPM on day 12, cells can be observed to express Tuj1 after about 3 days on day 15. Axon-like structures are also visible on day 15, as shown in the phase contrast image.

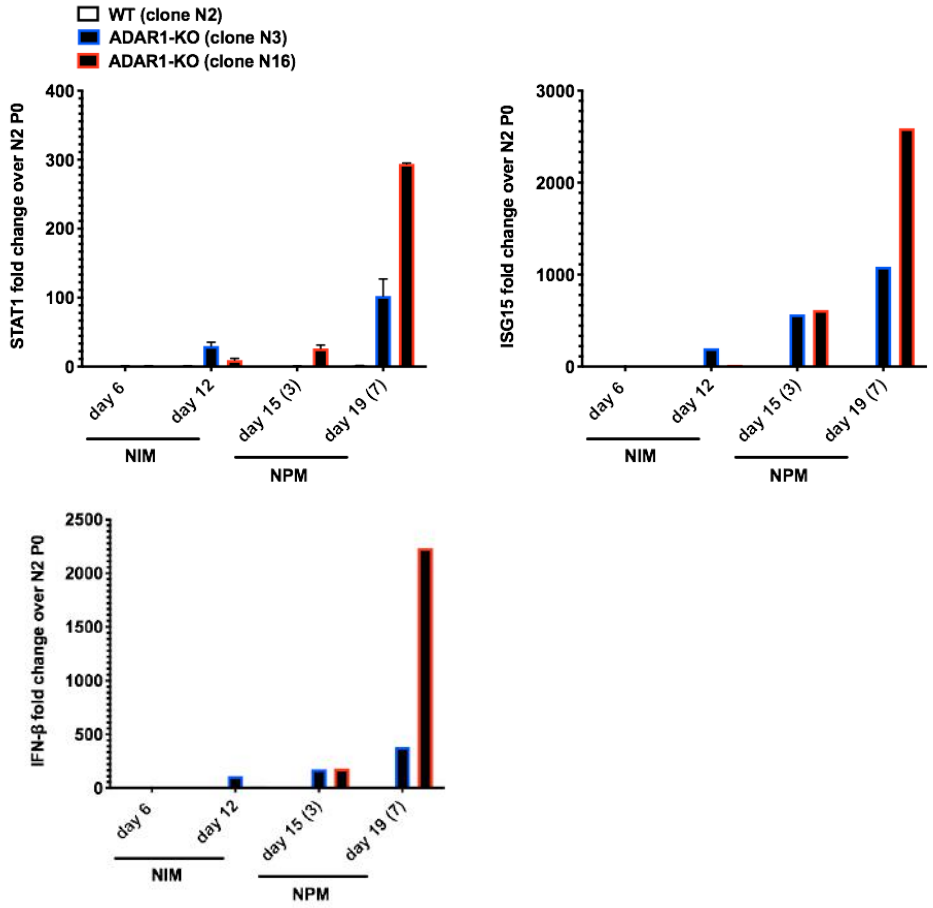


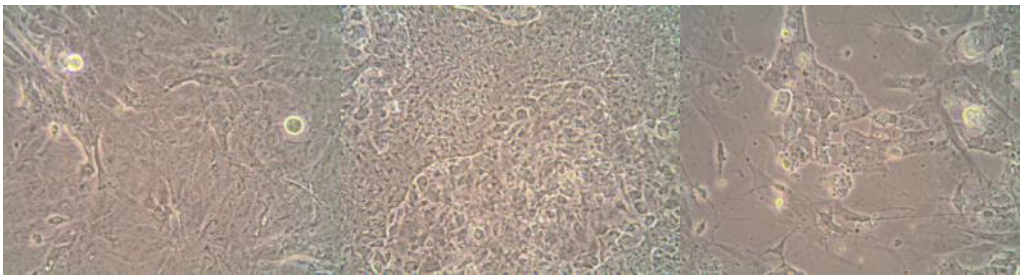
Figure 2.4.3 Spontaneous upregulation of IFN in ADAR1 KO NPCs
RNA was harvested from cell pellets at the indicated timepoints. For each pellet, 1ml of TRI-reagent was added to dissolve the cells. The mix was incubated at room temperature for 5min. Next, 200µl of chloroform per 1ml of TRI-reagent was added, and the mixture was vortexed for 10s and then incubated at room temperature for 3min. The mixture was then centrifuged at 12,000G for 15min at 4°C. The upper aqueous phase, which contains RNA, was removed and transferred, about 500µl, with care not to take off the white-colored layer in the middle. RNA extraction was then continued with the Direct-zol RNA MiniPrep kit,

starting with adding to the samples a volume of pure ethanol equal to the removed aqueous solution. Digestion of genomic DNA with DNase I was done at 28°C for 30min. The RNA was reverse-transcribed to cDNA for SYBR Green qPCR. Gene expression was normalized relative to RPS11. Upregulation of interferon-beta, ISG15, and STAT1 was observed in both ADAR1 KO clones, with no additional stimulus being provided to the cells. The qPCR primer sequences are listed in appendix 1.

Of note, the spontaneous upregulation of interferon-beta in the absence of ADAR1 and the concurrent stimulation of ISG expression in the NPCs resulted in changes in cell growth, viability, and morphology. There appeared to an inverse correlation between the growth rate of and the level of interferon upregulation, particularly after the third passage, or 19 days after start of culturing (7 days in NPM). Furthermore, evidence of cell death based on the emergence of halo-like features around cells as seen on phase contrast microscopy correlated with the level of interferon expression (Figure 2.4.4)

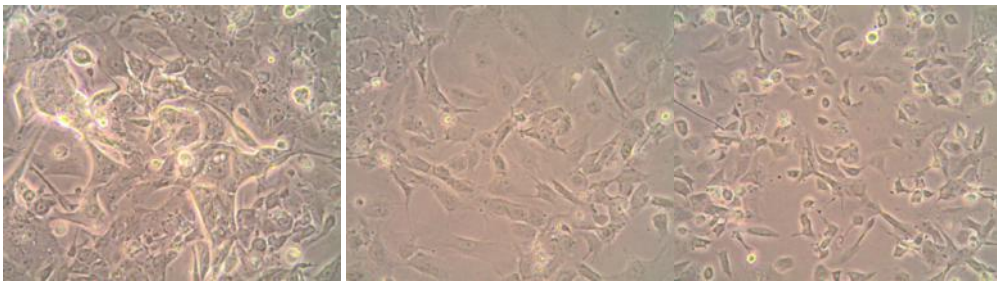
P2d7 (P3 or day 19):

N2 **N3** **N16**



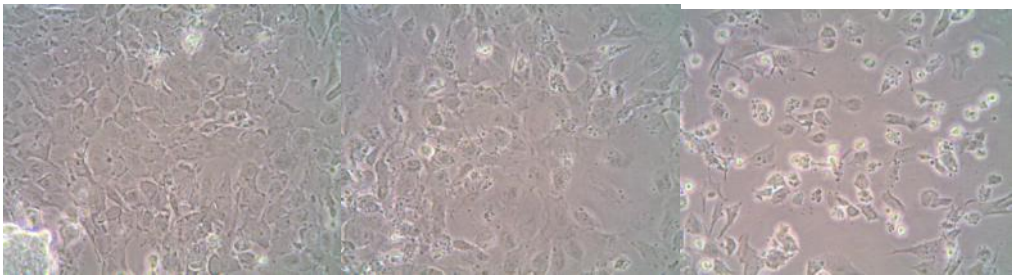
P3d1 (day 20):

N2 **N3** **N16**



P3d2 (day 21):

N2 **N3** **N16**



P3d3 (day 22):

N2 **N3** **N16**

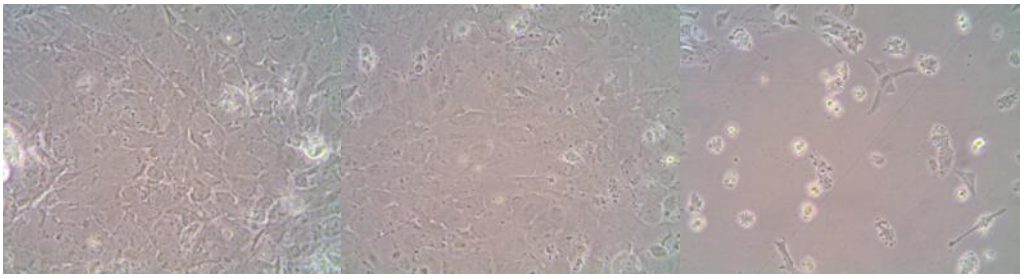


Figure 2.4.4 Cell death in ADAR1 KO NPCs

The NPCs in 6-well plates were cultured until day 19, after 7 days of incubation with NPM. At this timepoint, the WT and N3 ADAR1 KO clones were confluent and the N16 ADAR1 KO clone, which had higher upregulation of interferon compared to clone N3, was about 80% confluent. Cells were counted for passaging, and the number of viable cells were determined using Trypan blue staining. Viable cells were then plated at a density of $1.5x10^5$ viable cells/cm^2 and continued to be cultured with NPM media.

Phase contrast microscopy images were taken for the P3 timepoint and each additional day afterward for 3 days, as shown above. After the third passage, the WT clone continue to grow at a rate comparable to previous passages, but the N3 and particularly N16 clones appeared to grow slower, and there was visible evidence of cell death.

To examine if the observed interferon and ISG upregulation are also observed in other cell types, we examined how these genes are expressed in hepatocytes, also differentiated from the same three hESC clones: N2, N3, and N16. The hepatocytes were responsive to interferon, expressed albumin as expected, but there was no spontaneous upregulation of interferon-beta, ISG15, or STAT1 (Figures 2.4.5 and 2.4.6).

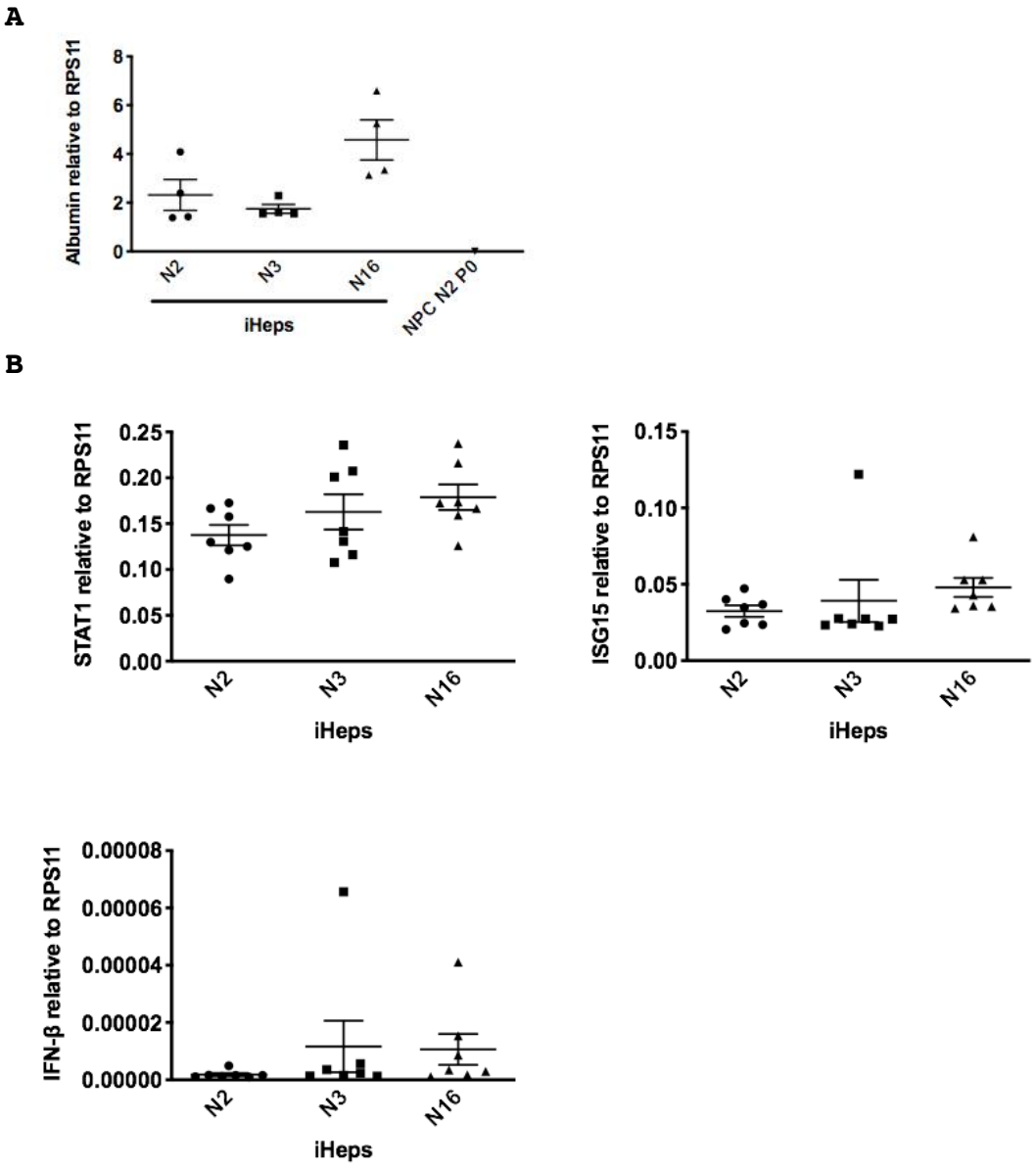


Figure 2.4.5 ADAR1 KO hepatocytes have normal ISG expression
A. In this experiment, hESCs were induced to become endoderm germ layer cells using the STEMdiff Definitive Endoderm Kit from STEMCELL Technologies. Once endoderm was established, differentiation down a hepatocyte lineage was done by culturing in basal media (KnockOut DMEM/F-12 CTS, 10% knockout serum replacement, 0.5% GlutaMAX, 0.5% non-essential amino acids) with hepatocyte growth factor (100ng/ml) for 8 days and then the same media but with addition of dexamethasone (40ng/ml) for another 3 days. Next, the culture media was switched to Lonza hepatocyte culture medium with oncostatin M (20ng/ml) for 5 days. RNA was then extracted from the induced hepatocyte-like cells (iHeps), and expression of albumin was examined to assess hepatocyte

differentiation. As a control, albumin expression was not detected in NPCs.
B. The same genes tested in NPCs (ISG15, STAT1, interferon-beta) were also tested in the iHeps, which showed no significant change in interferon or ISG expression levels in the ADAR1 KO clones (N3 and N16) compared to the WT clone (N2).

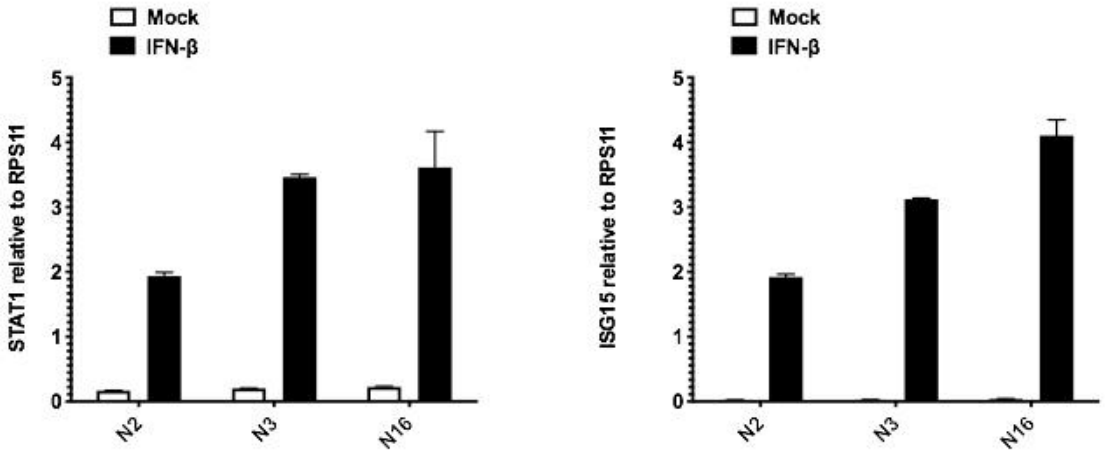


Figure 2.4.6 ADAR1 KO hepatocytes respond to interferon
The iHeps were treated with interferon-beta for 24 hours before RNA harvest and analysis of gene expression. The treated WT and ADAR1 KO cells all had the ability to upregulate ISG expression, indicating that they are responsive to interferon. Thus, the results in figure 2.4.5 are not due to lack of responsiveness to interferon.

We noticed, in the comparison between ADAR1 KO NPCs and hepatocytes, that the spontaneous upregulation of interferon and ISGs is intrinsic to the NPC population. Here, there is the possibility of a neuronal-specific RNA species that is an ADAR1 target, either as a binding target or editing target, or both. This RNA species, if expressed at all, appears to not be of consequence in hepatocytes with regard to interferon and ISG expression.

Next, based on previous studies suggesting that PKR (protein kinase R) works in tandem with RIG-I/MDA5 mediated interferon upregulation (119, 120), we used immunoblot to examine expression and phosphorylation of PKR. Classically, as part of the antiviral defense pathway and sensing of dsRNA, autophosphorylation of PKR dimers activates other kinase domains within the protein, enabling it to phosphorylate eIF2α (eukaryotic translation initiation factor 2α). As a result, translational activity is inhibited in cells, slowing down the ability of viruses to replicate in infected cells (121).

Of course, translational shutdown is not without consequence for cell biology, and prolonged shutdown may result in slower growing cells and even cell death, both of which seem to occur in the ADAR1 KO NPCs. On immunoblot, we detected phosphorylated PKR as early as day 12 after the start of NPC-differentiation in one of the ADAR1 KO clones (Figure 2.4.7)

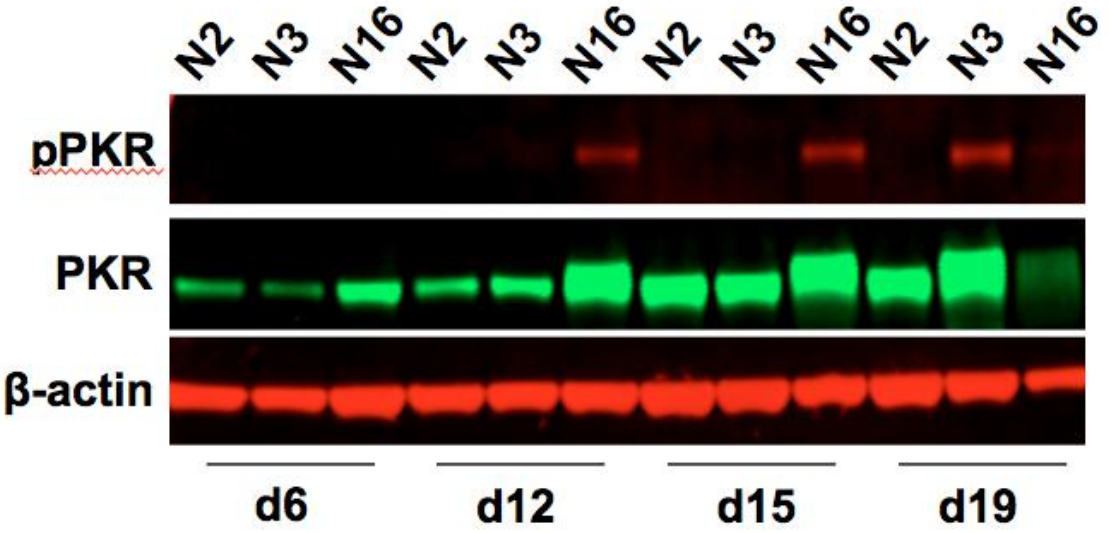


Figure 2.4.7 PKR activation in ADAR1 KO NPCs
Cell pellets from the indicated timepoints were prepared for immunoblot to detect PKR, using the rabbit anti-PKR antibody from Abcam that recognizes amino acids 500-600 of human PKR. The phosphorylated version of PKR (pPKR) was detected using the rabbit anti-pPKR antibody from Abcam that recognizes human PKR specifically when threonine-451 is phosphorylated. This version of PKR has a different mobility on the SDS-PAGE gel, as seen in the PKR lanes where pPKR is also present. Both ADAR1 KO clones (N3 and N16) showed activation of PKR, although the N16 clone experienced this earlier than the N3 clone. By day 19, the N16 colony was growing slower and also showing signs of cell death; here the PKR and pPKR bands both appear fainter, although the actin levels are lower also. The immunoblot was done as described previously, except after transfer to nitrocellulose, LI-COR Odyssey blocking buffer was used instead of 5% milk for blocking and dilution of antibodies. On the second day, secondary antibodies were conjugated to Odyssey-compatible infrared dyes rather than horseradish peroxidase, to prepare for immediate imaging on the LI-COR machine after incubation with secondary antibodies and washing.

In summary, ADAR1 appears to be important in neural progenitor cells for suppressing a spontaneous upregulation of interferon, perhaps in response to immuno-stimulatory RNA species. This effect was not observed in hepatocyte-like cells,

also differentiated from the same parental hESC WT and ADAR1 KO clones; the hepatocytes were capable of expressing ISGs in response to exogenous interferon stimulation.

The changes in cell growth and viability in the NPCs seem to parallel the phosphorylation of PKR and upregulation of interferon. The N16 ADAR1 KO clone seems to be more sensitive to lack of ADAR1 than the N3 KO clone, or it has a more responsive interferon-signaling pathway, as revealed by the significantly higher fold change in ISG expression compared to the N3 KO clone. Furthermore, by day 19 (7 days in NPM), the N16 clone may already have translational shutdown as seen by the decrease in PKR expression in figure 2.4.7.

Having revealed a particular function for ADAR1 in NPCs, and consistent with the neuro-inflammatory phenotype of AGS patients with defective ADAR1, we next wondered how ADAR1 isoforms evolved to be essential in humans. Could evolution of the longer ADAR1 isoforms, with more Z-DNA/RNA binding domains, be important for binding and regulating Alu element activity when these sequences began to copy themselves more actively, potentially creating immunogenic intermediates? It is known that Z-conformation sequence motifs are found in Alu elements that form inverted repeat structures, which is something that seems to be particularly abundant in primate genomes, including the human genome (122).

2.5 Evolution of ADAR1 isoforms

The evolutionary history of ADAR1 and its homologs, ADAR2 and ADAR3, is vast. In brief, adenosine deaminases acting on tRNAs (ADATs), which are found in prokaryotes and eukaryotes, are thought to be ancestors of the modern ADAR proteins, which gained their dsRNA-binding domains, possibly through gene duplication, around the time the animal and protist kingdoms split (123, 124). Organisms grouped in the animal kingdom are known to have ADAR proteins (125, 126). Examination of changes in the domain structures of ADAR proteins reveals that organisms placed later in the phylogenetic tree have longer isoforms, and in particular, more Z-DNA/RNA binding domains. For example, while invertebrates such as drosophila and squid have dsRNA binding domains in their ADAR proteins, there are no Z-DNA/RNA binding domains (57). Vertebrates are perhaps some of the more recent types of animals to have evolved, and we set out to examine the p150 and p110 coding potential of the ADAR1 locus of 100 different vertebrate species by using the Vertebrate Multiz Alignment (Figure 2.5.1).

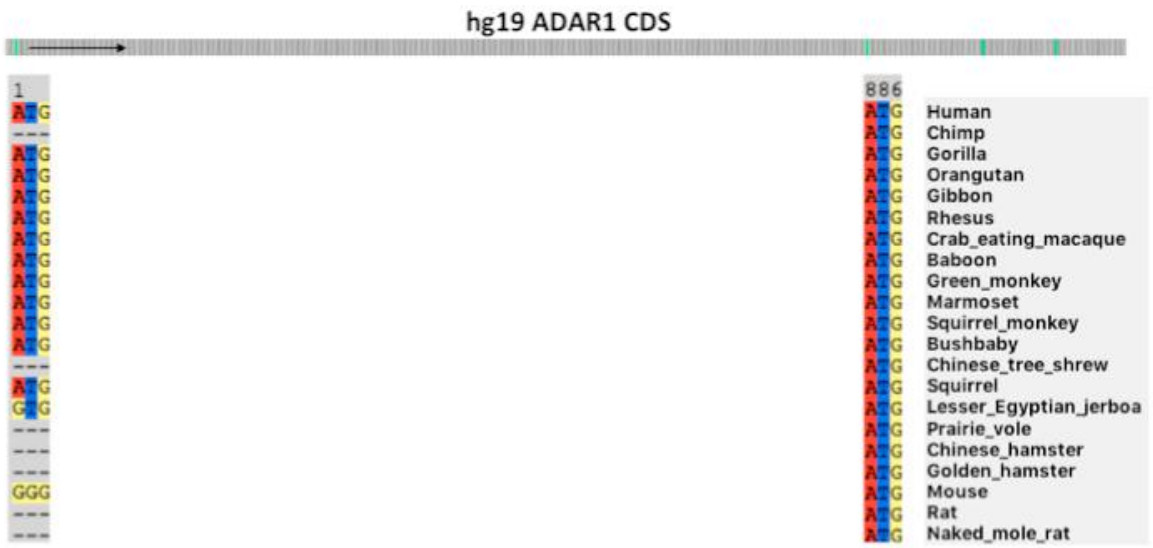
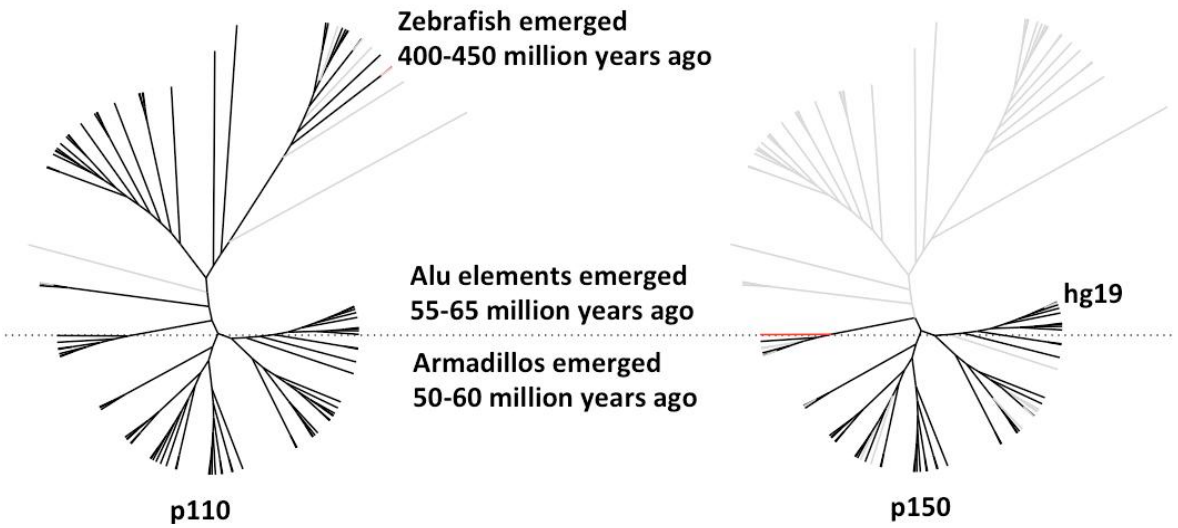


Figure 2.5.1 ADAR1 Vertebrate Multiz Alignment
The genomes of 100 different vertebrate species were aligned using the human hg19 sequence as a reference, and a snapshot of cDNA assemblies corresponding to the full-length open-reading frame in humans is shown above, generated using the SeaView software. The open-reading frame is shown with green bars to indicate the methionine codons. Position-1 is the p150-ATG position-886 is the p110-ATG. The ordering of species is ordered based on branching in the phylogenetic tree. For example, panTro4 (chimpanzee) is under hg19 (homo sapiens) and represents a single branching event.

Of note, there is high conservation in the p110-ATG position across the 100 vertebrate species but less for the p150-ATG. Placing the p150-ATG on the phylogenetic tree reveals a belated appearance of p150 compared to p110 (Figure 2.5.2). The first appearance of ADAR1 p150 is in the nine-banded armadillo, which emerged between 50-60 million years ago, around the same time that early primates appeared (127, 128). Of note, Alu elements also emerged around this time from fusions of the 7SL RNA, and they will eventually amplify throughout primate and human genomes (129, 130). The first appearance of ADAR1 p110 is in zebrafish, which is thought to have evolved 400-450 million years ago. Of note, armadillos have tRNA-derived short interspersed elements, different in origin from 7SL-derived Alu elements, but these SINE elements can form secondary structures that may be potential ADAR1 targets (131).

