2020

Characterizing the RNA Editing Specificityof ADAR Isoforms and Deaminase Domains In Vitro

Mariel Therese Bartley

CHARACTERIZING THE RNA EDITING SPECIFICITY OF ADAR ISOFORMS AND DEAMINASE DOMAINS *IN VITRO*

A Thesis Presented to the Faculty of

The Rockefeller University

in Partial Fulfillment of the Requirements for

the degree of Doctor of Philosophy

by

Mariel Therese Bartley

June 2020

CHARACTERIZING THE RNA EDITING SPECIFICITY OF ADAR ISOFORMS AND DEAMINASE DOMAINS *IN VITRO*

Mariel Therese Bartley, Ph.D.

The Rockefeller University 2020

Adenosine deaminases acting on RNA (ADARs) convert adenosine-to-inosine in double stranded RNA. Selectivity of editing sites depends on the sequence of the RNA as well as the secondary structure. Identification of sites of ADAR editing by editome analysis is skewed due to the different abundance of each adenosine-containing triplet, as well as the presence of complex RNA structures.

To determine the editing specificity of each ADAR protein, a high throughput sequencing based assay was developed to measure editing in a synthetic dsRNA substrate with one of each of the 16 different adenosine-containing triplets, so that each possible editing site was equally represented.

The ADAR1- and ADAR2-deaminases, as well as the full-length ADAR2 and isoforms ADAR1-p150 and ADAR1-p110, were purified and activity was measured for each, so that the inherent activity of the deaminase domains could be characterized and then compared to the editing patterns seen for the longer proteins containing dsRNA binding domains and Z-RNA binding domains.

The ADAR1 deaminase was found to skew slightly to favoring 5'A editing sites, while ADAR2-deaminase favored 5'U. From homology modelling of ADAR1 onto the ADAR2 crystal structure, this difference in editing specificity could be due to the ADAR1 protein having a weaker interaction with the orphan base of the RNA substrate.

Characterization of full-length ADAR1-p110 and ADAR2 found that each full-length protein had less editing of the UAG triplet than the respective deaminases, while ADAR1-p110 increased editing of the AAG triplet and ADAR2 increased editing of the CAG triplet. Comparing ADAR1 isoforms p110 and p150 found no significant differences in editing specificity.

The *in vitro* assay was also used to confirm the inactivity of the ADAR3 deaminase, and probe and characterize the editing specificity of the ADAR3 deaminase mutant A389V, which rescued editing activity. The pattern of editing was similar to that of the ADAR1-deaminase, despite ADAR3 sharing sequence similarity with ADAR2.

Insights into the different patterns of substrate selectivity by ADAR deaminases and the likely causes of these differences, can provide insight for future development of ADAR deaminase constructs for site-directed RNA editing.

A pilot experiment for characterizing *in vitro* editing in complex RNA was also performed, using HEK 293T total RNA and reovirus T1L RNA as substrates for the purified ADAR constructs. An adapted sequencing library preparation was successfully used to identify and count individual editing events in reovirus T1L RNA, and further improvements are required to generate enough editing sites to compare editing frequencies in both substrates to the 50bp *in vitro* substrate. The goal of characterizing editing in these complex substrates will be to identify RNA secondary structures which are differentially edited by the full-length ADAR1-p110 and ADAR2, when compared to the dsRBD-lacking deaminase constructs.

For George and Jane

A course more promising

Than a wild dedication of yourselves

To unpath'd waters, undream'd shores

The Winter's Tale, William Shakespeare

ACKNOWLEDGEMENTS

Thanks to Charlie, for steering me and this project in the right direction, and for building a lab of the best people. To all the members of the Rice lab, thanks for making every day interesting.

Thank you to Meigang Gu and Hachung Chung, for large scale advice on the direction of this project, and Jorg Calis and Brad Rosenberg, for specific assistance with assay development.

Many thanks to Joe Luna, Mohsan Saeed and Stephanie Sarbanes, for constant advice and support, for both scientific and non-scientific issues.

Christine Lai, Kate Rozen-Gagnon, Connie Zhao, John Fak, Caryn Hale, and Elisabeth Murphy all assisted with RNA sequencing, Dom Olinares and Brian Chait collaborated on mass spectrometry, Deena Oren provided advice and instrumentation for protein purification, Ji-Dung Luo performed bioinformatic analysis, and Pradeep Ambrose aided with coding.

Thanks to Danica Sutherland, Pavithra Aravamudhan and Alison Ashbrook, for reovirus protocols and reagents.

And to my committee, for the advice through the years that led to this thesis.

To my family, for unconditional support from half a world away.

Stephanie, Josh, Mel and Dan, I wouldn't have survived New York without you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter I – Introduction

Post-transcriptional and post-translational regulation

The complexity of an organism is dependent on both the total number of gene products available and the regulation of expression of those genes. To this end, organisms require ways of increasing the variation in the transcriptome, and consequently the proteome, without increasing the amount of information in the genome or risking damage to the DNA (Athanasiadis et al. 2004; Garncarz et al. 2013).

This increase in variation can be due to post-translational modifications at the protein level, such as the addition of phosphate or ubiquitin moieties, or through post-transcriptional modification at the RNA level. Modifications of RNA affect both the number of gene products available, through mechanisms such as alternative splicing and editing, but also affect the non-coding RNAs involved in regulation (Keegan et al. 2001; Barraud & Allain 2012).

Post-transcriptional regulation of non-coding RNAs are a significant source of organismal complexity, as even though only ~1.5% of the genome is translated into a functional protein, the majority of the genome is associated with at least one RNA transcript (ENCODE Project Consortium et al. 2007), implying a huge pool of RNAs involved in non-protein-coding roles within cells.

The combination of increased protein diversity and complex gene regulation allows for time-dependent and site-specific modulation of cellular function (Avesson & Barry 2014). Individual cells have a specific milieu of protein and RNA to define their role in the organism, without the risk of damaging the information source: as regulation at the protein and RNA level allows conservation at the DNA level.

RNA editing, a specific post-transcriptional modification

One form of post-transcriptional regulation is RNA editing, which involves the insertion, deletion or, more commonly, alteration of a nucleotide within a polynucleotide RNA substrate (Keegan et al. 2001). Editing of RNA can affect the coding potential, stability and localization of RNA (Serra et al. 2004; Garncarz et al. 2013), which then causes time-dependent and site-specific changes in the cell. The primary modifications made to RNA are conversion of single bases from cytosine to uracil (C-to-U) or adenosine to inosine (A-to-I), with A-to-I the most widespread editing type in higher eukaryotes (Keegan et al. 2001). This widespread A-to-I editing is primarily carried out by proteins Adenosine Deaminase Acting on RNA 1 (ADAR1) and ADAR2, a pair of proteins that specifically edit adenosines in double-stranded RNA. Editing within coding regions can lead to changes in the translated protein, such as an amino acid substitution or introduction or deletion of a premature stop codon, as C-to-U edits swap out one RNA base for another, and A-to-I edits introduce a non-standard base which is recognized as guanosine (G) by translation and reverse-transcription machinery (Avesson & Barry 2014). One such example is the amber/W site in hepatitis delta virus (HDV) that requires editing by ADARs to change the UAG sequence into UGG, which is read through to produce a longer isoform of the viral protein HDAg (Casey et al. 2006). The shorter isoform S-HDAg is required for replication of HDV, and the longer isoform L-HDAg is required for formation of HDV particles; editing by ADAR is a required component of the HDV viral lifecycle to shift from replication to packaging.

Outside of a small number of species, such as octopus and squid, that have majority editing in coding regions (Alon et al. 2015), editing-induced coding changes are a relatively rare outcome, with up to 85% of editing events occur in non-coding regions for most species (Athanasiadis et al. 2004; Ramaswami & Li 2014). Outcomes of editing in non-coding RNA include disruption of

RNA secondary structure by introduction of I:U or U:G mismatches – which can affect protein binding interactions and potentially lead to RNA degradation – as well as altered pre-mRNA splicing and miRNA processing and targeting –due to sequence changes in splice site donor, acceptor or regulatory sequences or within miRNA sequences (Valente & Nishikura 2005; Toth et al. 2006; Hogg et al. 2011).

The discovery and characterization of ADAR proteins

Although the term 'RNA editing' was first coined following the observation by Benne et al. (1986) of the insertion of four uridines into the trypanosome mitochondrial cytochrome oxidase subunit II (*COXII*) transcript, the majority of RNA editing events are through base conversion, not insertion. Shortly after this discovery in trypanosomes, the conversion of cytosine to uracil was observed by Powell et al. (1987) in apolipoprotein B (*APOB*) transcripts.

In that same year, two groups (Bass & Weintraub 1987; Rebagliati & Melton 1987) working with *Xenopus laevis* oocyte extracts identified what was initially thought to be a helicase protein, causing the unwinding of double stranded RNA (dsRNA). The unwinding was not due to helicase activity, but due to conversion of adenosines in the duplex to inosines (Bass & Weintraub 1988). The newly introduced inosines formed I:U wobble pairs, destabilizing the duplex and leading to the aforementioned unwinding activity. Polson et al. (1991) then determined that the mechanism of adenosine to inosine conversion was through hydrolytic deamination. The solved crystal structure of the *Escherichia coli (E. coli)* cytidine deaminase (*cdd*) by Betts et al. (1994) illustrated that cytidine-to-uridine conversion also uses a mechanism of hydrolytic deamination. The mechanism for both A-to-I and C-to-U deamination is shown in **figure 1.1** (Gerber & Keller 2001).

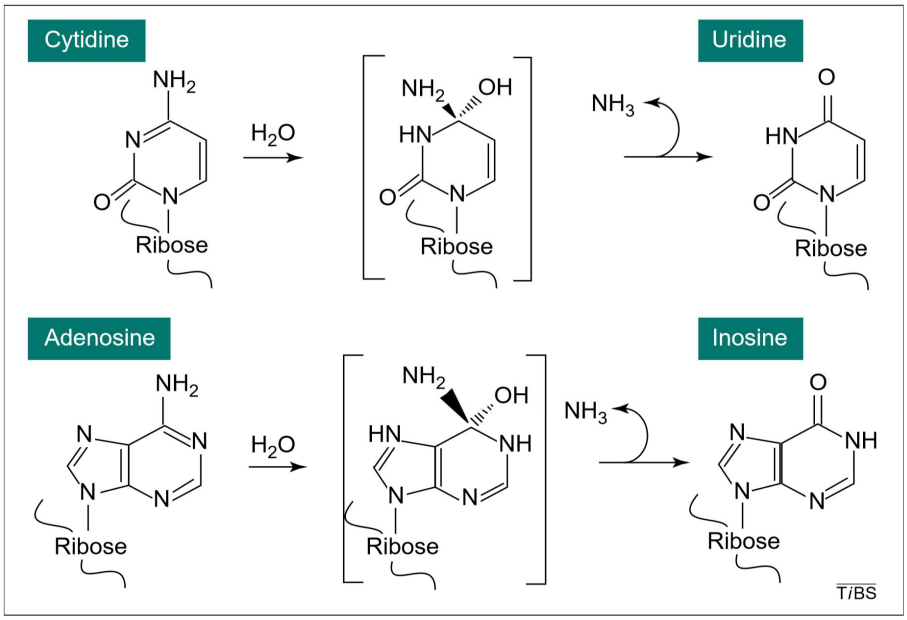

Figure 1.1. Deamination mechanism used by AID/APOBEC and ADAR proteins. Both families utilize hydrolytic deamination to remove the amine group, with AID/APOBEC deaminating the C4 of the cytidine base to produce uridine, and ADAR deaminating C6 of the adenosine base to produce inosine. Inosine is recognized as guanine by splicing and translation machinery (Gerber & Keller 2001).

Following the initial observation of A-to-I editing in 1987, the family of proteins responsible was identified and cloned in the years that followed. The human homolog of the protein responsible for the observed A-to-I editing in *Xenopus* extracts was cloned in 1994 by Kim et al. and dubbed double-stranded RNA adenosine deaminase (*DRADA*). Shortly after, Patterson and Samuel (1995) observed that the protein encoded by *DRADA* exists in two main size isoforms. As more members of the ADAR family were discovered the field developed a standardized nomenclature, detailed by Bass et al. (1997). At this stage *DRADA* was renamed *ADAR*, and the expressed protein ADAR1. Shortly thereafter, Melcher et al. (1996a) identified and cloned a second A-to-I editing

gene from the rat genome, dubbed dsRNA-Specific Editase 1 (*Red1*); the human *RED1* was subsequently cloned by O'Connell et al. (1997) and this gene was later renamed *ADARB1*, encoding protein ADAR2 (Bass et al. 1997).

Concurrent to the identification of the two active enzymes in the ADAR family, two inactive proteins were also discovered. First, in 1995, Schumacher et al. cloned Testis Nuclear RNA-Binding Protein (*TENR*) from mice, a testes-specific protein that lacks the active site residues. Patterson and Samuel (1995) observed that the mouse *TENR* shared sequence homology with the human *DRADA*, indicating that the proteins were likely in the same family. *TENR* was later renamed Adenosine Deaminase Containing 1 (*ADAD1*) to match the standardized nomenclature previously set out (Bass et al. 1997; Keegan et al. 2004). Lastly, the brain specific *Red2* was first cloned from rats by Melcher et al. (1996b) and then the human *RED2* was cloned by Chen et al. (2000) and renamed *ADARB2*, encoding protein ADAR3. Both groups observed that ADAR3 is restricted to specific regions of the brain and has no apparent editing activity. From sequence alignment, the active site residues are still present, so the mechanism of inactivity must be different to that of ADAD1, but as of yet has not been determined.

The protein responsible for the C-to-U editing in apoB transcripts observed in 1987 by Powell et al., was later identified as Apolipoprotein B mRNA Editing enzyme, Catalytic polypeptide-like 1 (*APOBEC1*) (Teng et al. 1993). Alignment of the APOBEC1 sequence, along with other members of the Activation Induced cytidine Deaminase (AID, encoded by *AICDA*) / APOBEC family discovered later, show similar deaminase motifs by sequence alignment (Conticello et al. 2005). Although the AID/APOBEC family have a wider range of editing substrates, with some members of the family editing DNA rather than RNA, all members use the same catalytic mechanism as the ADAR family (Conticello et al. 2005, Salter et al. 2016).

Evolution of the CDA superfamily

Although the ADAR and AID/APOBEC families target different nucleotides for editing, all proteins have been determined to belong to the Cytidine Deaminase-like superfamily (Hogg et al. 2011), supported by the observation that all of the proteins use the same mechanism of hydrolytic deamination to convert bases within polynucleotides. Sequence alignment identifies similar catalytic domains, both to each other and to the mononucleotide deaminating enzyme Cytidine Deaminase (*CDA*), which is a single-domain protein that converts free cytosine nucleosides into uridines (Jin et al. 2009). Some primordial form of CDA is the most likely ancestor of the modern editing enzymes. Multiple analyses by Jin et al. (2009), Keegan et al. (2011), Grice & Degnan (2015), and Kohn et al. (2015) have determined a likely evolutionary development pathway for the AID/APOBEC/ADAR families within the CDA superfamily.

From a primordial CDA, a likely duplication event – and mutations in the residues in the active site excluding cytosine and allowing adenosine – led to a split between the cytidine-editing CDA and the newer adenosine-editing protein. This new adenosine-specific protein was an ADAT2-like protein. The modern Adenosine Deaminase Acting on tRNA 2 (*ADAT2*) is a eukaryotic tRNA editing enzyme that is similarly present in bacteria as the *tadA* protein and in yeast as Tad2p (*TAD2)*, indicating the appearance of ADAT2-like proteins occurred before the split of eukaryotes and prokaryotes (Grice & Degnan 2015). Further expansion of the tRNA-editing enzymes likely occurred early in eukaryotic development, with mirrored expansion in both protozoa and metazoa: in yeast the Tad2p-like protein later duplicated into Tad1p and the inactive Tad3p, while in metazoa the *ADAT2*-like gene duplicated into *ADAT1* and the inactive *ADAT3* (Gerber & Keller 2001). The last metazoan ADAT2-like protein to appear, ADAT1, is theorized to be the ancestor of the ADAR family.

ADAR genes have not been identified in plant, fungal or yeast genomes (Jin et al. 2009), indicating their emergence after the metazoa/protozoa split. However, ADAR proteins must have developed shortly after this, as ADARs are present in cnidarians, including anemones and hydrozoans, as well as sponges and ctenophores, which represent one of the earliest splits in the metazoan lineage (Grice & Degnan 2015, Kohn et al. 2015). This implies that ADAR proteins came into existence before the split of Eumetazoa from sponges and ctenophores.

At this stage all of the RNA editing enzymes were still single domain proteins, with only a catalytic domain. The emergence of ADAR proteins likely occurred following a duplication of an *ADAT1*-like gene and the acquisition of a double-stranded RNA binding domain (dsRBD) to produce the ancestral ADAR2. The most likely split of the ancestral ADAR2 and ancestral ADAR1 also occurred before the Eumetazoa/sponge split, as an ADAR1-like protein is present in sponges and anemones. The ancestral ADAR1-like protein was distinguished from the ADAR2-like protein by the acquisition of a Z-DNA binding domain (ZBD) (Grice & Degnan 2015).

Later, in early vertebrates, the ADAR2-like protein likely duplicated again, gained a ssRNA binding patch and lost enzymatic activity to produce the ADAR3 protein (Jin et al. 2009). A second duplication of ADAR2 during early vertebrate development, and subsequent loss of activity and restriction to the testes led to the appearance of ADAD1.

The maintenance of both inactive proteins in vertebrates – ADAR3 is conserved in all tested vertebrates from fish to human (Jin et al. 2009), and ADAD1 has been identified in mice, rats, macaques and humans – indicates that these proteins have a necessary regulatory role in these species. It is more likely that ADAR3 and ADAD1 represent two different duplication events from ADAR2, rather than a single duplication event followed by the proteins becoming tissue-specific and then losing catalytic activity through different means.

The presence of ADAR proteins throughout metazoan development indicates that ADAR-mediated editing could have played a role in gene regulation during this development, potentially contributing to the emergence of complex tissue organization. The high levels of ADAR1 observed during early stages of development (Wang et al. 2004; Samuel 2012) and the specific expression in neural tissue and editing of brain-specific transcripts implies a role for ADAR in the development of the nervous system (Hoopengardner et al. 2003).

Particularly, the emergence of ADAR proteins in metazoan development mirrors the emergence of a complex nervous system, and continued expression in neural tissue implies a continued role in nervous system function.

In more recent evolutionary times, ADAR1 has been lost in some species, including some insects and crustaceans, while ADAR2 is typically maintained (Keegan et al. 2011). For example, *Drosophila melanogaster* (*D. melanogaster*) only has a single ADAR2-like gene, and no ADAR1-like equivalent. A small number of species have ADAR-like proteins that cannot be defined as either ADAR1 or ADAR2, specifically nematodes and flatworms. This includes *Caenorhabditis elegans*, which has genes *adr-1* and *adr-2* that are not distinctly identified as homologs of either ADAR1 or ADAR2.

The CDA-like enzymes which still recognized cytosine did not diversify until much later. The ancestral AID-like protein only appears in vertebrates, followed by duplication into APOBEC2 and APOBEC4 soon after. In mammals, two more APOBECs appeared: APOBEC3 from an AID-like ancestor and APOBEC1 from an AID-like or APOBEC4-like ancestor. APOBEC3 then underwent a rapid expansion into APOBEC3A-H specifically in primates (Salter et al. 2016). The much later evolution of the APOBEC family relative to the ADAR family, could be due to the role of APOBECs as a correction mechanism. Increased gene sequence variation and organismal

complexity require a compensation mechanism such as DNA/RNA editing by APOBECs to maintain genomic stability (Brennicke et al. 1999). The expansion of the APOBEC3 family in primates is especially interesting, considering that the majority of ADAR1 editing targets are in Alu elements, which are also primate specific (Daniel et al. 2014; Levanon & Eisenberg 2015).

Characteristics of the ADAR Family Members

The three members of the modern ADAR family in humans are ADAR1, ADAR2, and ADAR3, encoded by *ADAR, ADARB1,* and *ADARB2,* respectively. As seen in **figure 1.2** (Tomaselli et al. 2013) all three proteins contain a catalytic deaminase domain at the C-terminal end, two or three double-stranded RNA binding domains (dsRBDs), and a nuclear localization signal (NLS). Some members of the family also contain Z-DNA/Z-RNA binding domains (ZBDs), a nuclear export signal (NES) and an R domain that can bind ssRNA/ssDNA (Samuel 2001; Bass 2002).

ADAR1 has two significant size isoforms, ADAR1-p150 and ADAR1-p110, generated by alternative promoter usage and alternative splicing (Patterson & Samuel 1995; George & Samuel 1999. The longer isoform, ADAR1-p150, is 1226 amino acids (aa) and expressed from an interferon-inducible promoter. ADAR1-p110, the shorter isoform, starts from M296 of the longer isoform, generating a 931aa protein from a constitutive promoter (Kim et al. 1994; O'Connell et al. 1995; Herbert et al. 1997). ADAR2 and ADAR3 are also expressed from constitutive promoters. Whilst ADAR1-p150 and -p110 are named for their apparent size when run on a sodium dodecyl sulfate-polyacrylamide gel (Patterson and Samuel 1995; George et al. 2005), that is 150kDa and 110kDa respectively, the proteins are actually smaller when calculated from the amino acid sequence: ADAR1-p150 is ~133kDa and ADAR1-p110 is ~103kDa. The smaller ADAR2 (701aa) and ADAR3 (739aa) proteins are calculated to be ~77kDa and ~80kDa respectively.

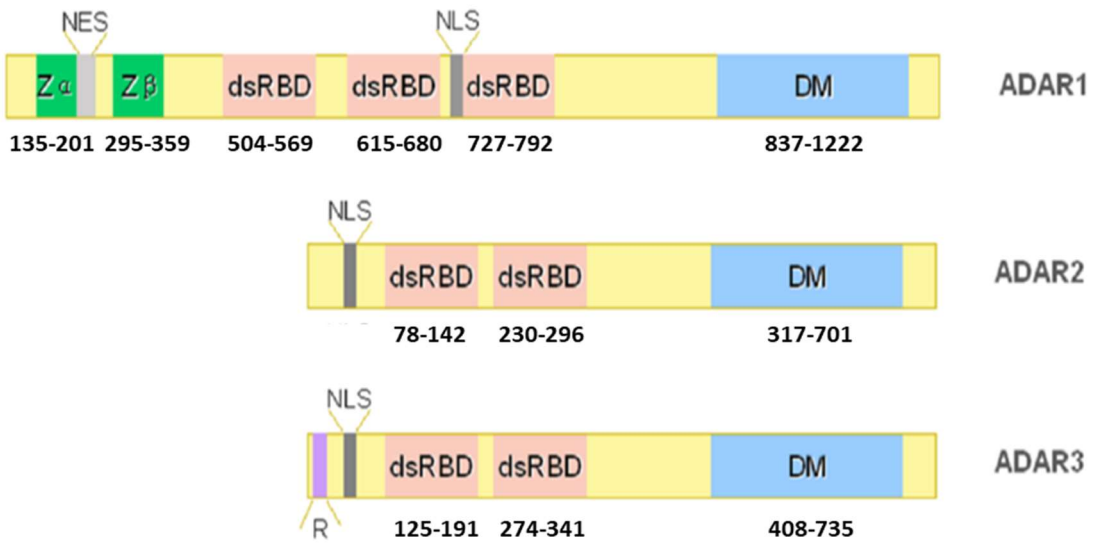

Figure 1.2. Structure of ADAR family proteins: ADAR1, ADAR2, and ADAR3. The ADAR enzymes contain a C-terminal conserved catalytic deaminase domain (DM), two or three dsRBDs in the N-terminal portion. ADAR1 full-length protein also contains a N-terminal Zα domain with a nuclear export signal (NES) and a Zβ domain, while ADAR3 has a R-domain. A nuclear localization signal is also indicated (Tomaselli et al. 2013). Domain boundaries are shown below each protein domain.

Expression across cell types

ADAR1-p110 is ubiquitously expressed across many tissue types, with highest mRNA levels observed in the brain and lungs (Kim et al. 1994; O'Connell et al. 1995; Patterson & Samuel 1995). ADAR1-p150 is similarly ubiquitously expressed across many tissue types, albeit at much lower levels than ADAR1-p110, and has increased levels following infection. Interestingly, George *et al.* (2005) found that in the mouse, upregulation of ADAR1-p150 after infection led to increased expression in many tissues, but not in the brain. Similar to ADAR1-p110, ADAR2 is ubiquitously expressed, with low levels in most tissues and highest expression in the brain (Melcher et al. 1996a). ADAR3 is exclusively expressed in the brain, with highest levels observed in the thalamus

and amygdala (Chen et al. 2000). As previously mentioned, the expansion of the ADAR family of proteins paralleled the development of complex nervous systems, and the specific expression of ADAR proteins in the brain further illustrates the point that these proteins are likely playing an as-yet-to-be-determined role in the complex post-transcriptional regulation occurring in the brain.

Localization within cells

All three ADAR proteins possess a putative nuclear localization signal (NLS) (Eckmann et al. 2001; Strehblow et al. 2002) and have been observed to be exclusively nuclear, with the exception of the longer ADAR1-p150 isoform that also contains a nuclear export signal (NES) in the N-terminal sequence absent in ADAR1-p110 (Poulsen et al. 2001). Due to having both an NLS and NES, ADAR1-p150 has been observed to be primarily cytoplasmic, with the ability to shuttle between the cytoplasm and nucleus (Patterson and Samuel 1995; Poulsen et al. 2001). Both ADAR2 and ADAR3 have a "classic" NLS (cNLS), that interacts with members of the importin alpha family to facilitate nuclear localization (Maas & Gommans 2009), with the cNLS sequence in the unstructured N-terminal region of each protein. The ADAR1 NLS is non-canonical, with the motif split into two regions of the protein. Half of the residues are upstream of the 3rd dsRBD and the remaining residues are immediately downstream of the dsRBD. An additional α-helix present in this dsRBD brings the N- and C-terminal flanking sequences into close proximity to form the full NLS motif. Nuclear localization of ADAR1, unlike the other two family members, is facilitated by transportin 1 (Barraud et al. 2014). Strehblow et al. (2002) observed that nuclear localization was impeded when RNA was bound to the protein, and NMR structures of the 3rd dsRBD by Barraud et al. (2014) identified the mechanism causing the inhibition. The group observed that binding of dsRNA to the domain is a steric hindrance to binding of transportin 1, allowing either RNA binding or nuclear localization but not both.

Characteristics of the protein domains

Double-stranded RNA binding domains

As shown in **figure 1.2**, ADAR2 and ADAR3 possess two dsRBDs, while ADAR1 contains three. ADAR3 also contains an R-rich domain, which has been shown to bind to ssRNA and ssDNA (Melcher et al. 1996b; Chen et al. 2000).

Double-stranded RNA binding domains (dsRBD) are found in a number of viral, prokaryotic and eukaryotic proteins with diverse functions and proteins with dsRBDs generally have multiple copies of the motif with varying affinities for dsRNA (Chang & Ramos 2005; Lunde et al. 2007). They have a conserved structural motif of an α-helix followed by three anti-parallel β-sheets and then a second α-helix. This structural conservation allows the domain to recognize the RNA sugar-phosphate backbone, which gives dsRBD the ability to distinguish between RNA and DNA but not to recognize different sequences (Stefl et al. 2005). Shown in **figure 1.3** (Barraud & Allain 2012) is a sequence alignment of a number of dsRBDs from ADAR family members from human, *d. melanogaster* and *c. elegans*. A trio of lysine residues form the conserved binding motif KKxxK, highlighted in red in helix α2, and these residues project into the major groove of the RNA (Chang & Ramos 2005; Poulsen et al. 2006). Mutation of the motif from KKxxK to EAxxA has been used by multiple groups to decrease RNA binding affinity from the nM range to mM (Chilibeck et al. 2006; Valente & Nishikura 2007; Ota et al. 2013; Barraud et al. 2014). The Interaction is stabilized by contact between the minor grooves on either side with less conserved residues in the α1 helix (LxxLxxL) and loops 2 (PxHxP) and 4 (T/S) of the dsRBD. Overall, the dsRBD binding footprint covers ~20bp of the RNA, across two minor grooves.

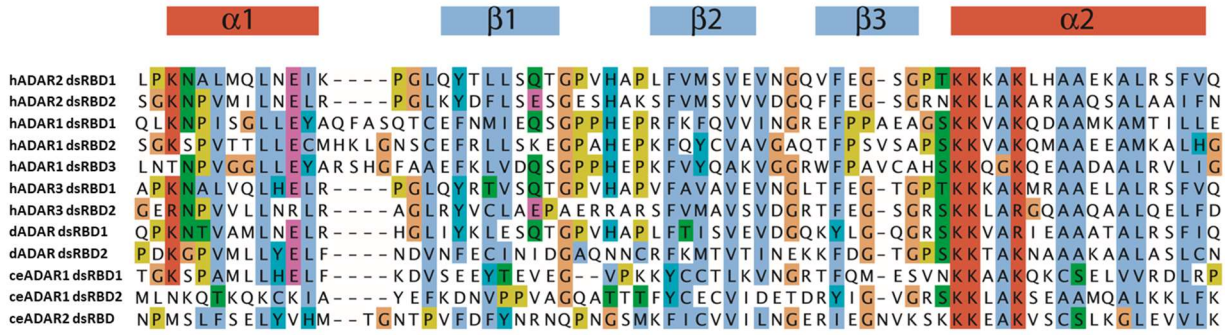

Figure 1.3. Sequence homology of dsRBD of ADAR in human, d. melanogaster and c. elegans, with conserved residues in color. The location of the two α-helices and three β-sheets is shown above the sequences, as are the 4 unstructured loops. The conserved KKxxK motif can be seen highlighted in red in the α2 helix at the C-terminal end (Barraud & Allain 2012).

There is some evidence that if the RNA substrate is distorted from a perfectly helical shape, the dsRBD domains can interact with the RNA in a sequence-specific manner, as opposed to the majority of interactions between dsRBDs and RNA, which occur along the RNA backbone in a non sequence-specific manner. Interactions between dsRBDs and the bases in RNA have been shown to interact directly in hairpin loop structures (Ramos et al. 2000; Wu et al. 2004; Tian et al. 2004). Stefl et al. (2006, 2010) showed, using NMR solution structures, binding of the ADAR2 dsRBD2 to a well-characterized ADAR2 substrate, *GLURB* R/G, with a number of interactions between the protein and specific bases of the substrate.

Z-DNA binding domains

ADAR1-p150 contains two Z-DNA/Z-RNA binding domains (ZBDs) at the N-terminal end of the protein, that can bind to left-handed Z-form nucleic acid (Athanasiadis et al. 2005). Z-form nucleic acid is an alternative, higher energy, form of nucleic acid, which is left-handed as opposed to A-

and B-form nucleic acid which are right-handed. The formation of Z-form nucleic acid can be induced and stabilized following binding of a ZBD (Lee et al. 2012; Ng et al. 2013), which is mediated through a combination of hydrogen bonding to the backbone, van der Waals interactions and coordination of water molecules (Schwartz et al. 1999). ZBD binding is not necessary for RNA editing by ADAR1, but editing is enhanced when Zα is bound to Z-RNA (Koeris et al. 2005; Herbert & Rich 2001). The roles of the ZBD in ADAR function have not yet been determined.

ZBDs have a conserved structural motif of three α-helices followed by a β-hairpin. Ha et al. (2009) crystallized the ADAR1 Zα domain in complex with Z-DNA, showing the binding interaction and identifying residues involved in the interaction. Mutation of these residues to K169A and Y177A is able to disrupt Z-form binding, confirming their role in binding (Ng et al. 2013). Several of the important binding residues in Zα are missing in Zβ, which is not able to bind Z-form nucleic acid, and the role of Zβ is not yet known (Schwartz et al. 1999; Barraud et al. 2012; Athanasiadis et al. 2005). The shorter ADAR1-p110 isoform also contains the Zβ domain but, as the domain is non-functional, the ADAR1-p110 protein does not have the ability to bind Z-DNA/Z-RNA.

The specific role of the Zα domain in ADAR1-p150 has yet to be determined but it has been determined that the domain is required for normal ADAR1 function, as some of the mutations associated with the Type I interferonopathy Aicardi-Goutières Syndrome (AGS) are found in the Zα domain (Rice et al. 2012). A potential role for the Zα domain could be to localize ADAR1-p150 to sites of active transcription or other sites of complex left-handed RNA structure (Herbert & Rich 2001). Interestingly, the E3L protein of vaccinia virus also contains a ZBD, and is a potent inhibitor of ADAR1 editing activity (Liu et al. 2001).

Deaminase domain

The CDA superfamily is defined by the C-terminal deaminase domain present in all members of the family (Schaub et al. 2002). From sequence alignment – and based on crystal structures for some members of the family (Betts et al. 1994; Macbeth et al. 2005) – all family members contain a distinct zinc-dependent deaminase (ZDD) motif: a histidine and two cysteines, often separated into two or three separate regions of the domain. Some members of the family have the cysteines in close proximity to each other, while for others the residues are distant in the sequence but brought together in the secondary structure. A fourth residue, a glutamate that forms a HxE motif with the histidine, is believed to mediate proton transfer between a coordinated H_2O molecule and the adenosine (or cytosine) base, as illustrated previously in **figure 1.1** (Gerber & Keller 2001).

The editing mechanism for ADAR1 is presumed, based on sequence alignments the crystal structures available for ADAR2. Mutating the ADAR1 active site glutamate (E912A) and histidine (H910Q) kills activity, evidence that these residues are components of the active site (Liu & Samuel 1999; Valente & Nishikura 2007). In 2005, Macbeth et al. first crystallized the deaminase domain of ADAR2, illustrating the zinc ion coordinated by the ZDD, which in ADAR2 is formed by residues H394, E396, C451 and C516. This also was the first observation that ADAR2 has a cofactor: an inositol hexakisphosphate (IP_6) molecule was found buried in the core, presumably as a stabilizer. IP_6 was found to be required for protein activity, and similarly is required for ADAT1 activity indicating that IP_6 is likely present in all ADAR/ADAT proteins.

Prior to 2016, the orientation of adenosine in the active site was theorized through modelling adenosine monophosphate (AMP) into the crystal structure (Pokharel et al. 2009; Mizrahi et al. 2012). For the adenosine to sit in the active site, it required the base to be free from the double-stranded substrate RNA. Sequence analysis of Xenopus and human ADARs by Hough and Bass

(1997) showed similarities between ADAR and DNA cytosine methyltransferases, including the presence of the ZDD residues. Shortly before this observation, a crystal structure by Cheng et al. (1993) of the bacterial methyltransferase *hhaI* showed the mechanism of action by these proteins: the targeted cytosine was flipped out of the RNA and embedded in the protein active site. Due to the similarity in sequences, Hough and Bass theorized that ADARs could be using the same base-flipping mechanism to access the target adenosine. This mechanism was finally observed for ADAR2 with the 2016 crystal structure by Matthews et al., which captured the ADAR2 deaminase in complex with a dsRNA substrate. The trapped conformation was achieved by replacing the adenosine with the analogue base 8-azanebularine, which cannot be fully deaminated and so gets trapped in complex with the glutamate and water molecule (Mizrahi et al. 2012). They also used an RNA substrate with an A:C mismatch that is easier to flip out than an A:U pair (Källman et al. 2003), and mutated a glutamate residue that interact with the orphaned cytosine to a glutamine, as the E488Q mutation has previously been shown to increase editing efficiency (Kuttan & Bass, 2012; Phelps et al. 2015).

Figure 1.4 shows the crystal structure of the ADAR2 deaminase domain in complex with a dsRNA substrate (Matthews et al. 2016). The targeted adenosine is visible in the active site, flipped out of the duplex RNA, and the orphaned cytosine is in close proximity to the E488Q residue of the protein. The loop of protein that interacts with the opposing RNA strand is theorized to be the mechanism for base-flipping: the protein intercalates into the RNA, forcing the two strands of RNA to loosen and allowing the adenosine to flip. This base-flipping loop is highlighted yellow in **figure 1.5**, which shows the alignment of the ADAR1, ADAR2 and ADAR3 deaminases. Interestingly, the glutamate residue that interacts with the orphan base (E488 in ADAR2, E1008 in ADAR1) is flanked by glycine residues, likely making the base-flipping loop highly flexible.

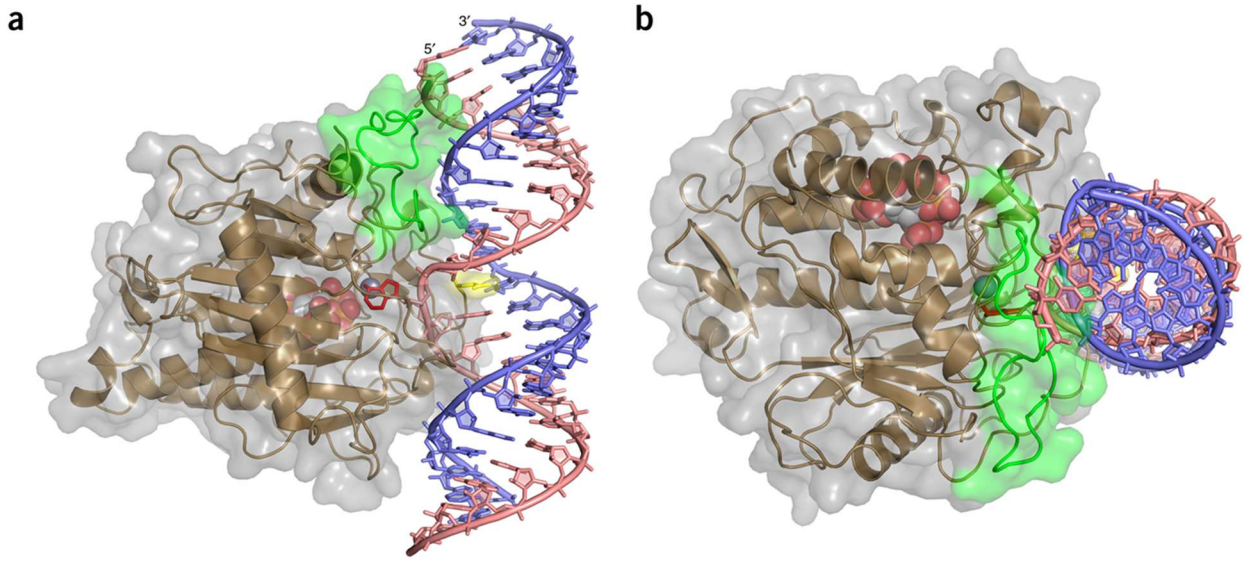

Figure 1.4. Crystal structure of ADAR2 deaminase in complex with dsRNA, from Matthews et al. (2016) [PDB 5ED1] from side (a) and top (b) view. The analogue base 8-azanebularine is shown in red, flipped out of the duplex RNA and coordinated by residues in the active site; the zinc ion is visible as a grey sphere behind the base in (a). The IP6 cofactor buried in the core of the protein is visible as a red space filling model. The 5' binding loop is shown in green, in contact with the 5' end of the RNA substrate. These residues were not visible in the previous crystal structure by Macbeth et al. (2005) [PDB 1ZY7]. The mutated residue E488Q is highlighted yellow and is in contact with the orphaned base.

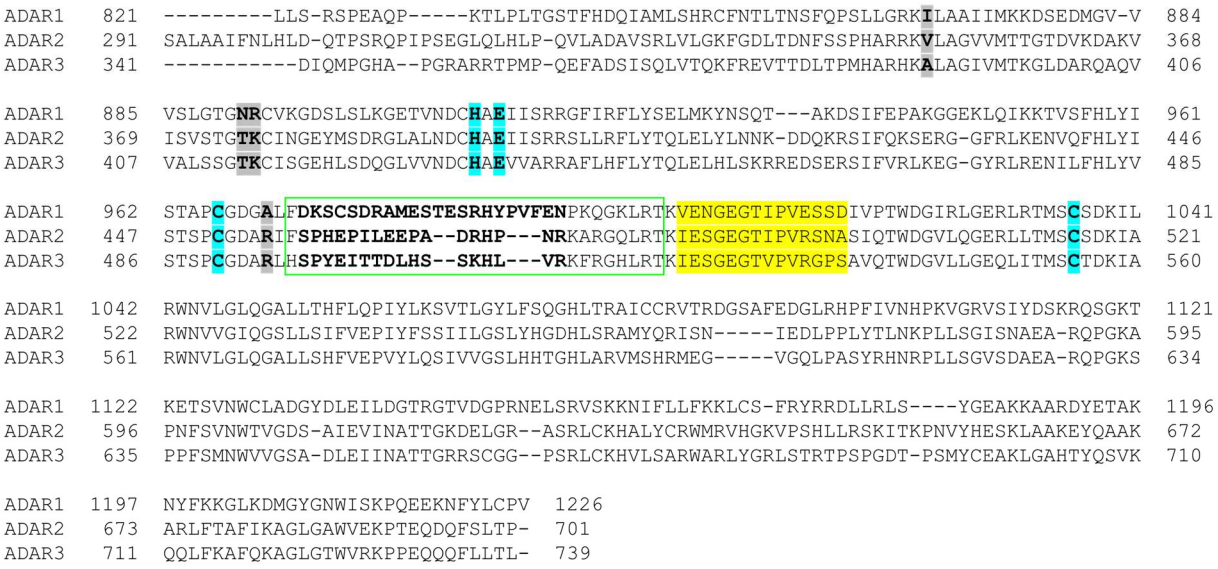

Figure 1.5. Sequence alignment of the deaminase domains from human ADAR1, ADAR2 and ADAR3. The zinc-dependent deaminase motif is highlighted in cyan, other residues that line the active site pocket are colored grey. The loop that intercalates into the RNA to interact with the orphaned base is highlighted yellow and the 5' binding loop is shown in the green box, matching the green residues in Figure 4. Bolded residues in the green box are those that are only structured when in contact with RNA.

The glycine immediately upstream of the glutamate in ADAR1 is the site of one of the most deleterious AGS disease mutations, G1007R, which decreases editing activity of the protein (Rice et al. 2012), implying that the interaction with the orphan base is highly linked to the activity level of the protein.

Figure 1.4 also illustrates a region of protein that was unstructured in the previous ADAR2 crystal structure (Macbeth et al. 2005) but becomes structured in the presence of substrate RNA: the 5' binding loop, so named as it interacts with the 5' end of the adenosine-containing RNA strand. In **Figure 1.5**, this region is highlighted by a green box, with the unstructured residues in bold. The ADAR1 loop has five additional residues and so the ADAR2 crystal structure cannot accurately predict the structure of this region for ADAR1.

The alignment of deaminase sequences from ADAR1, ADAR2 and ADAR3 shows the relationship between the proteins that has been used to determine their evolutionary distance. The ADAR1 sequence is equally similar to both ADAR2 (55% similar, 39% identical) and ADAR3 (54% similar, 39% identical), which follows if ADAR1 and ADAR2 split early in metazoan development. ADAR3, as a likely duplication of ADAR2 in vertebrates, shares 75% similarity and 56% identity with the ADAR2 protein. The ZDD residues that coordinate the active site zinc and H_2O are highlighted cyan and are present in ADAR3 despite the lack of activity. Other residues that line the active site are highlighted grey, with ADAR2 and ADAR3 sharing the same residues except for one: V351 / A389. The ADAR1 active site residues are more divergent. While ADAR2 and ADAR3 have residues V/A-T-K-R, ADAR1 has I-N-R-A. The most significant difference is A970, which in ADAR2 is R455. From crystal structures of ADAR2, the arginine interacts with the N7 of the purine adenosine. The smaller alanine residue in ADAR1 likely coordinates a H_2O molecule that bridges the gap between the residue and base (Mizrahi et al. 2012).

Overall, the deaminase has a footprint of ~20bp along the length of the substrate RNA, with three main regions of contact: the adenosine in the active site, the base-flipping loop that contacts the orphan base, and the 5' binding loop that interacts with the 5' end of the substrate, of differing lengths in ADAR1 and ADAR2.

ADAR editing substrates

General RNA characteristics

Early work characterizing ADAR editing found that RNA duplexes >100bp with perfectly or near-perfectly matched base pairs undergo hyper-editing, with promiscuous deamination of 50-60% of adenosines in the substrate (Bass & Weintraub 1988; Nishikura et al. 1991). As substrate length decreased, so did editing efficiency. Shorter RNAs with mismatches, bulges and internal loops decreased overall editing levels and increased the selectivity of editing sites, with <10% editing (Lehmann & Bass 1999, 2000).

Nishikura et al. (1991) determined that a 15bp RNA duplex was not sufficient for editing by ADAR1, but a duplex of 23bp was editable. This length is just beyond the expected footprint of the deaminase domain interaction, at 20bp. Modelling of the ADAR2 deaminase in complex with duplex RNA of this length by Thomas & Beal (2017) showed that editing of a substrate this short would be due to binding by the deaminase domain alone, as the RNA structure would be distorted by the interaction and not allow dsRBD to bind. Modelling the interaction for a 33bp duplex, they observed that both the deaminase and a single dsRBD would be able to bind this slightly longer substrate.

ADAR2 editing substrates

ADAR2 has several well characterized editing targets in coding regions of brain-specific transcripts. These targets include the Glutamate Receptor subunit B (GluR-2, encoded by *GLURB*) (Liu & Samuel 1999; Lehmann & Bass 2000) and the Serotonin 2C Receptor (5-HT_{2C}R, encoded by *5HT2C*) (Burns et al. 1997; Liu et al. 1999). The *5HT2C* transcript has at least five adenosines (designated sites A-E) which are edited to different efficiencies by ADAR1 and ADAR2. As many as 24 different transcript sequences have been found, that have different patterns of expression in different regions of the brain; the subsequent changes in protein residues affect alternative splicing and affinity for binding partners.