A



B

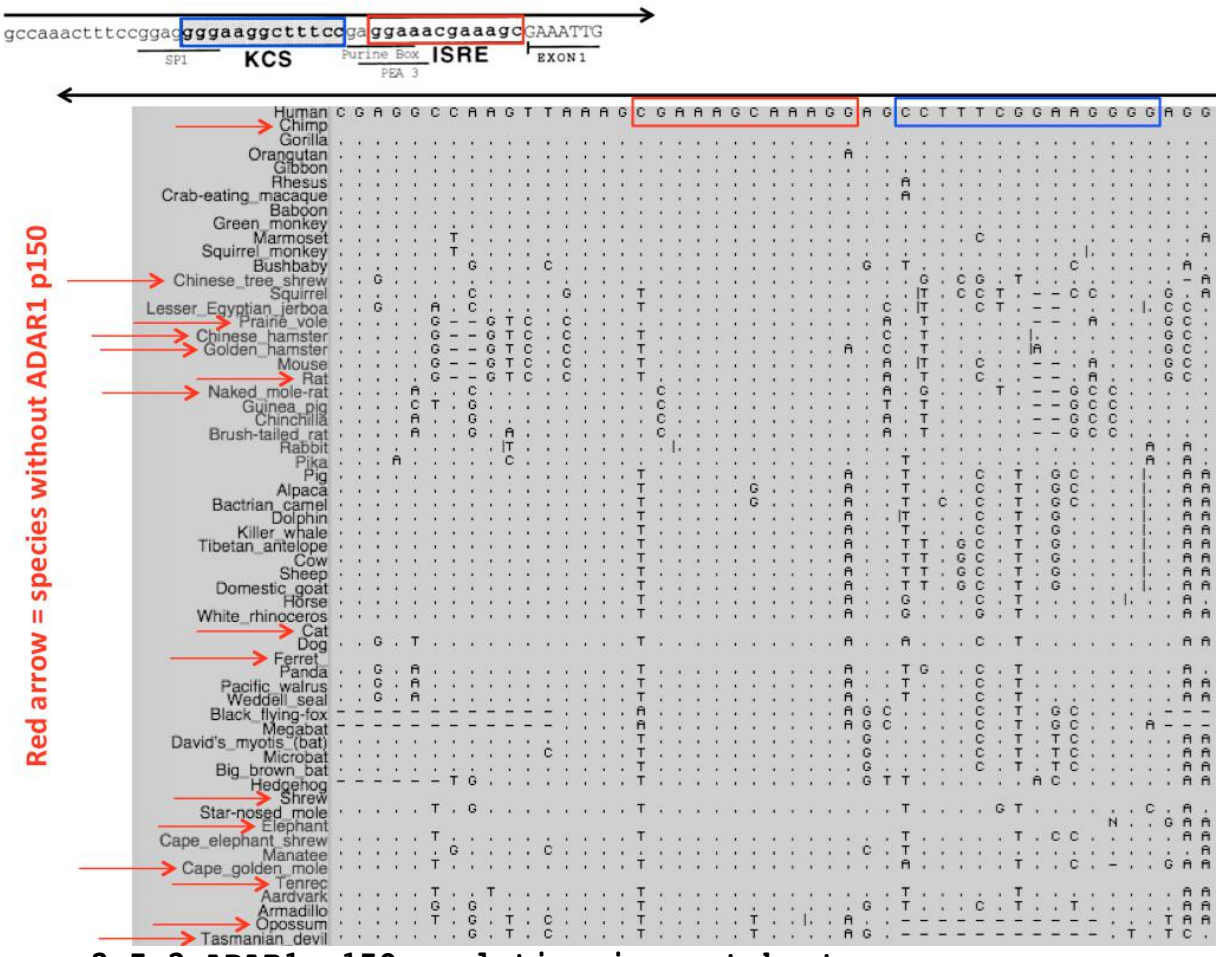


Figure 2.5.2 ADAR1 p150 evolution in vertebrates

A. Shown are the 100 species found in the Vertebrate Multiz Alignment, organized on a phylogenetic tree, generated using the FigTree software. Lines in gray show absence of either p110

(left) or p150 (right) for each species, and lines in red show the first organisms that had a coding sequence aligning with the human (hg19) version of either ADAR1 p110 or p150: zebrafish and armadillo, respectively. Here, alignment requires presence of the p110 or p150 start codons along with homology downstream. B. Shown here are all vertebrates that have sequences upstream of exon 1A (p150-encoding) aligning with the corresponding human sequence. All animals that have ADAR1 p150 also appear to have the interferon promoter (ISRE sequence boxed in red), with some single nucleotide variants in the alignment. Some animals that lack p150-coding potential also seem to have this ISRE, about half, and the other half without p150 also lack the ISRE. Chimpanzees somehow appear to lack both the p150 promoter and the p150-ATG. The blue box is a kinase-conserved sequence found upstream of PKR and ADAR, and the schematic is adapted from George and Samuel (1999).

The ADAR1 full-length p150 isoform, which has the additional Z-alpha-DNA/RNA binding domain in the N-terminal end, distinct from the second Z-beta-DNA/RNA binding domain, aligned to the human p150 version in 45 out of 100 vertebrates. By contrast, the p110 isoform aligned to all except 7 fish and 2 mammals: lamprey, spotted gar, Mexican tetra, stickleback, southern platyfish, Nile tilapia, yellowbelly pufferfish, platypus, and tammar wallaby. The p110 isoform appears to have predated the p150 isoform, and when we consider invertebrates such as squid, octopus, and drosophila, all of which have ADAR, the Z-alpha domain is also lacking in their ADAR proteins (57).

Fish are some of the oldest vertebrates, and although their genomes do contain transposable elements, they are not as numerous as the amount found in primate and human genomes (132, 133). The evolution of the p150 isoform, with its high-affinity Z-conformation binding domain, coincides with evolution of species that have Z-conformation Alu repeat elements embedded in their genomes, such as early primate species. Of note, in humans, specific loss of the Z-alpha domain in p150 due to genetic variations causing a frameshift upstream of the p110-ATG has been observed in AGS patients, even though p110, with its Z-beta domain, is still present and functional (134).

Of curiosity is the chimpanzee ADAR locus, which lacks the p150-ATG on the Mulitz alignment (Figure 2.5.3). Although other nonhuman primate species, such as gorilla and orangutan, have sequences with p150-coding potential, examination of the overall editing levels revealed lower A-to-I edit frequencies, at least in brain tissue, for nonhuman primates compared to humans (135). There is the possibility that Alu sequences and structures are different in primates in such a way that they are less of a target for the Z-conformation or dsRNA binding domains of ADAR1,

and this could account for the differences in editing levels between humans and nonhuman primates. Editing levels in other species, such as mouse and drosophila, are also lower than in humans, partly because the genomes of those organisms lack traditional Alu substrates. Of interest, out of the total number of editing sites in octopus, almost one-quarter are in protein-coding regions (57).

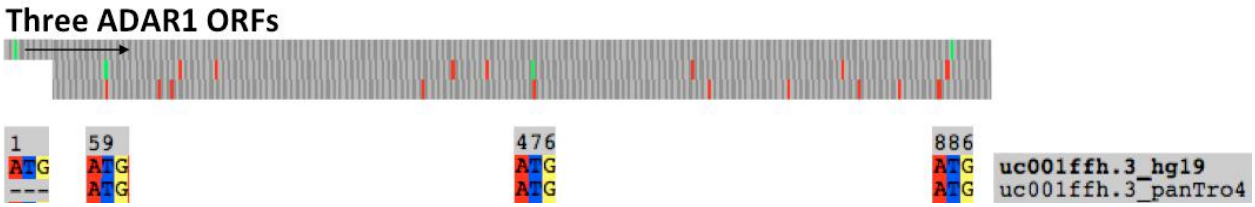


Figure 2.5.3 ADAR1 locus in chimpanzees
Shown above is a SeaView-generated image that corresponds to the coding sequence region of ADAR1, with the three possible reading frames displayed. Of note, there are two methionine codons in the second reading frame, but initiation of translation at these locations would result in production of truncated peptides because of nearby downstream stop codons. Although the chimpanzee (panTro4) sequence lacks the p150-ATG, it does have the p110-ATG and also the two ATG codons in the second reading frame, as aligned to the human ADAR coding sequence (hg19).

As a side note, ADAR1 is not alone in the battle against Alu invasion (to use a term from Alan Herbert) of primate genomes. The endoribonuclease Dicer is also important in preventing accumulation of Alu elements, as was observed in eye tissue, in which buildup of Alu RNA in the cells of retinal epithelium leads to macular degeneration (136).

Looking at the coding sequence of ADAR1 between the p150-ATG and p110-ATG across the different vertebrate species, there is variability in the number of ATG codons in this middle region, but for humans and nonhuman primates, the number is two. The consequences of this with respect to expression of p110 will be discussed in the following chapter.

CHAPTER 3. Translational regulation of ADAR1p110

Following up directly from chapter 2, this chapter aims to 1. describe the unique organization of start codons within the ADAR1 open-reading frame; 2. describe regulation of p150 and p110 translation from the full-length ADAR1 mRNA; and 3. show how p150 and p110 isoform-specific knockout cell lines can be generated without changing the coding sequence of either isoform.

3.1 ATG codon organization in ADAR1

As was discussed at the conclusion of chapter 2, the human ADAR1 coding sequence has only two methionine codons in between the p150-ATG and p110-ATG codons. This unique feature was conserved in nonhuman primates, although other vertebrate species had varying numbers of ATG codons in this middle region (Figure 3.1.1).

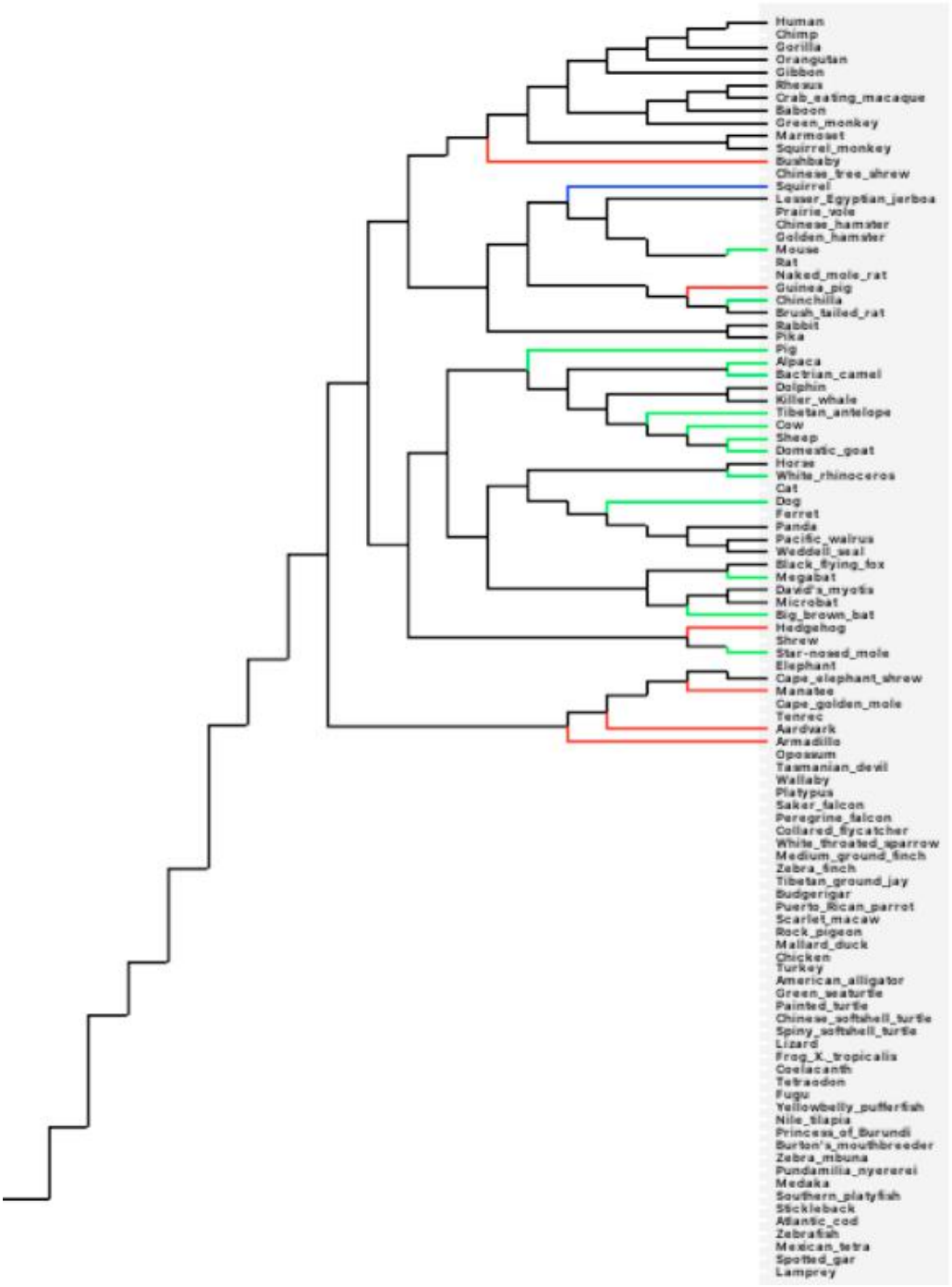
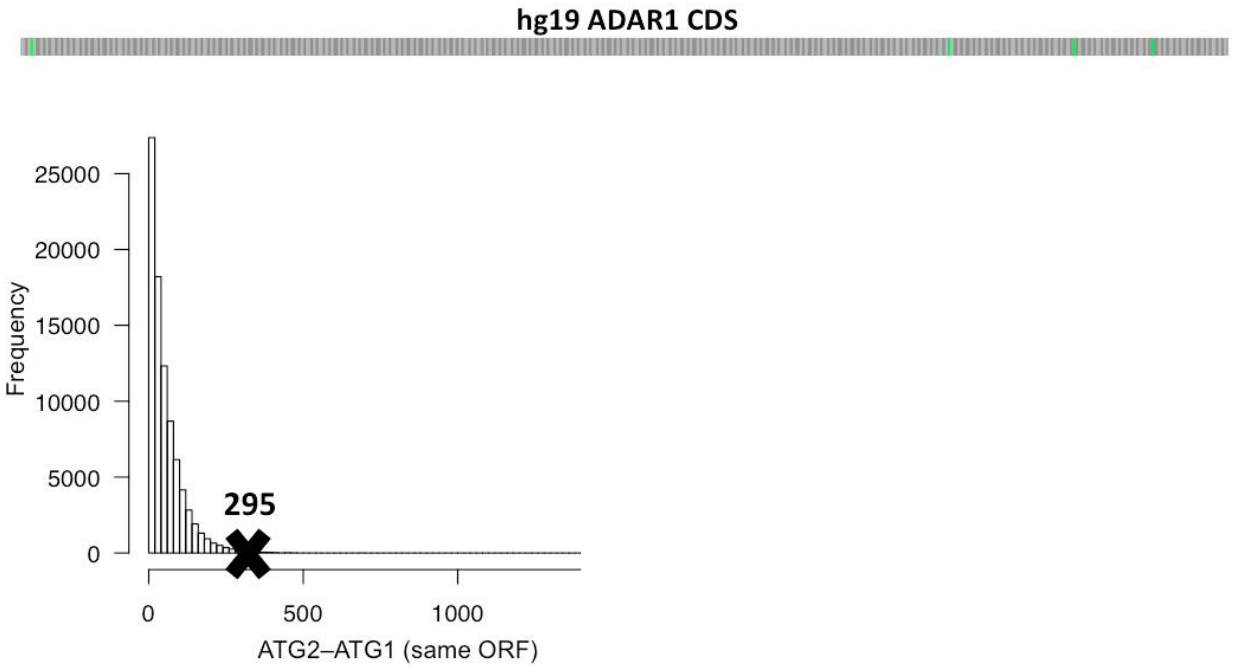


Figure 3.1.1 Vertebrate ATG codons between p150 and p110
Shown above is a phylogenetic tree listing the 100 vertebrate species included in the Multiz Alignment. Black lines point to vertebrates that have exactly 2 methionine codons in between the p150-ATG and p110-ATG; blue=1 or less; green=3; and red=4 or more. Species that lacked a p150-ATG, p110-ATG, or both were excluded (no lines).

The following nonhuman primates have exactly 2 ATG codons in between the p150 and p110 start codons: gorilla, orangutan, gibbon, rhesus, crab-eating macaque, baboon, green monkey, marmoset, and squirrel monkey. Notably, the first in-frame ATG after the p150 start codon is the p110-ATG, and the distance between these two codons: 885 bases, stands out among coding sequences in that most genes have lower numbers of bases between the first two methionine codons in the primary reading frame (Figure 3.1.2).

A



B

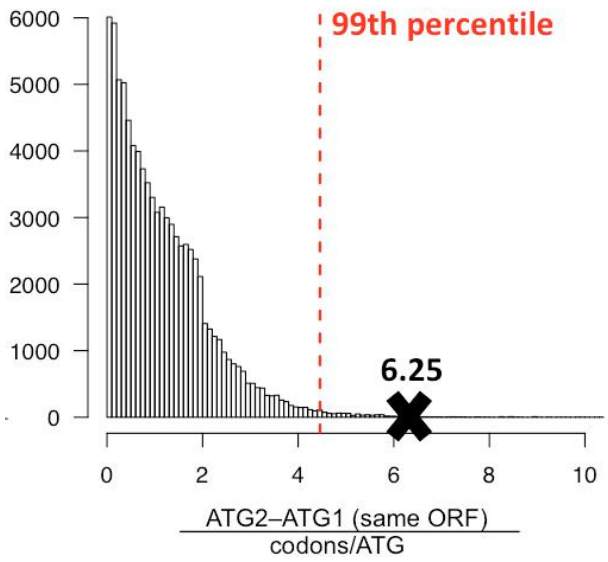


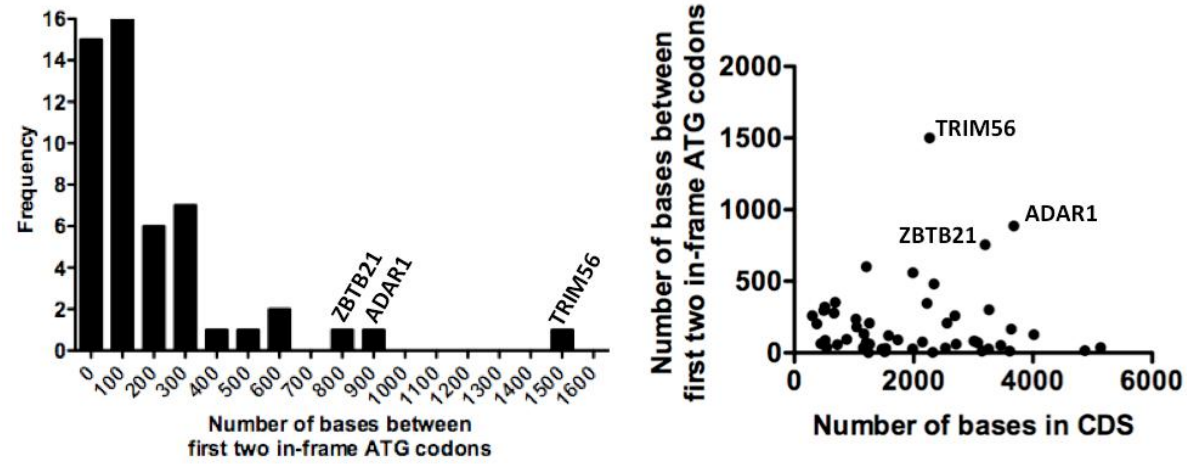
Figure 3.1.2 Distance between first two ATG codons
A. Human cDNA sequences (57,348 in total) compiled from GRCh38 using BioMart were analyzed using RStudio, first to split the sequences that start with ATG into triplets, corresponding to codons, and then to assign each codon a number starting from 1 and increasing by 1 for each codon; the difference in the assigned numbers for the first and second ATG codons (if applicable, as a minority of open-reading of frames have only one ATG) was calculated for each sequence, and then all values were plotted on a histogram, showing counts of genes with different codon distances between the first two ATG codons (ATG2-ATG1) in the same reading frame. As an example, marked by the "X" in the left histogram, the ADAR1 gene has an ATG at position 1, and the next ATG is at position 296, and 296-1=295. As the green bars (start codons) show in the ADAR1 reading frame schematic above the histogram, the ATG3 and ATG4 codons are spaced much closer to each other and to ATG2 than the ATG2 is to ATG1.
B. Because of variation of gene lengths, and specifically the possibility that this variation may impact the distribution of methionine codons and the average distances between them, we considered a normalization factor for the histogram: total number of codons per total number of ATG codons in the reading frame. This normalization factor essentially splits the gene length (in units of codons) by the number of ATG codons, and is testing for deviation away from a value of 1, which corresponds to an ATG2-ATG1 codon distance that equals the codon distance if all ATG codons are distributed evenly across the gene. As the normalized histogram shows, most genes have an ATG2-ATG1 codon distance less than what would be expected if all ATG codons are distributed evenly, and only 1% of genes have an ATG2-ATG1 codon distance that is about 4.5 times longer than what the distance would be if all ATG codons are distributed evenly; notably, ADAR1 has an ATG2-ATG1 codon distance that is 6.25 times longer than this normalization factor.

Examination of the primary ADAR1 reading frame revealed a unique organization of the first two ATG codons, which are separated by a distance of more than 6 times the distance if methionine codons are distributed evenly across the reading frame. We were curious to see how ADAR1 would appear when taken in context of a subset of similar genes rather than all genes. We chose a list of 51 ISGs, of which ADAR1 is one of them, from a previous screen of effector genes downstream of the type I interferon response (101). Even within this subset of genes with similar functions, ADAR1 stands out as having a long ATG2-ATG1 distance (Figure 3.1.3).

A

IFIH1/MDA5	ABCA9	B4GALT5
MOV10	SAT1	IFITM1
DDX60	DDX3X	RNF24
UBA7	SLFN12	ZNF107
FLJ11286	APOL6	CCL2
NOS2A	TRIM56	BUB1
MS4A4A	C9orf19	ZNF295
PABPC4	ADAR	DTX3L
N4BP1	PLEKHA4	BST2
APOBEC3G	CEACAM1	TLR7
OAS3	ANKFY1	TFEC
CNP	RAB27A	EIF2AK2
BLZF1	APOL3	FLT1
TLK2	VEGFC	RASGEF1
PCTK3	ZNF313	ZAP
CASP1	OAS1	RBM25
MX2	RARRES3	
FNDC3B		

B



C

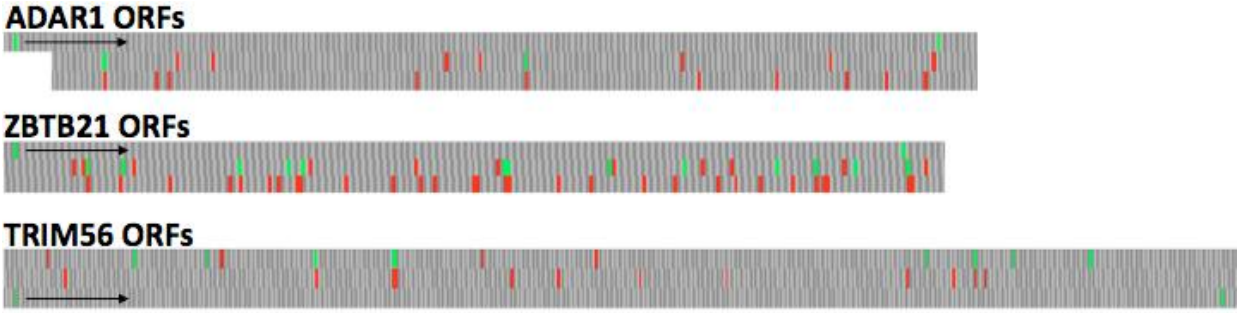


Figure 3.1.3 Distance between first two ATG codons in ISGs

A. Human cDNA sequences were extracted for the gene set shown, corresponding to effector genes involved in the type I interferon response.

B. The number of bases between the first and second ATG codons in the primary reading frame was calculated for all 51 genes listed in panel A, and the values were plotted on a histogram. To the right of the histogram is a plot showing number of bases between ATG2 and ATG1 as a function of the total number of bases in the ISG reading frames. Longer reading frames are not necessarily correlated with longer distances between the first two ATG codons. Three genes that have particularly long

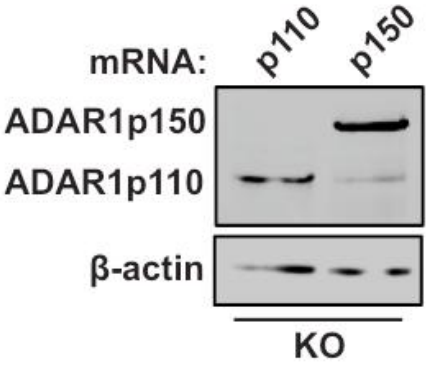
distances between ATG2 and ATG1 are highlighted in both charts: ZBTB21, ADAR1, and TRIM56.
C. The three reading frames of ZBTB21, ADAR1, and TRIM56 are shown, with green bars showing start codons and red bars showing stop codons. Of note, the ADAR1 gene has only two out-of-frame start codons between ATG2 and ATG1, while ZBTB21 and TRIM56 have 12 and 8, respectively.

3.2 Regulation of p150 and p110 translation

In chapter 2, the observation that p110 persisted following deletion of exons 1B and 1C suggested that exon 1A containing mRNAs could give rise to p110. This hypothesis was further supported when exon 1A and its promoter were deleted in the exon 1B/1C double-knockout background, resulting in ablation of p110 signal on immunoblot. Further, the observation that p110 is induced following interferon treatment could be explained by induction of the novel 1A-1C-2 splice isoform, which has only p110-coding potential, or alternatively, induction of the 1A-2 splice isoform, which has both p150- and p110-coding potential.

We therefore aimed to investigate if p110 could be produced from the canonical p150-encoding exon 1A mRNA. First, to examine if p150 transcripts can express p110, in vitro transcribed p150 mRNA was generated and transfected into exon 1B/1C/1A knockout cells (ADAR1 KO cells). 24 hours after transfection, immunoblot revealed presence of p110 in addition to p150, and RT-PCR using primers to amplify the coding region between the p150 and p110 start codons revealed a single band, suggesting that p150 and p110 are being produced from a single mRNA species (Figure 3.2.1).

A



B

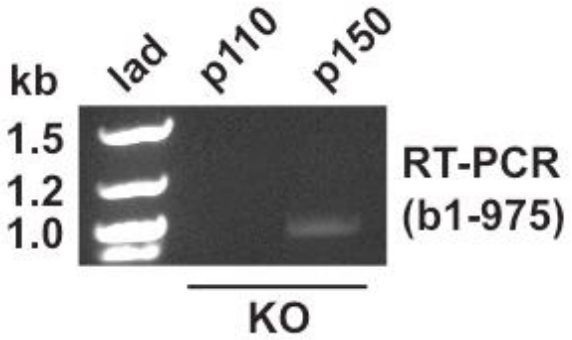


Figure 3.2.1 Expression of p110 from p150 mRNA
A. In vitro transcription of mRNA was done using the SCRPSY backbone vector, which has a T7 promoter and the Invitrogen Gateway attB1 and attB2 cloning sites downstream of the promoter for inserting cDNA sequences into the vector for transcription. There are no XhoI sites in the ADAR1 p150 or p110 sequences, so the XhoI enzyme was selected to linearize SCRPSY with the cloned p110 and p150 cDNA sequences. 10μg of DNA was used as the starting material, and following linearization and purification, the plasmids were diluted to 0.5 μg/μl to prepare for in vitro transcription, which was done using the Cellscript T7 transcription kit, following manufacturer guidelines for synthesizing RNA and adding the 5' cap and 3' polyA sequence. Purification of RNA was then done using the Qiagen RNeasy Mini kit, following manufacturer guidelines. Transfection of in vitro transcribed RNA (250ng) was done using Lipofectamine 2000 and OptiMEM, following manufacture guidelines. ADAR1 KO 293T cells were switched to transfection media (1.5% FBS in DMEM with NEAA), and the RNA/Lipo2000/OptiMEM mixes were added with centrifugation at 1,000G for 30min at 37°C. Media was not changed, and cells were harvested for immunoblot 24 hours later as described previously. Immunoblot with a C-terminal specific ADAR1 antibody shows that transfection of the p150 mRNA in ADAR1 KO cells gave rise to both p150 and p110 protein isoforms.