The *GLURB* pre-mRNA has two sites, Q/R and R/G, that are targeted by ADAR2; the residue changes due to editing at these sites decrease Ca^{2+} permeability of the Glutamate receptor ion pore. The brain specific nature of ADAR2 targets, coupled to the brain specific expression of ADAR3, implies that ADAR2/ADAR3 editing could be involved in transcriptional regulation in the brain. The *GLURB* Q/R site is especially notable, as editing by ADAR2 is ~100% efficient in the adult brain. The level of editing is lower, and variable across different regions of the brain, during development. ADAR3 has been shown to be able to bind to the *GLURB* pre-mRNA, suppressing the level of ADAR2 editing (Chen et al. 2000; Oakes et al. 2017b).

ADAR1 editing substrates

Unlike ADAR2, the majority of ADAR1 editing sites have been found in non-coding regions, split evenly between introns and 3'UTRs (Chung et al. 2018). Most of these events are localized to Alu repeat elements (Kim et al. 2004; Ramaswami & Li 2013; Daniel et al. 2014; Tan et al. 2017; Chung et al. 2018). Daniel et al. argues that editing is most likely occurring in inverted repeat Alu

elements, which form long duplexes due to the similar sequences of each Alu, rather than occurring in individual Alu elements, which form complex tertiary structures. The purpose of ADAR1 editing of non-coding RNA is likely to prevent self-RNA from activating innate immune responses. Absence of ADAR editing in mice was found to be embryonic lethal (Liddicoat et al. 2015), with the lethal phenotype rescued by concurrently knocking down the innate immune sensor MDA-5. Chung et al. (2018) found that ADAR$^{-/-}$ neural progenitor cells also exhibited MDA-5 dependent immune activation, as well as activation of the dsRNA sensor PKR. ADAR1 may be preventing activation of PKR and MDA-5 by decreasing the pool of double-stranded self-RNA.

Structure of RNA editing substrates

The difference in substrate pools for ADAR1 and ADAR2 implies that there are distinct interactions between the deaminase domain, dsRBDs and substrate RNA for each proteins. Work by Thomas and Beal (2017) and Wang et al. (2018) to analyze the RNA substrates favored by ADAR1 and ADAR2 have found that the RNA structure and presence of loops and bulges is a strong determinant for substrate selection – especially considering that dsRBDs primarily bind the RNA backbone and do not have much sequence specificity. A number of RNA structures favored by each ADAR protein are shown in **figure 1.6** (Thomas & Beal 2017; Wang et al. 2018).

Both the full-length ADAR1 and the ADAR1 deaminase favor RNA with loops or more complex structures 3 to 6bp 5' of the editing site, such as in human transcripts *GLI1*, *AZIN1*, and *NEIL1*, and yeast transcripts *HER1* and *GSY1* (**1.6 a-e**). The full-length ADAR2 favors substrates with long duplex RNA 3' of the active site, with mismatches occurring at least 9-10bp downstream, such as the two *GLURB* substrates Q/R (**1.6 f**) and R/G (**1.6 g**). The solution structures of ADAR2

dsRBDs binding to the *GLURB* R/G substrate by Stefl et al. (2006, 2010) illustrated this specificity: the two dsRBDs of ADAR2 bound to the region of the duplex containing the mismatches, downstream of the editing site.

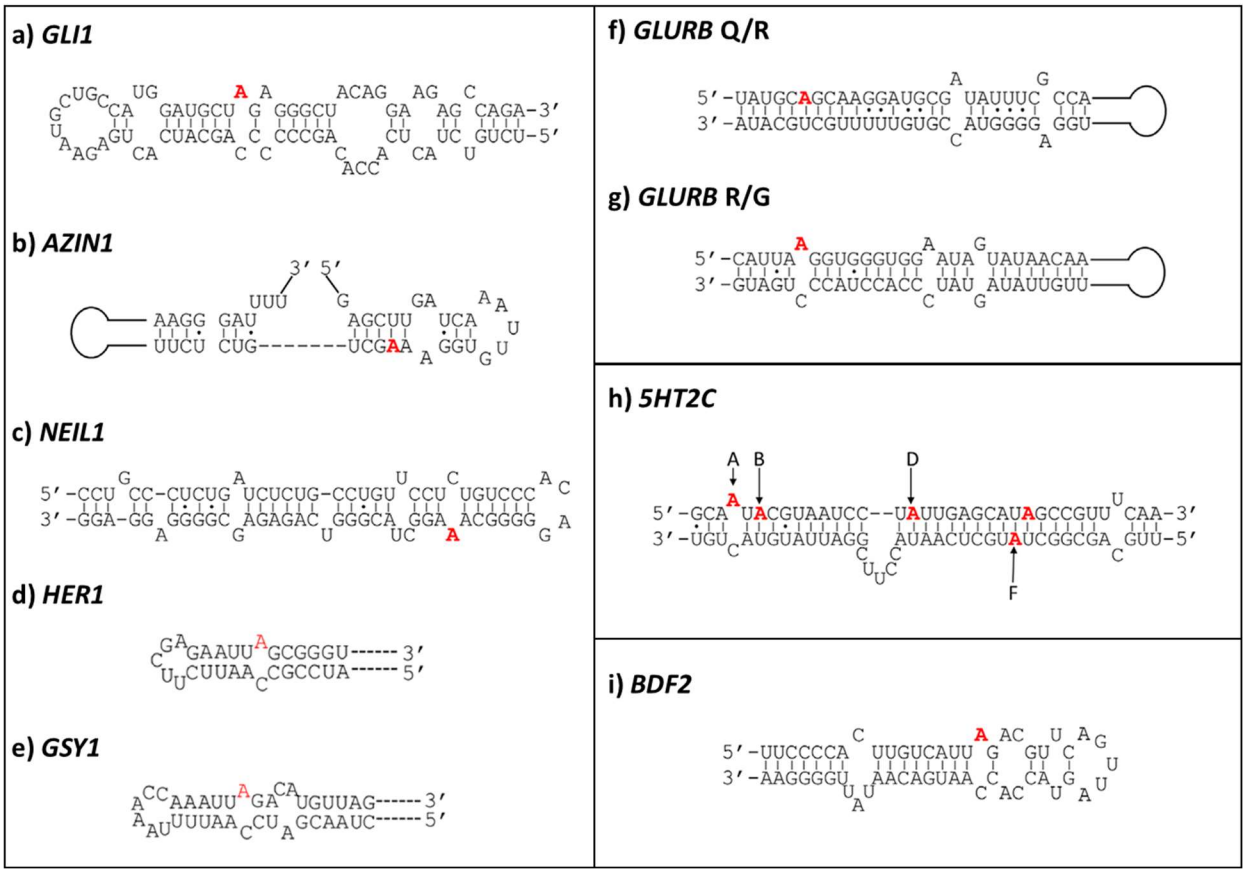

Figure 1.6. Structures of RNA substrates edited by ADAR1 (a-e), ADAR2 (f-g), both (h) and ADAR2-deaminase (i), with targeted adenosines highlighted in red. ADAR1 targets include the human transcripts *GLI1* (a), *AZIN1* (b), and *NEIL1* (c) and the yeast transcripts *HER1* (d) and *GSY1* (e). ADAR2 targets are the *GLURB* transcripts at the Q/R (f) and R/G (h) sites. The *5HT2C* transcript (h) is a target for both ADAR1 and ADAR2, while the yeast-derived *BDF2* (i) substrate is edited by the ADAR2 deaminase. Figure adapted from Thomas & Beal 2017, with yeast substrates from Wang et al. (2018).

As ADAR1 has specificity dependent on the RNA structure 5' of the editing site, and ADAR2 depends on the 3' structure, there is an RNA substrate that contains both of these characteristics: the human *5HT2C* transcript (**1.6 h**). Sites A and B are edited by ADAR1, and though not shown in the figure, there is a loop 4bp upstream of the editing sites, and other loop 7bp upstream. The two un-highlighted adenosines between sites B and D are edited by both ADAR1 and ADAR2 to low efficiency. Site D is edited specifically by ADAR2, and the expected mismatch appears 15bp downstream.

The ADAR2 deaminase has a different substrate specificity to the full-length protein, with observed editing in fully-duplex regions, such as site F in the *5HT2C* transcript (**1.6 h**), as well as favoring editing in sites at the 5' edge of loops, such as in yeast *BDF2* (**1.6 i**). This tendency to edit in loops had been previously observed by Phelps et al. (2015).

Development of methods for measuring catalytic activity

Adenosine deaminase editing activity on dsRNA was first observed by Bass and Weintraub (1988), following incubation of *Xenopus* nuclear extract with a 518bp dsRNA. When the RNA was digested and the bases separated by Thin Layer Chromatography (TLC), they observed loss of signal for adenosine (A) and an increased signal co-migrating with guanosine (G). Further TLC that could separate G from the similar base inosine (I) showed the newly generated base was co-migrating with I, not G. Whilst early measurements of ADAR activity utilized radiolabeled TLC to compare the signal intensity of A and I, this later transitioned over to using fluorescent substrates. One particular assay uses an RNA substrate with a UAG stop codon in a hairpin loop. Following ADAR editing of the site, there is readthrough of the hairpin and translation of a downstream open reading frame, generating the fluorescent signal (Oakes et al. 2017a).

To be able to identify specific A-to-I editing events, rather than measuring the overall activity level, Sanger sequencing has been used. Koeris et al. (2005) measured editing using this method by incubating cytosolic extracts from HEK 293T overexpressing ADAR1 with a synthetic dsRNA. Following RNA isolation, reverse transcription and cloning into a plasmid, individual clones were sequenced and the ratio of A:G at each adenosine position was counted to determine the editing percentage, which ranged from negligible to ~70% edited.

With the recent advances and availability of High Throughput Sequencing, identification of many editing sites in the transcriptome concurrently is now feasible (Ramaswami & Li 2014). 'Editomes' – the A-to-I sites within a transcriptome – have been generated for many different species, including *D. melanogaster* (Yablonovitch et al. 2017), domestic cows (Bakhtiarizadeh et al. 2018), sheep (Zhang et al. 2019a), pigs (Zhang et al. 2019b), wild boar (Yang et al. 2019), octopus, squid and cuttlefish (Liscovitch-Brauer et al. 2017). In humans, an editome 'atlas' has been generated, with editomes for many different tissues allowing in depth analysis of how editing patterns change in different tissues (Tan et al. 2017).

Characterising editing site sequence preferences

Along with selection of RNA substrates by structure, the sequence surrounding the targeted adenosine also plays a role in how editable specific adenosines are. By identifying which adenosines are edited to high or low efficiency and then observing the sequence context, general rules for which 5' and 3' neighbors are allowed have been determined. The neighbor preferences initially determined for ADAR1 by Polson and Bass (1994) were a 5' neighbor of A = U > C > G, and a 3' neighbor of G = C > U > A. Using a sequencing-based method to count editing events in a 795bp synthetic dsRNA, Eggington et al. (2011) updated the neighbor preference, determining the 5' neighbor favors U > A > C > G and the 3' neighbor favors G > C > A > U for both ADAR1

and ADAR2. Both Eggington et al. and Bahn et al. (2012) extended the analysis to sites -5 to +5 of the editing site but found no discernible pattern beyond the immediate neighbors.

From the crystal structure of ADAR2, Matthews et al. (2016) modelled different bases into the 5' and 3' positions of the substrate RNA and found that a 5'C or 5'G would have a steric clash with the protein, explaining why these 5' neighbors are disfavored. The 3' neighbor position can accommodate any of the four bases, but only a 3'G can act as a hydrogen bond donor to stabilize the active site conformation, thus why a 3'G is far more prevalent than the other three options. Targeting of editing sites is not only dependent on what is accommodated by the active site however, as the *GLURB* Q/R site – edited to ~100% in the adult brain – is a C<u>A</u>G, which would be sterically unfavorable. The high level of editing of this site is specifically dependent on strong binding of the dsRBDs (Öhman et al. 2000).

Additional complications arise when neighbor preferences are determined from editome data. Counting all editing events in an editome does not consider transcript abundance, the percentage of editing at each location or even the relative abundance of each triplet in the transcriptome. This was illustrated by Wang et al. (2018) when they introduced human ADAR1 into *S. cerevisiae* to identify RNA structures that favor editing. The resulting editome was almost entirely skewed to U<u>A</u>G, whereas in a human editome the ADAR1 neighbor preferences show similar levels of A<u>A</u>G, U<u>A</u>G and C<u>A</u>G editing, as well as most other possible combinations (Chung et al. 2018). This could imply that the yeast transcriptome has a high abundance of UAG relative to all other possible 5' and 3' neighbors for A, and that determining ADAR neighbor preferences from editome data can be fraught.

Use of ADAR proteins for site-directed RNA editing

In recent years, there has been much exploration in the use of CRISPR/Cas9 systems to correct genetic disorders, but risks remain in making permanent changes in genetic material. Utilizing ADAR proteins to target changes in the transcript rather than in the genome could avoid these issues, as the genetic information is untouched, and the edited transcripts degrade over time. The effects of editing can also be dialed in; the level of editing can be controlled by how much guide-RNA is added, and this guide-RNA can be targeted to specific cell types (Jain et al. 2019). This method of Site-Directed RNA Editing (SDRE) is being explored for both A-to-I editing by ADARs (Cox et al. 2017; Vogel et al. 2018; Merkle et al. 2019; Qu et al. 2019) and C-to-U editing by APOBECs (Vu & Tsukahara 2017; Salter & Smith 2018).

There are several different avenues being tested for SDRE systems, but all use a guide-RNA with complementary sequence to the editing target. The guide forms a duplex with the sequence, making the site double-stranded and thus editable by ADAR proteins. Initially designed similar to the CRISPR-Cas system, where a Cas fused to an ADAR-deaminase was used as the effector (Cox et al. 2017), newer methods are being developed to avoid use of Cas proteins.

The two main methods of SDRE development are addition of both exogenous ADAR and a guide-RNA (Vogel et al. 2018), or addition of just a guide-RNA to target endogenous ADARs to the editing sites (Merkle et al. 2019; Qu et al. 2019). Utilizing endogenous ADAR further decreases potential side effects, as the only foreign addition to cells is the guide-RNA, but this method has the downside of only being able to improve targeting efficiency through changes to the guide-RNA. Both Merkle et al. and Qu et al. use the A:C mismatch at the targeted site (Källman et al. 2003) to improve efficiency, but Qu et al. only use guides that form duplexes with the target sequence.

This strategy depends on the A:C mismatch for improved efficiency and decreases off-target effects by use of U:A mismatches next to other adenosines in the sequence, which decreases editing of those sites. Merkle et al. uses a more complicated guide-RNA, with a single-stranded sequence that binds to the target site, followed by a double-stranded segment that mimics the *GLURB* R/G substrate, in an attempt to increase dsRBD binding of the target. Unfortunately, both groups only investigate UAG editing sites, which are known to be easily targeted by ADAR proteins. How these systems would fare against adenosines in difficult-to-edit sequences, such as GAN or CAU, remains to be seen.

The second method of SDRE currently being pursued is the use of guide-RNAs in combination with exogenous ADAR proteins (Vogel et al. 2018). This method has the advantage of being able to modify both the RNA and ADAR to improve editing of the targeted sites. Vogel et al. utilized simple guide-RNAs similar to Qu et al., which form perfect duplexes with the targeted strand, except for a A:C mismatch at the editing site. The exogenous proteins used were the ADAR1 and ADAR2 deaminase domains, conjugated to the guide-RNA by SNAP tags. This group also utilized the E488Q mutation to increase editing by both proteins (Kuttan & Bass, 2012; Phelps et al. 2015). Vogel et al. observed that while the deaminase mutants increased editing at all sites tested relative to the wildtype proteins – with 60-80% editing achieved for adenosines with a 5'A or 5'U, the editing efficiency at adenosines with a 5'G neighbor still had <20% editing. Further understanding of the targeting by ADAR deaminases is required to be able to efficiently edit these sites. The use of targets of RNA-binding proteins identified by editing (TRIBE) has the additional freedom of engineering both the ADAR deaminase specificity and the RNA-binding specificity (McMahon et al. 2016; Xu et al. 2018), as this method utilizes fusions of ADAR catalytic domains to RNA-binding domains of other proteins.

Statement of purpose

The selection of editing targets by ADAR proteins depends on both the sequence and structure of the RNA substrate. Identification of editing sites in transcriptomes, to generate an editome, can be used to identify RNA structures that are favored by ADAR proteins. ADAR specificity for substrate sequences are more difficult to define in this manner. The 5' and 3' neighbors of the targeted adenosine have been shown to strongly restrict which adenosines are edited, with 5'G preventing most editing, and 3'A/U being much less common than 3'G. However, charactering the neighbor specificity can be skewed depending on how editing sites are identified.

Characterising 5' and 3' neighbor preferences using transcriptome data has the tendency to be skewed, as the editome does not contain information about the editing efficiency at each site or abundance of the transcripts. The abundance of each adenosine-containing triplet across the whole transcriptome is also not taken into account. For example, AAA and GAG are highly prevalent in the transcriptome: could editing of these sites be amplified due to the sheer number of sites available.

Work by Wang et al. (2018) to identify RNA structures favored by ADAR1 and ADAR2 introduced these proteins in *S. cerevisiae* and succeeded in identifying RNA structures that each protein favored (**fig1.6**). However, when they plotted the 5' and 3' neighbor frequencies almost all editing events occurred at the U**A**G triplet, which is indicative of the high prevalence of UAG in the yeast transcriptome, and not necessarily indicative of the editing substrate specificity of ADAR1 or ADAR2.

Current developments in SDRE also provide impetus to identify the specific catalytic activity of ADAR proteins, as site-directed editing of specific sites in the transcriptome would require an

ADAR that can efficiently edit that site. Targeting endogenous ADARs to these sites requires those proteins to be able to edit, and difficult-to-edit sites such as 5'C/G sites cannot be efficiently edited with endogenous proteins. To be able to target these sites for SDRE, engineered ADARs with improved editing efficiency for these sites are required. However, this requires knowing the specific editing preferences of each ADAR protein, as well as the editing specificity of the deaminase domains. The specific editing preferences of each ADAR can be characterized *in vitro*, to be able to determine the 5' and 3' neighbor specificity of each proteins in the absence of complicating factors, such as complex RNA structures or transcripts of different abundance.

In this thesis, an *in vitro* editing assay was developed using an RNA substrate with a single instance of each adenosine-containing triplet, so that the relative editing of each site could be directly compared. The specificity for each triplet was characterized for deaminase domains of both ADAR1 and ADAR2, as well as for the full-length ADAR1-p150, ADAR1-p110 and ADAR2, to identify any changes in substrate specificity due to the dsRBDs.

The *in vitro* assay was also used to probe and characterize the editing specificity of the ADAR3 deaminase mutant A389V, to identify the cause of ADAR3 inactivity and determine which residues are important for substrate selectivity by comparing the activity of the ADAR3 mutant with the ADAR1 and ADAR2 deaminases.

A pilot experiment to develop the editing assay for more complex substrates was then performed, measuring low levels of editing activity by the purified ADAR constructs against substrates HEK 293T total RNA and the dsRNA reovirus T1L genome. Improvements in assay conditions are required to generate enough editing sites to compare editing frequencies to the 50bp *in vitro* substrate, and to characterize the role of dsRBDs on substrates with increased length and complexity.

Chapter II – Expression and purification of ADAR proteins

To characterize the editing specificity of the ADAR deaminase domains, and to compare to the specificity of the full-length proteins, the deaminase-only and full-length constructs needed to be expressed and purified. To make these comparisons, the full-length human ADAR1 isoforms ADAR1-p150 (Accession no. P55265-1) and ADAR1-p110 (Accession no. P55265-5), and the full-length human ADAR2 protein (Accession no. P78563) needed to be generated, as did the ADAR1-deaminase (ADAR1-D) and ADAR2-deaminase (ADAR2-D). For analysis of the inactivity of human ADAR3 (Accession no. Q9NS39), the ADAR3-deaminase domain (ADAR3-D) also needed to be generated, as did the ADAR3-deaminase point mutant A389V (ADAR3-D A389V).

Choice of expression system

After initial isolation from cellular extracts (Bass & Weintraub 1987; Polson & Bass 1994), attempts to express ADAR1 from *Escherichia coli* failed, with the isolated protein found to be insoluble (unpublished observations). ADAR proteins have since been successfully expressed from insect cells (Cho et al. 2003), yeast (Macbeth et al. 2004), and mammalian cells (Koeris et al. 2005). The insolubility in bacteria is likely due to the presence of IP_6 in the deaminase core (Macbeth et al. 2005), a molecule that is present in eukaryotes but not in bacteria. For this project the *Spodoptera frugiperda* (Sf9) insect cell expression system was chosen, utilizing the Bac-to-Bac Baculovirus Expression System (Invitrogen) to generate baculoviruses for expression of each ADAR construct (Luckow et al, 1993).

Choice of protein tags

All constructs were generated with an N-terminal SUMOstar, a mutated form of the Smt3 tag. The mutations render the tag uncleavable by the naturally occurring Ulp1 protease in Sf9, only allowing specific cleavage by the mutated Ulp1-sgs protease (Peroutka et al. 2008).

Cleavage of SUMOstar by Ulp1-sgs occurs after the C-terminal Gly-Gly motif of the SUMO, leaving the native N-terminal of the expressed protein intact. The presence of SUMOstar also assists in improved protein folding, and cleavage of the Gly-Gly motif is dependent on the tertiary structure of the protein not the sequence, and therefore observation of tag cleavage is a good indicator that the expressed protein is well folded.

Tags for purification were inserted N-terminal to the SUMOstar, with several different tags tested to see which produced the best level of protein expression for each construct. Tags tested were: 8His-SUMOstar, 8His-3xFLAG-SUMOstar and 8His-Strep-SUMOstar. All constructs were purified by binding of the 8His tag to nickel nitrilotriacetic acid (Ni-NTA) resin, with elution in high imidazole buffer.

Cloning of ADAR constructs

All constructs were cloned into the SpeI/NotI site of the pFASTBAC vector, immediately following the SUMOstar cleavage site. When expressed, following SUMOstar cleavage, the N-terminal of each protein will have a Thr-Ser before the ADAR sequence due to the presence of the SpeI site. Full-length ADAR1-p150 (aa. 1-1226) and ADAR2 (aa. 1-701) ORFs were inserted into a pFASTBAC vector containing an upstream 8His-Strep-SUMOstar tag. The shorter ADAR1-p110 isoform (aa. 296-1226) was subcloned from the ADAR1-p150 sequence and inserted into a pFASTBAC vector containing an upstream 8His-SUMO tag. The ADAR1-D (aa. 832-1226),

ADAR2-D (aa. 299-701) and ADAR3-D (aa. 355-739) were cloned into the pFASTBAC vector containing an upstream 8His-3xFLAG-SUMOstar tag. The ADAR3-D point-mutant A389V was generated by site-directed mutagenesis, replacing the alanine GCG with valine GTG. A schematic of all constructs is shown in **figure 2.1**.