B. The agarose gel shows reverse transcription (RT) PCR products generated with primers that amplify the region corresponding to bases 1-975 of the p150 open reading frame. The p110 open reading frame begins at nucleotide position 886.

Expression of the in vitro transcribed p150 mRNA, which lacks the 5′ UTR regions found in the native mRNA isoforms, resulted in not only p150, but also p110 protein. Of interest next was the mechanism of this p110 translation. To investigate whether an internal ribosome entry site (IRES) is located upstream of the p110 initiation codon, a bicistronic reporter plasmid (137), a gift from the Eran Segal lab at Weizmann Institute, was used in which translation of enhanced green fluorescent protein is facilitated only if the upstream cloned sequence has IRES activity. The cDNA sequence between the p150-ATG and p110-ATG (bases 1-885) was cloned into the reporter plasmid and then transfected into WT 293T cells for 24 hours. The ADAR1 cDNA sequence in question showed an eGFP signal similar to that of the random (empty) sequence, in contrast to the eGFP signal (88% of mRFP+ transfected cells) seen with the encephalomyocarditis virus (EMCV) IRES sequence (Figure 3.2.2). Data from this reporter plasmid suggest that internal ribosome initiation mechanisms resulting from sequences upstream of the p110-ATG do not contribute significantly to p110 translation.

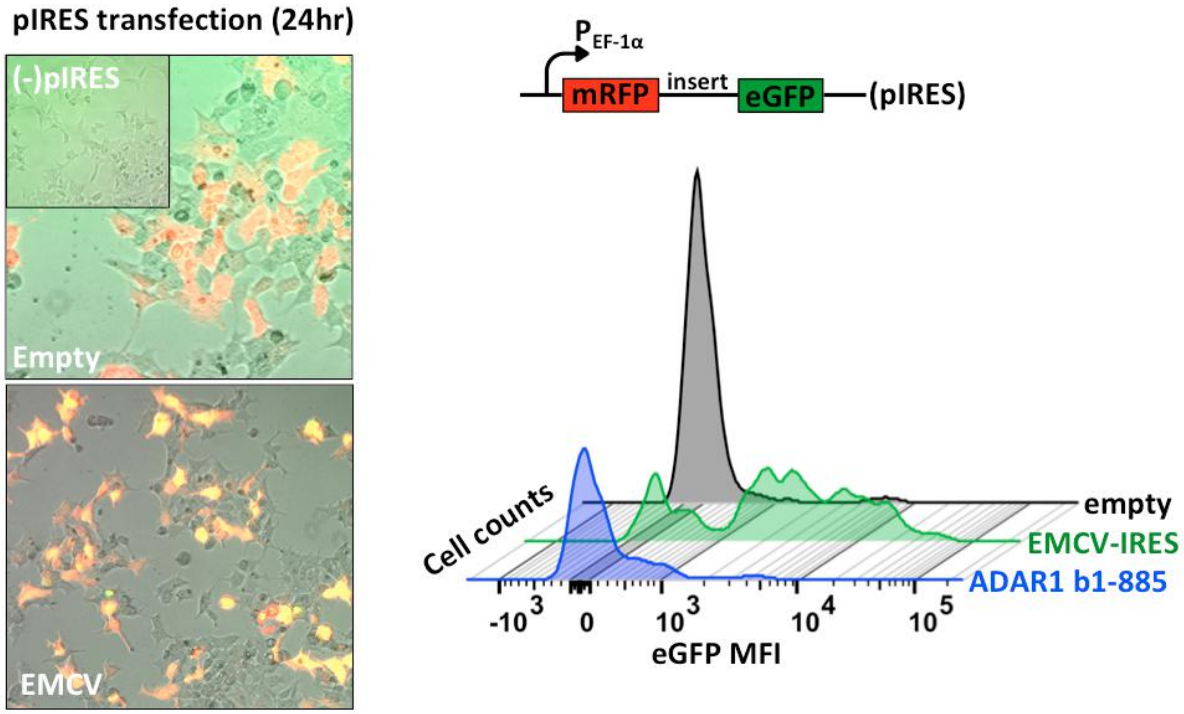


Figure 3.2.2 Internal ribosome entry site analysis
WT 293T cells were transfected with a bicistronic reporter plasmid with a human elongation factor-1 alpha (PEF-1α)

constitutive promoter that encodes for monomeric red fluorescent protein (mRFP) and a sequence of interest upstream of enhanced green fluorescent protein (eGFP). The images on the left show a snapshot of cells in culture imaged with blue light, green light, and phase contrast, overlayed, for the untransfected, empty, and EMCV-IRES conditions. Note the co-localization of red and green signal in cells that were transfected successfully and were expressing both the mRFP via the constitutive promoter and the eGFP via internal initiation via the EMCV-IRES sequence. The histogram on the right shows counts of eGFP mean fluorescent intensity (MFI) values for a 46-base random sequence (empty), the 575-base encephalomyocarditis virus internal ribosomal entry site sequence (EMCV-IRES), and the 885-base sequence upstream of the p110-ATG (ADAR1 b1-885).

There has been an interest to express ADAR1 p150 without p110, particularly in the context of seeing how the A-to-I RNA editomes compare between when only individual isoforms are expressed and when both are expressed. However, as had been mentioned in previous studies, the challenge lies in being able to remove p110 without also removing p150 (61, 138, 139). We devised strategies to reduce translation initiation at M296, the p110 start codon, starting with changing the p110-ATG to CTC (M296L), which, not surprisingly, decreased p110 protein levels but yet also resulted in production of smaller ADAR1 isoforms. These smaller isoforms could result from translation initiation downstream of M296 in the same reading frame as p150 and p110.

Next, we reasoned that the sequences surrounding the p110 methionine codon would influence translation initiation. How typical (optimal) are the sequences surrounding the p110 start codon? Looking at the same group of 51 ISGs (including ADAR1) as listed in figure 3.1.3, we calculated and plotted the frequencies that A, C, G, and T appear at the -3, -2, -1, +4, and +5 nucleotide positions surrounding the annotated translational start sites for those genes (Figure 3.2.4).

There are distinct preferences for the sequences surrounding the annotated translational start site, as expected (140—143), with frequencies ranging from 6% for the atypical (weak/non-optimal) nucleotides, to 47% for the typical (strong/optimal) nucleotides. Notably, the p110-ATG is surrounded by optimal consensus nucleotides, specifically guanosines at position -3 and +4. We modified the sequences surrounding the p110-ATG to test how a weaker (most atypical) translation initiation context would affect production of p110. Immunoblot revealed that these mutations resulted in a protein phenotype similar to that of mutating the p110 start codon: significantly reduced p110 protein production and production of ADAR1 isoforms smaller than p110 (Figure 3.2.3).

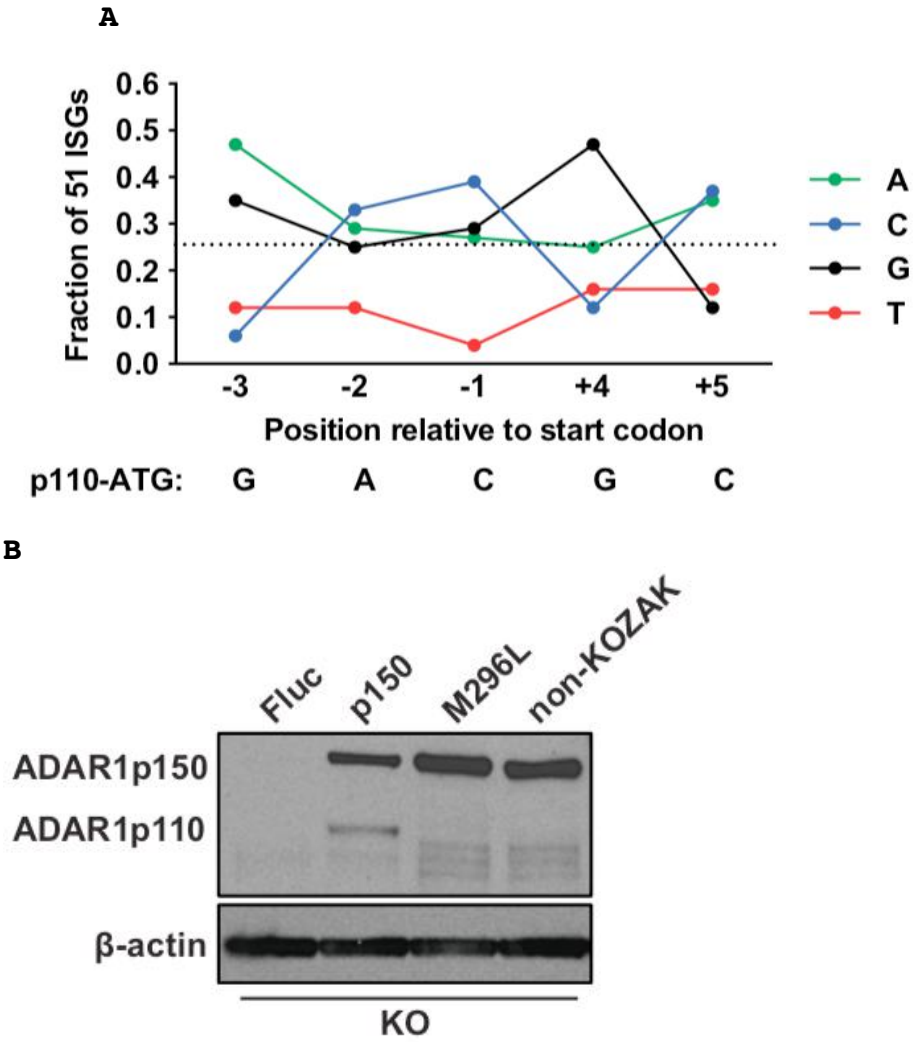


Figure 3.2.3 Mutations to reduce p110 translation

A. The graph shows an analysis of the consensus sequences surrounding the first ATG codon of fifty other genes with similar properties (nucleic acid binding and induction by interferon) to ADAR1. The number of times that A, C, G, and T appear at the -3, -2, -1, +4, and +5 nucleotide positions surrounding the first ATG codon was tallied up and divided by 51 (total number of genes) to determine the frequencies shown on the vertical axis. For comparison, the bases at the same positions surrounding the p110-ATG are shown below. Of note, if bases occurred at random at each position, one would expect the frequencies for each base at each position to approach 0.25 since there are four possibilities.

B. The immunoblot shows ADAR1 KO cells with stable expression of various integrated cDNA constructs driven by a CMV-promoter: firefly luciferase (Fluc), WT p150 (p150), p150 sequence with p110-ATG mutated to CTC (M296L), and p150 sequence with non-synonymous mutations (GACATGGC→TTTATGCT) in the consensus

sequences surrounding the p110-ATG (non-KOZAK). Mutations were introduced into the WT sequence using primers that contained the desired mutations and the WT p150 expression plasmid as a template for the initial PCR reactions. Overlap PCR was then used to assemble the fragments with mutations to be cloned back into the expression plasmid, which was packaged into lentivirus for transduction. At the moment, we only mutated the bases surrounding the p110 start codon, although creating an A-to-G mutation at the +4 position of the p150 start codon would also be an interesting experiment.

With regard to the sequences surrounding ATG codons, we were also interested to examine how the pattern of bases around the first ATG compares to the pattern of bases around subsequent ATG codons in the same reading frame, and not only for the list of 51 ISGs, but for all compiled human protein-coding sequences. Nearly the same pattern was observed when all coding sequences were included in the analysis as when only considering the 51 ISGs, and examination of sequences surrounding downstream ATG codons revealed a different pattern from the sequences surrounding the first ATG, primarily in the -5, -4, -3, -2, and +5 positions (Figure 3.2.4).

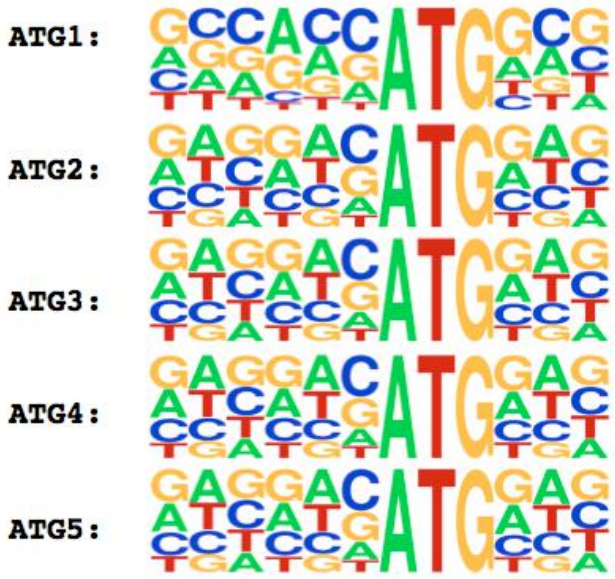


Figure 3.2.4 Analysis of Kozak consensus sequences
For this analysis in RStudio, 57,348 human cDNA sequences starting from 6 bases upstream of the annotated translational start codon were extracted using Biomart, and then split into letters grouped in threes (codons). Each triplet was assigned a

number consecutively starting from 1. The numbers corresponding to ATG triplets were extracted and the letters (bases) surrounding them corresponding to the -6, -5, -4, -3, -2, -1, +4, +5, and +6 positions were extracted for analysis. These sequences were entered into the WebLogo software (UC Berkeley), and the output frequency plots shown are size-proportional representations of the number of times that A, C, G, and T appear at the -6, -5, -4, -3, -2, -1, +4, +5, and +6 nucleotide positions surrounding each ATG codon, divided by 57,348, the total of genes included in the analysis. The first ATG codon has a unique pattern: G(C/G)CA(C/A)C for the -6 to -1 positions and GCG for the +4 to +6 positions. The ATG2 and ATG5 positions have nearly identical patterns of surrounding bases: (G/A)(A/T)(G/C)(G/A)(A/T)(C/G) for the -6 to -1 positions and (G/A)(A/T)(G/C) for the +4 to +6 positions, and the frequencies of the bases approach 0.25, suggesting there is perhaps less of a preference for specific bases at these surrounding positions. The corresponding sequences surrounding the ATG1 (p150-ATG) and ATG2 (p110-ATG) of ADAR1 are shown below for comparison. Of note, the p150-ATG is surrounded by a sub-optimal consensus sequence as compared to the p110-ATG, even though the annotated translational start codon for this open-reading frame is the p150-ATG.

The smaller proteins in the M296L and non-KOZAK mutants, which are not seen in the WT p150 sample, are likely produced by initiation of translation by ribosomes that would not have arrived downstream (or would have arrived at low levels) if translation started earlier at the M296 position. This observation, coupled with examination of the Kozak consensus sequences surrounding the p110 and p150 start codons, suggest leaky ribosome scanning as a possible mechanism that contributes to downstream translation of p110 from the p150 mRNA.

ADAR1 M296L and non-KOZAK mutants successfully suppressed p110 protein expression from the p150 mRNA, but these ADAR1 variants have non-synonymous mutations that may interfere with ADAR1 protein function. Furthermore, smaller ADAR1 variants were produced, as detected by the C-terminal specific ADAR1 antibody, and these smaller variants may also have dsRNA binding and editing abilities similar to that of p110, given that they differ, putatively, only in the beta-Z-DNA/RNA domain, since the first dsRNA binding domain does not appear until the very end of exon 2 and beginning of exon 3.

Of note, within exon 2, there are 6 other ATG codons aside from the p110-ATG in the same reading frame, and these could all be potential translation initiation sites for scanning 40S subunits that did not assemble translation complexes at upstream ATG codons. We therefore aimed to introduce synonymous mutations

in the region between the p150 and p110 start codons that could reduce leaky scanning and translation initiation at the p110-ATG and other downstream start codons.

Within the 885-base stretch of nucleotides between the first two ATGs in the main reading frame, there are two out-of-frame ATGs in a single alternate reading frame (Figure 3.2.5). During the process of leaky ribosome scanning on the p150 mRNA, there is a possibility that the full initiation complex assembles at a downstream methionine codon, which could be in the same or a different reading frame from the first AUG. Of note, AUG codons in alternate reading frames are possible start sites because scanning from the 5′ cap of mRNAs proceeds not in triplet fashion, as had been revealed by studies looking at the pattern of 40S-protected mRNA footprints (144). We reasoned that increasing the chance of translation initiation in an alternate reading frame would decrease the number of scanning 40S subunits that reach and assemble initiation complexes at the p110-AUG. To increase the chance of translation initiation in an alternate reading frame, synonymous mutations were introduced in the region upstream of the p110 start codon.

The changes are as follows: 1. create an additional start codon in the alternate reading frame in optimal Kozak context; 2. change the bases surrounding the two ATG codons in the alternate reading frame to increase likelihood of translation initiation at those start codons, and alter the bases surrounding the p110-ATG codon to decrease likelihood of translation initiation at that start codon; and 3. remove stop codons that would terminate early translation of the peptide in the alternate reading frame. Removing stop codons in the alternate reading frame can reduce likelihood of 40S scanning re-initiation when there is imperfect disassembly of translating monosomes upon encountering a stop codon (145). With stop codons up to the p110-AUG removed in the alternate reading frame, the monosomes will continue assembling the alternate peptide until a STOP codon is encountered downstream of the p110-ATG.

Immunoblot revealed a significant decrease in p110 protein levels when the three groups of mutations mentioned above were introduced together into the WT open-reading frame (Figure 3.2.5). Here, the synonymous mutations used to create the KOZAK-mut clone are limited by the availability of degenerate codons, and these changes appear to be less effective in attenuating production of p110 compared to what was observed in the non-KOZAK clone in figure 3.2.3, in which the most atypical sequences were introduced around the p110-ATG, necessitating non-synonymous mutations.

The sequence shown below (on the next page) was ordered from Gene Universal, and it arrived in a pUC57 plasmid, as was the case for the GFP-KI donor plasmid mentioned in chapter 2.

For this gene insert, the unique restriction sites SpeI and SalI (shown bolded in black) present in the recipient expression plasmid (pTRIP) were included along with flanking sequences so the insert, which includes the entire ADAR1 open-reading frame plus surrounding sequences, can be digested from the pUC57 plasmid and cloned into pTRIP for packaging into lentivirus. Of note, lentiviral particles used to transduce ADAR1 KO cells were made using WT 293T cells transfected with plasmids encoding retroviral envelope (VSVG) and structural/functional (Gag-Pol) genes, along with the pTRIP expression plasmid with the sequence of interest. The expression plasmids give rise to mRNA sequences with long terminal repeat sequences at the ends, and these mRNAs are packaged into lentiviral particles, which can reverse transcribe the mRNA into cDNA that will randomly integrate into the host cell genome during the infection (transduction) cycle.

The positions of synonymous mutations designed to reduce p110 expression are highlighted in bold and underlined, and also color-coded according to the mutation category. The p150-ATG and p110-ATG codons are shown bolded in green.

5'-CCTCTGCTAACCATGTTCATGCCTTCTTCTTTTTCCTACAG**ACTAGT**CCAGTGTGGTGGAATTCTGCAGATATC
AACAAGTTTGTACAAAAAAGCAGGCTTGGAAGGAGTTCGAACC**ATG**AATCCGCGGCAGGGGTATTCCCTCAGCGGAT
ACTACACCCATCCATTTCAAGGCTATGAACACAGACAGCTCAGATACCAGCAGCCTGGGCCAGGATCTTCCCCCAGT
AGTTTCCTGCTTAAGCAAATTGAATTTCTCAAGGGGCAGCTCCCAGAAGCACCGGTTATTGGAAAGCAGACACCGTC
ACTGCCACCTTCCCTCCCAGGACTCCGGCCAAGGTTTCCAGTACTACTTGCCTCCAGTACCAGAGGCAGGCAAGTGG
ACATCAGGGGTGTCCCCAGGGGCGTGCATCTCGGAAGTCAGGGGCTCCAGAGAGGGTTCCAGCATCCTTCACCACGT
GGCAGGAGTCTGCCACAGAGAGGTGTTGATTGCCTTTCCTCACATTTCCAGGAACTTAGTATCTACCAAGATCAGGA
ACAAAGGATCCTTAAGTTCCTGGAAGAGCTTGGGGAAGGGAAGGCCACCACAGCCCATGACCTGTCTGGGAAACTTG
GGACTCCGAAGAAAGAAATCAATCGAGTTTTATACTCCCTGGCAAAGAAGGGCAAGCTACAGAAAGAGGCAGGAACA
CCCCCTTTGTGGAAAATCGCGGTCTCCACTCAGGCTTGGAACCAGCACAGCGGAGTGGTTAGACCCGATGGTCATAG
CCAAGGAGCCCCAAACTCAGACCCGAGTTTGGAACCGGAAGACAGAAACTCCACATCTGTCTCAGAAGATCTTCTTG
AGCCTTTTATTGCAGTCTCAGCTCAGGCTTGGAACCAGCACAGCGGAGTGGTTAGACCAGACAGTCATAGCCAAGGA
TCCCCAAACTCAGACCCAGGTTTGGAACCTGAAGACAGCAACTCCACATCTGCCTTGGAAGATCCTCTTGAGTTTTT
GGATATGGCAGAGATCAAGGAGAAAATCTGCGACTATCTCTTCAATGTGTCTGACTCCTCTGCCCTTAATTTGGCTA
AAAATATTGGCCTTACCAAGGCCCGAGATATAAATGCTGTGCTAATTGACATGGAAAGGCAGGGGGATGTCTATAGA
CAAGGGACAACCCCTCCCATATGGCATTTGACAGACAAGAAGCGAGAGAGGATGCAAATCAAGAGAAATACGAACAG
TGTTCCTGAAACCGCTCCAGCTGCAATCCCTGAGACCAAAAGAAACGCAGAGTTCCTCACCTGTAATATACCCACAT
CAAATGCCTCAAATAACATGGTAACCACAGAAAAAGTGGAGAATGGGCAGGAACCTGTCATAAAGTTAGAAAACAGG
CAAGAGGCCAGACCAGAACCAGCAAGACTGAAACCACCTGTTCATTACAATGGCCCCTCAAAAGCAGGGTATGTTGA
CTTTGAAAATGGCCAGTGGGCCACAGATGACATCCCAGATGACTTGAATAGTATCCGCGCAGCACCAGGTGAGTTTC
GAGCCATCATGGAGATGCCCTCCTTCTACAGTCATGGCTTGCCACGGTGTTCACCCTACAAGAAACTGACAGAGTGC
CAGCTGAAGAACCCCATCAGCGGGCTGTTAGAATATGCCCAGTTCGCTAGTCAAACCTGTGAGTTCAACATGATAGA
GCAGAGTGGACCACCCCATGAACCTCGATTTAAATTCCAGGTTGTCATCAATGGCCGAGAGTTTCCCCCAGCTGAAG
CTGGAAGCAAGAAAGTGGCCAAGCAGGATGCAGCTATGAAAGCCATGACAATTCTGCTAGAGGAAGCCAAAGCCAAG
GACAGTGGAAAATCAGAAGAATCATCCCACTATTCCACAGAGAAAGAATCAGAGAAGACTGCAGAGTCCCAGACCCC
CACCCCTTCAGCCACATCCTTCTTTTCTGGGAAGAGCCCCGTCACCACACTGCTTGAGTGTATGCACAAATTGGGGA
ACTCCTGCGAATTCCGTCTCCTGTCCAAAGAAGGCCCTGCCCATGAACCCAAGTTCCAATACTGTGTTGCAGTGGGA
GCCCAAACTTTCCCCAGTGTGAGTGCTCCCAGCAAGAAAGTGGCAAAGCAGATGGCCGCAGAGGAAGCCATGAAGGC
CCTGCATGGGGAGGCGACCAACTCCATGGCTTCTGATAACCAGCCTGAAGGTATGATCTCAGAGTCACTTGATAACT
TGGAATCCATGATGCCCAACAAGGTCAGGAAGATTGGCGAGCTCGTGAGATACCTGAACACCAACCCTGTGGGTGGC
CTTTTGGAGTACGCCCGCTCCCATGGCTTTGCTGCTGAATTCAAGTTG**GTCGAC**CAGTCCGGACCTCCTCACGAGCC
CAAGTTCGTT

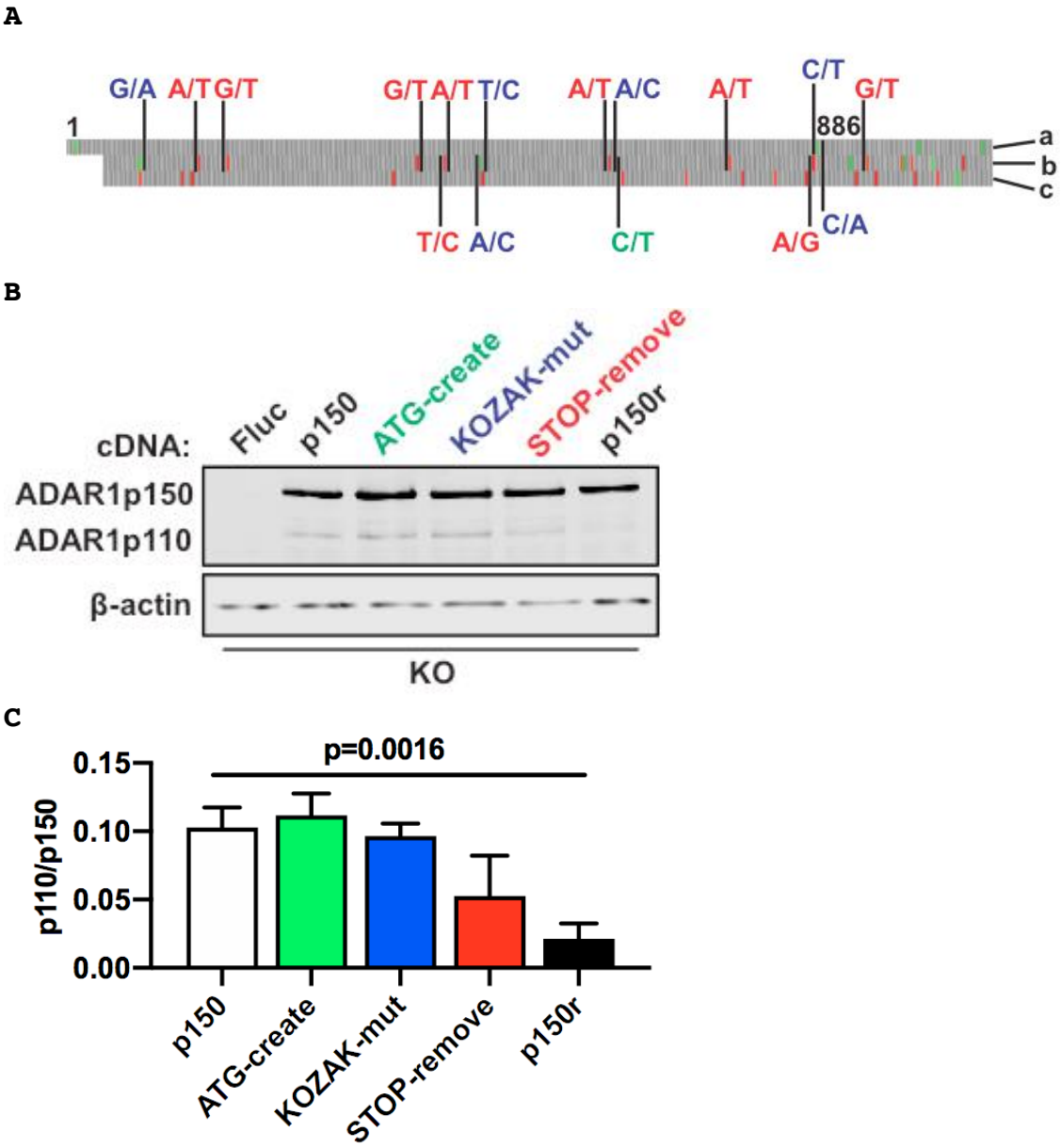


Figure 3.2.5 Synonymous mutations to reduce p110 levels
A. The start of the ADAR1 p150 sequence is shown labeled with nucleotides 1 (p150-ATG) and 886 (p110-ATG). Letter "a" refers to the primary ADAR1 open reading frame, while letters "b" and "c" refer to the two alternate reading frames. The mutation shown color-coded in green creates an additional start codon in the "b" reading frame. Mutations shown color-coded in blue weaken the consensus nucleotides surrounding the p110 start codon in the "a" reading frame and strengthen the consensus nucleotides surrounding the start codons in the "b" reading frame. Mutations shown color-coded in red remove the seven stop codons upstream of the p110-ATG and one stop codon downstream of the p110-ATG for the "b" reading frame. All mutations preserve

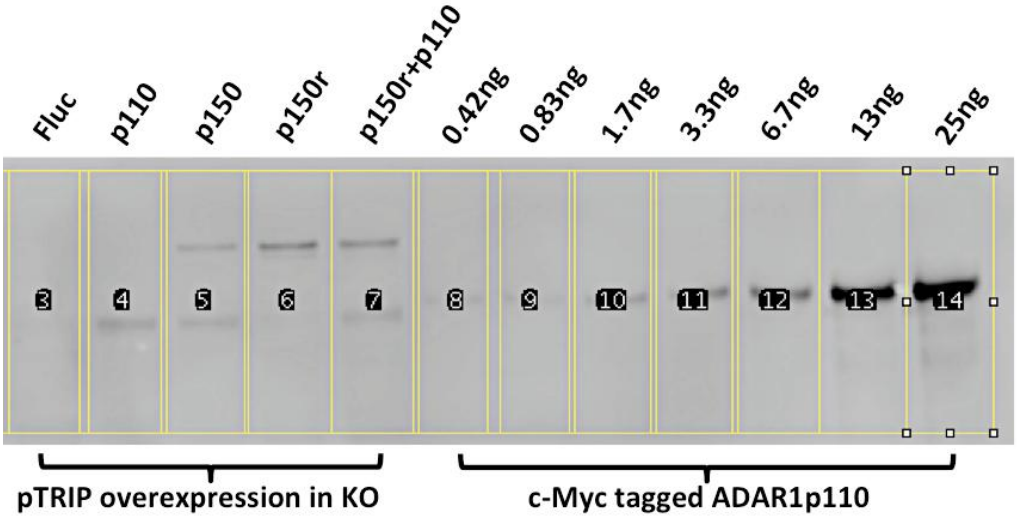
the protein coding sequence for the primary ADAR1 "a" reading frame.
B. The immunoblot shows ADAR1 KO cells that were transduced with lentivirus to create cell clones that have stable expression of various constructs: the firefly luciferase sequence (Fluc), the WT p150 sequence (p150), the sequence with the green color-coded mutation (ATG-create), the sequence with blue color-coded mutations (KOZAK-mut), the sequence with red color-coded mutations (STOP-remove), and the sequence with combined green/blue/red mutations (p150r).
C. This graph shows relative band intensities of p110 and p150 quantified in ImageJ. The p-value was calculated using an unpaired t-test comparing p150 and p150r.