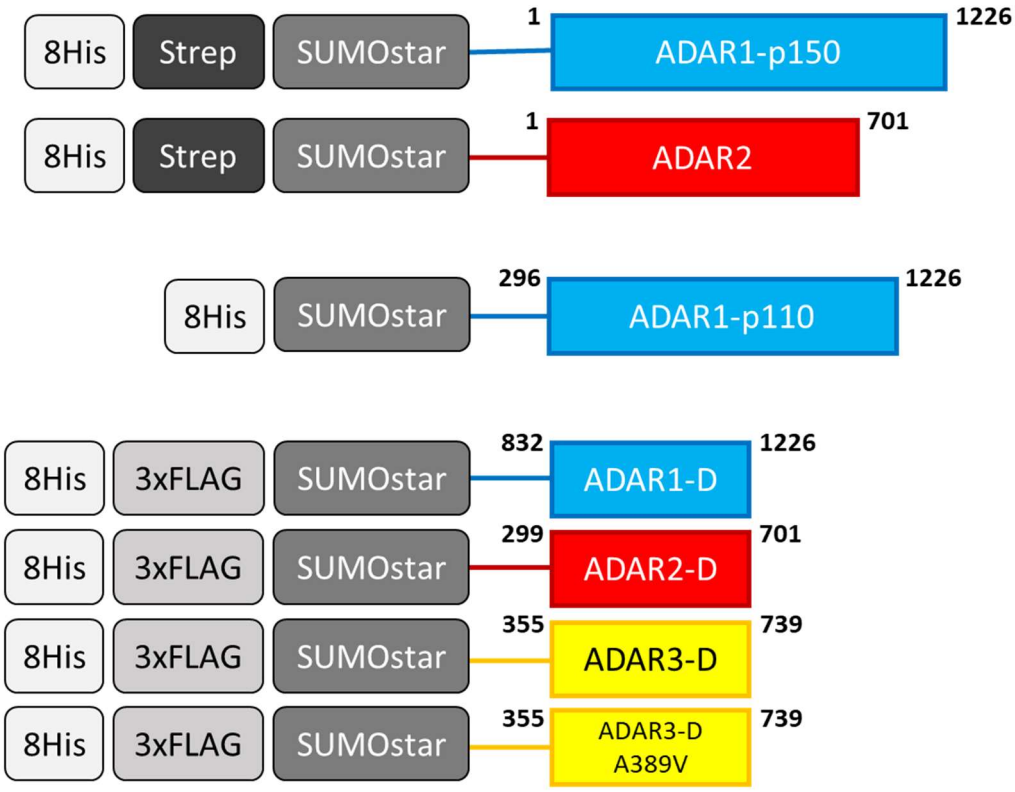

Figure 2.1. Schematic of ADAR constructs. Full-length ADAR1-p150 and ADAR2 were cloned with an N-terminal 8His-Strep-SUMOstar tag. The shorter isoform ADAR1-p110 was cloned with an N-terminal 8His-SUMOstar tag, and the four deaminase-only constructs – ADAR1-deaminase (ADAR1-D), ADAR2-deaminase (ADAR2-D), ADAR3-deaminase (ADAR3-D) and ADAR3-deaminase point mutant A389V (ADAR3-D A389V) were cloned with an N-terminal 8His-3xFLAG-SUMOstar tag.

Determination of deaminase construct boundaries

The ADAR2-D construct was generated with the same domain boundaries, aa. 299-701, as that used by Macbeth et al. (2005) to generate the crystal structure of the domain. Domain boundaries of the ADAR1-D construct, aa. 832-1226, were determined by limited proteolysis. Following generation of a longer construct of residues 809-1226 and treatment with trypsin, the new N-terminal was identified using MALDI-TOF Mass Spectrometry, in collaboration with Dominic Olinares of the Chait Lab. The N-terminal of the ADAR3-D construct was determined by sequence alignment to the ADAR1 sequence, with residue T355 aligning to the ADAR1 T832, and the final construct having aa. 355-739.

Generation of baculovirus for expression in Sf9

Once constructs were cloned into pFASTBAC plasmids, the Bac-to-Bac Baculovirus Expression System was used to transpose each expression cassette from the plasmid onto a bacmid (Luckow et al. 1993). Bacmid DNA was then transfected into adherent Sf9 cultures (Thermo Fisher) using Insect GeneJuice (Novagen), and the p1 baculovirus was collected from the supernatant at 6 days post-transfection. Each baculovirus was then amplified by infecting $2x10^8$ Sf9 cells in suspension culture at an MOI of 0.01, collecting the p2 baculovirus from the supernatant at 4 days post-infection.

Expressing ADAR constructs from Sf9

Sf9 were grown in suspension culture to reach a density of $5x10^6$ cells/ml in 600ml, and then infected by addition of 12ml p2 baculovirus harboring an ADAR construct, at an estimated MOI of ~2. Following incubation for 48 hours at 27°C, with shaking at 120rpm, cell pellets were harvested using a refrigerated centrifuge, spinning at 600g for 15 min. Once supernatant was

discarded, cell pellets were resuspended in 15ml phosphate storage buffer (20% sucrose, 1.5x Phosphate-Buffered Saline (PBS) pH 7.2) and snap-frozen in LN_2 before storage at -80°C.

Purification of ADAR constructs from Sf9

Initial purification conditions were adapted from Macbeth and Bass (2007), with alterations to improve protein solubility and stability. PBS was added to the cell pellet storage buffer, and glycerol and Na/K PO_4 pH 8.2 were added to purification buffers to increase the fraction of soluble protein. When ADAR constructs were purified in the presence of reducing agent 2-mercaptoethanol (β-ME), loss of editing activity was observed over a period of 3-4 days. In collaboration with Dominic Olinares (Chait Lab), Native Electrospray Ionization mass spectrometry identified that β-ME was forming covalent adducts on the protein, most likely with the unpaired cysteines in the active site. Due to this, alternate reducing agents were used during purification, with tris (2-carboxyethyl) phosphine (TCEP) added for initial steps up to incubation with Ni-NTA resin, and dithiothreitol (DTT) added for all post Ni-NTA steps. The stability issue was not completely solved, with protein activity extended from ~4 days to 1 month. Due to this, all protein assays were performed within 4-5 days of purification.

Cell lysis and clarification

All ADAR constructs underwent the same initial lysis protocol through to incubation with Ni-NTA resin, at which point the conditions for deaminase constructs and full-length constructs split due to different Ni-NTA wash conditions. All steps were carried out on ice, or at 4°C. For each construct, a cell pellet generated from 600ml Sf9 was thawed and diluted into 80ml lysis buffer: 500mM NaCl, 10mM Tris-HCl pH8.5, 20mM Na/K PO_4 pH8.2, 5% glycerol, 0.5% IGEPAL (Nonidet P-40), 1mM PMSF, 2µg/ml DNase I, 2µg/ml RNase A, 15mM imidazole, 1mM TCEP.

Cell disruption and shearing of DNA was achieved through sonication, with 6x15sec pulses separated by 5min on ice. The insoluble material was then separated out by centrifugation at 19000rpm (29600g) for 60min in a refrigerated centrifuge. The supernatant was mixed with 3ml washed Ni-NTA slurry (50% resin), and incubated on a nutator for 1hr at 4°C. The supernatant-resin mixture was then loaded onto a column, allowing any unbound material to flow through by gravity, and then washed with 25ml high-salt wash buffer: 500mM NaCl, 10mM Tris-HCl pH8.5, 20mM Na/K PO_4 pH8.2, 10% glycerol, 20mM imidazole, 1mM TCEP. At this point, the conditions for purification of deaminase constructs and full-length constructs separated.

Purification of ADAR deaminase constructs

For deaminase constructs ADAR1-D, ADAR2-D, ADAR3-D, and ADAR3-D A389V, following the high-salt wash of the Ni-NTA column an additional wash was performed with 25ml low-salt wash buffer: after washing the Ni-NTA with high-salt buffer, an additional wash with 25ml low-salt buffer was done: 100mM NaCl, 10mM Tris-HCl pH8.5, 20mM Na/K PO_4 pH8.2, 10% glycerol, 20mM imidazole, 1mM TCEP. Bound protein was then eluted from the Ni-NTA resin in 10ml low-salt, high-imidazole elution buffer: 100mM NaCl, 10mM Tris-HCl pH8.5, 20mM Na/K PO_4 pH8.2, 10% glycerol, 250mM imidazole, 1mM TCEP. The 10ml elution was then loaded directly onto a HiTrap Q Fast protein liquid chromatography (FPLC) column, eluting over a NaCl gradient of 0.1 – 1.5M (0 – 50%) in 8 column volumes (CV), collecting 2ml fractions at a rate of 0.4 ml/min. The column buffers used were composed of NaCl, 10mM Tris-HCl pH8.5, 20mM Na/K PO_4 pH8.2, 10% glycerol, 2mM DTT. Each deaminase construct eluted in fractions 13-16, which can be seen in the traces on the left-hand side of **figure 2.2** (for ADAR1-D and ADAR2-D) and **figure 2.3** (for ADAR3-D and ADAR3-D A389V). Fractions 13-16 were pooled, and 100µg Ulp1-sgs protease was added to cleave the tag, incubating >16hr at 4°C.

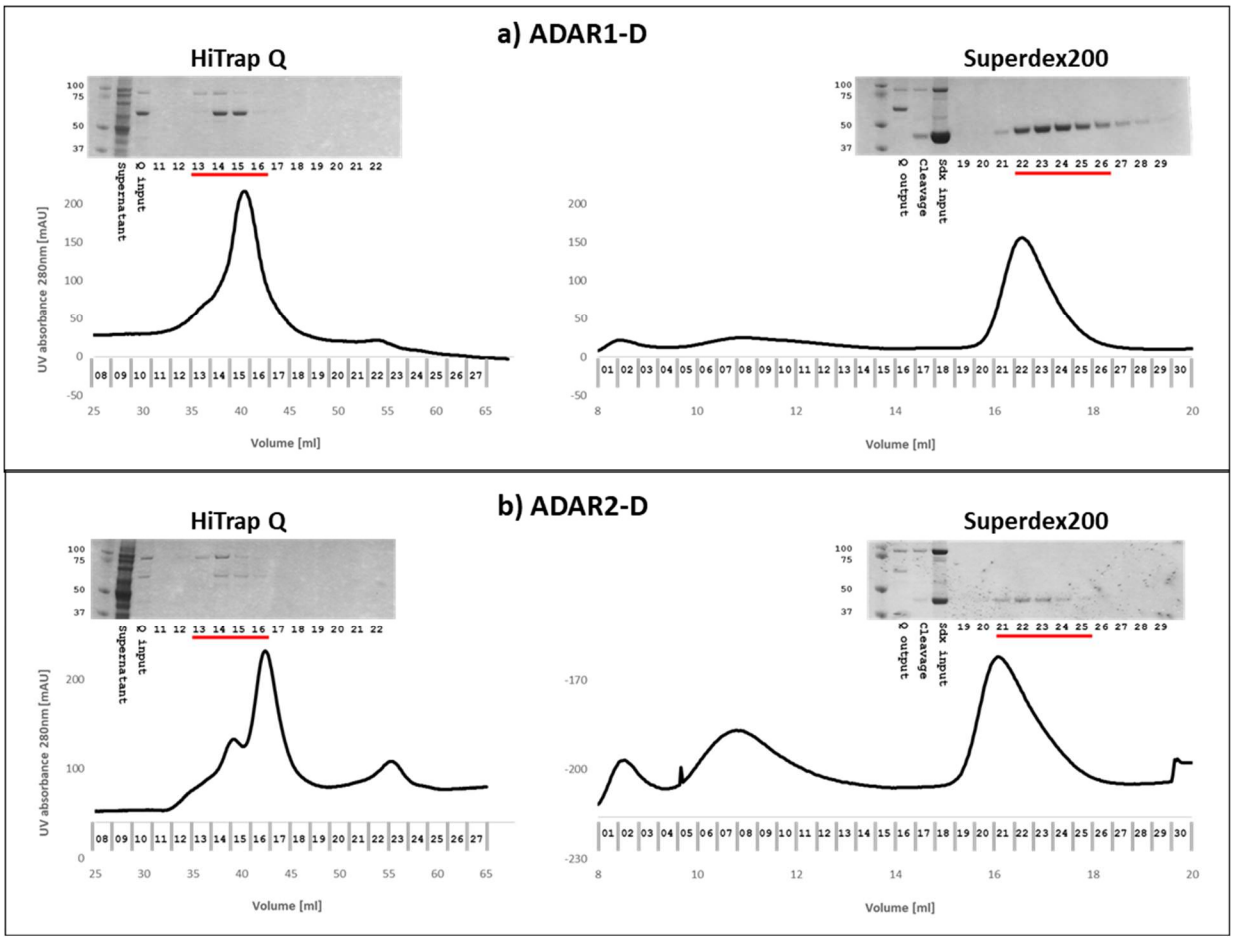

Figure 2.2. Purification details for ADAR1-D (a) and ADAR2-D (b). Each protein was isolated from Sf9 cell lysates by Ni-NTA binding of the His-tag, with sample eluted from the Ni-NTA column shown at the Q input lane in the HiTrap Q protein gels, visualized using Coomassie staining. The UV absorbance trace [measuring mAU at 280nM] is shown for the HiTrap Q columns at left and the Superdex200 10/300 columns at right. Tagged samples, with size ~65kDa, eluted from the HiTrap Q column in fractions 13-16. In the Superdex 200 gels shown at right, the tag was cleaved – visible as a drop in size from the Q output lane to cleavage lane at ~45kDa – with sample eluting from the Superdex column in fractions 21/22-26.

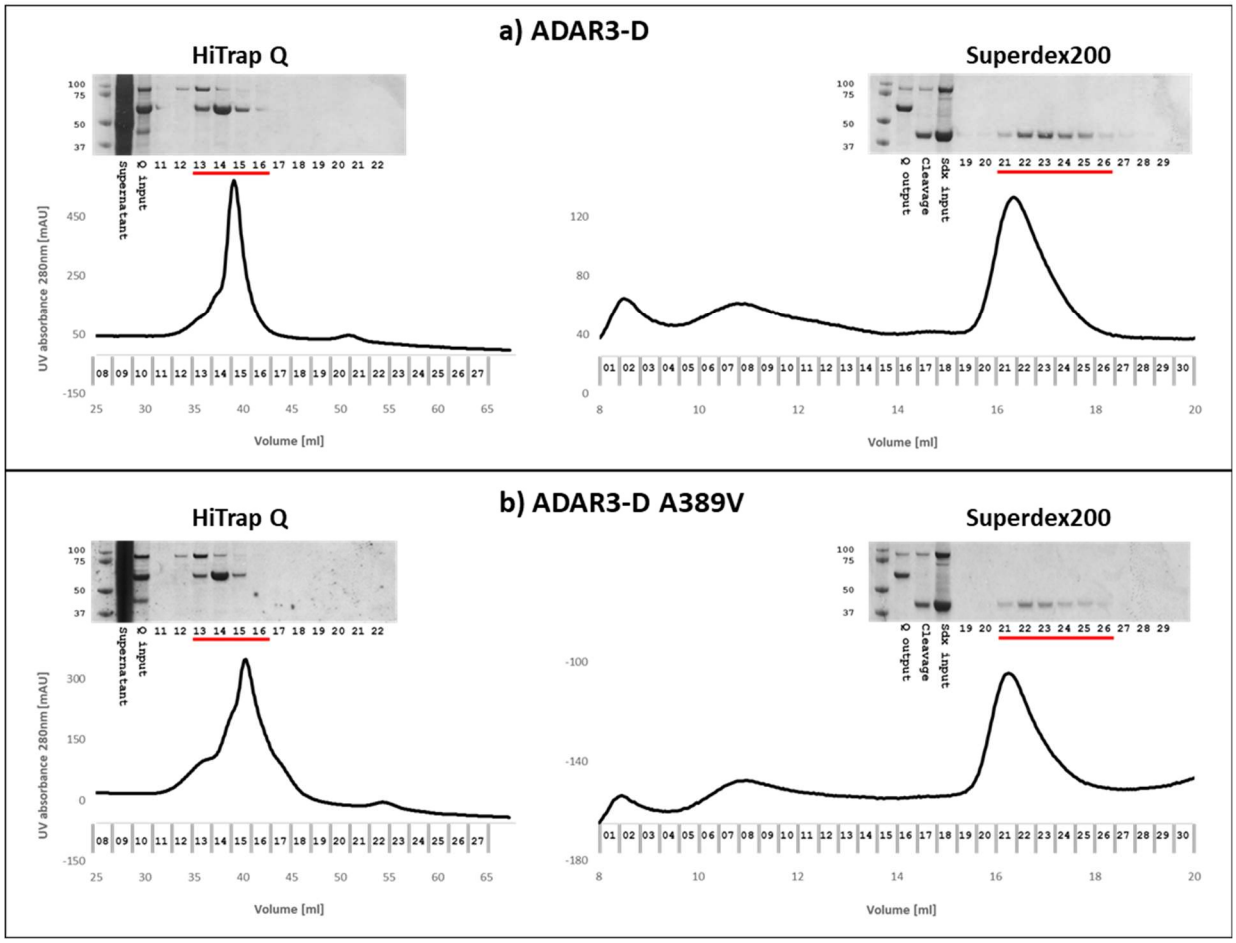

Figure 2.3. Purification details for ADAR3-D (a) and ADAR3-D A389V (b). Each protein was isolated from Sf9 cell lysates by Ni-NTA binding of the His-tag, with sample eluted from the Ni-NTA column shown at the Q input lane in the HiTrap Q protein gels, visualized using Coomassie staining. The UV absorbance trace [measuring mAU at 280nM] is shown for the HiTrap Q columns at left and the Superdex200 10/300 columns at right. Tagged samples, with size ~65kDa, eluted from the HiTrap Q column in fractions 13-16. In the Superdex 200 gels shown at right, the tag was cleaved – visible as a drop in size from the Q output lane to cleavage lane at ~45kDa – with sample eluting from the Superdex column in fractions 21-26.

After tag cleavage, 5mM imidazole was added to the sample before flowing it through 1ml washed Ni-NTA resin. The flow-through was collected and the cleaved tag should have remained bound to the resin. The sample was then concentrated to <1ml using a 30kDa MWCO spin filter (Amicon) and loaded onto a Superdex 200 10/300 FPLC column, in buffer 300mM NaCl, 10mM Tris-HCl pH8.5, 20mM Na/K PO_4 pH8.2, 10% glycerol, 2mM DTT. The sample was eluted over 1 CV (24ml), collecting 0.4ml fractions at a rate of 0.22 ml/min. All four deaminase constructs eluted with similar profiles, centered around fraction 23, as shown in the traces on the right-hand side of **figure 2.2** (for ADAR1-D and ADAR2-D) and **figure 2.3** (for ADAR3-D and ADAR3-D A389V. Fractions 22-26 were pooled for ADAR1-D, 21-25 for ADAR2-D, 21-26 for ADAR3-D, and 21-26 for ADAR3-D A389V. Each protein was concentrated to ~30µl using a 30kDa MWCO spin filter, in buffer 300mM NaCl, 10mM Tris-HCl pH8.5, 20mM Na/K PO_4 pH8.2, 5% glycerol, and 5mM DTT. Final protein concentration was measured by Qubit, with values shown in **table 2.1**. Proteins were then snap-frozen in LN_2 and stored at -80°C.

Table 2.1. Purification details of ADAR1, ADAR2 and ADAR3 deaminase constructs, with expected protein sizes, fractions collected from HiTrap Q ("Q") and Superdex 200 10/300 ("Sdx") FPLC columns, and final sample concentration.

Construct	Size [kDa]	Q fractions	Sdx fractions	Concentration [mg/ml]	Concentration [µM]
ADAR1-D	45	13 – 16	22 – 26	2.50	55.47
ADAR2-D	45	13 – 16	21 – 25	1.28	28.37
ADAR3-D	43	13 – 16	21 – 26	2.38	55.42
ADAR3-D A389V	43	13 – 16	21 – 26	1.45	33.75

Purification of full-length ADAR constructs

For full-length constructs ADAR1-p150, ADAR1-p110, and ADAR2, following the high-salt wash of the Ni-NTA column a more stringent wash was performed with 25ml high-salt, mid-imidazole wash buffer: 500mM NaCl, 10mM Tris-HCl pH8.5, 20mM Na/K PO_4 pH8.2, 10% glycerol, 100mM imidazole, 1mM TCEP. Bound protein was then eluted from the Ni-NTA resin in 10ml high-salt, high-imidazole elution buffer: 500mM NaCl, 10mM Tris-HCl pH8.5, 20mM Na/K PO_4 pH8.2, 10% glycerol, 250mM imidazole, 1mM TCEP. 100µg Ulp1-sgs protease was added to the eluted sample to cleave the tag, incubating >16hr at 4°C.

After tag cleavage, the sample was run through 1ml washed Ni-NTA resin. The flow-through was collected and the cleaved tag should have remained bound to the resin. The sample was then concentrated to <1ml using a 30kDa MWCO spin filter and loaded onto a Superdex 200 10/300 FPLC column, in buffer 300mM NaCl, 10mM Tris-HCl pH8.5, 20mM Na/K PO_4 pH8.2, 10% glycerol, 2mM DTT. The sample was eluted over 1 CV (24ml), collecting 0.4ml fractions at a rate of 0.22 ml/min. The three full-length ADAR proteins eluted from the size exclusion column with different profiles, which can be seen in the traces in **figure 2.4.** For ADAR1-p150 fractions 7-10 were pooled, for ADAR1-p110 fractions 10-15 were pooled, and for ADAR2 fractions 13-18 were pooled. Each protein was concentrated to ~30µl using a 30kDa MWCO spin filter, in buffer 300mM NaCl, 10mM Tris-HCl pH8.5, 20mM Na/K PO_4 pH8.2, 5% glycerol, and 5mM DTT. Final protein concentration was measured by Qubit, with values shown in **table 2.2**. Proteins were then snap-frozen in LN_2 and stored at -80°C.

pg. 42

Figure 2.4. Purification details for ADAR1-p150 (a), ADAR1-p110 (b), and ADAR2 (c). Each protein was isolated from Sf9 cell lysates by Ni-NTA binding of the His-tag, with sample eluted from the Ni-NTA column shown in the Elution lane on the protein gel, visualized by Coomassie staining. The tag is cleaved, visible as a drop in size for each protein, to ~150kDa for ADAR1-p150, ~110kDa for ADAR1-p110, and ~75kDa for ADAR2. The cleaved sample was loaded onto a Superdex 200 10/300 size exclusion column, with UV absorbance trace [measuring mAU at 280nM] shown below the protein gels. ADAR1-p150 eluted from the column in fractions 7-10, ADAR1-p110 eluted in fractions 10-15, and ADAR2 eluted in fractions 13-18. The ADAR1-p150 (a) band was very faint, and the inset at the top-right corner zooms in on fractions 7-11, with the black arrow pointing at the ~150kDa band.