Collectively, the data presented in this section suggest that leaky ribosome scanning on the p150 mRNA results in translation of p110 protein, and this mechanism can be significantly inhibited by making a series of changes to the sequence between the p150 and p110 start codons. Putatively, these changes result in increased ribosome activity in an alternate reading frame, and decreased desirability of the p110-ATG with regard to how likely a monosome is to assemble there. With the ability now to significantly reduce p110 expression and retain expression of WT p150, we next aimed to generate cell lines for editing analysis.

3.3 ADAR1 p150 and p110 isoform-specific cell lines

For experiments involving stable expression of transduced ADAR1 constructs, the exon 1B/1A/1C KO clone (ADAR1 KO) described in chapter 2 was used as the starting cell line. First, we were interested in developing an assay for measuring ADAR protein levels in the stably transduced cell lines. Recombinant ADAR1 protein (p110 isoform) was used to set up a standard curve to relate ADAR1p110 immunoblot band intensity levels as a function of protein mass, and this standard curve was adapted for both p110 and p150. Fitting a linear equation to the data points allows calculation of the approximate mass of protein that corresponds to any band intensity level (Figure 3.3.1).

A



B

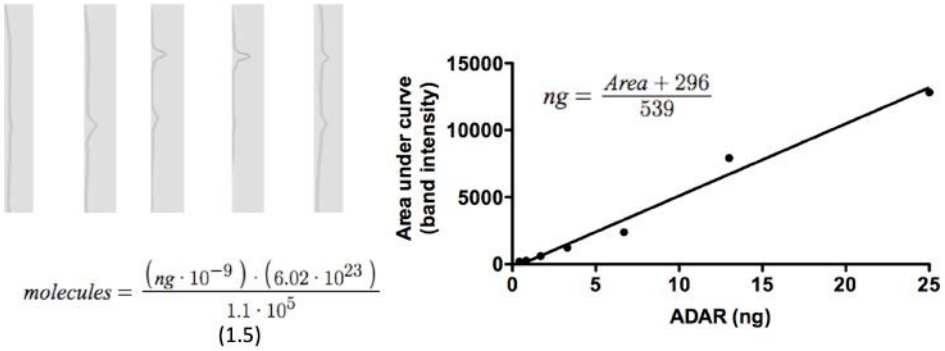


Figure 3.3.1 Standard curve to quantify ADAR1 protein levels
A. The immunoblot shows lysates from ADAR1 KO cells either stably transduced with various cDNA constructs or mixed with different amounts of recombinant c-Myc tagged p110 protein ranging from 0.42ng to 25ng. The yellow boxes around each lane were drawn using the ImageJ software, which then plots the signals from the top to the bottom of the boxes as lines, as shown in panel B.
B. The area of each peak (in arbitrary units) was calculated in ImageJ, and for peaks corresponding to the recombinant p110 protein lanes, the integral (area) values were plotted on the graph shown on the right side of panel B. A best-fit linear regression was applied, resulting in the equation shown inside the graph. The area values for the peaks corresponding to the overexpression clones could then be entered into the equation, and the mass of protein determined. The mass could then be converted to number of molecules using the molar mass and basic dimensional analysis. Finally, because each well was loaded with known numbers of cells (approximately 1/10 the volume of a 120,000-cell lysate mixture), we can approximate the ADAR1 molecules per cell for each sample.

Of note, one of the complications with using overexpression to produce ADAR1 and then doing A-to-I editing analysis is that the levels of protein may be different from that in WT cells, or the ratios of the levels of p110 and p150 isoforms may be off. We thus aimed to examine how the protein levels of ADAR1 isoforms compare between WT 293T cells and the ADAR1 KO cells stably transduced with various expression constructs to reconstitute ADAR1.

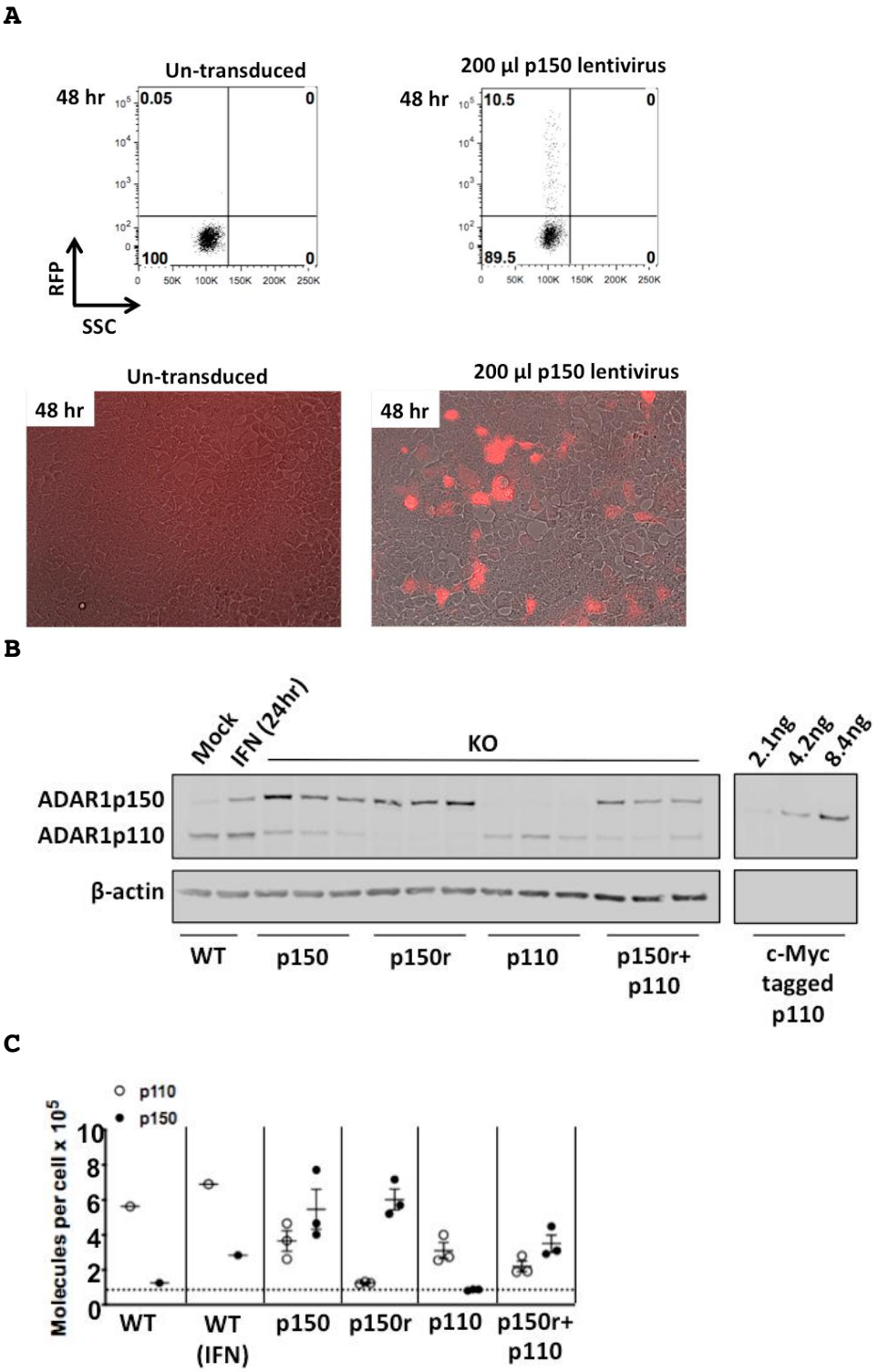


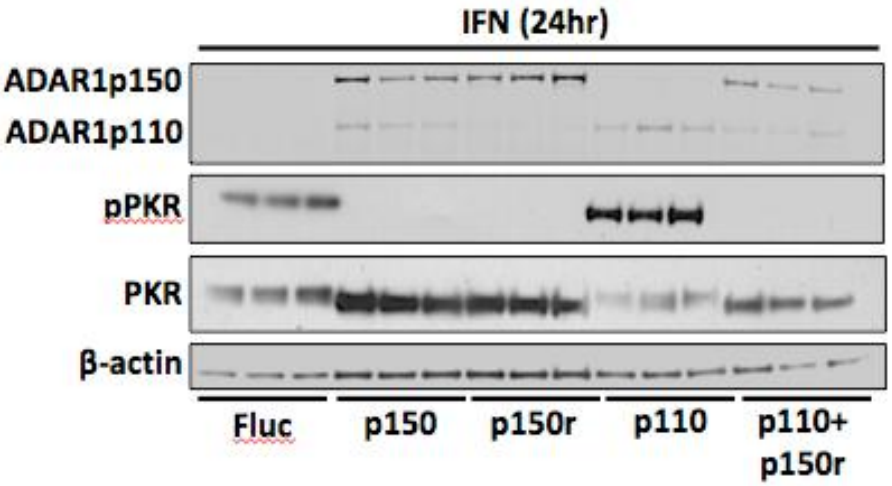
Figure 3.3.2 Generation of clones for editing analysis

A. During the process of generating ADAR1 overexpression clones, lentivirus titers were selected to create about one integration event per cell, which correlates to approximately less than 20% transduced in a population of 293T cells, although this number

can vary depending on the transduction conditions and cell type (146–148). Shown here is quantification of transduced RFP+ cells (the pTRIP expression system has an EMCV-IRES driving expression of RFP reporter protein), showing about 10% transduction for the indicated volume of p150-carrying lentivirus. This population of cells was then sorted, and 3 single-cell colonies were selected following immunoblot analysis of sorted single-cell clones. The process was then repeated for the other ADAR1 cDNA variants, and clones with similar levels of protein expression were selected to represent each of the four groups that will be included in the editing analysis.
B. The immunoblot shows ADAR1 levels in WT 293T cells with and without interferon, and ADAR1 KO cells transduced with p150, p150r, p110, and p150r+p110, all driven by the CMV-promoter in the pTRIP expression system. The recombinant p110 protein lanes shown in the right side of panel B were prepared as described in figure 3.3.1 and was used to generate the standard curve for determining the molecules per cell values for the lanes shown to its left.
C. ADAR protein levels are plotted, showing the comparison between endogenous and exogenous expression, and between the different overexpression clones. The p150 and p150r clones had, on average, a p150 expression level about twice that of the WT cells with interferon treatment. On average, the p150 and p110 clones had a p110 level about half that of WT cells.

As a final note to chapter 3, and as an extension to the work described in section 4 of chapter 2, the p150r group was analyzed in terms of the ability of p150 to suppress activation of PKR in the presence of interferon. Ribosomal subunits and polysomes were also profiled to determine the extent of translational activity in these cells with and without interferon treatment (Figure 3.3.3).

A



B

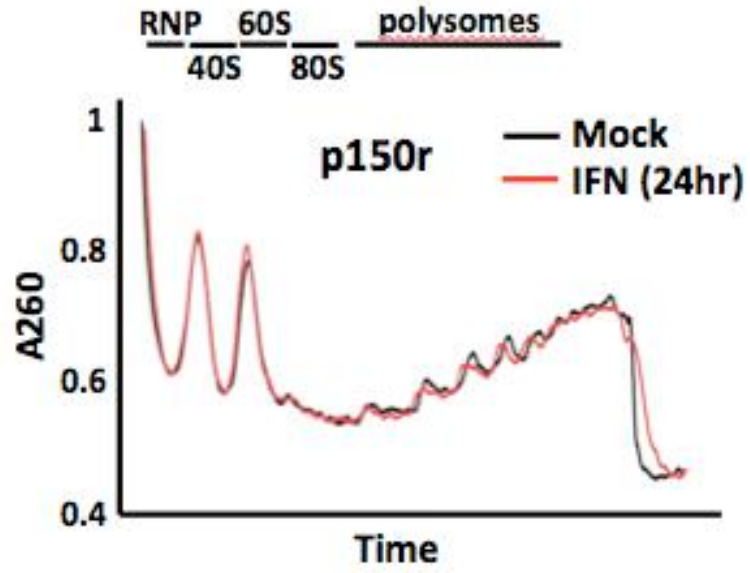


Figure 3.3.3 Interferon response in 293T cells
A. The immunoblot shows expression of ADAR, PKR, and phosphorylated-PKR in the 5 groups of clones with interferon treatment. Of note, activation (phosphorylation) of PKR occurs only in the clones that are lacking p150.
B. The graph shows absorbance values at 260nm corresponding to monosome and polysome fractions in one p150r clone with and without interferon treatment. Briefly, for polysome profiling, cells were harvested and lysed using polysome lysis buffer (Tris-HCl, NaCl, $MgCl_2$, DTT, Triton X-100, cycloheximide), and 10% and 50% sucrose gradients were prepared using Tris-HCl, KCl, and $MgCl_2$, DTT, cycloheximide, and SUPERase RNase inhibitor. Cell lysate supernatants were added to the top of the sucrose

gradient, and the top layer was passed through the 10% and 50% sucrose gradients, respectively, by ultracentrifugation for 2 hours at 38,000RPM. Following the centrifugation, the cell lysate supernatants were spread out in the sucrose gradient, and the mixture was slowly displaced into the polysome-profiling machine using 60% sucrose solution, and the 260nm absorbance values for the different fractions were plotted as a function of time using the TracerDAQ software. The absorbance peaks correspond to the different fractions present in the supernatant/sucrose mixture: free ribonucleoprotein (RNP) components, 40S subunits, 60S subunits, 80S monosomes, and finally polysomes.

Of note, we observed that p150 is necessary and sufficient to suppress PKR activation and translational shutdown during the interferon response, at least in 293T cells. It is possible that p150- and p110-targeted sites are both unique and shared, and the induction of both isoforms during the interferon response leads to optimal editing in the cytoplasm and nucleus, in which the p110 isoform is present predominately.

As one study recently suggested, the formation of potentially immunogenic dsRNA structures are often between intron-intron and intron-exon intramolecular (or even intermolecular) pairings, which often result from pairings of inverted Alu elements: for example, an intron or 3'UTR exon sequence that includes an Alu element, an intervening sequence, and then the same (or very similar) Alu element, but in reverse-complement orientation (149). These pairings could be destabilized by p110 prior to splicing, or the splicing ribonucleoproteins themselves could bind intron sequences and reduce double-strandedness prior to excising the introns.

Exon-exon dsRNA pairings that persist after splicing and that remain unedited by p110 would be potential targets for p150, especially if the dsRNA has strong Z-conformation motifs, which the p150 Z-alpha domain binds with high affinity (122). Although p110 is the predominant isoform in the nucleus, p150 can be present in both nucleus and cytoplasm; the p150 amino acid sequence contains both nuclear localization and export signals (150, 151).

Classifying the sites that are shared by both isoforms, or ones in which editing is done selectively by one isoform, will be a starting point to better understanding whether p110 and p150 play unique or redundant roles. To this end, the next chapter will present a potential taxonomy for the global ADAR1 editome.

CHAPTER 4. Taxonomy of ADAR1 A-to-I edit sites

The fourth chapter of this thesis aims to 1. describe how the cell clones selected for editing analysis will expand on knowledge within the A-to-I editing field; 2. describe the editing analysis workflow following RNA sequencing and show the proposed taxonomy of putative A-to-I edit sites; 3. discuss biological implications associated with classifying A-to-I edit sites.

4.1 Clones for A-to-I editing analysis

With the ability to express p150 with significantly reduced levels of p110, we now have the ability to start addressing questions such as: 1. Is the total A-to-I editome more or less the sum of the p110 and p150 editomes, with respect to edit positions? 2. To what extent do p110 and p150 play redundant roles with regard to A-to-I editing? Are there unique roles for p110 and p150 with regard to RNA binding and/or editing at baseline conditions, under certain states of cellular stress (152), and during the interferon response? We have already observed the necessity and sufficiency of p150 with regard to suppressing the beta interferon-induced translational shutdown mediated by PKR activation and eIF2α phosphorylation, although this experiment was done in 293T ADAR1 KO cells that are overexpressing high levels of p150.

Using a human cytomegalovirus (CMV) promoter-driven expression system, ADAR1 constructs (p150, p110, and p150r) and firefly luciferase (fluc) control constructs were stably integrated into ADAR1 KO cells using lentiviral delivery, as described in the final section of chapter 3. Of curiosity, to investigate how A-to-I editomes would potentially differ when p150 and p110 are translated from separate mRNA molecules (as compared to p150, in which one mRNA has the potential to give rise to both protein isoforms), one p150r clone was selected for a second round of transduction with p110 to create a cell line that also expresses p150 and p110 protein, but from separate mRNA molecules (Figure 4.1.1).

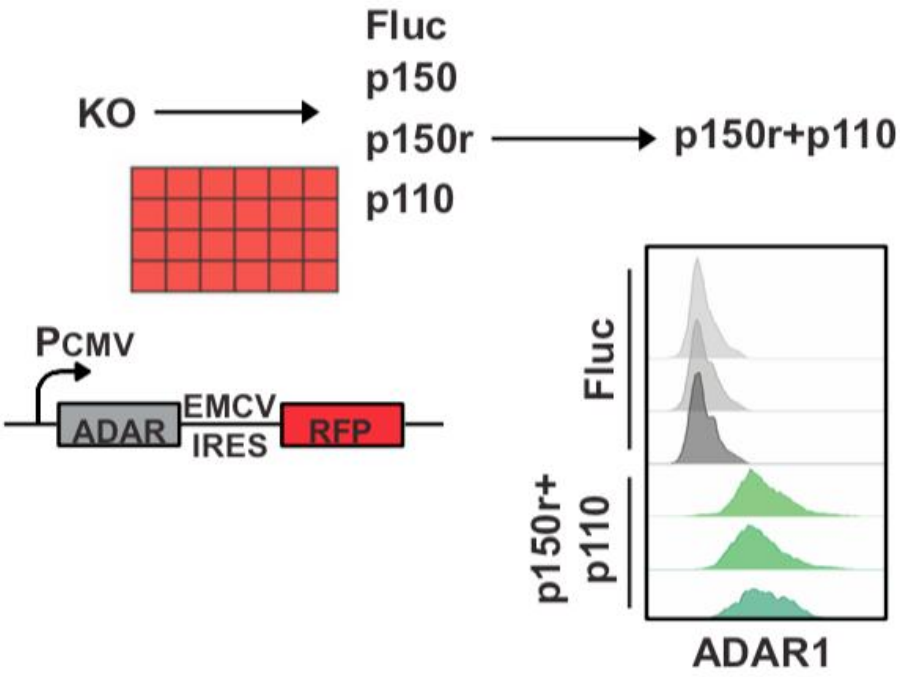


Figure 4.1.1 Overexpression system in ADAR1 KO cells
Four different groups were created to help classify the A-to-I editome: p150, to describe the editome when both isoforms are present and have the potential to be expressed from the same mRNA; p150r, to describe the editome when p150 is present with significantly reduced p110; p110, to describe the editome when only p110 is present; and p150r+p110, to describe the editome when both isoforms are present but translated from separate mRNA molecules. The CMV promoter drives transcription of stably integrated transgenes, which encode the ADAR1 isoform and an RFP reporter protein via IRES-mediated translation initiation. The RFP enables single-cell sorting using fluorescence signals. The 1B/1A/1C KO clone (ADAR1 KO) as described in chapter 2 was used as the starting cell for transduction of each of the four cDNA constructs. Only one group, the p150r group, went through a second round of transduction and single-cell cloning to create the p150r+p110 group, in which cells have both p150r and p110 transgenes. To analyze ADAR1 expression levels independent of RFP signal (an indirect representation of ADAR1 protein), cells from the p150r+p110 and fluc groups were incubated with rabbit anti-ADAR1 antibody optimized for flow cytometry, and the anti-rabbit secondary antibody mean-fluorescent intensity histograms are shown.

For the groups mentioned above, clones were selected in triplicate, partly to address the issue of error rates during reverse transcription of RNA in first step of library preparation (barring the use of error-correction techniques such

as rolling-circle RT). In one study, the error rate of murine leukemia virus RT was found to be 1/37,000 (153). The RT used in all reverse transcription reactions, unless other stated, is the Invitrogen Superscript III version, which is a genetically modified version of the Moloney murine leukemia virus RT (154). The error rate of Superscript III has been determined to be between 1/15,000 to 1/32,000, depending on temperature and other variables (155).

Downstream of the RT step, PCR errors (rare), sequencer base-calling errors, and unequal amplification of cDNA fragments are other concerns that can be addressed either in the variant calling workflow or by incorporating unique molecular tags during the RT step, as will be discussed as an ongoing effort in chapter 5.

4.2 Editing analysis workflow and taxonomy of sites

Libraries were prepared using the Illumina TruSeq Stranded Total RNA kit, with two modifications to the standard protocol to enrich for larger fragment sizes: reduction of fragmentation time to 2min at the 94°C incubation step with fragmentation buffer (likely some type of NaOH-containing solution); and following adaptor ligation, reducing the volume of AMPure XP PCR cleanup beads to 60% of the standard amount. RNA libraries were pooled, diluted, and sequenced using the 150-base paired-end sequencing option on the Illumina NextSeq 500 High Output and NovaSeq S1 flow cells. Base calling data stored in BCL files were converted into FASTQ files and demultiplexed using bcl2fastq Conversion Software at The Rockefeller University Genomics Resource Center. The NextSeq and NovaSeq sequencing results were merged and aligned to hg19 using STAR 2.5.4b with 2-pass mapping (156–158). The alignment algorithm allows up to 3 mismatches for each 22-nucleotide region and considers the mean insert sizes of the RNA library fragments when determining if a read pair is concordant.

Unique and concordant read pair mappings were selected for variant calling. For each sample, Picard 2.18.1 was used to calculate total read counts, Phred quality scores, and alignment percentages. Aligned SAM files were converted into BAM files using SAMtools 1.9. Mutect2 from GATK 4.0 was used to identify reference-read mismatches for each of the 15 samples (5 groups in biological triplicates: firefly luciferase, p150, p110, p150r, and p150r+p110).

Genomic positions with a unique mismatch type (one of 12 possibilities amongst the four bases) were considered for editing analysis. As a starting point, positions with a total read count of greater than or equal to 5 in all 15 samples were selected for ADAR1-editing analysis. A mismatch site was called for a group if the mismatched nucleotide read count at that site

is greater than or equal to 2 in each of the biological triplicates. A value of 2 was selected here based on both the distribution of read depths across selected genomic positions and the average error rate of Illumina sequencing platforms (Figure 4.2.1). For example, one study found the Illumina HiSeq had an error rate of 3/1000 in terms of mismatches (94). Variations on the "total depth=5" requirement change the total numbers of sites identified, as will be discussed in chapter 5.

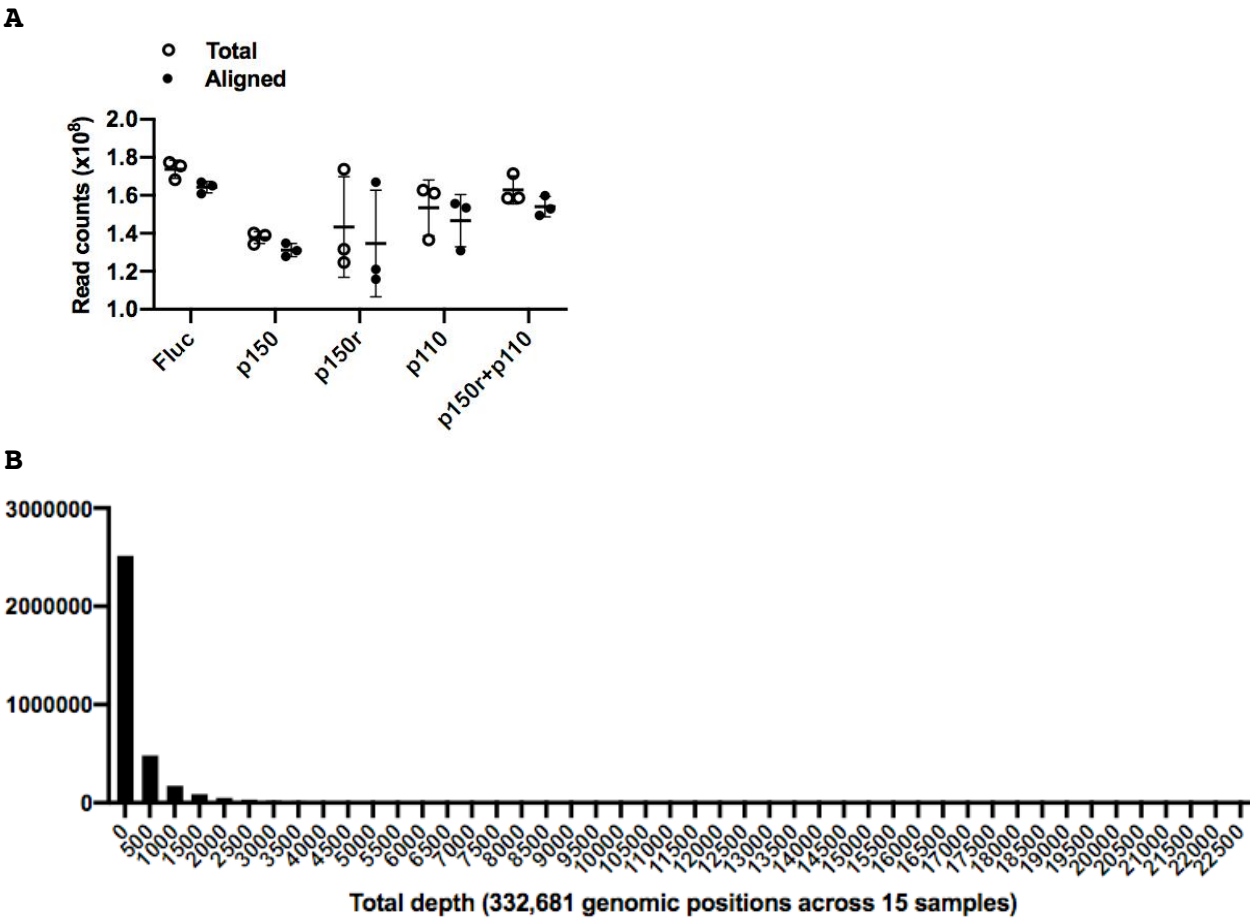


Figure 4.2.1 Alignment and mismatch calling
A. Following total RNA extraction, library preparation, and sequencing, alignment and variant identification were done using STAR and GATK, respectively. On average, between $1x10^8$ and $2x10^8$ reads per sample successfully aligned to hg19 with STAR two-pass alignment.
B. Following identification of genomic positions with single mismatches using GATK, the read depth at each position was extracted, and a histogram was generated showing the numbers of positions that have read depths corresponding to the range of depths across all extracted genomic positions.