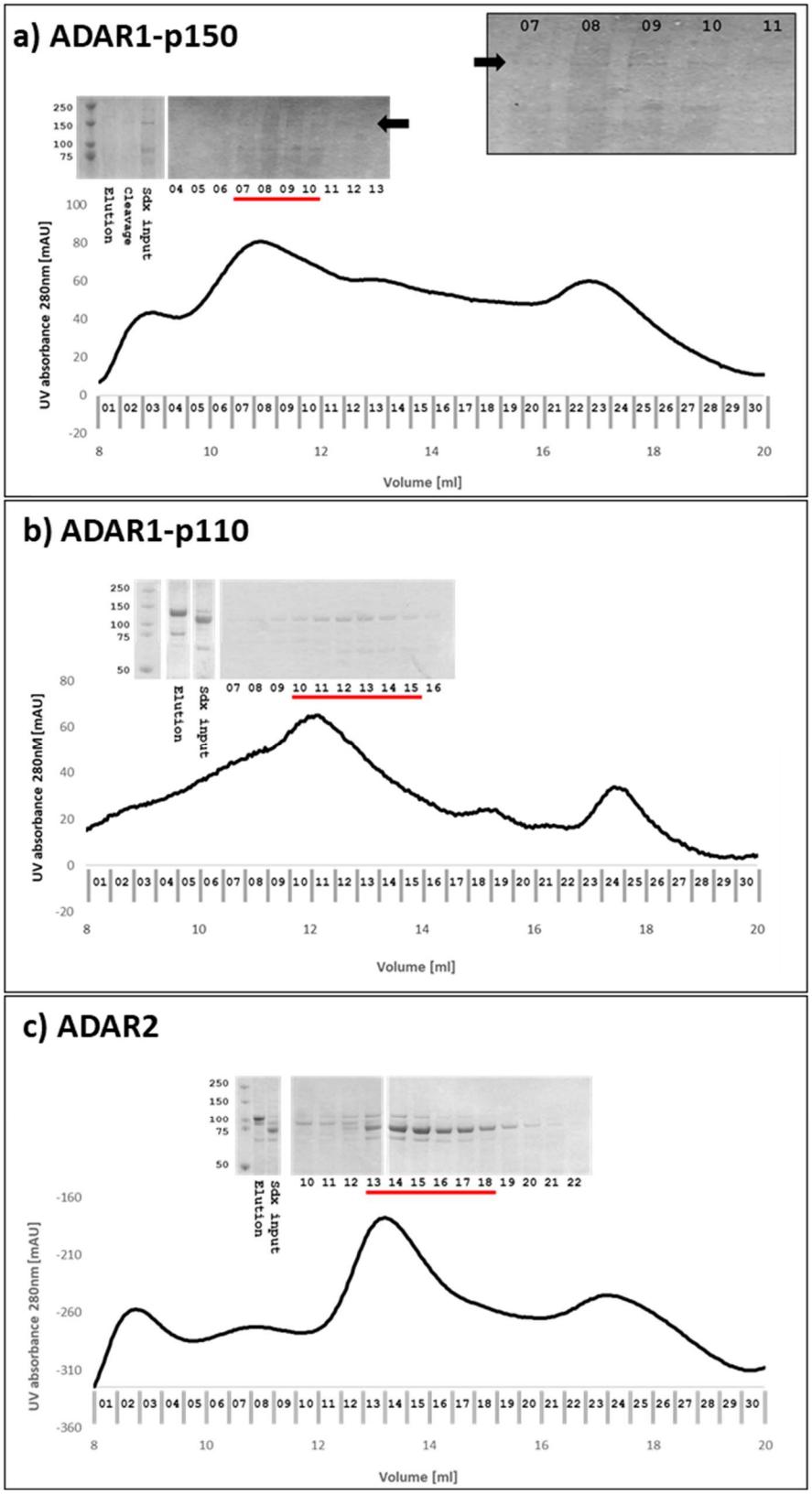
a) ADAR1-p150
07 08 09 10 11
250
150
100
75
Elution
Cleavage
Sdx input
04 05 06 07 08 09 10 11 12 13
UV absorbance 280nm [mAU]
Volume [ml]
b) ADAR1-p110
250
150
100
75
50
Elution
Sdx input
07 08 09 10 11 12 13 14 15 16
UV absorbance 280nM [mAU]
Volume [ml]
c) ADAR2
250
150
100
75
50
Elution
Sdx input
10 11 12 13 14 15 16 17 18 19 20 21 22
UV absorbance 280nm [mAU]
Volume [ml]

Table 2.2. Purification details of full-length ADAR1-p150, ADAR1-p110, ADAR2 proteins, with expected protein sizes, fractions collected from Superdex 200 10/300 ("Sdx") FPLC columns, and final sample concentration.

Construct	Size [kDa]	Sdx fractions	Concentration [mg/ml]	Concentration [µM]
ADAR1-p150	133	7 – 10	0.63	4.77
ADAR1-p110	103	10 – 15	1.03	10.04
ADAR2	77	13 – 18	0.71	9.24

Protein yield, purity and stability

For each ADAR construct, the total yield of protein from a 600ml Sf9 culture was calculated in [µg], from the measured concentration and sample volume, with values shown in **table 2.3**. Protein yield was low, but as the *in vitro* editing assay only required 2pmol of protein for each condition, this was enough protein for analysis purposes.

Table 2.3. Yield of each ADAR construct from a 600ml Sf9 culture, and measured purity of each sample after purification.

Construct	Concentration [mg/ml]	Sample volume [µl]	Yield [µg / 600ml Sf9]	Sample purity [%]
ADAR1-p150	0.63	25	15.75	50
ADAR1-p110	1.03	30	30.9	89
ADAR2	0.71	40	28.4	76
ADAR1-D	2.50	55	137.5	>98
ADAR2-D	1.28	45	57.6	93
ADAR3-D	2.38	40	95.2	93
ADAR3-D A389V	1.45	40	58	>98

Also shown in **table 2.3** is the estimated purity of each construct. Purity was determined by measuring the relative intensity of protein bands on the Coomassie gels shown for each construct in figure **2.2**, **2.3**, and **2.4**, using ImageJ2 (Rueden et al. 2017). The deaminase constructs were highly pure, with ADAR1-deaminase and ADAR3-deaminase A389V having no other visible bands on the gel, assuming purity of >98%. The ADAR2-deaminase and ADAR3-deaminase did have a second visible band, at ~35kDa, and were found to have purity in the range of 93%.

The full-length proteins were not generated to high purity, with the ADAR1-p110 and ADAR2 estimated to be 89% and 76% pure respectively, while the ADAR1-p150, with the lowest level of expression, was only purified to 50%.

When measuring editing activity of ADAR constructs, it was observed that the overall level of editing decreased over time, indicating the proteins were possibly losing stability. Attempts to improve protein stability by exchanging reducing agents improved longevity from 4-5 days (with β-ME) to 1 month (with DTT). After 1 month, proteins generated with DTT began to lose activity as well, with 3-month-old proteins having ~50% of the editing activity of new proteins. As the stability issue was not solved, all editing assays were carried out within 4-5 days of purification.

Following purification of full-length and deaminase-only ADAR constructs, the editing activity needed to be measured for each, and the specificity for different RNA sequences determined. To this end, a method using high throughput sequencing to characterize *in vitro* editing activity was developed, as detailed in Chapter III.

Chapter III – Development of an *in vitro* RNA editing assay

To characterize the editing specificity of each ADAR isoform and deaminase domain, a high throughput sequencing based method was developed to be able to measure the percentage of editing for each A site in a perfectly double-stranded RNA substrate, that contains one adenosine for each combination of 5' and 3' neighbor.

The method, adapted from Koeris et al. (2005) involves incubating ADAR protein and dsRNA *in vitro*, with no other factors involved. After this, the RNA was isolated and RNAseq libraries generated with Illumina P5/P7 adapters, so that thousands of sequences for each condition could be sequenced. Sample conditions were chosen so that a majority of RNA molecules only had a single edit, so to remove complications due to multiple edits occurring on a single substrate. Counting the number of editing events for each of the 16 adenosines, the editing specificity of each ADAR construct could then be characterized.

RNA substrate design

There are sixteen different combinations of an RNA triplet with adenosine at the second position. With four possible 5' neighbors and four possible 3' neighbors, those triplets are N**A**N: (A/U/C/G)**A**(A/U/C/G). ADAR proteins have been shown to have selective 5' and 3' neighbor preferences, and so to characterize the editing efficiency against each of the different triplet combinations *in vitro*, a synthetic dsRNA was designed with all possible sites represented.

The 50bp dsRNA was based on the substrate used by Koeris et al. (2005), with bases 1 – 10 and 41 – 50 remaining constant, to be used for reverse transcription (RT) and PCR amplification, as well as sequence alignment. The central region of bases 11-40 was altered from the previously published substrate, with the sequence replaced by one with 16 adenosines, each with one of the

combinations of 5' and 3' neighbor. In triplets where the 5' or 3' neighbor is an A, the triplets are overlapping, such as CAAAC that contains triplets CAA, AAA, and AAC. **Figure 3.1** shows the substrate used, with the adenosines highlighted in red.

Koeris et al. (2005) found that a $(CG)_6$ repeat at the 5' end facilitated increased editing by ADAR1-p150, compared to a dsRNA sequence with $(CCGG)_3$, implying that the $(CG)_6$ may be forming a left-handed RNA structure that acts as a binding site for the ADAR1-p150 Zα domain. The 5' constant end of the newly designed substrate has a shorter $(CG)_4$ repeat that can also form a left-handed structure (Schwartz et al. 1999). Koeris et al. based their RNA substrate off of a previously published dsRNA substrate used by Polson and Bass (1994), which was designed so that the majority of purines were along one strand to decrease formation of intramolecular structures during synthesis, such as cross-strand stacking. The only component of this original substrate that remains is the 10bp at the 3' end, and the substrate no longer has one purine-rich strand and one pyrimidine-rich strand. Indeed, the new substrate has many sites that can form cross-strand stacking interactions between the two RNA strands, especially in the segment of the sequence with purine-pyrimidine repeats, AUAUACAC. A 50bp dsRNA can potentially facilitate interactions with both the deaminase and a dsRBD, with footprints of ~20bp and ~16bp respectively. It is unknown if more than one dsRBD could be bound to a substrate of this length.

5′ CGCGCGCGGG**A**C**A**G**AA**U**C**AAA**C**A**U**AA**G**A**U**A**U**A**C**A**C**U**A**G**AGGACAGGGACC 3′
3′ GCGCGCGCCC**UGUCUUAGUUUGUAUUCUAUAUGUGAUCU**CCUGUCCCUGG 5′

Figure 3.1. 50bp dsRNA substrate for characterization of ADAR editing. The central 30bp, in bold, contains adenosines, each with a different combination of 5' and 3' neighbor. Only the strand containing the highlighted adenosines – designated the positive-sense strand – was reverse transcribed and sequenced.

Preparation of RNA and protein samples

The dsRNA substrate and all primers for reverse transcription and PCR were purchased from IDT. The 50bp synthetic dsRNA substrate purchased as a duplex, having been annealed at 94°C for 2min before cooling and isolating duplex RNA by HPLC chromatography. The dsRNA was resuspended in nuclease-free H_2O, then diluted to 0.1µM in 50mM NaCl, 10mM Tris-HCl pH7.5. Each ADAR protein was diluted to 0.1µM in 100mM NaCl, 10mM Tris-HCl pH8.5, 15% glycerol, 0.2mM EDTA, 5mM DTT. 2µl of dsRNA was mixed with 2µl protein to make final 1:1 ratio at 0.05µM. Samples were incubated for 1hr at 37°C in a thermocycler with the lid set to 100°C. After incubation, the RNA was isolated from the protein by phenol: chloroform extraction and ethanol precipitation, with each sample then resuspended in 10µl nuclease-free H_2O. In parallel to the samples generated, the ADAR1-deaminase was also incubated with ssRNA and sequenced to identify if any editing was apparent with this substrate, and no significant editing was measured.

Sample Replicates

For each purified ADAR construct, three samples were generated, one on each of three different days. On each day, protein and RNA were incubated, RNA was isolated, and libraries were prepared, with each protein condition represented by n=1 on each day. Each pool of samples was also sequenced on a different run, with the results shown in chapter IV compiled from three sequencing runs, each containing n=1 for each sample, to make a final n=3 for each condition.

Generation of sequencing libraries

The positive-sense strand of each RNA sample was reverse-transcribed to generate cDNA using the SuperScript III (Invitrogen) protocol, using custom primers [2pmol] that attached a partial Illumina p5 sequence, 5 degenerate bases (5N) to identify individual RT events, and a 4bp barcode

(designated xxxx) to identify each sample: GAGATCTACACTCTTTCCCTACACGACGCTCTT CCGATCTNNNNNxxxxGGTCCCTGTCC. After reverse transcription, PCR was used to attach the full Illumina p7 and p5 adapters to the 5' and 3' ends of the sample, using forward primer p7 (CAAGCAGAAGACGGCATACGAGATCGCGCGCG) and reverse primer p5 (AATGATACGG CGACCACCGAGATCTACACTCTTTCCCTACACGACG). To generate libraries, 10ul of cDNA was mixed with 1x ThermoPol Reaction Buffer (NEB), 10mM dNTPs (Invitrogen), 10µM each primer, and 2U Deep Vent DNA polymerase (NEB). A two-step PCR was performed, with initial denaturation (95°C for 1min) followed by 30 cycles of amplification (95°C for 20s, 72°C for 20s) and a final elongation step at 72°C for 5min. Each sample was run on 1.5% Agarose for 40min at 110V, then gel purified to 30µl using QG buffer (Qiagen) and MinElute spin columns (Qiagen). Sample quality was measured by TapeStation, and concentration was measured by TapeStation or Qubit Fluorometric Quantification.

Preparation of libraries for sequencing

Libraries were pooled at equimolar amounts to generate a sample with final concentration of 10nM. The pool was then diluted to 2nM and denatured using 0.1N NaOH. The denatured sample was further diluted to 5pM, with PhiX control (Illumina) added to final concentration 5%. 600µl of this sample was then loaded onto a MiSeq cartridge, using MiSeq Reagent Kit v2 (Illumina). MiSeq runs generated FASTQ files of 75bp single-end reads sequenced in the p5 to p7 direction.

Sequencing analysis pipeline

All analyses were completed using the public Galaxy web platform at usegalaxy.org (Afgan et al. 2018), with a workflow schematic shown in **figure 3.2**.

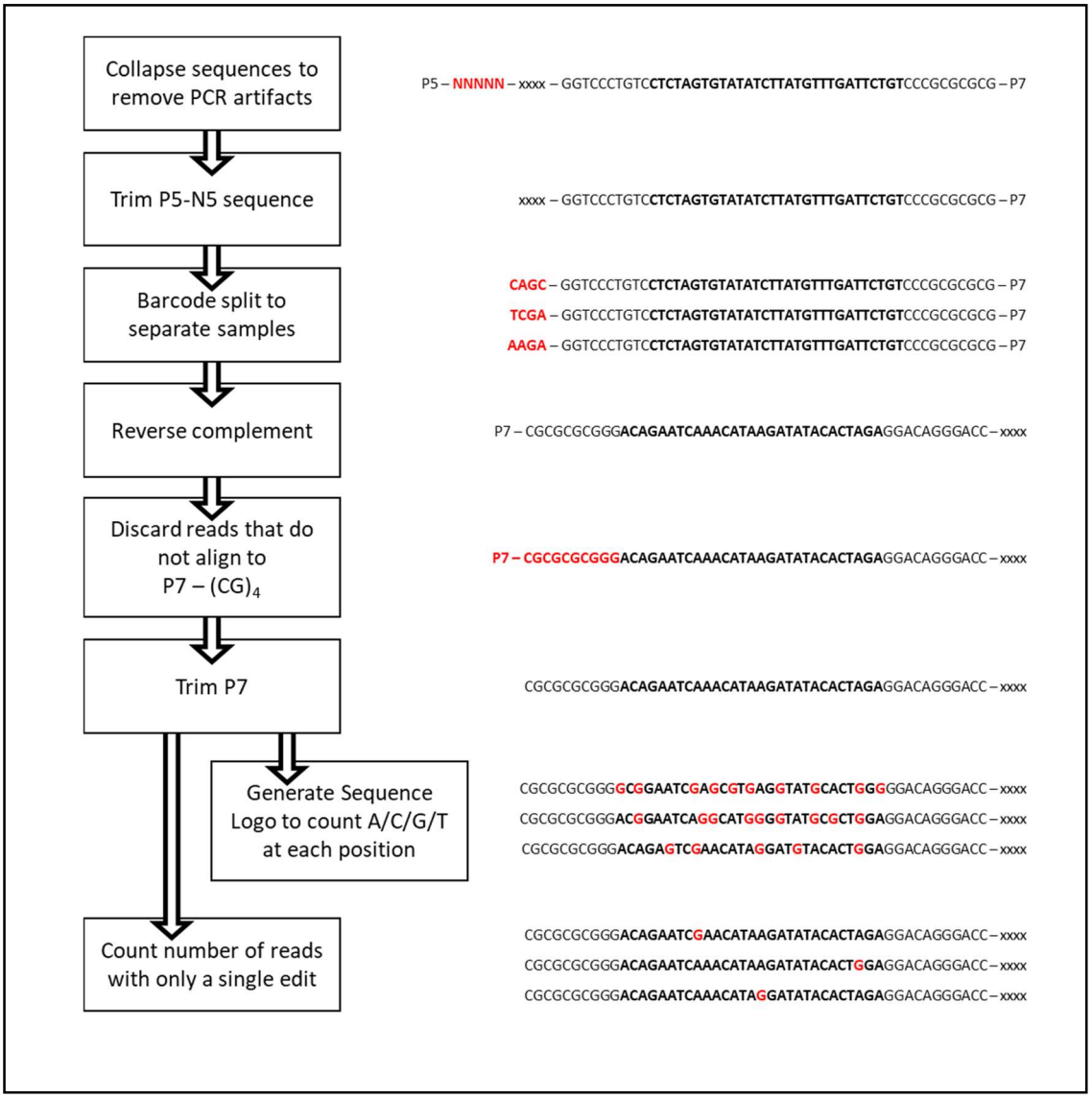

Figure 3.2. Sequencing analysis pipeline to align sequences and count number of A-to-G editing events. Analysis outputs both the total number of editing events at each position, and the number of reads with only a single edit. An example sequence is shown at right, with red highlighting the portion of the sequence being manipulated at each stage of the pipeline.

Raw FASTQ files were collapsed to remove identical sequences, leaving only one instance of each RT event, before trimming the 5N degenerate bases. A Barcode Splitter parsed the sequences into individual files for each barcode. Each sample was reverse complemented, aligned and trimmed to the CGCGCGCGGG sequence at the 5' end. At this point, each sample file had a collection of sequences, all of length 54bp (the 50bp original sequence plus a 4bp barcode now at the 5' end) with the first 10bp invariant. Two different workflows were then used to analyze this data.

Measuring total editing

First, the total amount of editing in the substrate was analysed. Using all sequences, a Sequence Logo plot (text format) was generated, with the number of times each nucleotide (A/C/G/T) was counted at each position. An example of a Sequence Logo plot is shown in **table 3.1**. The total editing at each adenosine position could then be calculated by counting the number of Gs at that position and dividing by the total number of reads. The overall editing percentage was calculated as the number of Gs counted at all 16 adenosine positions divided by the total number of A sites, that is: 16 times the number of reads. Values stated in figures are the average of n=3 replicates, with standard deviation. Significance between samples was calculated by the unpaired Student's t-test with two-tailed distribution, assuming unequal variance.

To compare the relative level of editing at each adenosine, the relative frequency of editing at each position was calculated. This was performed by counting the total number of observed editing events (total number of Gs across all A positions), and then dividing the number of Gs at each site by this value. Frequency was calculated for each replicate before combining to calculate the average, standard deviation, and significance. From the calculated relative frequencies, observations of which adenosine sites were more frequently edited could be made.

Table 3.1. Example of Sequence Logo output, for ADAR1-D (sample 20191012-11). Table shows positions 18-31 of the substrate, which corresponds to the bolded section of: CGCGCGCGGGAC AGAAU**C**AAA**C**A**U**AA**G**A**U**A**U**ACACUAGAGGACAGGGACC. Adenosine positions are red, both in the sequence and the table. It can be observed that there is a high number of Gs counted at most of the A sites.

#	A	C	G	T
18	88	25112	16	129
19	24130	41	1134	40
20	23326	19	1994	6
21	20141	36	5153	15
22	83	25138	26	98
23	24992	41	276	36
24	59	26	19	25241
25	19302	28	5982	33
26	17270	47	8004	24
27	609	44	24617	75
28	25167	24	131	23
29	65	53	166	25061
30	22666	25	2600	54
31	59	44	30	25212

Isolating sequences with single-editing events

After incubation with an ADAR protein, the pool of RNA is likely to contain a mix of substrates with only a single edit and substrates with more than one site edited. To analyze the frequency of editing at each adenosine without the added complication of having multiple editing events per substrate, the second analysis workflow was developed to only isolate and count substrates with a single edited adenosine. A Barcode Splitter was used to separate out sequences with a single A-to-G change, using barcodes for each of the 16 adenosine sites. The output file is a text file with the 16 different adenosine positions, and the number of sequences counted for each one. The sample incubation conditions of 0.05μM RNA and 0.05μM protein were chosen, as decreasing the protein concentration increased single-editing events as a fraction of all substrate. The concentration was not lowered below this value as the total number of editing events continued to

decrease. At these low protein and RNA concentrations, the editing reaction was sub-saturating when incubated for 1 hour at 37°C. Except for initial calculations of total editing frequency for each sample, the single-editing counts were used for all analysis.

A comparison of ADAR1-p150 samples generated from two different protein preparations (20170808 and 20181009) was used as a quick test of the reproducibility of the assay. **Figure 3.3** shows the total editing measured for ADAR1-p150 from those two different preparations, as well as comparing two different samples both generated at different times after purification from the 20181009 preparation. In **figure 3.3a** it can be seen that the 20170808 sample has the same level of editing activity as one of the 20181009 samples (no. 46) while the other 20181009 sample (no. 17) has a higher level of overall activity, due to being tested shortly after purification while sample 46 was generated ~1 month later, evidence of the loss of activity over time. **Figure 3.3b** shows that for both protein preparations of ADAR1-p150 the relative frequency of each triplet remains relatively constant, independent of total editing levels.