The distribution of read depths was considered along with the Illumina sequencing error rate (about 1/1,000, a Q score of 30, for up 75% of all called bases), to decide on a cutoff value for read depth of the alternate base in order to call a mismatch at that genomic position. Most of the identified 332,681 positions (coverage of at least 5 in all 15 samples) have read depths of less than 1000. For the positions with a read depth of greater than 1000, there is an increased chance of having a mismatch called at that site because of a sequencing error. This is a broad generalization, because there is the same 1/1,000 chance of calling a wrong base for the first cDNA fragment containing the position in question as the 1000th cDNA fragment containing that position.

We reasoned that the chance of having two mismatches called in error at the same position is low, particularly for sites with coverage of less than 1,000 total reads, taking into account the average sequencing error rate of 1/1,000. Thus, we settled on a value of "2" for read depth of the alternate base for calling a mismatch. Of note, genomic positions are discussed in reference to cDNA sequences, as the RNA samples were treated with DNase I for 30 minutes to remove as much genomic DNA contamination as possible prior to the RT step. We decided to align reads to genome rather than transcriptome to increase the range of possible alignments, including possible 3' extended RNA isoforms and independently transcribed repetitive elements that may not be well annotated in transcriptomic datasets. Of note, strand information can be added later in the workflow (based on P5/P7 read information contained in the stranded library, as described in figure 2.1.7).

Additional filters were applied to identify potential ADAR1-edit sites from the list of mismatches. For each called site, a mean mismatch frequency=$\Sigma_G/\Sigma_{(A+C+G+T)}$ was calculated based on summed read counts from the biological triplicates in each group. Next, to create a list of potential ADAR1-edit sites, mismatch sites were selected from any one of the four groups: p150, p110, p150r, or p150r+p110, and their edit frequencies were compared with edit frequencies of the same sites in the fluc group. To pass filter here, the selected site must have a mean mismatch frequency greater than 2 standard deviations above the mean mismatch frequency of that site in the fluc group. The goal of this step is to account for both background editing in the fluc group (from other RNA-editing enzymes such as APOBECs) and also to filter out SNPs (Figure 4.2.2).

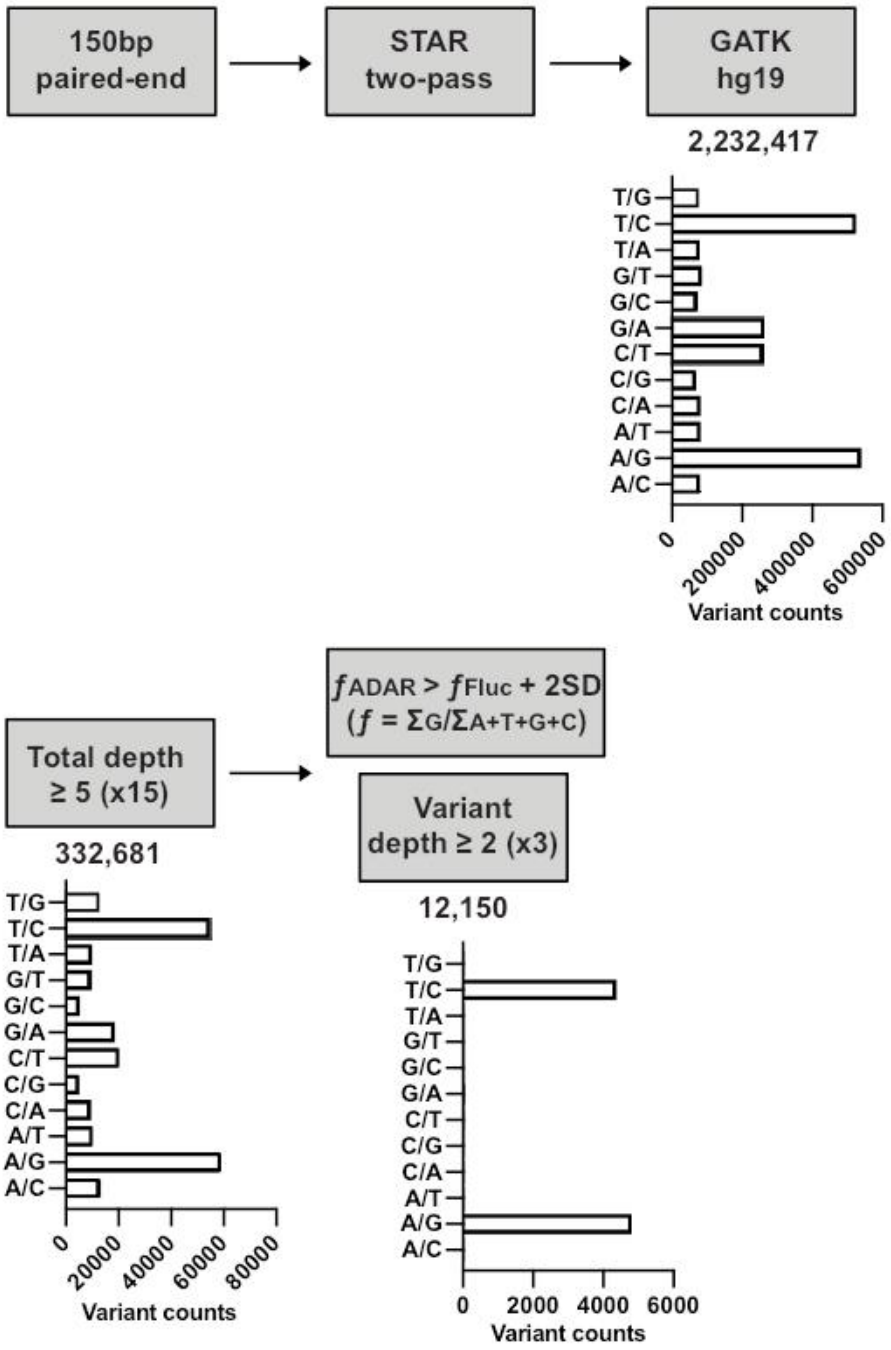


Figure 4.2.2 Identification of potential ADAR1-edit sites
Shown here is a schematic of the analysis workflow, starting with 150bp paired-end sequencing, alignment, and variant calling, which identified over 2 million unique mismatches between the cDNA and hg19 human genome build. To narrow down the list of mismatches to sites that can be compared across all 15 samples, only genomic positions with coverage in all samples were selected, and this narrowed the list of sites by about one log-fold (332,681), when a minimum depth of 5 was required in all samples. Next, selected mismatches were further filtered by

the requirement that they are present in all biological triplicates of a group with a mismatch depth of at least 2, to minimize the chance of calling sequencing errors. Next, the mismatch frequencies for the sites from the ADAR1 groups were calculated and compared with the mismatch frequencies of the same sites in the fluc group, to filter out background editing and SNPs. At this point, 12,150 potential ADAR1-edit sites were identified, and nearly all mismatches were of the A-to-G and T-to-C types, as expected before the addition of strand information.

There remains the possibility that in the case of heterozygous SNPs, only one chromosome, either containing or not containing the SNP, is expressed in some but not all samples in the groups of biological triplicates. Such variants would not be filtered out if they happened to not be expressed in the fluc samples, but they would not represent true ADAR1 edit sites. Therefore, genomic positions found in the dbSNP138 database were excluded from further analysis. Furthermore, sites that mapped to the ADAR1 coding sequence region were excluded, because many of these mismatches corresponded to the p150r mutations created to reduce p110 expression (Figure 4.2.3).

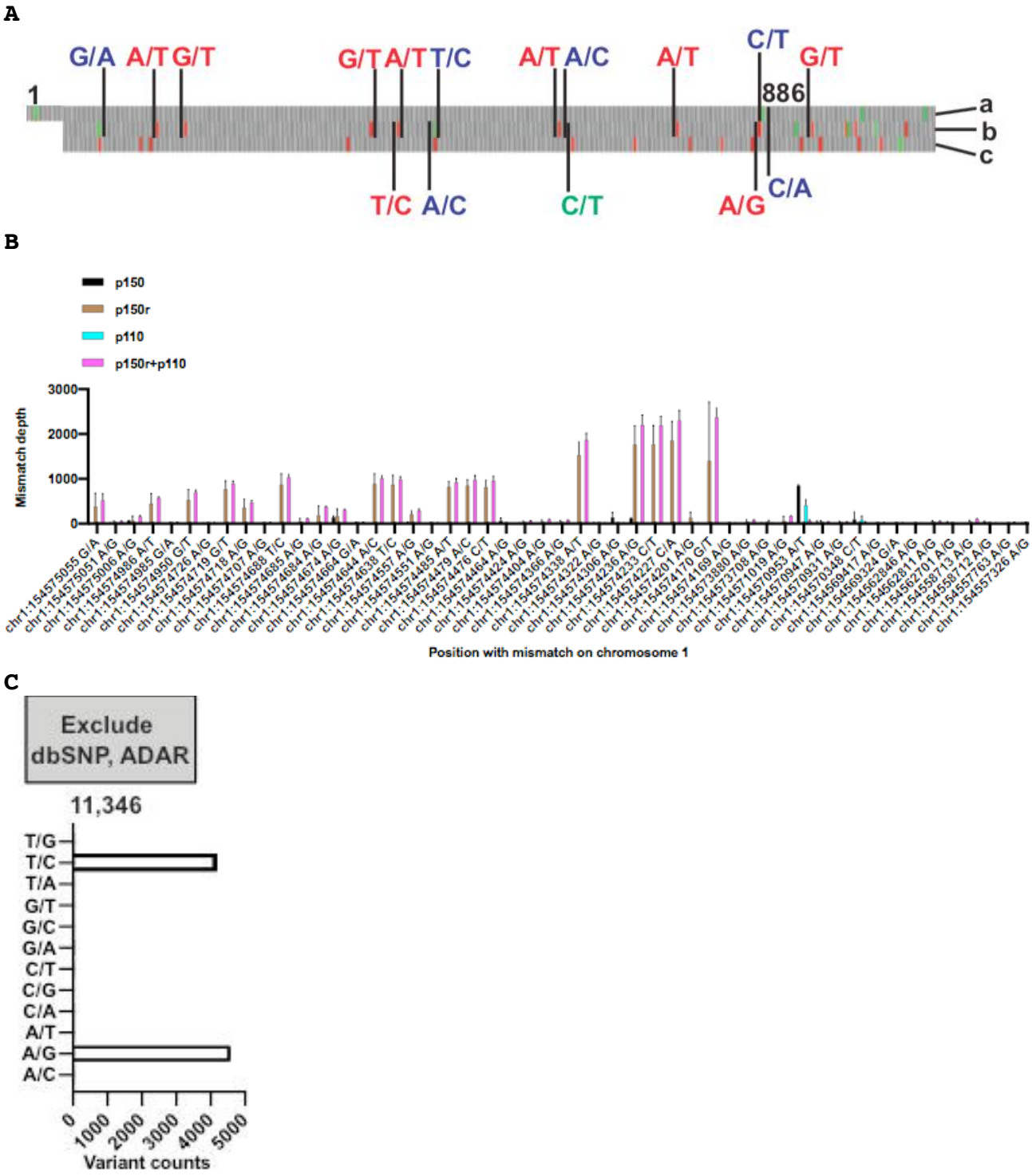


Figure 4.2.3 Additional filtering criteria

A. The schematic here, as seen previously in chapter 3, shows the mutations introduced to reduce p110 expression from the p150 reading frame.

B. Panel B shows read counts tallied up for the mismatch sites taken from a subset of the 12,150 sites identified from the

filtering strategy as described in figure 4.2.2. The subset contains all sites that mapped to the ADAR1 locus, and the graph shows their read depths. Mismatches with the highest read counts correspond to nucleotides on the mRNA molecules that were expressed from the stably integrated p150r transgene, as can be seen for the p150r and p150r+p110 samples. Because these sites are not true ADAR1-edit sites, they were excluded from the set of mismatches.
C. Sites from the dbSNP138 database were also excluded from further analysis, further narrowing the list of potential ADAR1-edit sites to 11,346. At this point, strand information has not been added to the genomic positions.

Importantly, strand information must be added when finalizing the list of A-to-I edit sites because there are A-to-G(I) genomic variants located in gene regions where transcription occurs in the antisense direction; these sites would therefore be T(U) instead A on the mRNA and would not be ADAR1 targets. By the same logic, T-to-C genomic variants located in a gene region where transcription occurs in the sense direction are also not true ADAR1-edit sites. Applying the strand information resulted in identification of 9,044 A-to-G(I) ADAR1-selective sites.

To classify the list of putative ADAR1 editing sites, a site was called as “PRESENT” for a group if the read count for G at that site is greater than or equal to 2 in each of the biological triplicates, and “ABSENT” for a group if the read count for G at that site is equal to 0 in each of the biological triplicates. As a result of this requirement, some sites were unable to be classified, because the read counts for G at those sites were greater than or equal to 2 in only one or two of the biological replicates in a particular group (Figure 4.2.4).

Using these classification filters, we identified 1,365 potential sites that are shared by both isoforms (present in both the p150r and p110 groups); 1,733 potential sites that are p150-selective (present in the p150r group and absent in the p110 group); 138 potential sites that are p110-selective (present in the p110 group and absent in the p150r group); and 277 potential sites that are p150/p110-selective (absent in both the p110 and p150r groups). The sites were annotated with information from Repeatmasker about repetitive elements, including Alu elements. Of note, Alu derives its name from the Arthrobacter luteus bacterial restriction endonuclease, which recognizes a target site in a class of repetitive elements found in human (about 11% of the human genome) and nonhuman primate genomes (159–162). Edit sites were also annotated with information regarding the function of the sequences that contain the sites.

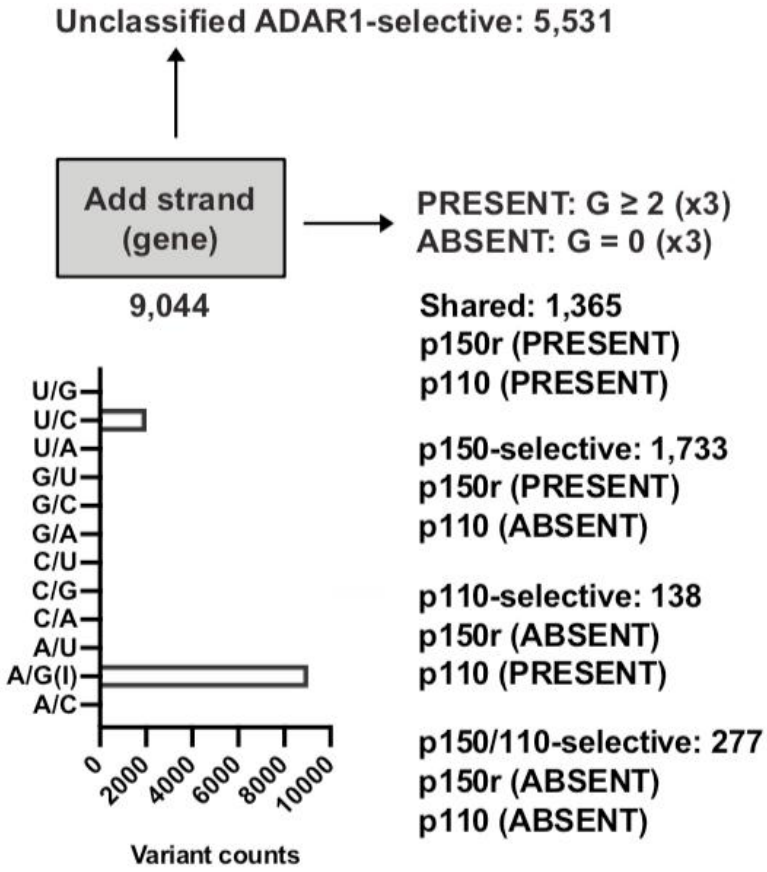


Figure 4.2.4 Classification of potential ADAR1-edit sites
Strand information was added to the list of mismatches, resulting in identification of 9,044 potential ADAR1-edit sties, which were further classified as sites shared between p110 and p150, p150-selective sites, p110-selective sites, and p150/110 double-selective sites. Of the sites that could be classified, most are either shared between p110 and p150 (39%) or p150-selective (49%). A minority of sites, putatively, are p110-selective (4%) and p150/p110-selective (8%), or in other words, dependent on p110 and p150 to both be present in order for editing to occur.

Of note, reliability in assigning ABSENT for a site in a particular group depends on a low false-negative rate, which can be hard to ascertain with bulk RNA sequencing, bulk alignment, and bulk variant calling. For instance, the combination of potential PCR bias and low coverage at repetitive Alu elements, where most edit sites are located, makes very possible that certain regions with sites that have low A-to-I edit frequencies might be determined to be entirely unedited in the final workflow. As discussed in chapter 2 (figure 2.1.7), the reduced coverage at Alu repeat regions is thought to result from a

combination of the RT-based library preparation method, lowered sequencing quality at homopolyer regions, and challenges in alignment.

With regard to the percentage of sites that classify in the four groups, we can create one possible scenario as a thought-experiment: an arbitrary false-negative rate of 4%, which could mean that 4% of sites determined to ABSENT in a group are falsely classified as ABSENT. This could mean that all the p110-selective and p150/p110-selective sites (which have a double ABSENT requirement) are mistakenly assigned to their respective groups. Keeping the same scenario, of the sites classified as p150-selective (49% of all classified sites), about 4% would be expected to be misclassified. The shared sites, because of their double PRESENT requirements, depend solely on a low false-positive rate. Because most edit sites have edit frequencies well below 50%, the chance of a false positive call is likely much lower than a false negative call, and in particular when taken into consideration the issue of low coverage at many putative edit sites.

One way to increase coverage on a lower-throughput level is to do amplicon sequencing, and this can be coupled with degenerate molecular tags added to gene-specific primers during the RT step. During the analysis workflow, reads corresponding to PCR duplicates will have the same barcode and can be counted as a single read that corresponds to a single original RNA molecule. This method still remains vulnerable to RT errors and RT stalling at structured Alu elements, but it will allow more insight to be gained about how much the issues of coverage and PCR bias contribute to false negative rates with respect to identifying and classifying A-to-I edit sites. Ongoing work using this method will be presented in Chapter 5.

Looking at the A-to-I editome across the samples, a few trends appear. Globally, the median length of edited genes is significantly longer than that of unedited genes. It is possible that longer genes have longer 3’UTRs and introns, both of which contain Alu elements, the primary target for A-to-I editing. Additionally, the samples that included p150, including p150 alone, had overall more target sites compared to when p110 is alone. However, median edit frequencies (defined as depth of G divided by total depth) remained similar across the samples, suggesting that p110 alone is as capable of editing efficiently as p150 alone (Figure 4.2.5).

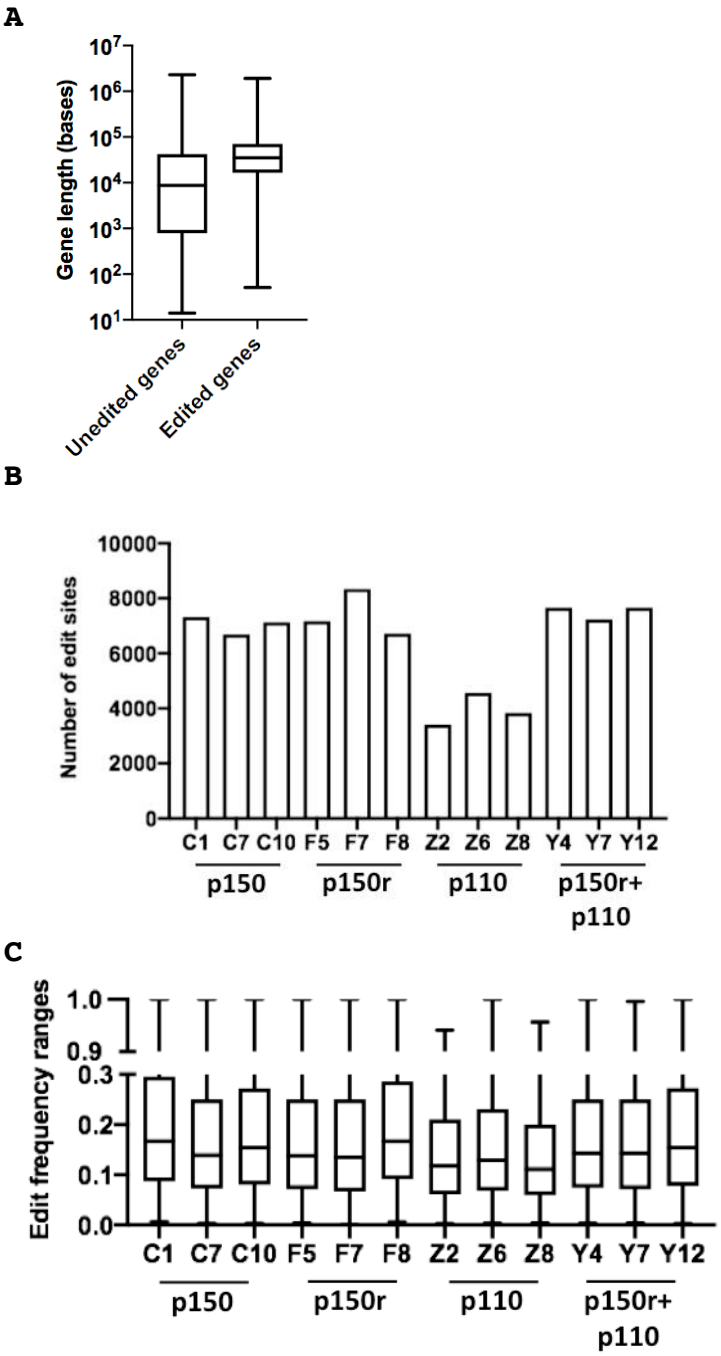


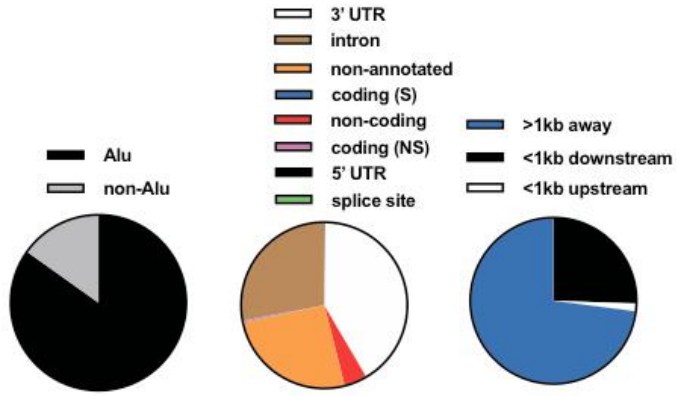
Figure 4.2.5 Global trends in editing

In panel A, gene lengths, including exons and introns, were extracted from Biomart for two lists of genes, one with ADAR1 edits and one without. Edited genes have a significantly longer median length compared to genes not targeted ADAR1. Panel B shows counts of all edit sites tallied up across the 15 samples, and panel C shows the median edit frequencies of all sites counted in panel B.

With regard to annotation, the majority of sites (85%) are located in Alu repeat elements, and a large proportion of sites map to intergenic (non-annotated RNA) regions (26%), introns (28%), and 3′ UTRs (41%). Of note, 1,389 sites were unable to be annotated because they were located in clusters of long genes in the same orientation, and the overlapping of sequenced cDNA inserts during alignment was insufficient to determine which gene the reads containing those sites originally came from. Techniques such as single-molecule sequencing would enable resolution of this issue.

A predicted independent structure of the AluSq2 repeat element located in the 3′ UTR of the ATM gene reveals that edit sites tend to cluster near loops adjacent to double-stranded regions rather than be located within long-stretches of perfectly complementary strands. It is possible that the binding domains attach to dsRNA structures adjacent to loops, allowing the deaminase domain to then access the adenosines found in the loops (61). Of course, molecules are fluid from a thermodynamic perspective, so RNA may shift back and forth from one predicted structure to another depending on the conditions in cells, and this could influence which sites are edited and when (Figure 4.2.6).

A



B

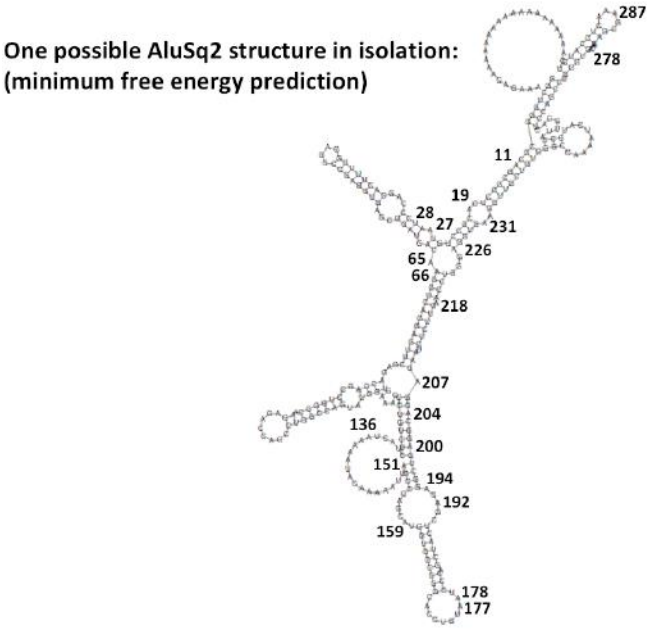


Figure 4.2.6 Annotation of global ADAR1-edit sites

A. Pie charts show annotation of the 9,044 sites, using Repeatmasker for the Alu annotation and RefSeq-defined gene functions in ANNOVAR for the annotation of other gene features. 84% of sites are located in Alu repeats; 16% are located in non-Alu regions, and could either be in other types of repetitive elements or non-repetitive elements. 42% of sites map to 3'UTR regions; 28% to intron regions; 4% to non-coding RNA; 0.2% to coding regions that lead to non-synonymous (NS) mutations and 0.1% to coding regions that lead to synonymous (S) mutations; and 0.02% to known splice acceptor sites in introns (of note, ADAR1 has the potential to create a GU splice donor site or create/remove an AG splice acceptor site). The remaining 26% of sites map to genomic regions that lack RefSeq annotations, although the extensive Ensembl database may have more genomic annotations corresponding to predicted RNA gene products that would allow further annotation of the remaining sites. The non-

annotated portion shown in orange was analyzed further by looking at their locations relative to neighboring genes. Of the 2,356 non-annotated sites, 26% are located less than 1kb downstream, 1% are located less than 1kb upstream, and 73% are located more than 1kb away from a neighboring gene, either upstream or downstream. Of note, the 26% of sites located within 1kb downstream of a gene could be from gene isoforms with un-annotated 3'UTR extensions (163).
B. Shown here is the AluSq2 sequence from the 3'UTR region of the ATM gene, with a structure prediction using the minimum free energy function from the Vienna RNAfold package. Edit sites are labeled with numbers corresponding to bases 1-323 in the Alu sequence, and they tend to cluster in or adjacent to loop structures. Of note, the 3'UTR of the ATM gene has an AluSc8 element in reverse-complement orientation about 1kb downstream of this AluSq2 element. While the independent structure shown above could exist within the 3'UTR, another possibility is formation of an inverted repeat structure between the two Alu elements. The inverted repeat is another classic target for ADAR1 editing and often has loops present in it because of imperfect complementarity between the Alu elements.