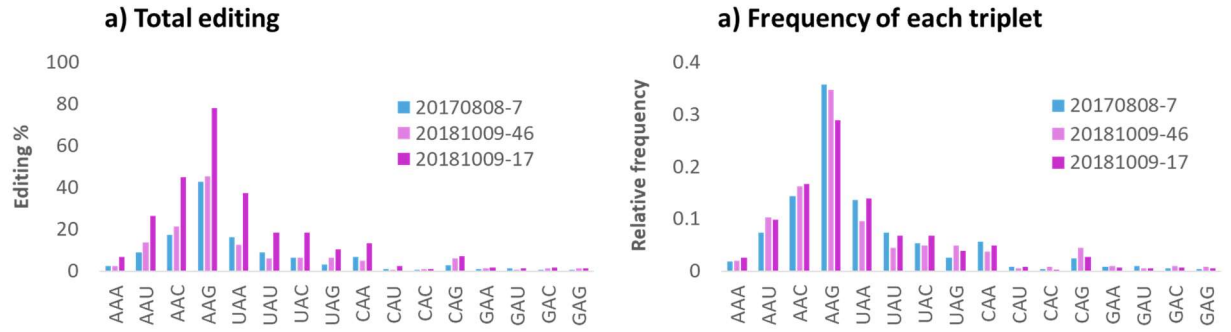

Figure 3.3. Total editing percentage at each triplet in the 50bp dsRNA substrate (a) and relative frequency of each triplet (b) for ADAR1-p150 samples generated from 20170808 protein stock (20170808-7) and 20181009 stock (20181009-17 and -46). For each ADAR construct tested, the sum of the frequencies across all 16 triplets will add to one.

Additional methods

Homology models of ADAR1 and ADAR3 deaminases

For comparisons between the ADAR1, ADAR2 and ADAR3 deaminases, homology models of the ADAR1 and ADAR3 deaminases were generated to compare to the ADAR2 crystal structure, for additional insight into differences found in deaminase activities. Models were generated using SWISS-MODEL (Waterhouse et al. 2018) in Alignment Mode, inputting the ADAR1 and ADAR3 deaminase sequences (shown in **figure 1.5**) and aligning to the ADAR2 crystal structure PDB 5ED2. Each deaminase model was then superimposed over the ADAR2-D crystal structure and aligned to active site residues V351, T375, K376, H394, E396, C451, and C516, so to align the active site of each to the RNA substrate also included in the crystal structure. All other residues were then observed for structural differences.

Identification of editing sites by ADAR proteins in complex substrates

After characterizing *in vitro* editing specificity of ADAR constructs against a simple 50bp dsRNA substrate, editing specificity of those same constructs was tested against more complex substrates. The substrates chosen were human HEK 293T total RNA and the reovirus T1L RNA genome (reoT1L).

Total RNA, containing single- and double-stranded RNA as well as RNA with more complex structures, was isolated from double knockout HEK 293T ADAR1/2$^{-/-}$cells, which should lack endogenous ADAR editing. This ADAR1/2$^{-/-}$ cell line was used by Chung et al. (2018) to characterize the 293T editome; cells were provided by Hachung Chung (formerly Rice Lab, currently Columbia University).

reoT1L packages its genome in 10 perfectly double-stranded RNA segments, which, while not a natural ADAR target, could potentially be a good substrate for non-specific editing (Hood et al. 2014). Isolated reoT1L virions were provided by Danica Sutherland and Pavithra Aravamudhan (Dermody Lab, University of Pittsburgh).

Isolating HEK 293T total RNA and reovirus T1L dsRNA

For both HEK 293T and reovirus T1L, RNA was isolated by phenol: chloroform extraction, and the aqueous phase was then mixed 1:1 with 70% ethanol and loaded onto a PureLink RNA Mini Kit spin column (Invitrogen). RNA was resuspended in 30µl nuclease-free H_2O. From a 293T cell pellet of $3x10^6$ cells, the typical yield was ~30µg (1µg/ul, 30nM per 50bp) and for 15µl reovirusT1L supernatant, the typical yield was ~300ng (10ng/µl, 0.3nM per 50bp).

Generating RNAseq libraries for 293T RNA and reovirus T1L edited by ADAR *in vitro*

100ng 293T RNA or 50ng reoT1L RNA was mixed with ADAR protein at a molar ratio of 1 protein : 50bp RNA and incubated at 37°C for 1hr. Samples were generated in triplicate for each ADAR1-p150, ADAR1-p110, ADAR1-deaminase, ADAR2, ADAR2-deaminase, and ADAR3-deaminase A389V, as well as for 293T-WT for a positive editing control, and 293T-KO and reoT1L-WT as negative editing controls. Libraries were prepared with the TruSeq Stranded Total RNA kit (Illumina), using 100ng input RNA for each 293T library, and 50ng input RNA for each reoT1L library. The 293T libraries were first depleted of rRNA using the NEBNext rRNA Depletion Kit (Human/Mouse/Rat), all other steps of the library protocol were as written. The reoT1L library protocol had two modifications. First, the protocol was started from the Clean Up RCP step, as no rRNA depletion was required. Second, the random hexamers used for reverse transcription were replaced with UMI-linked hexamers that contained a N6 unique molecular identifier 5' of the

hexamer, to be able to identify individual RT events. This was achieved by replacing the Illumina FSA mix with: 250ng UMI-hexamer, 0.5mM dNTPs, 1x First Strand Buffer, 10mM DTT, 1U RNaseOUT, 1x Actinomycin D, 1U Superscript II. For each 293T and reoT1L, the libraries were pooled at an equimolar ratio and sequenced on the Illumina NovaSeq SP (300 cycles, 1 lane per library) with read length of 150nt in paired end configuration.

RNA sequencing data analysis

Analysis of 293T and reoT1L libraries was done in collaboration with Ji-Dung Luo of the Bioinformatics Resource Center, The Rockefeller University. FASTQ files were aligned to hg19 reference genome (293T) or reovirus T1L genome (reoT1L) following STAR (2.7.1a) 2-pass mapping protocol. The aligned SAM files were converted into BAM files using Picard (2.8.1).

The variations in each sample were identified by GATK-Mutect2 (4.0.8.0) and then annotated with ANNOVAR (20170717). A mismatch site was called if the total read depth at that site is greater than or equal to 5 and the alterative nucleotide read depth at that site is greater than or equal to 2. To eliminate SNPs in 293T cells, the allelic frequency of each position in 293KO cell line was applied to calculate threshold (mean ± SD). Only the variants whose allelic frequency larger than threshold were selected.

Next, mismatches were selected that are present in any one of the groups: ADAR1-p150, ADAR1-p110, ADAR1-deaminase, ADAR2, ADAR2-deaminase, ADAR3-Deaminase A389V, or 293T-WT, but not in the 293T-KO or reoT1L-WT group. Finally, mismatch sites present in the dbSNP (snp138, hg19) database were excluded from further analysis. Finally, A-to-G and T-to-C reference-to-read mismatches were selected for annotation.

Determining triplet frequency in the 293T editome

The relative frequency of each adenosine triplet in the 293T editome was calculated by counting all identified editing sites in the published editome of HEK 293T (Chung et al. 2018) and parsing into the number of sites counted for each adenosine triplet. These counts were generated for both the ADAR1-p150 specific editome and the ADAR1 (p150+p110) editome. Counts were then further separated into the number of editing sites in Alu elements and the number of sites in nonAlu sequences, to generate relative frequency for Alu and nonAlu, in the p150- and ADAR1-editomes.

The relative frequency of all adenosine triplets in the transcriptome (hg19, cDNA and ncRNA annotations) were counted using an adapted seqinR code (analysis performed by Joseph Luna), to count the number of times each triplet appears in the transcriptome and generate a relative frequency for each.

The frequency of each adenosine triplet in Alu repeats was counted in the same manner, using sequences compiled by Konkel et al. (2015) that characterized active human Alu elements. The set of active Alu elements was used rather than the full ensemble of Alu elements, as only the active elements are transcribed, making them potential editing substrates.

Chapter IV – Characterising *in vitro* editing activity of ADAR constructs

For each of the ADAR constructs expressed and purified, the newly developed *in vitro* editing assay was used to measure the level of editing activity and the specificity for different adenosine-containing triplets. First, the total editing level was measured for each of ADAR1-p150, ADAR1-p110, ADAR1-deaminase, ADAR2, and ADAR2-deaminase, to observe the level of activity for each protein, and if ADAR deaminase constructs had decreased activity due to the loss of the other domains. Then, the editing events were separated into each of the different adenosine positions, showing the relative frequency of each 5' and 3' neighbor. The editing frequencies characterized for each construct were then compared, both to those previously characterized (Eggington et al. 2011) and to each other.

A comparison of the editing specificity of the ADAR-1 and ADAR-2 deaminases was performed to inform how the structure of each deaminase is involved in substrate selection.

Each deaminase construct was then compared to its respective full-length protein, ADAR1-p110 or ADAR2, to observe if there was additional selectivity due to the presence of dsRBDs, and how this differed for the two ADAR family members.

The ADAR1 isoforms p150 and p110 were also compared to identify if the longer isoform had additional specificity due to the ZBD.

The *in vitro* editing assay was then utilized to probe the inactivity of ADAR3, first to confirm the lack of editing by the wildtype ADAR3-deaminase, and then to probe the active site differences between ADAR2 and ADAR3 by characterizing editing of the ADAR3-deaminase mutant A389V.

Measuring total editing levels

Before measuring the editing level for each triplet individually, the total editing activity was measured for each of ADAR1-p150, ADAR1-p110, ADAR1-deaminase, ADAR2 and ADAR2-deaminase, with values shown in **table 4.1**.

Table 4.1. Overall editing of the 50bp dsRNA substrate by each ADAR construct, following incubation for 1 hour with 0.05μM protein and 0.05μM dsRNA. Values are n = 3 ± SD.

	Total editing [%]	**± SD**
ADAR1-p150	11.1	3.31
ADAR1-p110	4.0	0.22
ADAR1-deaminase	10.2	1.39
ADAR2	14.6	1.82
ADAR2-deaminase	4.2	0.93
Negative control	0.8	0.32

The overall editing level ranged from 4 – 15 %, with the negative control having <1% editing across all adenosine sites in the substrate. This low level of total activity agrees with that previously seen by Eggington et al. (2011), as they achieved ~20% editing after 4 hours incubation of protein and RNA at the same ratio as that used here. ADAR2 had the highest total activity, with 14.6 ± 1.82%, followed by similar levels of editing by the long ADAR1-p150 isoform, with 11.1 ± 3.31%, and the ADAR1-deaminase, with 10.2 ± 1.39%. Interestingly, the third full-length protein, ADAR1-p110, had much lower activity, with 4.0 ± 0.22%, and the ADAR2-deaminase had the lowest activity of all constructs tested, with 4.2 ± 0.93%.

It is interesting that the ADAR2-deaminase had much lower editing than the ADAR1-deaminase, considering that both proteins were isolated to similarly high purity. Similarly, the ADAR1-p110 had the highest purity of the full-length constructs at ~80%, compared to the ~70% for ADAR2 and ~50% for ADAR1-p150, yet the p110 protein had much lower activity. The full-length ADAR1-p150 protein was expected to have higher levels of editing than the ADAR1-deaminase construct, due to the former being a full-length functional protein and the latter being a single domain. The similar levels of editing seen could be due to the deaminase being isolated at a higher purity than the ADAR1-p150 protein and therefore having a higher actual concentration of protein in the editing reaction.

Measuring total editing levels for each adenosine triplet

After measuring the total editing for each construct, the editing events were then separated into the 16 different adenosine triplets, to measure the level of editing for each combination. **Figure 4.1** shows the percentage of editing for each triplet, for each protein construct tested. For each triplet, the measured editing level was compared to that measured for the negative control, and significance was calculated by unpaired T-test (two-tailed distribution, assuming unequal variance), with **table 4.2** detailing which triplets were found to have significant levels of editing over background. Sites without significance are colored grey in **table 4.2**, and a number of sites that had $P<0.05$, but the negative control had a higher level of measurement are also greyed out.

It can be seen in **figure 4.2** that all ADAR constructs follow the published trend of editing at sites with a 5'A or 5'U, and less editing of sites with a 5'C or 5'G sites, with the exception of CAG. The background level of editing at the GAG site was much higher than the other triplets, due to the sequencing alignment requirement that the 3'G is always fixed, but no ADAR proteins had significant editing of this site.

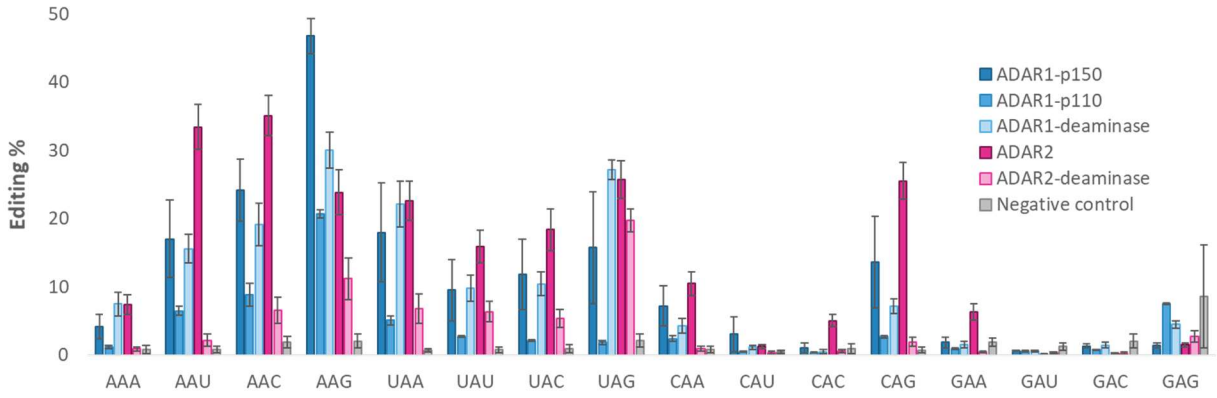

Figure 4.1. Total editing at each triplet in the 50bp dsRNA substrate, measuring percentage as number of A-to-G changes counted at each location over the total number of reads. Showing editing levels for ADAR1-p150, -p110, and -deaminase, and ADAR2 and ADAR2-deaminase. Grey bars represent background A-to-G noise in the sequencing, from samples without protein added. Each bar is n = 3 ± SD.

ADAR2, which had the highest overall editing, could be seen to have high levels of editing across many sites, shown in **figure 4.1**, and had the most sites with significant editing (**table 4.1**).

ADAR1-deaminase was found to have significant editing at all 5'A and 5'U sites, and most 5'C sites. ADAR1-p150 and -p110 as well as ADAR2-deaminase had fewer sites with significant editing. For ADAR1-p110 and ADAR2-deaminase this could possibly be due to the lower level of overall activity of each protein (**table 4.1**). ADAR1-p150 had a low number of sites with significant editing (**table 4.2**) but the overall activity measured was higher than ADAR1-p110 and ADAR2-deaminase (**table 4.1**).

From **figure 4.1**, it can be observed that the higher level of overall editing is likely due to a high level of editing at AAG, with ~45% editing of that site, and low editing of most other triplet sites. All of the proteins tested except ADAR1-p110 showed significant editing of the adenosine in the UAG triplet. UAG is the sequence of the amber/W site in hepatitis delta virus (HDV); editing of the amber/W site is required for HDV to shift from replication to packaging and complete the viral lifecycle (Casey et al. 2006).

Table 4.2. Identification of triplets with significant levels of editing over background. Significance is calculated by T-test (unpaired, two-tailed, unequal variance), comparing each sample (n = 3) to the negative control. Non-significant sites (ns) and significant sites where the negative control was higher than the sample are shaded grey. * [0.05> P > 0.01], ** [0.01 > P > 0.001], *** [P < 0.001]

	ADAR1-p150	**ADAR1-p110**	**ADAR1-D**	**ADAR2**	**ADAR2-D**
AAA	ns	ns	*	**	ns
AAU	*	***	**	**	ns
AAC	**	**	**	**	*
AAG	***	***	***	**	*
UAA	*	**	**	**	*
UAU	ns	**	**	**	*
UAC	*	ns	**	**	*
UAG	ns	ns	***	**	***
CAA	*	*	*	**	ns
CAU	ns	ns	*	*	ns
CAC	ns	ns	ns	**	ns
CAG	ns	**	**	**	*
GAA	ns	ns	ns	*	*
GAU	ns	ns	ns	*	*
GAC	ns	ns	ns	*	*
GAG	ns	ns	ns	ns	ns

Comparing total editing frequencies to single editing frequencies

To avoid complications due to multiple editing events on a single substrate, all reads with only a single editing event were isolated, and the number of reads for each triplet were counted. The relative frequency of each triplet was then calculated for each site, and frequencies were also calculated for the total editing counts from **figure 4.1** as a comparison. **Figure 4.2** shows the relative frequency of editing for each triplet site, for total (**4.2a**) and single (**4.2b**) editing events.

Overall, the frequency of editing for each adenosine-containing triple looked similar for the total and single events, and all further analysis was completed using only the single event frequencies.

For each **4.2a** and **4.2b**, if we rank the triplets from most to least frequent, the order of editing is similar for most of the triplets, but the UAA triplet drops in the ranking for all constructs except ADAR2. Most of the sites that have lower rank in the single editing frequency are those that have multiple As in a row – such as UAA, as well as AAA, AAG, and AAU. This implies that in the total editing data, there are substrates with multiple As in a row edited, and using the total editing data to measure the frequency of UAA, for example, would be a combination of editing actually occurring at the adenosine in the UAA triplet as well as editing occurring for UAI, if the 3'A was already edited.

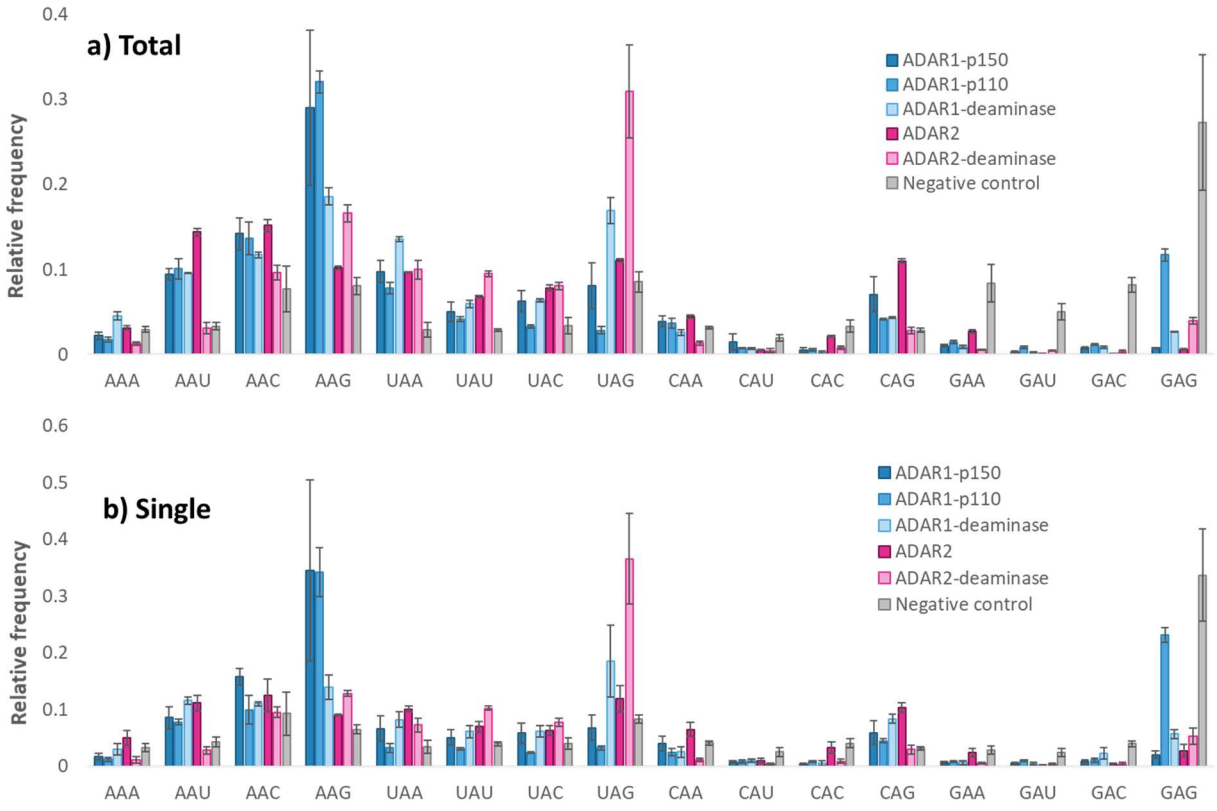

Figure 4.2. Relative frequency of editing for each adenosine-containing triplet, for all editing events counted in each sample (a) and for the pool of substrate with only a single editing event (b). For each ADAR construct tested, the sum of the frequencies across all 16 triplets will add to one. n = 3 ± SD.

Specificity in the deaminase domain: comparing ADAR1 and ADAR2 deaminases

To identify differences in editing specificity, each of the ADAR constructs were compared one-on-one. First, the two deaminase constructs were compared. These proteins lack dsRBDs and editing specificity of each will be due to the inherent substrate selectivity of the deaminase domain alone. The relative editing frequency for each triplet is shown in **figure 4.3** for the comparison of ADAR1-deaminase and ADAR2-deaminase.

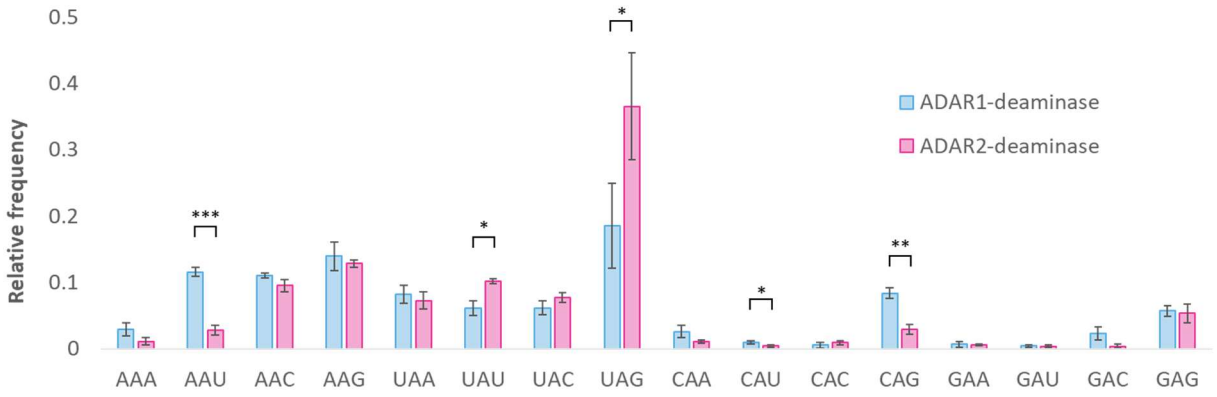

Figure 4.3. Comparison of the relative frequency of NAN triplets (single edits only) for ADAR1-deaminase and ADAR2-deaminase. Triplets with significant differences in frequency are indicated by asterisks for * [0.05> P > 0.01], ** [0.01 > P > 0.001], *** [P < 0.001]. n = 3 ± SD.