Despite the challenges associated with classifying A-to-I edit sites using bulk RNA sequencing, we propose in the final section of this chapter to discuss some trends with regard to how the sites classify, and also the biological implications of isoform-selective and isoform-shared editing.

4.3 Implications of an A-to-I editing taxonomy

Within the sorting scheme shown in figure 4.2.4, p150-selective sites were found to be largely located in 3' UTRs (69%), while p110-selective sites were mostly found in introns (71%). Considering that p110 is predominately found in the nucleus, its selectively edited sites would be expected to be located in intronic Alu repeat elements, which predominate in the nucleus. One study concluded that the average pre-mRNA has 8.8 exons and 7.8 introns, and about 80% of the exons have less than 200 bases. Most introns are between 20 and 11,000 bases in length, and less than 10% have more than 11,000 bases (164).

Although p150 has the potential to shuttle between nucleus and cytoplasm, the relative amounts of p150 and p110 in the nucleus seem to favor p110-mediated editing. Consistent with this, 71% of p110-selective sites are in introns and only 11% of p150-selective sites map to introns. After splicing, Alu elements present in 3'UTRs would be targets for p150-editing in the cytoplasm. Consistent with this, 69% of p150-selective sites are located in 3'UTRs. Of the sites that are potentially shared independently by p110 and p150, 51% map to 3'UTRs and 30% map to

introns (Figure 4.3.1). These shared sites could be present on both pre-mRNAs and spliced mRNAs, and the sites not located in intron regions that remain unedited by p110 in the nucleus could become targets for p150 in the cytoplasm.

For editing in the nucleus, one can imagine potential competition between the p110 and p150 isoforms for the shared target sites. The relatively higher levels of p110 compared to p150 would seem to favor p110 editing, but p150 has the unique high-affinity Z-alpha binding motif, allowing p150 to bind Alu repeats in Z-conformation with higher affinity than p110. These factors may allow a balance to be achieved between p110 and p150 in terms of editing within the nucleus.

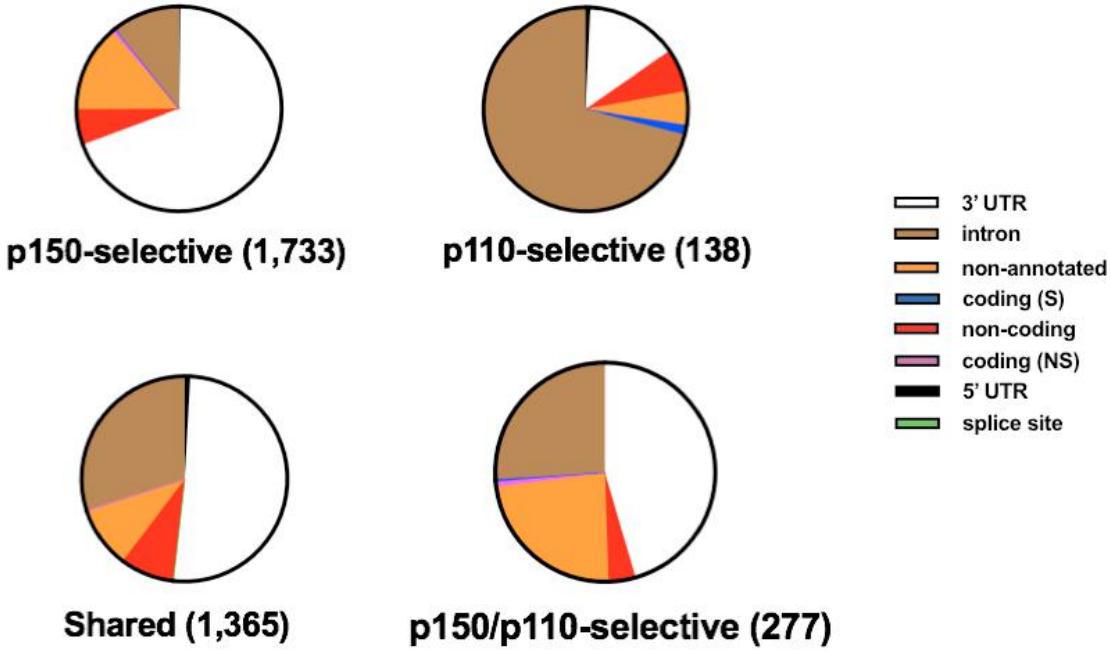


Figure 4.3.1 Annotation of grouped ADAR1-edit sites
The pie charts show proportions of annotations for the ADAR1-edit sites grouped in the four different categories. As mentioned earlier, some sites were unable to be annotated because they are located in gene clusters. For p150-selective sites, 0.2% are located in 5'UTR; 69% in 3'UTR; none in known splice sites; 6% in non-coding RNA; 14% in non-annotated RNA; 0.5% in coding regions; and 11% in introns. For p110-selective sites, 0.8% are located in 5'UTR; 15% in 3'UTR; none in known splice sites; 7% in non-coding RNA; 5% in non-annotated RNA; 2% in coding regions; and 71% in introns. For p110 and p150 shared sites, 0.9% are located in 5'UTR; 51% in 3'UTR; 0.07% in known splice sites; 8% in non-coding RNA; 10% in non-annotated RNA; 0.1% in coding regions; and 30% in introns. For p150/p110 double-selective sites, 0.2% are located in 5'UTR; 41% in 3'UTR;

0.02% in known splice sites; 4% in non-coding RNA; 26% in non-annotated RNA; 0.4% in coding regions; and 28% in introns.

When we consider the possibility that some sites might require the presence of both p110 and p150 for efficient editing to occur, two possible mechanisms (not mutually exclusive) come to mind. One scenario is that selective editing of certain sites by one isoform changes the structure of the RNA in a way that makes other sites now accessible by the other isoform. Another scenario is that potential formation of heterodimers between p150 and p110 could influence selectivity for certain RNA targets. The formation of p150 and p110 homodimers has been shown to be required for editing under various in vitro conditions, but the formal examination of whether p150 and p110 can interact in vivo remains lacking (165—167). Under specific immunoprecipitation conditions using our GFP-tagged p150 (discussed in chapter 2), we observed low levels of p150 and p110 that co-precipitated with the GFP-tagged p150 (Figure 4.3.2).

Preparation of lysates for immunoprecipitation differed from the standard immunoblot protocol, with modifications designed to preserve protein-protein interactions (in the absence of cross-linking steps). Notably, all steps should be done on ice or as close to 4°C as possible. Cells were first scraped onto ice-cold PBS containing phenylmethylsulfonyl fluoride (PMSF), a protease inhibitor, and then centrifuged for 3min at 1,500RPM. Next, cells were lysed using a mixture 10mM HEPES, 150mM KCl, 3mM $MgCl_2$, 0.5% NP-40 non-ionic surfactant, Roche proteinase inhibitor cocktail, and 0.1mM PMSF 0.1 M, all diluted in PBS. To shear genomic DNA, cell lysates were passed through a 26-gauge needle 20 times and then centrifuged at 15,000G for 10min.

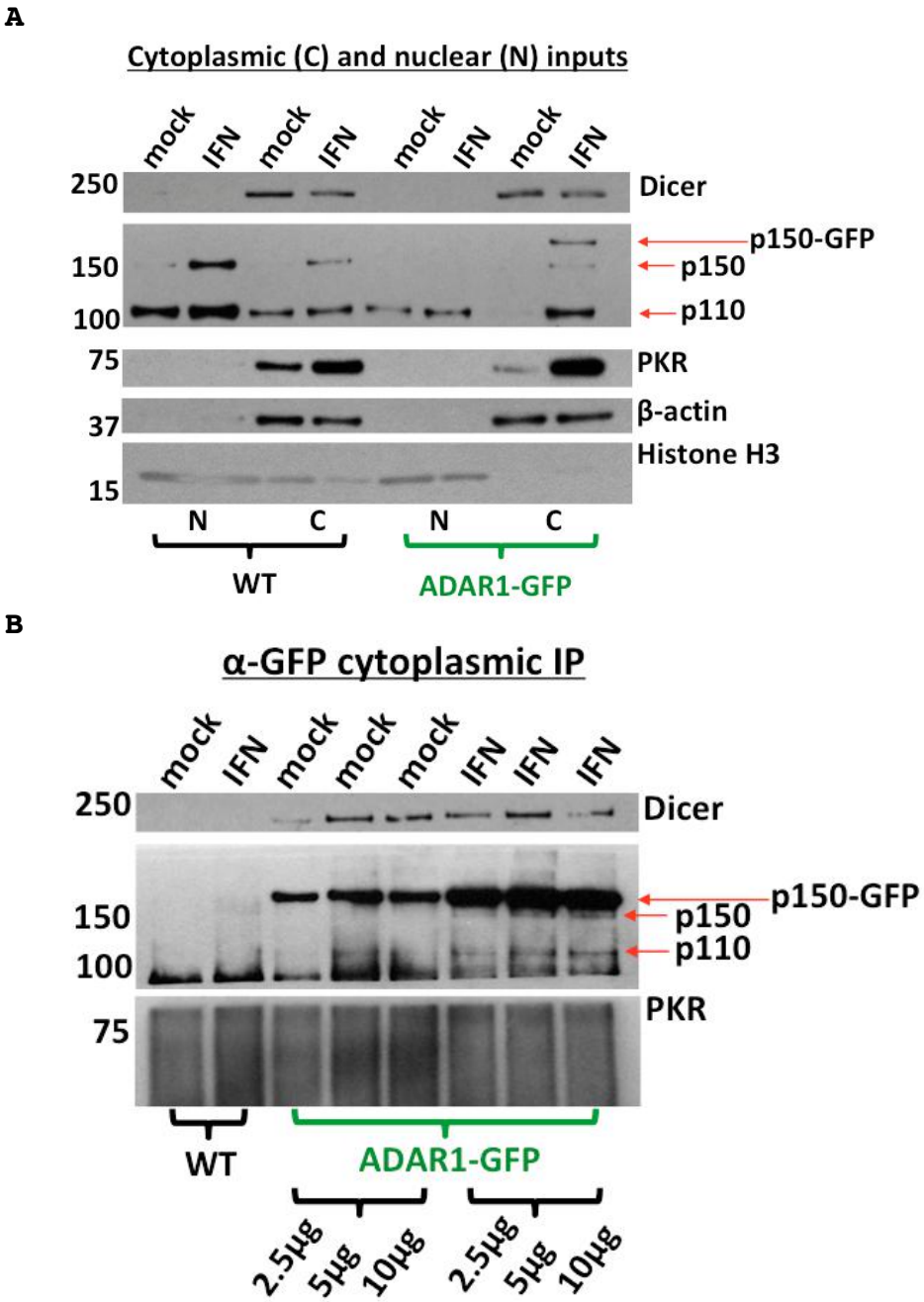


Figure 4.3.2 Immunoprecipitation of ADAR1p150

A. The immunoblot shows nuclear and cytoplasmic fractions of WT 293T cells and a heterozygous GFP-KI 293T clone described in chapter 2. During the preparation of lysates for immunoprecipitation, the supernatant was split for immunoblot and immunoprecipitation, and pelleted nuclei were also saved for immunoblot. Dicer is a cytoplasmic protein known to interact with ADAR1 during microRNA processing (168). PKR is a familiar dsRNA binding protein that has been shown to interact with ADAR1 under certain conditions, such as HIV-1 infection (169). Beta-actin and histone H3 are controls to assess the purity of the

cytoplasmic and nuclear fractions. The p110 isoform appears predominantly nuclear during baseline conditions, as can be seen comparing the nuclear (N) and cytoplasmic (C) mock-treated lanes; here, some nuclear materials appear to contaminate the cytoplasmic fractions, as histone H3 was detected in the WT mock and IFN "C" lanes. During interferon treatment, the GFP-p150-KI clone appeared to show presence of p110 protein in the cytoplasm. The p150 isoform is predominately cytoplasmic and is induced following interferon treatment, and the heterozygous ADAR1-GFP clone shows induction of both WT p150 and the GFP-tagged p150. The cytoplasmic fractions shown in the immunoblot were used as the starting material for the anti-GFP immunoprecipitation experiment.
B. A rabbit polyclonal antibody, previously purified on a GFP affinity column (gift from Peggy MacDonald), was added to Invitrogen Protein G Dynabeads, following manufacturer guidelines to bind back ends of the anti-GFP antibodies to the protein G. Next, magnetic bead-bound anti-GFP antibodies of various input amounts (2.5μg, 5μg, and 10μg) were incubated overnight with cytoplasmic lysates at 4°C on a rotating platform. Several wash steps were done the next day. The first set of washes were done 3 times, 5min each, using a mix of 10mM HEPES, 150mM KCl, 3mM $MgCl_2$, and 0.5% NP-40. The second set of washes were done 2 times, 5min each, using the same mixture except with 0.05% NP-40. Finally, bound proteins were eluted with immunoblot sample buffer with DTT, and the eluate was boiled and loaded for immunoblot, probed with antibodies that recognize Dicer, the C-terminal end of ADAR1, and PKR. As expected, the cytoplasmic input lanes from WT cells did not show evidence of Dicer, ADAR1, or PKR with anti-GFP immunoprecipitation. Dicer appeared in the mock and IFN treated lanes, and the GFP-tagged p150 increased in intensity in the IFN treated lanes compared to mock treated. Low levels of WT p150 and p110 isoforms appear to have co-precipitated with the GFP-tagged p150. PKR did not appear to co-precipitation with GFP-tagged ADAR1 in this experiment.

Of note, the stringency of final wash steps in all immunoprecipitation experiments influences what remains bound to the antibody-bead complexes at the end to be eluted for examination by immunoblot. At this point, more experiments are needed to determine whether p150 and p110 form heterodimers in vivo, and whether this complex could influence the selectivity of RNA targets for binding and editing.

A possible editing model, based on the proportions of annotated edit sites in the different categories, is that p150-selective editing occurs more in the cytoplasm while p110-selective editing occurs in the nucleus, where introns are located (Figure 4.3.3). Although almost no edit sites were

located in annotated splice acceptor (AG) sites, the possibility exists that p110-editing in introns can create new splice donor (AU-to-GU) and acceptor sites (AA-to-AG) in regions surrounded by motifs such as the poly-pyrimidine tract commonly found upstream of splice acceptor sites.

Formation of inverted repeat structures in the nucleus can occur between intron-intron, intron-exon, and exon-exon sequences, and all these are potential substrates for p110 and p150 binding and editing. Following splicing, remaining exon-exon dsRNA structures would be targeted by p150 following export to the cytoplasm. Binding by p150 is likely to be more high-affinity than p110 because of the Z-alpha domain unique to p150 (122).

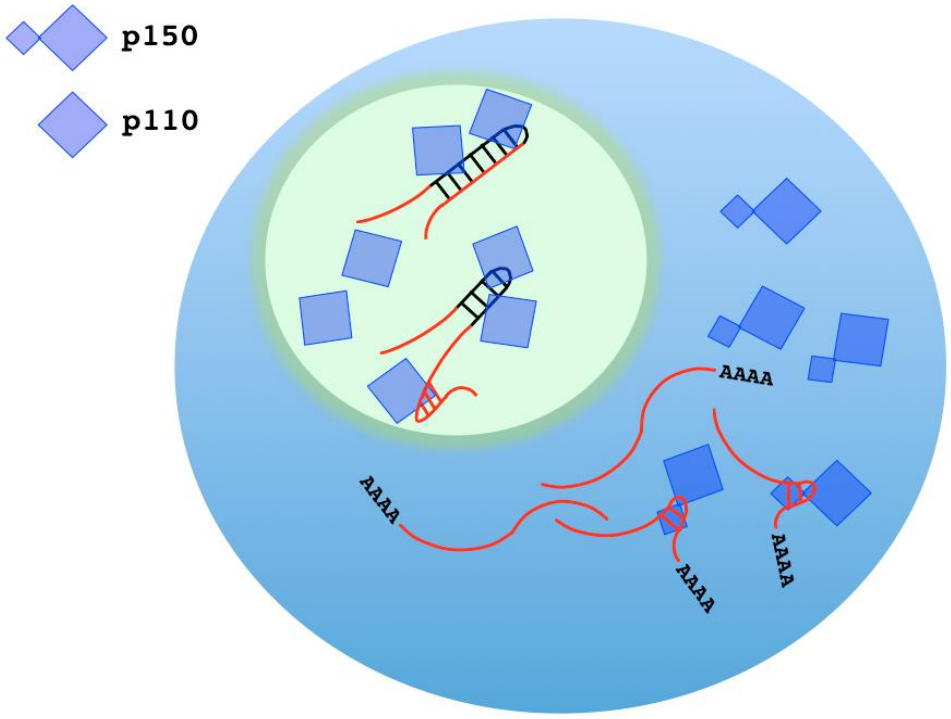


Figure 4.3.3 Model of ADAR1 p110 and p150 editing
In this simplified schematic, p110 is shown as a nuclear protein and p150 as a cytoplasmic protein. This distribution represents the general baseline distributions of the two isoforms, although p150 does has the ability to enter the nucleus, and p110 has the ability to enter the cytoplasm under certain conditions. Intron regions of RNA targets are shown in black while exon regions are shown in red.

Finally, our classification provides clarification on results from a prior study in which A-to-I editing analysis was performed on wild-type, ADAR1 knockout, and ADAR1 p150 KO cells (61). In that study, editing frequencies of both p150-dependent

sites and total ADAR1-dependent sites (including sites not dependent on p150 for editing to occur) were increased following interferon treatment. As expected, editing frequencies of p150-dependent sites were increased because total p150 protein levels go up after adding interferon. Less was clear about the other ADAR1-dependent sites, specifically whether they are p110-selective, shared, or require both p110 and p150 to be present for editing to occur.

Prior studies were unable to taxonomically divide ADAR1-dependent sites into isoform-independent (shared), p110-selective, and p150/p110-selective (synergistic) sites because of the inability to express p150 without p110. At the current read depth, we find that of these ADAR1-dependent sites, excluding p150-selective sites (49% of sites), 39% are shared by both isoforms. And as mentioned in chapter 2, we observed that p110 is induced along with p150 during the interferon response. Thus, it makes sense that not only the edit frequencies of p150-dependent sites would increase following interferon treatment, but the edit frequencies of all ADAR1-dependent sites would increase, as observed in the study mentioned above. We propose that increased expression of both p110 and p150 contribute to optimal editing of RNA targets in the nucleus and cytoplasm during the interferon response.

CHAPTER 5. Future directions and concluding remarks

From a semantics perspective, to describe “why” something happens in biology often seems more difficult than describing “how” something happens, or “what” is happening. One could ask: Why do human cells have both ADAR1p110 and ADAR1p150 isoforms? Alternatively, one could ask: How are p110 and p150 produced in human cells? One could ask: Why does A-to-I editing occur? An alternative question could be: What sites on RNA show A-to-I editing? The work presented in this thesis has tried to address the “how” and “what” questions related to ADAR1 expression and A-to-I editing. The final chapter of this thesis aims to discuss future directions related to the themes of p110 and p150 regulation and editing.

5.1 Endogenous expression of p110 and p150 in different cells

Our work focused on the HEK 293T cell as a model, largely because of the conveniences with regard to maintenance, and the efficiency of genetic modifications. We have deleted using CRISPR-Cas9 the first three splice donor exons along with their promoters, revealing the sum total of ADAR1 splice variant contributions to ADAR1 protein production, at least in 293T cells.

Of note, experiments done in other cell types have revealed some promoter activity within exon 2 and also alternative splicing activity within exon 2 in a region that can act as an intron (170, 171). Our cell line showed undetectable ADAR1 protein levels at chemiluminescence-level sensitivity and ablation of editing at known ADAR1 edit sites following deletion of exons 1B, 1A, and 1C. In our ADAR1 1B/1A/1C KO cell line, direct transfection of the p150-encoding mRNA variant resulted in production of p150 and p110 with observation of a single-sized band by reverse transcriptase (RT) PCR.

Different cell lines likely have variations in transcription factor expression, promoter activities, and alternative splicing events. Nevertheless, it would be interesting to examine formally these findings from other cell types in 293T cells, possibly using ADAR1 exon-2 specific probes to look for different sized variants by Northern blot.

We identified multiple RNA variants that have p110-coding potential, including the 1B-2 variant, 1C-2 variant, 1A-1C-2 variant, and also the 1A-2 variant. Synonymous mutations introduced into the coding region between the p150 and p110 start codons that aimed to increase translation initiation rates in an alternate reading frame resulted in significantly decreased p110 expression from the p150 mRNA. This enabled expression of p150 in isolation, making possible now the study and comparison of editomes when each isoform is present in isolation.

We selected a system that involves stable expression of ADAR1 isoforms in a knockout background, and comparison with a control expressing firefly luciferase. One of the advantages of this system is that p150 and p110 levels can be kept constant during certain perturbations such as interferon treatment. This would be helpful in addressing questions such as how interferon-induced co-factors could influence ADAR1 editing. In an endogenous context, treatment with interferon not only stimulates expression of potential ADAR1 trans-acting factors but also upregulates ADAR1 expression, making it difficult to separate the relative contributions of either one in isolation to the interferon editome. Of note, many protein trans-regulators of ADAR editing have been identified recently (172).

While an overexpression system has its merits, several drawbacks include the artificial nature of the mRNAs, which lack UTR sequences and also contain an IRES for RFP expression, and perhaps more significantly, the differences in protein levels between endogenous and exogenous expression contexts. Efforts have been made to knock-in the p150r sequence with its synonymous mutations into the endogenous context (ADAR1 exon 2), to create a p110-knockout cell line. Ongoing work involves using nucleofection of CRISPR ribonucleoprotein along with a p150r antisense ssDNA donor sequence. The background cell line would be an exon 1B/1C knockout cell line, which has the other p110-enconding mRNA variants removed (Figure 5.1.1). Using a WT background cell line could also be interesting, because a homozygous knock-in here would only be able to express p110 and p150 from separate mRNA molecules (with significantly reduced p110 leaky expression on the 1A-2 transcript with the p150r mutations).

We had considered the possibility that if p110/p150-selective sites do occur, they could potentially be further classified in terms of sites that are uniquely edited only when p110 and p150 have the potential to be translated from the same mRNA species. During any given time, one mRNA may be coated with clusters of ribosomes, allow assembling of multiple polypeptides in tandem from one mRNA. As part of a previous thought experiment, we entertained the possibility that co-translation of p110 with p150 from one mRNA might increase the chance of heterodimer formation (due to spatial proximity), while translation of p110 and p150 from spatially separated mRNAs might favor homodimer formation over heterodimer formation. Perhaps p110 and p150 homodimers and heterodimers have different preferred RNA targets.

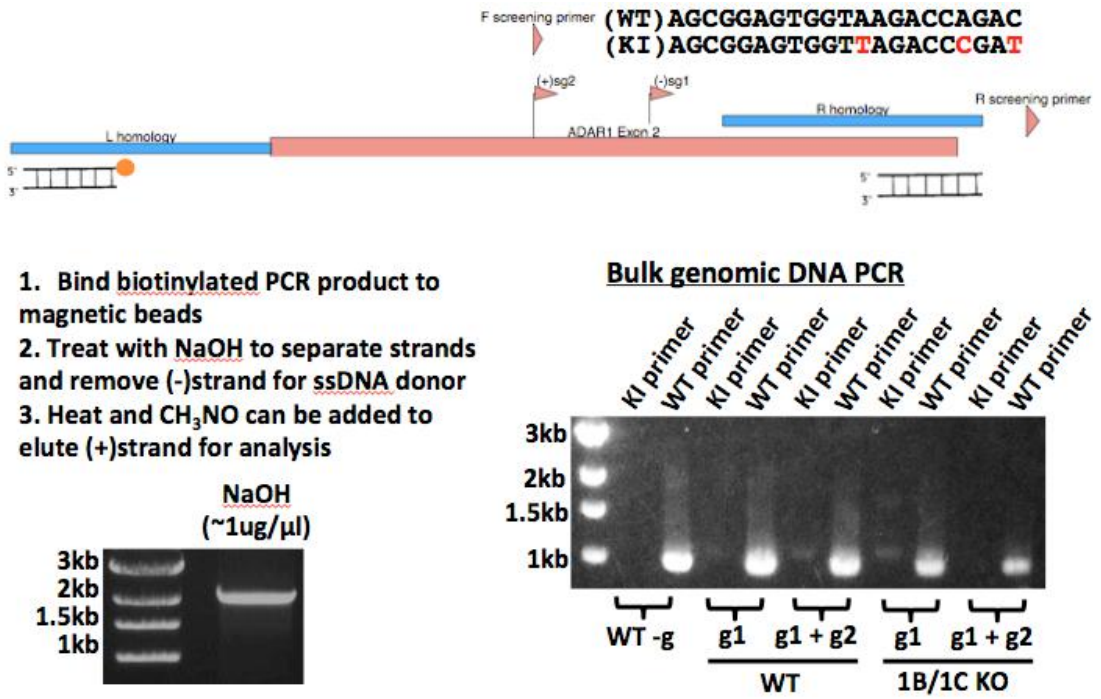


Figure 5.1.1 Ongoing work related to p150r knock-in
In this experiment, a gene block containing the p150r mutations was ordered from Gene Universal. The sequence includes homology arms for homology-directed repair during the CRISPR knock-in process. The sequence was amplified by PCR using a biotinylated primer that primes synthesis of the sense sequence and allows separation of the sense and antisense strands using streptavidin. The antisense ssDNA was isolated and electroporated into WT and exon 1B/1C KO cells along with a CRISPR ribonucleoprotein complex prepared from recombinant Cas9 protein with guide RNAs. The electroporation was done using the Lonza 4D-Nucleofector system. Analysis of bulk genomic DNA with the indicated primer pairs (KI screening primers have bases shown in red, and importantly, one at the 3' priming end, that correspond to the edited genomic sequence) suggests that there may be some degree of knock-in with use of the guides tested, such as guide#1 in the 1B/1C KO cell, although the PCR product sizes unexpectedly differed between the WT and KI pairs of primers. To reduce the background signal, the reverse primer was designed to bind a region outside the end of the antisense ssDNA sequence. Therefore, PCR products should correspond only to genomic DNA as the starting template material. Excising the gel bands and building a deep sequencing library from these PCR products would be an appropriate next step.

Other possibilities for knock-in include non-cleavage based methods that use fusions between Cas9 and deaminase proteins or other single-base editing proteins (173–176). Of note, many of

these fusion proteins can be introduced into cells by plasmid-based transient expression systems, and these plasmids are now commercially available on Addgene.

Of note, knock-in of the p150r sequence in cell types such as embryonic stem cells would allow experiments to be done looking at how p150 and p110 individually shape differentiation, and also how the editomes of individual isoforms compare between different cell types.

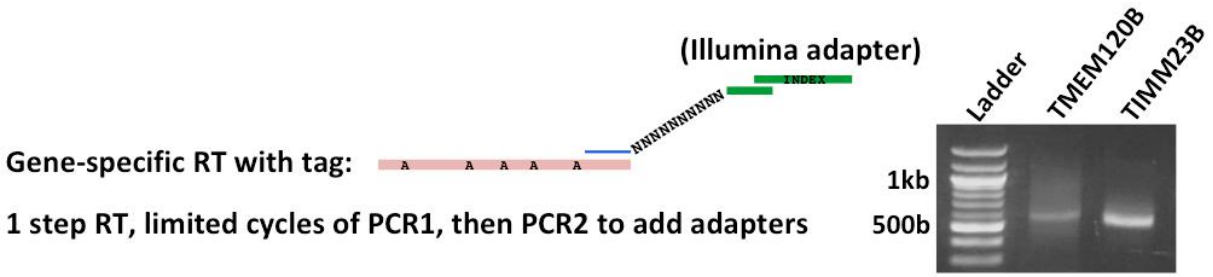
An additional thought, overexpression of the p150 Z-alpha domain in trans with p110 and analysis of A-to-I edit sites would highlight Z-alpha domain-dependent editing substrates amongst the p150-selective sites.