From **figure 4.3**, it can be seen that ADAR2-deaminase had a relatively high frequency of editing at the adenosine in triplet UAG, while the ADAR1-deaminase was spread more evenly across most 5'A and 3'U sites, as well as some editing at the CAG site. Significant difference was seen in the frequency at CAA, CAU and GAG, but these sites did not have significant levels of editing. Of the sites with measurable editing and significant difference between the two deaminases, ADAR1-deaminase had higher frequency at AAA, AAU and CAG, while ADAR2-deaminase had higher frequency for UAU and UAG.

Observing the trend of which deaminase has higher frequency for specific neighbors, it can be seen that ADAR1-deaminase had higher frequency for all 5'A sites, while ADAR2-deaminase had higher frequency for most of the 5'U sites, except UAA. It appears that the ADAR1 and ADAR2 deaminases have a very subtle difference in editing preference, with ADAR1 favoring 5'A and ADAR2 favoring 5'U.

Modelling the ADAR1-deaminase on the ADAR2 crystal structure

To further investigate the differences seen in 5' neighbor selectivity, the ADAR1-deaminase was modelled onto the ADAR2 crystal structure (PDB 5ED2, Matthews et al. 2016) that was generated in complex with a dsRNA substrate. The ADAR3-deaminase was also modelled for comparison, and all three structures can be seen in **figure 4.4**, with the ADAR1 homology model (**4.4a**), ADAR2 crystal structure (**4.4b**) and ADAR3 homology model (**4.4c**).

From initial observations, the ADAR2 and ADAR3 structures are very similar, with the secondary structures of the ADAR3 model mapping closely onto the equivalent ADAR2 structures, as expected of protein domains with 75% similarity by sequence. The ADAR1 deaminase, being only 55% similar to each ADAR2 and ADAR3, has a similar organization of secondary structures, but the entire domain appears to be tilted slightly. The catalytic domains of all members of the CDA superfamily contain a distinct five-stranded beta-sheet core, with an alpha helix aligned perpendicular to the strands, which is visible for all three deaminases in **figure 4.4**, but the helix in ADAR1 is lower than those for ADAR2 and ADAR3, with the deaminase domain rotated slightly clockwise in relation to the RNA substrate.

Figure **4.4b** and **4.4c** show that the 5' RNA binding loop – a stretch of flexible residues that interact with the 5' end of the RNA substrate, marked in the figure by a black arrow – is similarly located for the ADAR2 and ADAR3 deaminases, although the ADAR3 loop is possibly closer to the substrate. The flexibility of this region makes it difficult to model accurately, and the ADAR2 loop was only identified in the crystal in complex with RNA, in the earlier crystal structure by Macbeth et al. (2005) this stretch of residues was not visible in the structure due to high flexibility.

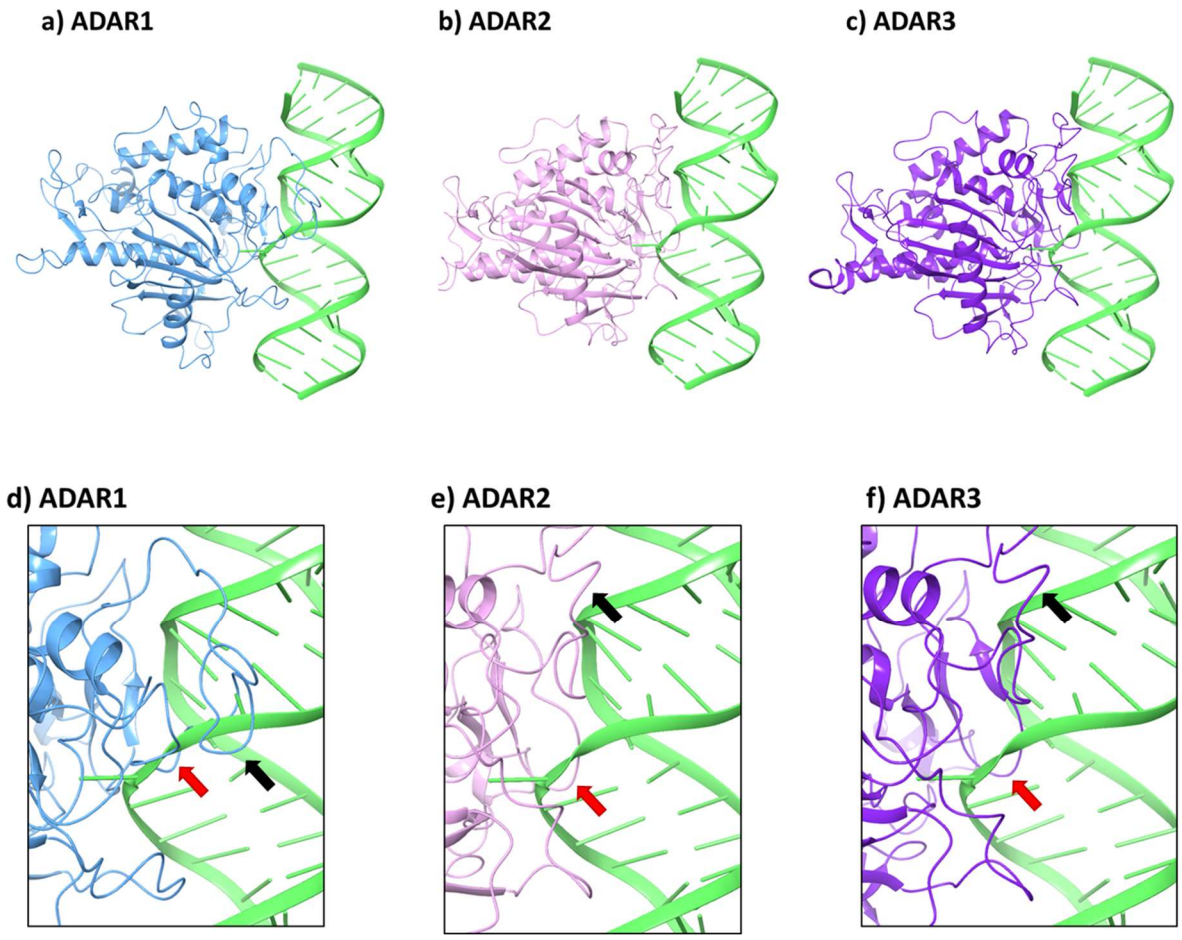

Figure 4.4. Homology models of ADAR1 and ADAR3 deaminases mapped onto the ADAR2 crystal structure (PDB 5ED2, Matthews et al. 2016), generated in complex with double-stranded substrate RNA. Each model was aligned to the active sites residues and modelled with the RNA substrate from the ADAR2 crystal, which is shown with the edited strand oriented with the 5' end of the substrate at the top of the image. The ADAR1 (a) and ADAR3 (c) models flank the ADAR2 (b) crystal structure, with closer images of the orphan base binding loop shown for ADAR1 (d), ADAR2 (e) and ADAR3 (f), with binding loop marked by a red arrow for each. The 5' RNA binding loop is also shown for each protein in (d, e, f) marked by a black arrow.

The slight clockwise rotation of the ADAR1 domain, as well as the additional five residues in the ADAR1 5' RNA binding loop, brings the loop closer to the targeted editing site, as seen in figure **4.4d**, with more overlap between the protein loop and the RNA duplex than either of the other deaminases. Whilst the ADAR2 and ADAR3 5' RNA binding loops are interacting with bases one full turn above the editing site, the ADAR1 binding loop appears to be passing directly above the 5' neighbors of the flipped base.

Again, it is difficult to determine the actual location of the binding loop without a crystal structure of the ADAR1 domain, but this closer interaction to the active site does correlate with RNA substrate structures identified by Thomas and Beal (2017) and Wang et al. (2018).

While the ADAR2 deaminase favored sites at the immediate 5' end of loops – which could also mean favoring perfect double-stranded substrate 5' of that loop – the ADAR1 deaminase favored editing sites with a loop 3-6bp upstream. Perhaps if the RNA substrate had a loop 5' of the editing site, rather than being a perfect duplex, then the ADAR1 5' RNA binding loop would intercalate into that loop, rather than the unlikely overlap of protein and RNA seen in **4.4d**.

Modelling the orphan-base binding loop

The orphan base binding loop, which intercalates into the RNA duplex to interact with the orphan base and free the targeted adenosine for base-flipping and editing, can be seen in **figure 4.4d-f**, marked by a red arrow for each deaminase. The loops look to be in a similar position for ADAR2 (**4.4e**) and ADAR3 (**4.4f**), but the ADAR1 loop is shifted further away from the orphan base (**4.4d**).

The orphan base binding loop is composed of residues (S/N)G**E**GT(V/I)PV, with the highlighted glutamate interacting with the orphan base to stabilize the flipped conformation. The presence of flanking glycines is necessary for flexibility of the loop, as the ADAR1 disease mutation of

G1007R, the glycine upstream of the glutamate, kills activity of the protein (Rice et al. 2012). Highlighting just the orphan base binding loop residues in the homology models generated in **figure 4.4**, the location of each deaminase loop with respect to the RNA duplex is shown in **figure 4.5**, with ADAR1 in blue, ADAR2 in pink and ADAR3 in purple.

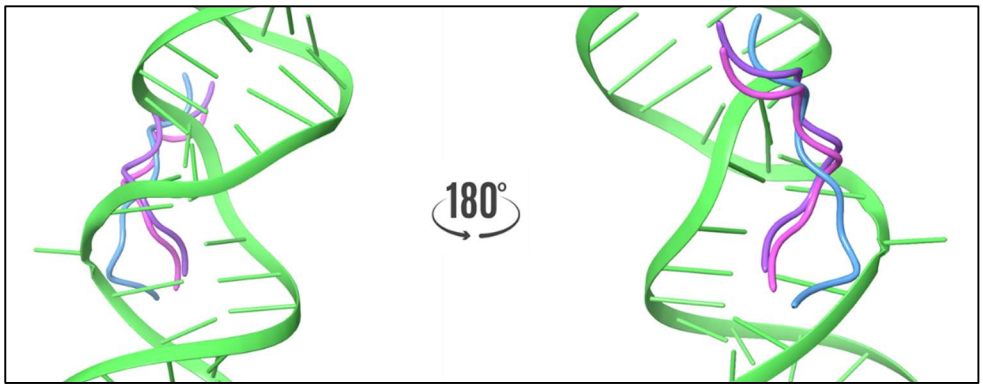

Figure 4.5. The orphan base flipping loop of ADAR1 (blue), ADAR2 (pink) and ADAR3 (purple) modelled onto an RNA duplex. The duplex at left is in the same orientation as all models in figure 4.4, with the duplex at right rotated 180° to better illustrate the protein intercalating into the RNA duplex. The duplex is oriented so that the top of the image is the 5' end of the edited strand.

As seen in **figure 4.5**, the ADAR1 orphan-base binding loop deviates from the ADAR2 and ADAR3 loops, despite the similar sequences. The variation is occurring at the N-terminal end of the binding loop, which has asparagine (N) in ADAR1 and serine (S) in ADAR2 and ADAR3. This residue forms hydrogen bond interactions with the immediate 3' neighbor to the editing site, most often guanosine, and both N and S are able to forge this bond, although serine is a stronger interacting partner.

Comparing the two active ADAR deaminases, it can be seen that the ADAR2 loop pushes much further into the RNA duplex, forming a close interaction with the orphan base, while the ADAR1 loop sits much closer the RNA backbone of the editing strand, implying a more distant, weaker interaction with the orphan base. It could be that ADAR2 pushes further into the RNA duplex and competes strongly for the orphan base, flipping the now unpaired adenosine into the active site. If the ADAR1 does not have the ability to compete for the orphan base, could ADAR1 instead select for editing sites with weaker RNA duplex interactions.

The trend for ADAR1 targeting sites with 5'A and ADAR2 favoring sites with 5'U could be explained by the difference in orphan-base interactions. Multiple pyrimidines in a row are less stable than alternating purine-pyrimidines and are easier to flip out of a duplex structure (Colizzi et al. 2019). This has previously been seen for ADAR editing, with Herbert and Rich (2001) finding that having multiple pyrimidines 5' of the editing site increase the level of editing. Alternating purine-pyrimidine sequences form cross-strand stacking of pyrimidines that increases the stability of the duplex and so increases the energy required to break the structure. The AAG triplet, favored by ADAR1-deaminase, and UAG, favored by ADAR2-deaminase, are schematized in **figure 4.6**, which illustrates the cross-strand stacking of the UAG triplet with the non-editing RNA strand.

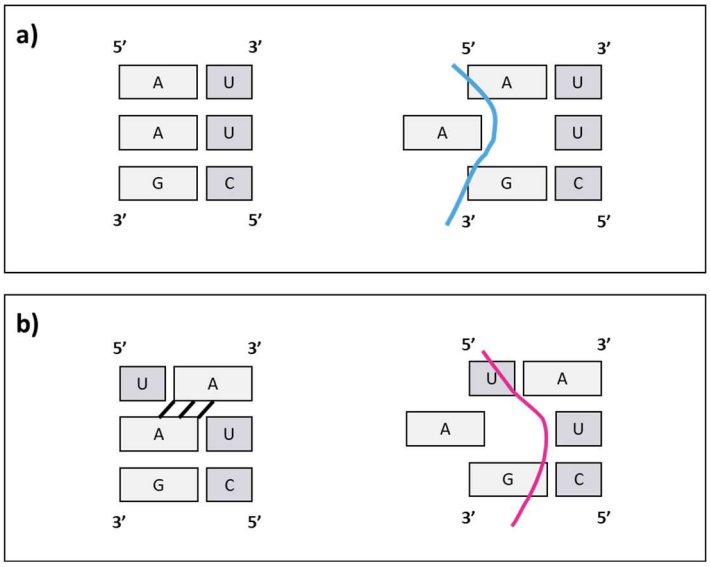

Figure 4.6. Schematic of AAG (a) and UAG (b) triplets in an RNA duplex, with cross-strand stacking of pyrimidines visible for UAG (b). Base-flipping of the adenosine is shown on the right for each triplet, with the relative location of the orphan-base binding loop of ADAR1 (a, blue) and ADAR2 (b, pink), illustrating that the more stable duplex of UAG (b) requires the protein loop to interact much closer to the orphan base, while the weaker duplex of AAG (a) does not require the same level of interaction.

Figure 4.6a illustrates the weaker RNA duplex of AAG, and the more distant interaction between the orphan base and the ADAR1 protein loop, while the stronger UAG duplex (**4.6b**) requires the ADAR2 loop to push much closer to the orphan base to disrupt the cross-stacked pyrimidines.

Comparing the full-length ADAR1-p110 and ADAR2

Having observed that the ADAR1 and ADAR2 deaminases have slightly different editing specificities, the full-length proteins were then compared to see if the same patterns were observed or if the addition of dsRBDs provided additional specificity. **Figure 4.7** shows the relative editing frequency for ADAR1-p110 and ADAR2. All relative frequencies were plotted for 15 adenosine sites, excluding the GAG site, which was negative for editing for all proteins but had high background signal for proteins with low overall editing, such as ADAR1-p110.

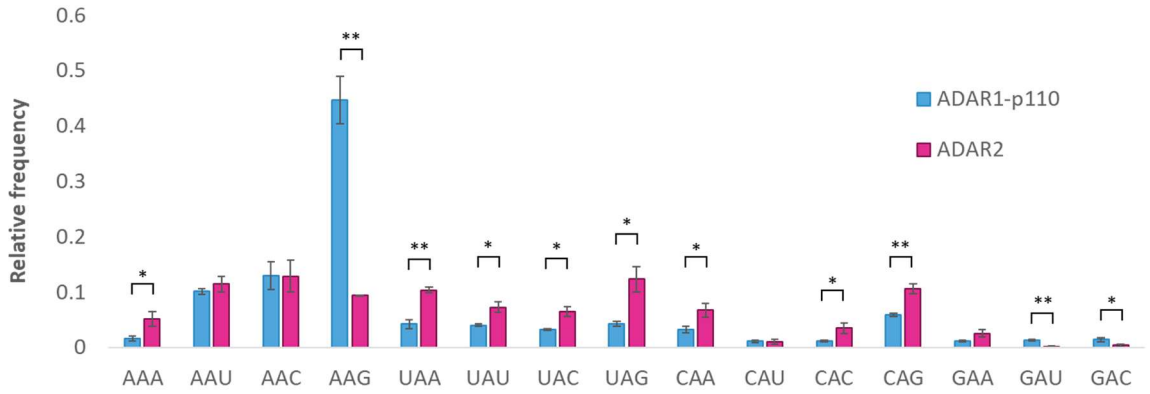

Figure 4.7. Comparison of the relative frequency of NAN triplets (single edits only) for ADAR1-p110 and ADAR2, relative frequency counted across 15 triplets, with GAG removed due to high noise. Triplets with significant differences in frequency are indicated by asterisks for * [0.05> P > 0.01], ** [0.01 > P > 0.001], *** [P < 0.001]. n = 3 ± SD.

As seen in **figure 4.7**, ADAR2 had higher frequency for most adenosine-containing triplets, with the exception of the 5'A sites AAU and AAC, which had no significant difference between the two proteins, and AAG, which was highly favored by ADAR1-p110, with 3.2 fold higher frequency than ADAR2 at this site. ADAR2 had similar frequency for most sites, with no single site at more than ~10% of all edits, while the AAG site represented ~45% of ADAR1-p110 edits.

The editing pattern observed in **figure 4.7** was very different to that seen for the deaminase domains (**fig 4.3**), implying that the dsRBDs do provide additional substrate selectivity. If ADAR1-p110 was utilizing binding by the dsRBDs to target the AAG site so specifically, the binding could be occurring upstream or downstream of the site, as the AAG site is right in the middle of the 50bp dsRNA substrate, at position 26.

The role of dsRBDs: comparing full-length ADARs to deaminase-only

Having observed that the full-length ADAR1-p110 and ADAR2 have very different patterns of editing frequency for adenosine-containing triplets, each full-length ADAR was then compared to its respective deaminase-only construct.

Comparing ADAR1-p110 to ADAR1-deaminase

Figure 4.8 shows the relative editing frequency for ADAR1-p110 and ADAR1-deaminase.

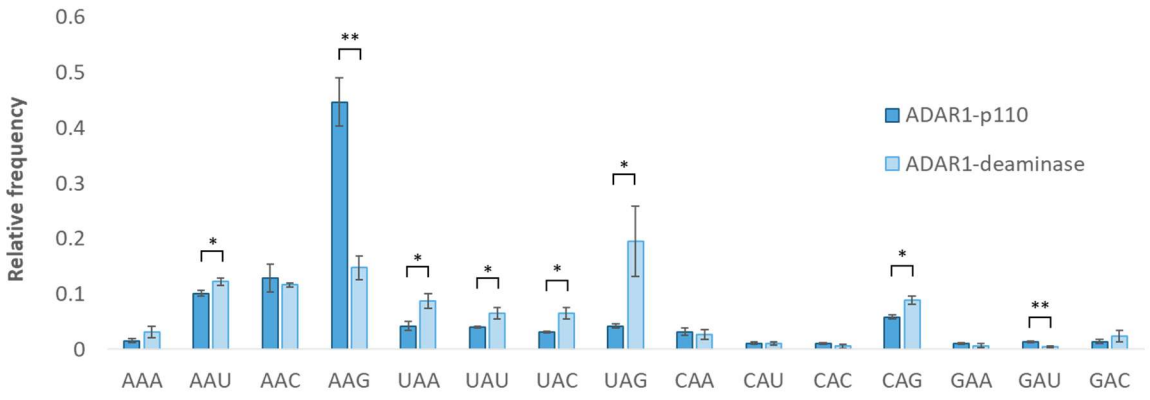

Figure 4.8. Comparison of the relative frequency of NAN triplets (single edits only) for ADAR1-p110 and ADAR1-deaminase. Triplets with significant differences in frequency are indicated by asterisks for * [0.05> P > 0.01], ** [0.01 > P > 0.001], *** [P < 0.001]. n = 3 ± SD.

Similar to the comparison of ADAR1-p110 to ADAR2, p110 had significantly higher frequency at AAG, 2.6-fold higher than ADAR1-deaminase, while ADAR1-deaminase had significantly higher frequency at all other sites: AAU, UAN, and CAG, with highest frequency at UAG with 3.2-fold higher frequency than p110. When comparing ADAR1-deaminase to ADAR2-deaminase, the ADAR1-deaminase had higher frequency of 5'A sites, while ADAR2-deaminase favored 5'U. Here, ADAR1-p110 had higher frequency of 5'A than the deaminase, indicating that while the deaminase favored 5'A, the addition of dsRBD pushed that to an even higher frequency.

Comparing ADAR2 to ADAR2-deaminase

Figure 4.9 shows the relative editing frequency for ADAR1-p110 and ADAR1-deaminase.

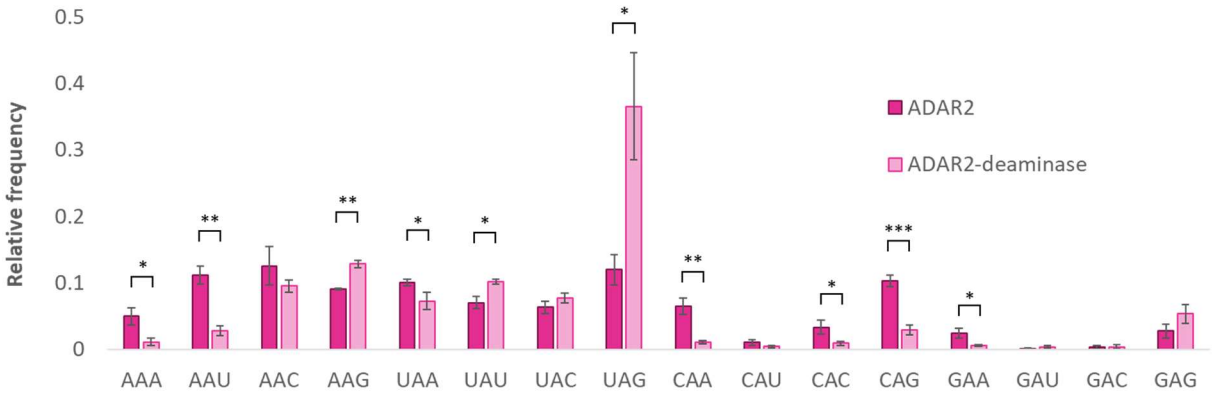

Figure 4.9. Comparison of the relative frequency of NAN triplets (single edits only) for ADAR2 and ADAR2-deaminase. Triplets with significant differences in frequency are indicated by asterisks for * [0.05> P > 0.01], ** [0.01 > P > 0.001], *** [P < 0.001]. n = 3 ± SD.