5.2 Amplicon sequencing with unique molecular tags

During bulk RNA sequencing, coverage at repetitive elements in particular is limited, possibly due to a combination of secondary structures impacting RT processivity during library preparation, lower quality base calls near homopolyer regions, and challenges involved with alignment. Ongoing efforts to quantify editing at selected sites include use of gene-specific RT primers to make amplicons for sequencing. Furthermore, molecular tags consisting of a random sequence of nucleotides were added to RT primers so that each primer creates a first-strand cDNA that contains a unique barcode corresponding to the original RNA molecule (Figure 5.2.1). The methods for amplicon sequencing library preparation and analysis were adapted from several sources (177—182).

As a side note, unique dual-matched indexes were chosen to increase the accuracy of assigning read to samples by 100 times compared to single index combinations (183). Index-hopping is the term used to describe a process that can cause misassignment of reads to samples during demultiplexing. One proposed mechanism of index-hopping during library preparation is the carry-over of adapters with index sequences during pooling of samples, either because of excess adapters or incomplete cleanup. During subsequent steps, the retained adapters with index sequences can prime PCR reactions and result in mixing of indexes between the pooled sample fragments. In the case of building libraries for amplicon sequencing, the inclusion of index sequences within primer sequences in the final PCR step before pooling already makes improbable the possibility of index-hopping based on the proposed mechanism of hopping. Informatic analysis of unexpected (unmatched) combinations of dual indexes present in the pooled libraries should reveal a value of zero or near zero.

A



B

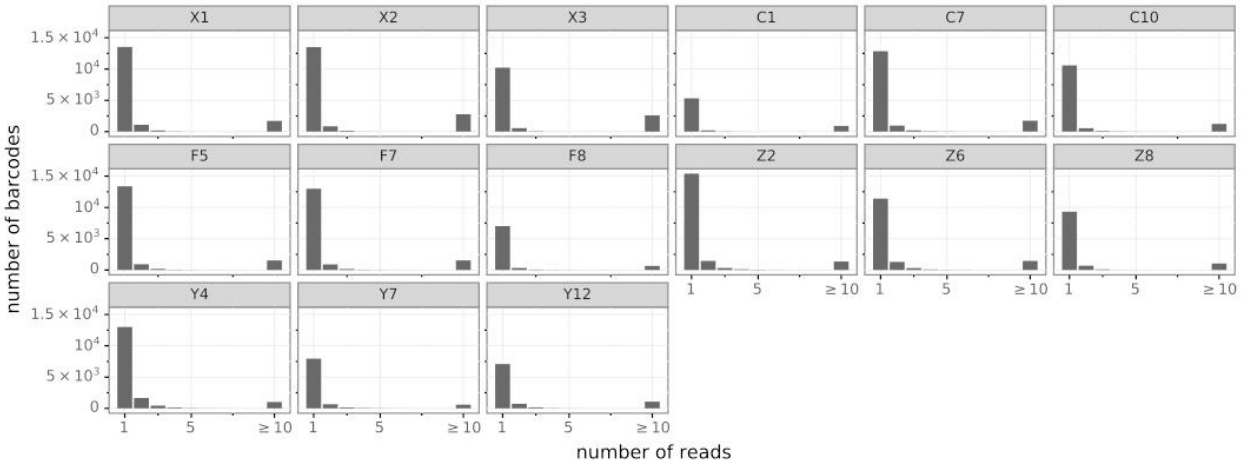


Figure 5.2.1 Amplicon sequencing with molecular barcodes
A. As an example of ongoing amplicon sequencing work, two examples are showing corresponding to the 3'UTR regions of TMEM120B and TIMM23B, which have ADAR1-edit sites. The amplicons were created first with a one-step RT reaction, in which the antisense gene-specific primer has a 10-base degenerate sequence that serves as a molecular tag for each RNA molecule. Second-strand cDNA synthesis using a sense gene-specific primer was coupled with the first round of PCR (12 cycles) to add half of the Illumina adapters to both ends of the cDNA fragment. A second round of PCR was done to add the rest of the Illumina adapter sequences along with a unique dual index corresponding to each of the 15 samples, allowing pooling of samples in one sequencing reaction.
B. After amplicons were built for each sample, they were pooled and normalized for sequencing on MiSeq. Analysis was done with an adapted version of the dms_tools2 software developed by the Jesse Bloom lab. The histograms here show tallied up counts of barcodes with different read counts; as shown, most barcodes were sequenced only once, suggesting PCR duplicates were minimal in our library preparation method. The data comes from barcoded amplicons in the 15 pooled samples: X1-X3 are fluc clones; C1, C7, and C10 are p150 clones; F5, F7, and F8 are p150r; Z2, Z6,

and Z8 are p110 clones; and Y4, Y7, and Y12 are p150r+p110 clones.

PCR biased duplications during library preparation is a potential concern, due to intrinsic differences in templates (184, 185). However, given that most unique barcodes were sequenced only once, PCR duplicates appear be a small contributor to final bias in the counting of edited and unedited sites on transcripts. As expected, the coverage is significantly increased in the amplicon sequencing compared to bulk RNA sequencing (Figure 5.2.2).

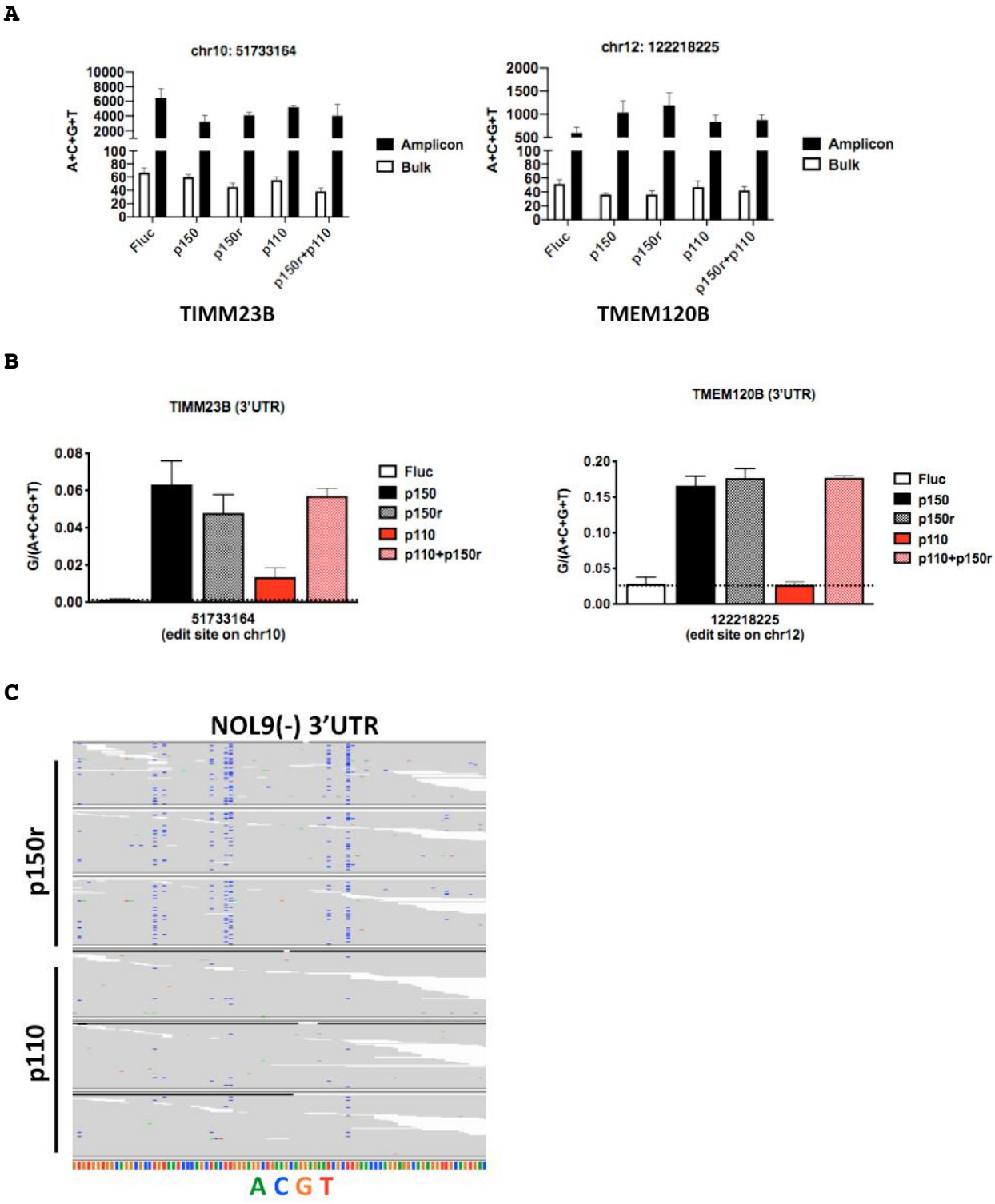


Figure 5.2.2 Amplicon sequencing provides high coverage
A. The challenge of getting unbiased high coverage at repetitive Alu elements can be addressed by locus-specific amplicon sequencing. Because duplicate reads with the same barcode can be collapsed down to a single consensus read, the base counts from these collapsed reads accurately reflect that of the original

RNA sample. For barcodes with only one read, RT and sequencing errors are still possible; barcodes with multiple reads allow assembly of a consensus sequence, as was the original intent of the dms_tools2 workflow: to find deep sequencing errors and identify true mutations.
B. Shown here, according to the results of amplicon sequencing, are edit frequencies of a site in TIMM23B shared by p110 and p150 (although p150 alone seems to edit this site with a higher frequency than p110), and a p150-selective site in TMEM120B.
C. Shown here is part the NOL9 gene viewed in IGV that corresponds to an independent alignment of the bulk RNA sequencing data with HISAT2 rather than STAR. The region shown in the 3'UTR reveals multiple T-to-C (antisense mRNA) genomic mismatches that appear predominately in the p150r samples. Of note, selecting regions with hyper-editing to build amplicons for sequencing will allow extracting more information from a single amplicon.

Given that global median edit frequencies are around or below 0.2, producing an unbiased and reproducible taxonomy of ADAR1-edit sites in which sites are accurately categorized requires in particular high coverage at all positions to reduce the chance of false negatives. Amplicon sequencing provides a way to gain more coverage at select regions. Globally, in the bulk RNA sequencing analysis, the original cutoff depth requirement of 5 can also be modified, say to 50, and also 100. When this is done, the number of sites that pass the filters described in chapter 4 decreases, as expected, although the trends remain the same, in the sense that p150-selective and shared edit sites far outnumber the putative p110-selective and p150/p110 double-selective sites (Figure 5.2.3).

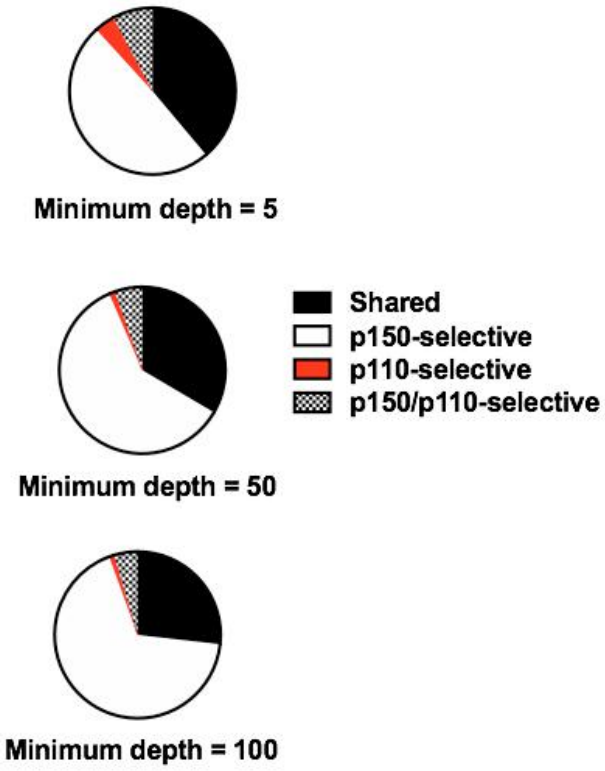


Figure 5.2.3 Changing depth requirement in bulk analysis
The pie charts above show the proportion of sites that classify in the four categories with different read depth requirements. With a minimum depth of 5, as already shown in chapter 4, 9,044 A/G mismatches were identified, of which 3,513 could be classified. With a minimum depth 50, the number of A/G mismatches passing filters is reduced to 2,447, of which 1,256 could be classified. And with a minimum depth of 100, 1,181 A/G mismatches were identified, of which 646 could be classified. The pattern that p150-selective sites outnumber shared sites, which outnumber p110-selective and p150/p110-selective sites, holds for all three conditions, but what seems to increase slightly as the depth requirement increases is the proportion of p150-selective sites. Of note, the minimum depth requirement was changed while holding other filters in the analysis workflow constant, such as the minimum mismatch depth of 2 in each triplicate. Here, there could be another filter applied to only consider sites that pass a certain minimum edit frequency threshold.

Amplicon sequencing could reveal sites that are rigidly edited in a specific manner by only p150 or p110, but it could also reveal sites that are more fluid and harder to place clearly in one of the defined categories, for example, the site in TIMM23B (chr 10: 51733164) shown in panel B of figure 5.2.2. By doing an unpaired t-test comparing the p150r and p110 groups, we learn that the p150r group edits this site at a frequency

significantly higher than that of the p110 group. But does this mean we place this site in the p150-selective group, or should it be placed in the shared group, because p110 also edits this site at a frequency significantly higher than that in the fluc group? To a degree, these questions are most interesting not so much about the taxonomy but perhaps more the consequences of categorized editing. As an example, we have made the suggestion that editing of p150-selective sites is critical for suppressing activation of PKR and preventing translational shutdown during the interferon response.

Taxonomy is a valid exercise on its own, but part of the motivation to categorize in the first place is to learn what having p110-selective and p150-selective sites means for RNA, proteins, and cell biology. Thus, the ongoing work described above in this chapter help address but a fraction of potential questions related to ADAR1 and A-to-I editing. We have provided a starting point for future targeted analysis of edit sites. As brief examples, genes of interest with edit sites that create novel splice acceptor and donor sites could be checked for production of splice isoforms unique to cells with those sites edited; genes with editing in the 5′UTR, though rare, could be checked for changes in protein expression by a simple immunoblot.

In a broader sense, we wonder if there are situations in which inosine and guanosine are not redundant, in the sense that certain biological processes might not interpret an A-to-I edit in the same way as, say, an A-to-G mutation. With exciting developments in experimental techniques, future investigations into editing of biological information at the DNA, RNA, and protein levels will undoubtedly be fruitful.

Appendix 1: qPCR primer pair sequences

Exon 1B-exon 2 sense: 5′-GACTGAAGGTAGAGAAGGCTACG-3′
Exon 1B-exon 2 antisense: 5′-GCTGGTACCTGAGCTGTCTG-3′
Exon 1A-exon 2 sense: 5′-AAGCGAAATTGAACCGGAGC-3′
Exon 1A-exon 2 antisense: 5′-GGAGCTGCCCCTTGAGAAAT-3′
Exon 1C-exon 2 sense: 5′-CAGCACTTTGGGAGGCC-3′
Exon 1C-exon 2 antisense: 5′-GCTGGTACCTGAGCTGTCTG-3′
RPS11 sense: 5′-GCCGAGACTATCTGCACTAC-3′
RPS11 antisense: 5′-ATGTCCAGCCTCAGAACTTC-3′
STAT1 sense: 5′-AGGAAAAGCAAGCGTAATCTTCA-3′
STAT1 antisense: 5′-TATTCCCCGACTGAGCCTGAT-3′
ISG15 sense: 5′-GGCTGGGAGCTGACGGTGAAG-3′
ISG15 antisense: 5′-GCTCCGCCCGCCAGGCTCTGT-3′
IFN-beta sense: 5′-GTCAGAGTGGAAATCCTAG-3′
IFN-beta antisense: 5′-ACAGCATCTGCTGGTTGAAG-3′
RPS11 sense: 5′-GCCGAGACTATCTGCACTAC-3′
RPS11 antisense: 5′-ATGTCCAGCCTCAGAACTTC-3′
Albumin sense: 5′-TTTATGCCCCGGAACTCCTTT-3′
Albumin antisense: 5′-TGTTTGGCAGACGAAGCCTT-3′

Appendix 2: Guide RNA sequences

GFP-KI guide#1 (sense):
Target sequence with PAM:
5′-GGGCGCAATGAATCCGCGGC(AGG)-3′
ssDNA oligos for cloning into PX459:
5′-***CACC***GGGCGCAATGAATCCGCGGC-3′
    3′-CCCGCGTTACTTAGGCGCCG***CAAA***-5′
GFP-KI guide#2 (antisense):
Target sequence with PAM: 5′-TCGCGGGCGCAATGAATCCG(CGG)-3′
ssDNA oligos for cloning into PX459:
5′-***CACC***GCGCGGGCGCAATGAATCCG-3′
    3′-CGCGCCCGCGTTACTTAGGC***CAAA***-5′
Exon-1B-KO upstream guide (antisense):
Target sequence with PAM:
5′-CAAACTCTCATCTAGAGGCC(TGG)-3′
ssDNA oligos for cloning into PX458:
5′-***CACC***GCAAACTCTCATCTAGAGGCC-3′
    3′-CGTTTGAGAGTAGATCTCCGG***CAAA***-5′
Exon-1B-KO downstream guide (antisense):
Target sequence with PAM:
5′-AGAAGGACAGAGGCTCTTAC(CGG)-3′
ssDNA oligos for cloning into PX458:
5′***CACC***GAGAAGGACAGAGGCTCTTAC-3′
   3′-CTCTTCCTGTCTCCGAGAATG***CAAA***-5′
Exon-1C-KO upstream guide (sense):
Target sequence with PAM:
5′-GATACTGCTTAGTAGTGAAG(AGG)-3′
ssDNA oligos for cloning into PX458:
5′-*CACC*GATACTGCTTAGTAGTGAAG-3′
    3′-CTATGACGAATCATCACTTC*CAAA*-5′
Exon-1C-KO downstream guide (antisense):
Target sequence with PAM:
5′-GCTACTCTTGGCCCAAATC(TGG)-3′
ssDNA oligos for cloning into PX458:
5′-*CACC*GCTACTCTTGGCCCAAATC-3′
    3′-CGATGAGAACCGGGTTTAG*CAAA*-5′
Exon-1A-KO upstream guide (sense):
Target sequence with PAM:
5′-CGTAGTTCTCATGCAGCGGA(GGG)-3′
ssDNA oligos for cloning into PX458:
5′-*CACC*GCGTAGTTCTCATGCAGCGGA-3′
    3′-CGCATCAAGAGTACGTCGCCT*CAAA*-5′
Exon-1A-KO downstream guide (sense):
Target sequence with PAM:
5′-CTTGGACCTTCGCCGCCGTC(TGG)-3′
ssDNA oligos for cloning into PX458:
5′-*CACC*GCTTGGACCTTCGCCGCCGTC-3′
    3′-CGAACCTGGAAGCGGCGGCAG*CAAA*-5′

Appendix 3: NPC differentiation protocol

Day 0 (P0):

1. Coat 12-well (3.5 cm^2) plate with matrigel at 37°C for at least one hour. Coat 6 coverslips (for immunocytochemistry analysis) in 24-well plate (1.9 cm^2) with matrigel at 37°C for at least one hour. Also coat 24-well plate without coverslips for harvesting of cells for qPCR.
2. Detach and plate hESCs ($2x10^5$ viable cells/cm^2)
    a. Remove and save old media.
    b. Wash once with 1X PBS.
    c. Add 500µl Accutase gentle dissociate medium + 10µM Y-27632 per well.
    d. Incubate at 37°C on rocker for 5min or until cells have lifted.
    e. Add old media back.
    f. Dissociate cells fully and collect cells in 15mL Falcon tube.
    g. Centrifuge for 1min at 500-1000RCF.
    h. Aspirate media and mix cells with 500µl NIM + 10µM Y-27632, making sure to dissociate cells fully.
    i. Plate cells at a density of $2x10^5$ viable cells/cm^2 with 1.5mL NIM + 10µM Y-27632 per well.

Day 6 (P1):

1. Coat 6-well plate (9.6 cm^2) and 6 coverslips in 24-well plate (1.9 cm^2) with matrigel at 37°C for at least one hour.
2. Heat up NIM + 10µM Y-27632 at 37°C.
3. Detach and plate NPCs ($1.5x10^5$ viable cells/cm^2) for NPC monolayer passage 1 (P1).
4. Collect $4.8x10^5$ cells for immunoblot and the rest for qPCR (this timepoint is called passage 1 or P1).
    a. Add 1mL 1X PBS.
    b. Centrifuge for 5min at 1000RCF.
    c. Aspirate supernatant (use pipette if worried about aspirating pellet) and freeze P1 pellets at -80°C.

Day 12 (P2):

1. Coat 6-well and 24-well plates with matrigel at 37°C for at least one hour. Coat coverslips in 24-well plate with matrigel at 37°C for at least one hour.
2. Heat up NIM + 10µM Y-27632 at 37°C.
3. Detach and plate NPCs ($1.5x10^5$ viable cells/cm^2) for NPC monolayer passage 2 (P2).
4. Switch media from NIM to NPM
5. Collect $4.8x10^5$ cells for immunoblot and the rest for qPCR.

Day 15 (P2 day 3):

1. From the 24-well plate without coverslips, collect $4.8x10^5$ cells for immunoblot and the rest for qPCR.
2. From 24-well plate with coverslips, fix cells for ICC (Figure 2.3.2).

a. Aspirate media and wash once with 1X PBS.
b. Add 350µl of 4% PFA diluted in 1X PBS for 15min at room temperature.
c. Wash three times with 1X PBS and store at 4°C until ICC.

Day 19 (P2 day 7):

1. From the 24-well plate without coverslips, collect $4.8x10^5$ cells for immunoblot and the rest for qPCR.
2. From 24-well plate with coverslips, fix cells for ICC (Figure 2.3.2).

Works Cited

1. L. B. Carey, RNA polymerase errors cause splicing defects and can be regulated by differential expression of RNA polymerase subunits. *Elife* **4**, 1–10 (2015).
2. T. B. Johnson, R. D. Coghill, Researches on pyrimidines. C111. The discovery of 5-methyl-cytosine in tuberculinic acid, the nucleic acid of the tubercle bacillus. *J. Am. Chem. Soc.* **47**, 2838–2844 (1925).
3. G. R. WYATT, Recognition and estimation of 5-methylcytosine in nucleic acids. *Biochem. J.* **48**, 581–584 (1951).
4. R. Benne, *et al.*, Major transcript of the frameshifted coxll gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* **46**, 819–826 (1986).
5. B. Blum, N. Bakalara, L. Simpson, A model for RNA editing in kinetoplastid mitochondria: RNA molecules transcribed from maxicircle DNA provide the edited information. *Cell* **60**, 189–198 (1990).
6. J. D. Alfonzo, O. Thiemann, L. Simpson, The mechanism of U insertion/deletion RNA editing in kinetoplastid mitochondria. *Nucleic Acids Res.* **25**, 3751–3759 (1997).
7. L. M. Powell, *et al.*, A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell* **50**, 831–840 (1987).
8. K. Higuchi, *et al.*, Human apolipoprotein B (apoB) mRNA: Identification of two distinct apoB mRNAs, an mRNA with the apoB-100 sequence and an apoB mRNA containing a premature in-frame translational stop codon, in both liver and intestine. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 1772–1776 (1988).
9. A. V. Hospattankar, K. Higuchi, S. W. Law, N. Meglin, H. B. Brewer, Identification of a novel in-frame translational stop codon in human intestine ApoB mRNA. *Biochem. Biophys. Res. Commun.* **148**, 279–285 (1987).
10. P. P. Lau, W. Xiong, H. J. Zhu, S. H. Chen, L. Chan, Apolipoprotein B mRNA editing is an intranuclear event that occurs posttranscriptionally coincident with splicing and polyadenylation. *J. Biol. Chem.* **266**, 20550–20554 (1991).
11. J. T. Holt, R. L. Redner, A. W. Nienhuis, An oligomer complementary to c-myc mRNA inhibits proliferation of HL-60 promyelocytic cells and induces differentiation. *Mol. Cell. Biol.* **8**, 963–973 (1988).
12. T. G. Lawson, *et al.*, Influence of 5′ proximal secondary structure on the translational efficiency of eukaryotic mRNAs and on their interaction with initiation factors. *J. Biol. Chem.* **261**, 13979–13989 (1986).
13. D. A. Melton, Injected anti-sense RNAs specifically block messenger RNA translation in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **82**, 144–148 (1985).

14. M. R. Rebagliati, D. A. Melton, Antisense RNA injections in fertilized frog eggs reveal an RNA duplex unwinding activity. *Cell* **48**, 599–605 (1987).
15. B. L. Bass, H. Weintraub, A developmentally regulated activity that unwinds RNA duplexes. *Cell* **48**, 607–613 (1987).
16. B. L. Bass, H. Weintraub, An unwinding activity that covalently modifies its double-stranded RNA substrate. *Cell* **55**, 1089–1098 (1988).
17. R. W. Wagner, J. E. Smith, B. S. Cooperman, K. Nishikura, A double-stranded RNA unwinding activity introduces structural alterations by means of adenosine to inosine conversions in mammalian cells and Xenopus eggs. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 2647–2651 (1989).
18. A. G. Poison, *et al.*, The Mechanism of Adenosine to Inosine Conversion by the Double-Stranded rna Unwinding/Modifying Activity: A High-Performance Liquid Chromatography-Mass Spectrometry Analysis. *Biochemistry* **30**, 11507–11514 (1991).
19. K. Nishikura, *et al.*, Substrate specificity of the dsRNA unwinding/modifying activity. *EMBO J.* **10**, 3523–3532 (1991).
20. R. F. Hough, B. L. Bass, Purification of the Xenopus laevis double-stranded RNA adenosine deaminase. *J. Biol. Chem.* **269**, 9933–9939 (1994).
21. M. A. O'Connell, W. Keller, Purification and properties of double-stranded RNA-specific adenosine deaminase from calf thymus. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 10596–10600 (1994).
22. U. Kim, *et al.*, Purification and characterization of double-stranded RNA adenosine deaminase from bovine nuclear extracts. *J. Biol. Chem.* **269**, 13480–13489 (1994).
23. U. Kim, Y. Wang, T. Sanford, Y. Zeng, K. Nishikura, Molecular cloning of cDNA for double-stranded RNA adenosine deaminase, a candidate enzyme for nuclear RNA editing. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 11457–11461 (1994).
24. M. A. O'Connell, *et al.*, Cloning of cDNAs encoding mammalian double-stranded RNA-specific adenosine deaminase. *Mol. Cell. Biol.* **15**, 1389–1397 (1995).
25. P. Edman, Method for Determination of the Amino Acid Sequence in Peptides. *Acta Chem. Scand.* **4**, 283–293 (1950).
26. M. Higuchi, *et al.*, RNA editing of AMPA receptor subunit GluR-B: A base-paired intron-exon structure determines position and efficiency. *Cell* **75**, 1361–1370 (1993).
27. B. Sommer, M. Köhler, R. Sprengel, P. H. Seeburg, RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell* **67**, 11–19 (1991).
28. R. W. Wagner, *et al.*, Double-stranded RNA unwinding and modifying activity is detected ubiquitously in primary tissues and cell lines. *Mol. Cell. Biol.* **10**, 5586–5590 (1990).

29. J. Craig Venter, *et al.*, The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
30. H. U. G. Weier, C. X. George, K. M. Greulich, C. E. Samuel, The interferon-inducible, double-stranded RNA-specific adenosine deaminase gene (DSRAD) maps to human chromosome 1q21.1-21.2. *Genomics* **30**, 372–375 (1995).
31. Y. Wang, Y. Zeng, J. M. Murray, K. Nishikura, Genomic organization and chromosomal location of the human dsRNA adenosine deaminase gene: The enzyme for glutamate-activated ion channel RNA editing. *J. Mol. Biol.* **254**, 184–195 (1995).
32. J. B. Patterson, C. E. Samuel, Expression and regulation by interferon of a double-stranded-RNA-specific adenosine deaminase from human cells: evidence for two forms of the deaminase. *Mol. Cell. Biol.* **15**, 5376–5388 (1995).
33. M. A. Frohman, M. K. Dush, G. R. Martin, Rapid production of full-length cDNAs from rare transcripts: Amplification using a single gene-specific oligonucleotide primer. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 8998–9002 (1988).
34. M. B. Karpova, *et al.*, Raji revisited: Cytogenetics of the original Burkitt's lymphoma cell line [12]. *Leukemia* **19**, 159–161 (2005).
35. C. X. George, C. E. Samuel, Characterization of the 5'-flanking region of the human RNA-specific adenosine deaminase ADAR1 gene and identification of an interferon-inducible ADAR1 promoter. *Gene* **229**, 203–213 (1999).
36. M. R. M. & L. M. Bilezikjian, Binding of a nuclear protein to the cyclic-AMP response element of the somatostatin gene. *Nature* **328**, 175–8 (1987).
37. K. L. Kuhen, J. W. Vessey, C. E. Samuel, Mechanism of Interferon Action: Identification of Essential Positions within the Novel 15-Base-Pair KCS Element Required for Transcriptional Activation of the RNA-Dependent Protein Kinasepkr Gene. *J. Virol.* **72**, 9934–9939 (1998).
38. J. E. Darnell, I. M. Kerr, G. R. Stark, Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. *Science* **264**, 1415–1421 (1994).
39. K. L. Kuhen, C. E. Samuel, Isolation of the interferon-inducible RNA-dependent protein kinase Pkr promoter and identification of a novel DNA element within the 5'-flanking region of human and mouse Pkr genes. *Virology* **227**, 119–130 (1997).
40. T. Williams, R. Tjian, Characterization of a dimerization motif in AP-2 and its function in heterologous DNA-binding proteins. *Science* **251**, 1067–1071 (1991).
41. C. X. George, C. E. Samuel, Human RNA-specific adenosine deaminase ADAR1 transcripts possess alternative exon 1 structures that initiate from different promoters, one

constitutively active and the other interferon inducible. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 4621–4626 (1999).
42. W. Lee, A. Haslinger, M. Karin, R. Tjian, Activation of transcription by two factors that bind promoter and enhancer sequences of the human metallothionein gene and SV40. *Nature* **325**, 368–372 (1987).
43. G. Zubay, D. Schwartz, J. Beckwith, Mechanism of activation of catabolite-sensitive genes: a positive control system. *Proc. Natl. Acad. Sci. U. S. A.* **66**, 104–110 (1970).
44. S. T. Smale, D. Baltimore, The "initiator" as a transcription control element. *Cell* **57**, 103–113 (1989).
45. D. Leprince, *et al.*, A putative second cell-derived oncogene of the avian leukaemia retrovirus E26. *Nature* **306**, 395–397 (1983).
46. W. V. Shaw, *et al.*, Primary structure of a chloramphenicol acetyltransferase specified by R plasmids. *Nature* **282**, 870–872 (1979).
47. H. Klenow, I. Henningsen, Selective elimination of the exonuclease activity of the deoxyribonucleic acid polymerase from Escherichia coli B by limited proteolysis. *Proc. Natl. Acad. Sci. U. S. A.* **65**, 168–175 (1970).
48. C. X. George, C. E. Samuel, Human RNA-specific adenosine deaminase ADAR1 transcripts possess alternative exon 1 structures that initiate from different promoters, one constitutively active and the other interferon inducible. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 4621–4626 (1999).
49. S. R. Green, L. Manche, M. B. Mathews, Two functionally distinct RNA-binding motifs in the regulatory domain of the protein kinase DAI. *Mol. Cell. Biol.* **15**, 358–364 (1995).
50. S. C. McCormack SJ, Ortega LG, Doohan JP, Mechanism of interferon action motif I of the interferon-induced, RNA-dependent protein kinase (PKR) is sufficient to mediate RNA-binding activity. *Virology* **198**, 92–99 (1994).
51. Y. Liu, C. X. George, J. B. Patterson, C. E. Samuel, Functionally distinct double-stranded RNA-binding domains associated with alternative splice site variants of the interferon-inducible double- stranded RNA-specific adenosine deaminase. *J. Biol. Chem.* **272**, 4419–4428 (1997).
52. Y. Liu, C. E. Samuel, Mechanism of interferon action: functionally distinct RNA-binding and catalytic domains in the interferon-inducible, double-stranded RNA-specific adenosine deaminase. *J. Virol.* **70**, 1961–1968 (1996).
53. L. Mittaz, *et al.*, Cloning of a human RNA editing deaminase (ADARB1) of glutamate receptors that maps to chromosome 21q22.3. *Genomics* **41**, 210–217 (1997).
54. C. X. Chen, *et al.*, A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains. *Rna* **6**, 755–

767 (2000).
55. T. Melcher, *et al.*, RED2, a brain-specific member of the RNA-specific adenosine deaminase family. *J. Biol. Chem.* **271**, 31795–31798 (1996).
56. T. Melcher, *et al.*, A mammalian RNA editing enzyme. *Nature* **379**, 460–464 (1996).
57. A. L. Yablonovitch, P. Deng, D. Jacobson, J. B. Li, The evolution and adaptation of A-to-I RNA editing. *PLoS Genet.* **13**, 1–14 (2017).
58. C. E. Samuel, Adenosine deaminase acting on RNA (ADAR1), a suppressor of double-stranded RNA–triggered innate immune responses. *J. Biol. Chem.* **294**, 1710–1720 (2019).
59. F. L. Graham, J. Smiley, W. C. Russell, R. Nairn, Characteristics of a human cell line transformed by DNA from human adenovirus type 5. *J. Gen. Virol.* **36**, 59–72 (1977).
60. F. L. Graham, A. J. Van Der Eb, H. L. Heijneker, Size and location of the transforming region in human adenovirus type 5 DNA. *Nature* **251**, 687–691 (1974).
61. H. Chung, *et al.*, Human ADAR1 Prevents Endogenous RNA from Triggering Translational Shutdown. *Cell* **172**, 811-824.e14 (2018).
62. M. H. Tan, *et al.*, Dynamic landscape and regulation of RNA editing in mammals. *Nature* **550**, 249–254 (2017).
63. A. A. Stepanenko, V. V. Dmitrenko, HEK293 in cell biology and cancer research: Phenotype, karyotype, tumorigenicity, and stress-induced genome-phenotype evolution. *Gene* **569**, 182–190 (2015).
64. L. Bylund, S. Kytölä, W. O. Lui, C. Larsson, G. Weber, Analysis of the cytogenetic stability of the human embryonal kidney cell line 293 by cytogenetic and STR profiling approaches. *Cytogenet. Genome Res.* **106**, 28–32 (2004).
65. R. L. Binz, *et al.*, Identification of novel breakpoints for locus- and region-specific translocations in 293 cells by molecular cytogenetics before and after irradiation. *Sci. Rep.* **9**, 1–10 (2019).
66. F. A. Ran, *et al.*, Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
67. J. P. Zhang, *et al.*, Efficient precise knockin with a double cut HDR donor after CRISPR/Cas9-mediated double-stranded DNA cleavage. *Genome Biol.* **18**, 1–18 (2017).
68. Z. Liu, *et al.*, Systematic comparison of 2A peptides for cloning multi-genes in a polycistronic vector. *Sci. Rep.* **7**, 1–9 (2017).
69. J. Vieira, J. Messing, The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. *Gene* **19**, 259–268 (1982).
70. J. M. Celeste Yanisch-Perron, Jeffrey Vieira, Improved M13 phage cloning vectors and host strains: nucleotide sequences

of the M13mpl8 and pUC19 vectors. *Gene* **33**, 103–119 (1985).
71. Y. Zhang, *et al.*, CRISPR-Cpf1 correction of muscular dystrophy mutations in human cardiomyocytes and mice. *Sci. Adv.* **3** (2017).
72. Z. Gao, A. Harwig, B. Berkhout, E. Herrera-Carrillo, Mutation of nucleotides around the +1 position of type 3 polymerase III promoters: The effect on transcriptional activity and start site usage. *Transcription* **8**, 275–287 (2017).
73. S. Lin, B. T. Staahl, R. K. Alla, J. A. Doudna, Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *Elife* **3**, e04766 (2014).
74. S. Okamoto, Y. Amaishi, I. Maki, T. Enoki, J. Mineno, Highly efficient genome editing for single-base substitutions using optimized ssODNs with Cas9-RNPs. *Sci. Rep.* **9**, 1–11 (2019).
75. Z. Xie, *et al.*, Optimization of a CRISPR/Cas9-mediated Knock-in Strategy at the Porcine Rosa26 Locus in Porcine Foetal Fibroblasts. *Sci. Rep.* **7**, 1–12 (2017).
76. M. Liu, *et al.*, Methodologies for improving HDR efficiency. *Front. Genet.* **10**, 1–9 (2019).
77. M. A. Jordan, D. Thrower, L. Wilson, Effects of vinblastine, podophyllotoxin and nocodazole on mitotic spindles. Implications for the role of microtubule dynamics in mitosis. *J. Cell Sci.* **102**, 401–416 (1992).
78. C. D. Yeh, C. D. Richardson, J. E. Corn, Advances in genome editing through control of DNA repair pathways. *Nat. Cell Biol.* **21**, 1468–1478 (2019).
79. S. M. Ryu, J. W. Hur, K. Kim, Evolution of CRISPR towards accurate and efficient mammal genome engineering. *BMB Rep.* **52**, 475–481 (2019).
80. B. J. Raney, *et al.*, ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic Acids Res.* **39**, 871–875 (2011).
81. J. Zhu, *et al.*, Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642–654 (2013).
82. G. Liang, *et al.*, Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 7357–7362 (2004).
83. T. B. Morrison, J. J. Weis, C. T. Wittwer, Quantification of lowcopy transcripts by continuous SYBR® green I monitoring during amplification. *Biotechniques* **24**, 954–962 (1998).
84. X. Wu, *et al.*, Intrinsic Immunity Shapes Viral Resistance of Stem Cells. *Cell* **172**, 423-438.e25 (2018).
85. B.-S. Jo, S. S. Choi, Introns: The Functional Benefits of Introns in Genomes. *Genomics Inform.* **13**, 112 (2015).
86. Y. K. Kim, V. N. Kim, Processing of intronic microRNAs. *EMBO*

J. **26**, 775–783 (2007).
87. A. J. Berk, P. A. Sharp, Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. *Cell* **12**, 721–732 (1977).
88. T. E. Koralewski, K. V. Krutovsky, Evolution of exon-intron structure and alternative splicing. *PLoS One* **6** (2011).
89. S. M. Berget, C. Moore, P. A. Sharp, Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 3171–3175 (1977).
90. K. Ito, *et al.*, Identification of pathogenic gene mutations in LMNA and MYBPC3 that alter RNA splicing. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 7689–7694 (2017).
91. A. M. Roy-Engel, *et al.*, Active Alu element "A-tails": Size does matter. *Genome Res.* **12**, 1333–1344 (2002).
92. M. Heydari, et al. Illumina error correction near highly repetitive DNA regions improves de novo genome assembly. *BMC Bioinformatics* **20**, 1–13 (2019).
93. A. E. Minoche, J. C. Dohm, H. Himmelbauer, Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol.* **12** (2011).
94. M. G. Ross, *et al.*, Characterizing and measuring bias in sequence data. *Genome Biol.* **14**, R51 (2013).
95. M. G. Kearse, J. E. Wilusz, Non-AUG translation: A new start for protein synthesis in eukaryotes. *Genes Dev.* **31**, 1717–1731 (2017).
96. M. Jain, *et al.*, Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
97. R. E. Workman, *et al.*, Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* **16**, 1297–1305 (2019).
98. L. Flintoft, Transcriptomics: Revealing the extent of RNA editing. *Nat. Rev. Genet.* **10**, 426–427 (2009).
99. A. Gebhardt, B. T. Laudenbach, A. Pichlmair, Discrimination of self and non-self ribonucleic acids. *J. Interf. Cytokine Res.* **37**, 184–197 (2017).
100. H. Kato, *et al.*, Cell type-specific involvement of RIG-I in antiviral response. *Immunity* **23**, 19–28 (2005).
101. J. W. Schoggins, *et al.*, A diverse range of gene products are effectors of the type i interferon antiviral response. *Nature* **472**, 481–485 (2011).
102. J. Tao, X. Zhou, Z. Jiang, cGAS-cGAMP-STING: The three musketeers of cytosolic DNA sensing and signaling. *IUBMB Life* **68**, 858–870 (2016).
103. A. K. Mankan, *et al.*, Cytosolic RNA:DNA hybrids activate the cGAS –STING axis . *EMBO J.* **33**, 2937–2946 (2014).
104. D. L. Burdette, *et al.*, STING is a direct innate immune

sensor of cyclic di-GMP. *Nature* **478**, 515–518 (2011).
105. A. P. J. de Koning, W. Gu, T. A. Castoe, M. A. Batzer, D. D. Pollock, Repetitive elements may comprise over Two-Thirds of the human genome. *PLoS Genet.* **7** (2011).
106. A. Conti, *et al.*, Identification of RNA polymerase III-transcribed Alu loci by computational screening of RNA-Seq data. *Nucleic Acids Res.* **43**, 817–835 (2015).
107. C. Y. Yu, H. C. Kuo, The emerging roles and functions of circular RNAs and their generation. *J. Biomed. Sci.* **26**, 1–12 (2019).
108. S. Sekar, W. S. Liang, Circular RNA expression and function in the brain. *Non-coding RNA Res.* **4**, 23–29 (2019).
109. D. Van Rossum, B. M. Verheijen, R. J. Pasterkamp, Circular RNAs: Novel regulators of neuronal development. *Front. Mol. Neurosci.* **9**, 1–7 (2016).
110. W. R. Jeck, *et al.*, Circular RNAs are abundant, conserved, and associated with ALU repeats. *Rna* **19**, 426 (2013).
111. S. L. Mehta, R. J. Dempsey, R. Vemuganti, Role of circular RNAs in brain development and CNS diseases. *Prog. Neurobiol.* **186**, 101746 (2020).
112. G. I. Rice, *et al.*, Mutations in ADAR1 cause Aicardi-Goutières syndrome associated with a type i interferon signature. *Nat. Genet.* **44**, 1243–1248 (2012).
113. K. Motomura, *et al.*, A Rho-associated coiled-coil containing kinases (ROCK) inhibitor, Y-27632, enhances adhesion, viability and differentiation of human term placenta-derived trophoblasts in vitro. *PLoS One* **12**, 1–21 (2017).
114. S. W. Lee, *et al.*, Optimization of Matrigel-based culture for expansion of neural stem cells. *Animal Cells Syst. (Seoul).* **19**, 175–180 (2015).
115. D. Zeineddine, A. A. Hammoud, M. Mortada, H. Boeuf, The Oct4 protein: More than a magic stemness marker. *Am. J. Stem Cells* **3**, 74–82 (2014).
116. S. Suzuki, J. Namiki, S. Shibata, Y. Mastuzaki, H. Okano, The neural stem/progenitor cell marker nestin is expressed in proliferative endothelial cells, but not in mature vasculature. *J. Histochem. Cytochem.* **58**, 721–730 (2010).
117. S. Lee, *et al.*, TuJ1 (class III β-tubulin) expression suggests dynamic redistribution of follicular dendritic cells in lymphoid tissue. *Eur. J. Cell Biol.* **84**, 453–459 (2005).
118. S. R. Hutton, L. H. Pevny, SOX2 expression levels distinguish between neural progenitor populations of the developing dorsal telencephalon. *Dev. Biol.* **352**, 40–47 (2011).
119. A. M. Pham, *et al.*, PKR Transduces MDA5-Dependent Signals for Type I IFN Induction. *PLoS Pathog.* **12**, 1–27 (2016).

120. M. A. Langereis, Q. Feng, F. J. van Kuppeveld, MDA5 Localizes to Stress Granules, but This Localization Is Not Required for the Induction of Type I Interferon. *J. Virol.* **87**, 6314–6325 (2013).
121. Peter A. Lemaire, Eric Anderson, Jeffrey Lary, Mechanism of PKR Activation by dsRNA. *J Mol Biol* **381**, 351–360 (2008).
122. A. Herbert, Z-DNA and Z-RNA in human disease. *Commun. Biol.* **2**, 1–10 (2019).
123. Y. Jin, W. Zhang, Q. Li, Origins and evolution of ADAR-mediated RNA editing. *IUBMB Life* **61**, 572–578 (2009).
124. L. F. Grice, et al., The origin of the ADAR gene family and animal RNA editing. *BMC Evol. Biol.* **15**, 1–7 (2015).
125. H. T. Porath, *et al.*, A-to-I RNA editing in the earliest-diverging eumetazoan phyla. *Mol. Biol. Evol.* **34**, 1890–1901 (2017).
126. L. Keegan, A. Khan, D. Vukic, M. O'Connell, ADAR RNA editing below the backbone. *Rna* **23**, 1317–1328 (2017).
127. B. Shapiro, R. W. Graham, B. Letts, A revised evolutionary history of armadillos (Dasypus) in North America based on ancient mitochondrial DNA. *Boreas* **44**, 14–23 (2015).
128. A. F. Gombart, T. Saito, H. P. Phillip, Exaptation of an ancient Alu short interspersed element provides a highly conserved vitamin D-mediated innate immune response in humans and primates. *BMC Genomics* **10**, 1–11 (2009).
129. Y. Quentin, Origin of the alu family: A family of alu-like monomers gave birth to the left and the right arms of the alu elements. *Nucleic Acids Res.* **20**, 3397–3401 (1992).
130. J. Häsler, K. Strub, Alu elements as regulators of gene expression. *Nucleic Acids Res.* **34**, 5491–5497 (2006).
131. G. Churakov, A. F. A. Smit, J. Brosius, J. Schmitz, A novel abundant family of retroposed elements (DAS-SINEs) in the nine-banded armadillo (Dasypus novemcinctus). *Mol. Biol. Evol.* **22**, 886–893 (2005).
132. C. G. Sotero-Caio, R. N. Platt, A. Suh, D. A. Ray, Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biol. Evol.* **9**, 161–177 (2017).
133. F. Shao, M. Han, Z. Peng, Evolution and diversity of transposable elements in fish genomes. *Sci. Rep.* **9**, 1–8 (2019).
134. A. Herbert, Mendelian disease caused by variants affecting recognition of Z-DNA and Z-RNA by the Zα domain of the double-stranded RNA editing enzyme ADAR. *Eur. J. Hum. Genet.* **28**, 114–117 (2020).
135. N. Paz-Yaacov, *et al.*, Adenosine-to-inosine RNA editing shapes transcriptome diversity in primates. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 12174–12179 (2010).
136. H. Kaneko, *et al.*, DICER1 deficit induces Alu RNA toxicity in age-related macular degeneration. *Nature* **471**, 325–332

(2011).
137. S. Weingarten-Gabbay, *et al.*, Comparative genetics: Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science.* **351** (2016).
138. C. K. Pfaller, R. C. Donohue, S. Nersisyan, L. Brodsky, R. Cattaneo, Extensive editing of cellular and viral double-stranded RNA structures accounts for innate immunity suppression and the proviral activity of ADAR1 p150. *PLoS Biol* (2018).
139. J. Galipon, R. Ishii, Y. Suzuki, M. Tomita, K. Ui-Tei, Differential binding of three major human ADAR isoforms to coding and long non-coding transcripts. *Genes* **8**, 1–13 (2017).
140. M. Kozak, Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res.* **12**, 857–872 (1984).
141. S. Nakagawa, Y. Niimura, T. Gojobori, H. Tanaka, K. ichiro Miura, Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Res.* **36**, 861–871 (2008).
142. H. Xu, *et al.*, Length of the ORF, position of the first AUG and the Kozak motif are important factors in potential dual-coding transcripts. *Cell Res.* **20**, 445–457 (2010).
143. S. Lee, *et al.*, Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U. S. A.* **109** (2012).
144. B. B. and A. A. T. Jonathan Bohlen, Kai Fenzl, Günter Kramer, Selective 40S footprinting reveals that scanning ribosomes remain cap- tethered in human cells. *bioRxiv* (2019).
145. A. V. Kochetov, *et al.*, uORFs, reinitiation and alternative translation start sites in human mRNAs. *FEBS Lett.* **582**, 1293–1297 (2008).
146. O. W. Merten, M. Hebben, C. Bovolenta, Production of lentiviral vectors. *Mol. Ther. Methods Clin. Dev.* **3**, 16017 (2016).
147. S. Charrier, *et al.*, Quantification of lentiviral vector copy numbers in individual hematopoietic colony-forming cells shows vector dose-dependent effects on the frequency and level of transduction. *Gene Ther.* **18**, 479–487 (2011).
148. M. Geraerts, S. Willems, V. Baekelandt, Z. Debyser, R. Gijsbers, Comparison of lentiviral vector titration methods. *BMC Biotechnol.* **6**, 1–10 (2006).
149. M. Barak, *et al.*, Purifying selection of long dsRNA is the first line of defense against false activation of innate immunity. *Genome Biol.* **21**, 26 (2020).
150. H. Poulsen, J. Nilsson, C. K. Damgaard, J. Egebjerg, J. Kjems, CRM1 Mediates the Export of ADAR1 through a Nuclear

Export Signal within the Z-DNA Binding Domain. *Mol. Cell. Biol.* **21**, 7862–7871 (2001).
151. Alexander Strehblow, Martina Hallegger, Nucleocytoplasmic Distribution of Human RNA-editing Enzyme ADAR1 Is Modulated by Double- stranded RNA-binding Domains, a Leucine-rich Export Signal, and a Putative Dimerization Domain. *Mol. Biol. Cell* **13**, 4100–4109 (2002).
152. M. Sakurai, *et al.*, ADAR1 controls apoptosis of stressed cells by inhibiting Staufen1-mediated mRNA decay. *Nat. Struct. Mol. Biol.* **24**, 534–543 (2017).
153. J. Ji, L. A. Loeb, Fidelity of HIV-1 Reverse Transcriptase Copying RNA in Vitro. *Biochemistry* **31**, 954–958 (1992).
154. M. L. Kotewicz, J. M. D'Alessio, K. M. Driftmier, K. P. Blodgett, G. F. Gerard, Cloning and overexpression of Moloney murine leukemia virus reverse transcriptase in Escherichia coli. *Gene* **35**, 249–258 (1985).
155. R. J. Orton, *et al.*, Distinguishing low frequency mutations from RT-PCR and sequence errors in viral deep sequencing data. *BMC Genomics* **16**, 1–15 (2015).
156. G. A. Van der Auwera, *et al.*, From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics* 1-33 (2013).
157. A. Dobin, *et al.*, STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
158. B. A. Veeneman, S. Shukla, S. M. Dhanasekaran, A. M. Chinnaiyan, A. I. Nesvizhskii, Two-pass alignment improves novel splice junction quantification. *Bioinformatics* **32**, 43–49 (2016).
159. E. Jay, R. Wu, Arthrobacter luteus Restriction Endonuclease Recognition Sequence and its Cleavage Map of SV40 DNA. *Biochemistry* **15**, 3612–3620 (1976).
160. E. Ullu, S. Murphy, M. Melli, Human 7SL RNA consists of a 140 nucleotide middle-repetitive sequence inserted in an Alu sequence. *Cell* **29**, 195–202 (1982).
161. E. Ullu, C. Tschudi, Alu sequences are processed 7SL RNA genes. *Nature* **312**, 171–172 (1984).
162. P. Deininger, Alu elements: Know the SINEs. *Genome Biol.* **12**, 1–12 (2011).
163. P. Miura, S. Shenker, C. Andreu-Agullo, J. O. Westholm, E. C. Lai, Widespread and extensive lengthening of 39 UTRs in the mammalian brain. *Genome Res.* **23**, 812–825 (2013).
164. M. K. Sakharkar, V. T. K. Chow, P. Kangueane, Distributions of exons and introns in the human genome. *In Silico Biol.* **4**, 387–393 (2004).
165. D. S. C. Cho, *et al.*, Requirement of dimerization for RNA editing activity of adenosine deaminases acting on RNA. *J. Biol. Chem.* **278**, 17093–17102 (2003).

166. K. A. Chilibeck, *et al.*, FRET analysis of in vivo dimerization by RNA-editing enzymes. *J. Biol. Chem.* **281**, 16530–16535 (2006).
167. L. Valente, K. Nishikura, RNA binding-independent dimerization of adenosine deaminases acting on RNA and dominant negative effects of nonfunctional subunits on dimer functions. *J. Biol. Chem.* **282**, 16054–16061 (2007).
168. H. Ota, *et al.*, ADAR1 forms a complex with dicer to promote MicroRNA processing and RNA-induced gene silencing. *Cell* **153**, 575–589 (2013).
169. G. Clerzius, *et al.*, ADAR1 Interacts with PKR during Human Immunodeficiency Virus Infection of Lymphocytes and Contributes to Viral Replication. *J. Virol.* **83**, 10119–10128 (2009).
170. K. Kawakubo, C. E. Samuel, Human RNA-specific adenosine deaminase (ADAR1) gene specifies transcripts that initiate from a constitutively active alternative promoter. *Gene* **258**, 165–172 (2000).
171. S. Lykke-Andersen, S. Piñol-Roma, J. Kjems, Alternative splicing of the ADAR1 transcript in a region that functions either as a 5′-UTR or an ORF. *RNA* **13**, 1732–1744 (2007).
172. E. C. Freund, *et al.*, Unbiased Identification of trans Regulators of ADAR and A-to-I RNA Editing ll ll Unbiased Identification of trans Regulators of ADAR and A-to-I RNA Editing. *Cell Reports* **31**, 107656 (2020).
173. N. M. Gaudelli, *et al.*, Programmable base editing of T to G C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).
174. L. Qu, *et al.*, Programmable RNA editing by recruiting endogenous ADAR using engineered RNAs. *Nat. Biotechnol.* **37**, 1059–1069 (2019).
175. G. Aquino-Jarquin, Novel engineered programmable systems for ADAR-mediated RNA editing. *Mol. Ther. Nucleic Acids* **19**, 1065–1072 (2020).
176. A. C. Komor, *et al.*, Improved base excision repair inhibition and bacteriophage Mu Gam protein yields C:G-to-T:A base editors with higher efficiency and product purity. *Sci. Adv.* **3**, 1–10 (2017).
177. N. C. Wu, *et al.*, High-throughput profiling of influenza A virus hemagglutinin gene at single-nucleotide resolution. *Sci. Rep.* **4**, 1–8 (2014).
178. M. B. Doud, J. D. Bloom, Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. *Viruses* **8**, 1–17 (2016).
179. C. B. Jabara, C. D. Jones, J. Roach, J. A. Anderson, R. Swanstrom, Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 20166–20171 (2011).

180. I. Kinde, J. Wu, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 9530–9535 (2011).
181. T. H. Zhang, N. C. Wu, R. Sun, A benchmark study on error-correction by read-pairing and tag-clustering in amplicon-based deep sequencing. *BMC Genomics* **17**, 1–9 (2016).
182. J. B. Hiatt, R. P. Patwardhan, E. H. Turner, C. Lee, J. Shendure, Parallel, tag-directed assembly of locally derived short sequence reads. *Nat. Methods* **7**, 119–122 (2010).
183. L. E. MacConaill, *et al.*, Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics* **19**, 1–10 (2018).
184. Y. Fu, P. H. Wu, T. Beane, P. D. Zamore, Z. Weng, Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genomics* **19**, 1–14 (2018).
185. S. G. Acinas, R. Sarma-Rupavtarm, V. Klepac-Ceraj, M. F. Polz, PCR-induced sequence artifacts and bias: Insights from comparison of two 16s rRNA clone libraries constructed from the same sample. *Appl. Environ. Microbiol.* **71**, 8966–8969 (2005).