Much like the ADAR1 full-length versus deaminase comparison, the ADAR2-deaminase had significantly higher frequency at UAG relative to the full-length ADAR2. ADAR2 had significantly higher frequency at AA(A/U), UAA, and CA(A/C/G) while the deaminase had higher frequency at AAG and UA(U/G). Much like the ADAR1-deaminase, the full-length ADAR2 had

relatively similar frequency for many 5' and 3' neighbour combinations of the editing site, with no single site having more than ~15% of the editing events, while ADAR2-deaminase had ~40% of all editing at the UAG site. Comparing the full-length and deaminase editing frequencies for ADAR1 and ADAR2, there were three adenosine-containing triplets that stood out: AAG, CAG and UAG.

AAG had similar editing frequency by the deaminases, but ADAR2 had lower frequency than ADAR2-deaminase and ADAR1-p110 had higher frequency than ADAR1-deaminase. This indicates that while the two deaminases had similar efficiency with editing this site, the ADAR1 dsRBDs increased efficiency, while ADAR2 dsRBDs either decreased efficiency or increased editing of other sites, for an apparent drop in efficiency.

CAG had the opposite pattern to AAG, with ADAR2 having higher frequency than ADAR2-deaminase, and ADAR1-p110 having lower frequency than ADAR1-deaminase. This indicates that ADAR2 dsRBDs increased editing efficiency for this site, confirmed previously as this site is present in the highly edited *GLURB* Q/R site (Stefl et al. 2010), which requires dsRBD binding to the 3' region of the transcript to facilitate editing.

UAG had higher editing frequency by ADAR2-deaminase than by ADAR1-deaminase, but both deaminases had higher frequency than the full-length proteins. Along with AAG, UAG is one of the most favorable editing sites, due to the 5'A/U not clashing sterically with the protein, and the 3'G interacting favorably with the N/S residue as described on pg. 68. As such UAG is an easily editable site, but neither ADAR1 nor ADAR2 increases editing of this site, indicating that the role of the dsRBDs is to increase editing for difficult-to-edit sites. It is therefore interesting that the other easily editable site, AAG, had increased editing by the ADAR1-p110 protein. Perhaps AAG is common in the ADAR1-favored RNA structures, with loops immediately 5' of the editing target.

The role of the N-terminal domain: comparing ADAR1 isoforms p110 and p150

The relative frequency of editing at each triplet is shown for ADAR1-p150 and ADAR1-p110 in **figure 4.10**. As the ADAR1 isoforms only differ by the N-terminal region, with ADAR1-p150 possessing a functional ZBD to interact with left-handed DNA/RNA, any differences in editing frequency between the ADAR1 isoforms would likely be due to the ZBD.

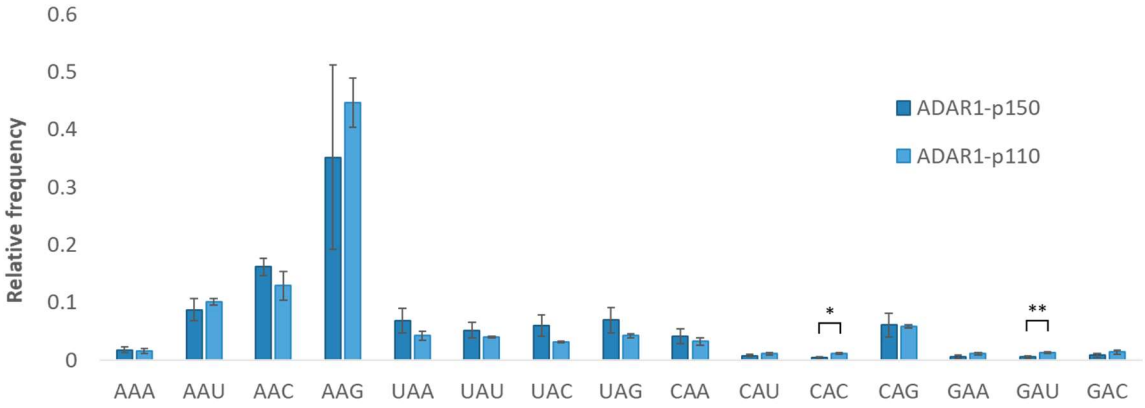

Figure 4.10. Comparison of the relative frequency of NAN triplets (single edits only) for ADAR1 isoforms p150 and p110, relative frequency counted across 15 triplets, with GAG removed due to high noise. Triplets with significant differences in frequency are indicated by asterisks for * [$0.05 > P > 0.01$], ** [$0.01 > P > 0.001$], *** [$P < 0.001$]. n = 3 ± SD.

From **figure 4.10**, it can be seen that the only editing sites with a significant difference in frequency are CAC and GAU, both sites that do not have significant editing over background. As such, there was no difference in editing by the ADAR1 isoforms on a perfectly double-stranded 50bp substrate. The highest editing frequency was for AAG, with ~35-40% of all edits occurring at this site. Both isoforms favored 5'A sites over other possible 5' neighbors.

The similar pattern of editing frequency by ADAR1-p150 and ADAR1-p110 could either mean that the ZBD has no effect on a substrate of this length, or that the role of the ZBD is not substrate-selectivity but could be to improve efficiency of editing for sites that are already good ADAR targets. The overall activity of the ADAR1-p150 isoform (11.1 ± 3.31%) was higher than the p110 isoform (4.0 ± 0.22%), which could be due to the presence of the ZBD, either increasing the editing efficiency of the protein or even stabilizing the protein so that a larger fraction of the purified protein is consistently active.

Work by Koeris et al. (2005) compared editing by ADAR1-p150 against a substrate that either did or did not contain a $(CG)_6$ repeat and found that the repeat sequence increased the editing level for sites that were already being edited, and had little effect on non-edited As, such as AAA repeats. This supports the notion that the ZBD in ADAR1-p150 increases the level of editing but does not necessarily have a role in selection of the editing site. As ADAR1-p110 is constitutively expressed and ADAR1-p150 is only induced following interferon stimulation, it is possible that the low level of editing seen for p110 could be the basal level of editing. A cell in an antiviral state may require a higher level of editing, thus the increased activity of the interferon-induced p150 isoform.

Probing the cause of ADAR3 inactivity

From sequence alignment to human ADAR1 and ADAR2, ADAR3 has all of the catalytic residues that form the catalytic core, but ADAR3 has never been shown to have activity and no target substrates have been identified. *In vitro*, ADAR3 has been shown to bind to, but not edit, RNA substrates of ADAR2 (Oakes et al. 2017b). Following the discovery of ADAR3, it was theorized that the lack of activity could be due to autoinhibition of the deaminase domain, by the dsRBDs of the protein (Cho et al. 2003).

Confirmation of ADAR3-deaminase inactivity

The deaminase domain of ADAR3 was expressed alone, to determine if the inactivity of the protein was due to the deaminase domain lacking catalytic activity, or due to inhibition by the other domains – the dsRBDs and R domain. In **table 4.3**, the overall level of editing measured for the ADAR3-deaminase is shown in bold, and at 1.1 ± 0.37% the value is similar to the negative control at 0.8 ± 0.32%. **Figure 4.11a** and **table 4.4** further illustrate that the ADAR3-deaminase is catalytically inactive, with **figure 4.11a** showing the negligible total editing measured for each adenosine-containing triplet, and **table 4.4** showing that all of the triplets had no significant difference to the negative control. As such, the inactivity of ADAR3 does not depend on the dsRBDs or R domain, as the deaminase domain alone lacks activity.

Table 4.3. Overall editing of the 50bp dsRNA substrate by the ADAR3-deaminase and the mutant A389V, following incubation for 1 hour with 0.05μM protein and 0.05μM dsRNA. Editing levels for other proteins previously shown in Table 4.1 as comparison. Values are n = 3 ± SD.

	Total editing [%]	**± SD**
ADAR1-p150	11.1	3.31
ADAR1-p110	4.0	0.22
ADAR1-deaminase	10.2	1.39
ADAR2	14.6	1.82
ADAR2-deaminase	4.2	0.93
Negative control	0.8	0.32
ADAR3-deaminase	**1.1**	**0.37**
ADAR3-deaminase A389V	**3.7**	**1.58**

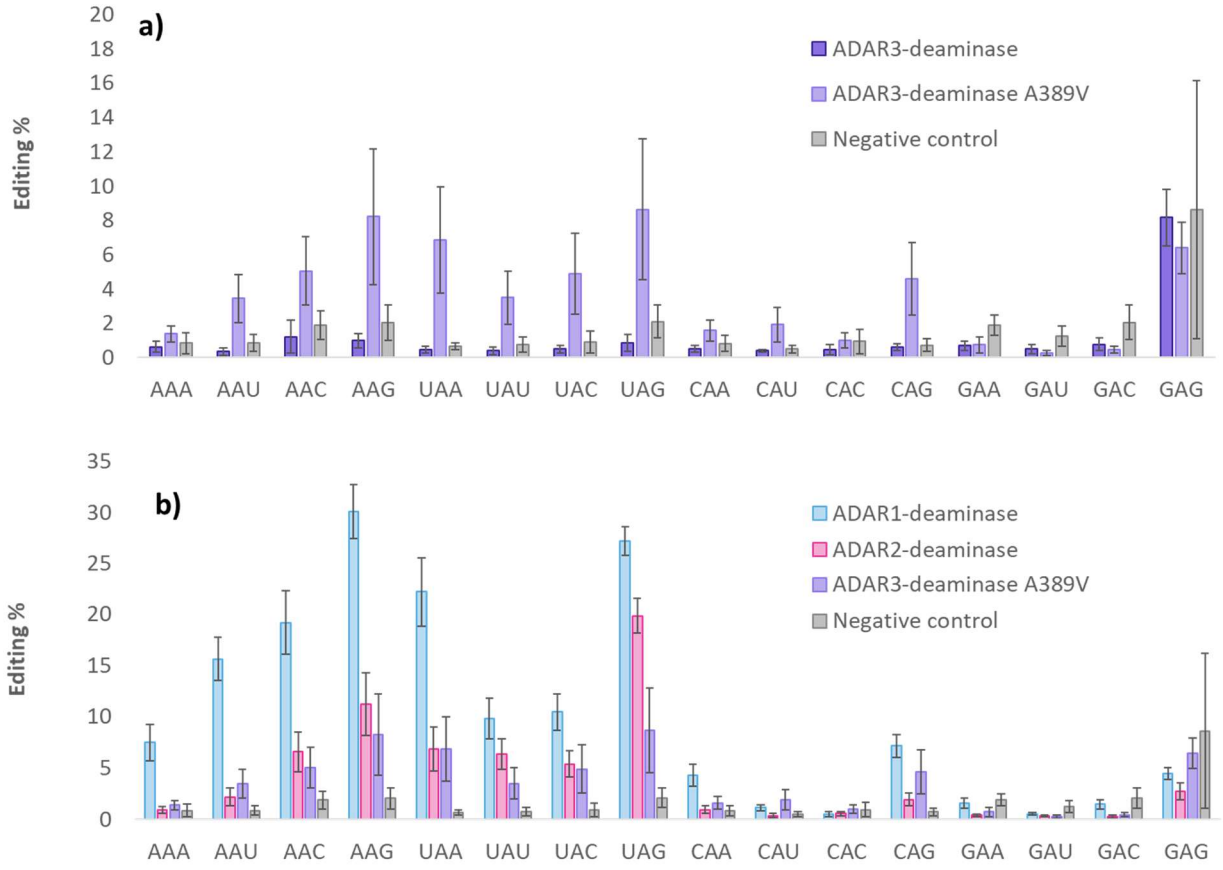

Figure 4.11. Total editing at each triplet in the 50bp dsRNA substrate, measuring percentage as number of A-to-G changes counted at each location over the total number of reads. Showing editing levels for ADAR1-p150, -p110, and -deaminase, and ADAR2 and ADAR2-deaminase. Grey bars represent background A-to-G noise in the sequencing, from samples without protein added. Each bar is n = 3 ± SD.

Table 4.4. Identification of triplets with significant levels of editing over background for ADAR3-deaminase and ADAR3-deaminase A389V. Significance calculated by T-test (unpaired, two-tailed, unequal variance), comparing each sample (n = 3) to the negative control (n = 3).

* [0.05> *P* > 0.01], ** [0.01 > *P* > 0.001], *** [*P* < 0.001]

	ADAR3-D	**ADAR3-D A389V**
AAA	ns	ns
AAU	ns	*
AAC	ns	ns
AAG	ns	*
UAA	ns	*
UAU	ns	*
UAC	ns	*
UAG	ns	*
CAA	ns	ns
CAU	ns	ns
CAC	ns	ns
CAG	ns	*
GAA	ns	ns
GAU	ns	ns
GAC	ns	ns
GAG	ns	ns

Design of the ADAR3-deaminase mutant A389V

From the sequence alignment of the three ADAR family members (**fig. 1.5**), all residues that form the active site pocket are the same for ADAR2 and ADAR3, except for one. A389 in ADAR3 is valine in ADAR2 and isoleucine in ADAR1. When the ADAR2 crystal structure was generated, it was theorized by Matthews et al. (2016) that this residue forms a hydrophobic 'floor' for the adenosine to sit on when coordinated in the active site. The presence of an alanine at this position in ADAR3 could play a role in the protein's inactivity, due to the shorter side chain length.

Measuring activity of the ADAR3-deaminase mutant A389V

The ADAR3-deaminase A389V mutant was generated and purified, and the level of editing was measured, with the overall activity shown in **table 4.3**. Unlike the wildtype ADAR3-deaminase, the A389V mutant was shown to have a low level of A-to-I editing. The low level of editing for the A389V mutant, 3.7 ± 1.58%, was similar to the level measured for ADAR1-p110 and ADAR2-deaminase.

Figure 4.11a shows the total editing measured at each triplet, with measurable editing by ADAR3-deaminase A389V at most of the 5'A and 5'U positions as well as CAG, similar to the ADAR1 and ADAR2 proteins. **Table 4.4** shows that adenosines in triplets AA(U/G), UAN and CAG had significant levels of editing compared to the negative control, confirming that the A389V mutation was able to rescue catalytic activity. The activity was still low, and the introduction of additional mutations could potentially increase the level of editing over that seen here.

Comparing ADAR3-deaminase A389V to ADAR1 and ADAR2 deaminases

Figure **4.11b** shows the overall level of editing for ADAR3 A389V in comparison to the ADAR1 and ADAR2 deaminases, showing that ADAR3-deaminase A389V had similar overall editing to ADAR2-deaminase at most sites. **Figure 4.12** shows relative frequency of single-editing events for all three deaminases. The ADAR3 mutant had similar frequencies to ADAR1 at most triplets, most apparent for editing sites UA(U/C/G), which all had significant differences to the ADAR2-deaminase but not to the ADAR1-deaminase. It is unexpected that the ADAR3-deaminase mutant behaves similar to ADAR1, considering the sequence similarity to ADAR2 and the homology model of ADAR3 also being more ADAR2-like than ADAR1-like (**fig 4.4, 4.5**). Perhaps the substrate selectivity of ADAR3 depends on the 5' RNA binding loop, which is the same length as

the ADAR2 loop, but the amino acid content is significantly different. The ADAR2 loop is **SP**H**E**PILEEPADR**H**PN**R**, and the ADAR3 loop is **SP**Y**E**ITTDLHSSK**H**LV**R**, with conserved residues in bold.

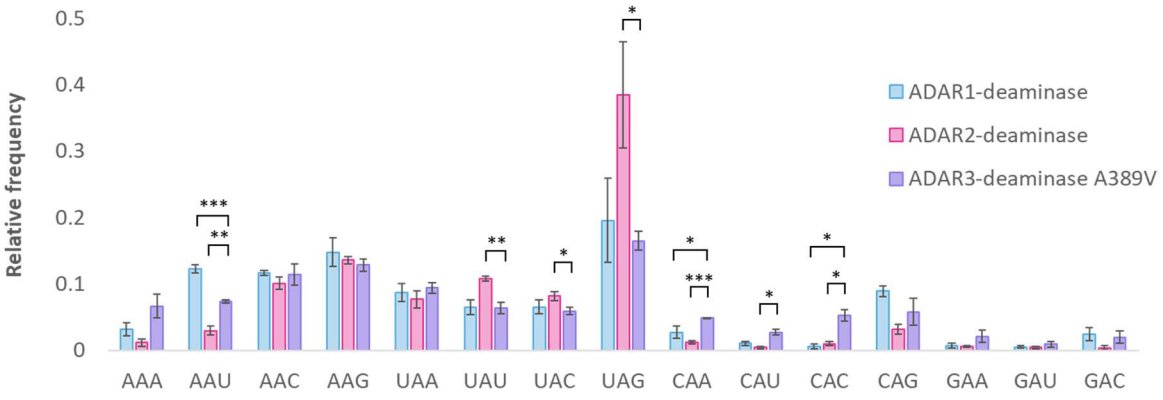

Figure 4.12. Comparison of the relative frequency of NAN triplets (single edits only) for deaminases ADAR1, ADAR2 and ADAR3 A389V., relative frequency counted across 15 triplets, with GAG removed due to high noise. Triplets with significant differences in frequency are indicated by asterisks for * [0.05> P > 0.01], ** [0.01 > P > 0.001], *** [P < 0.001]. n = 3 ± SD.

The only site with significantly different editing frequency to ADAR1 was AAU, where the mutant ADAR3 had higher frequency than ADAR2 and lower frequency than ADAR1. Although the 5'C sites CA(A/U/C) did not have significant editing over background, at all three of these sites the ADAR3-deaminase mutant had significantly higher frequency than both ADAR1 and ADAR2. Perhaps if additional mutations were found to increase the activity of the ADAR3-deaminase mutant, these sites may show signs of editing.

Comparisons to a previously published ADAR3 mutant

Concurrent with this work, Wang et al. (2019) recently published an ADAR3-deaminase penta-mutant A389V, V485I, E527Q, Q549R, Q733D that rescued editing activity. All mutations tested were to convert ADAR3 residues into the equivalent ADAR2 residues, and were tested on the background of E527Q, which is the ADAR3 equivalent of the ADAR2 E488Q mutation used to increase editing efficiency (Kuttan & Bass, 2012; Phelps et al. 2015). Alongside the A389V mutation, the three other mutations tested were V485I, which anchors the catalytic core glutamate in the active site, Q549R, which interacts with the RNA backbone of the non-edited strand near the orphan base, and Q733D, which sits adjacent to the inositol co-factor.

Initial tests were performed with an ADAR2 / ADAR3 chimera with ADAR3 residues 342-488. The chimera possessed the ADAR3 catalytic core, but the ADAR2 5' binding loop and orphan base binding loops. Wang et al. found that, for the chimera, E527Q was not able to rescue activity alone, but the double mutant A389V / E527Q was able to rescue. This likely means that the role of the E527Q mutation is to improve editing efficiency of already active ADARs but cannot itself rescue activity. Unfortunately, Wang et al. did not test the A389V mutation without the E527Q background in the chimera, so it is unknown if the A389V mutation alone would have been able to rescue activity for that construct.

Moving from the chimera to the full ADAR3 deaminase sequence, Wang et al. was unable to rescue activity with the quadruple mutant A389V, V485I, E527Q, Q733D, but the addition of Q549R to make the penta-mutant was able to generate editing activity. The lack of editing by the quadruple mutant is interesting, considering that **figure 4.11a** above demonstrates a single mutant A389V that is able to rescue activity. Possibly, the 50bp perfectly double stranded RNA substrate used in this work is an easier editing target for the ADAR3-deaminase, and the hairpin substrate

used by Wang et al. required more adaptations in the protein to be able to efficiently edit the site. The hairpin used had a 16bp stem, with the targeted site on a loop, similar to the sites favored by ADAR2. As the ADAR3-deaminase mutant characterized above in **figure 4.11** had similar editing patterns to ADAR1, not ADAR2, perhaps the choice of substrate was the reason for no observed activity by the A389V mutant.

Summary of Chapter IV results

A high throughput sequencing assay has been developed that can measure the editing level for adenosines with all possible combinations of 5' and 3' neighbour, characterizing differences between ADAR proteins and domains. ADAR2 was the most active protein, with 14.6 ± 1.82% overall editing. The low overall level was due to high editing of specific sites and very low editing of other sites.

The ADAR1 isoforms p150 and p110 had similar editing frequencies for each adenosine-containing triplet, indicating that on a 50bp substrate the ZBD provided no additional selectivity. However, the p150 protein had 3-fold more editing than p110, indicating that the ZBD may have assisted in increasing editing levels or stabilizing the protein, rather than having a role in substrate selectivity.

Comparing the ADAR1 and ADAR2 deaminases, it was observed that ADAR1-D favors 5'A and ADAR2-D favors 5'U. From homology modelling of ADAR1 onto the ADAR2 crystal structure, the orphan-base binding loop of ADAR1 did not intercalate as far into the RNA duplex and likely is less able to disrupt the duplex than ADAR2, which intercalates and interacts closely with the

orphan base. As such, it could be that ADAR1-deaminase favors 5'A sites as they are less rigid than 5'U sites, and more easily released into the flipped-out conformation.

Comparing the full-length ADAR1-p110 and ADAR2 proteins to their respective deaminase domains, it was observed that both full length proteins had decreased editing frequency of UAG relative to the deaminases. For site AAG, ADAR1-p110 had increased frequency while ADAR2 had a decrease, and the opposite was seen for CAG, with ADAR2 having increased frequency over the deaminase. This implies that the dsRBDs are providing additional editing site selectivity over the deaminase domains, and that ADAR1-p110 and ADAR2 are targeting different sites.

The deaminase domain of ADAR3 was confirmed to be catalytically inactive, and the single point mutant A389V rescued activity to a low level, similar to the overall editing level of ADAR2. Despite the sequence similarity to ADAR2, the ADAR3-deaminase mutant showed similar triplet frequencies to ADAR1, implying that there is something different in the ADAR3 interaction with substrate RNA. A likely candidate is the 5' RNA binding loop, with only four of 17 residues conserved between the two proteins.

Although the level of editing at 5'C sites by the ADAR3-deaminase mutant was not significant, the mutant did have significantly higher frequency at 5'C sites over both ADAR1 and ADAR2 deaminases. The ADAR3 deaminase could be a candidate for engineering an enzyme for site-directed editing of 5'C sites. Combining the A389V mutant with the E527Q mutant to increase activity further may generate an ADAR protein with 5'C editing capabilities, perhaps building a chimeric protein with ADAR2 dsRBDs, which were shown to increase editing frequency of CA(A/C/G) over the ADAR2-deaminase.

Chapter V – ADAR editing specificity from simple to complex substrates

Alongside characterization of ADAR protein activity *in vitro*, ADAR editing sites have also been identified as A/G mismatches in transcriptomic data. The identification of edited sites in a transcriptome – designated the editome – does not necessarily contain information about how efficiently each site is edited, or the abundance of transcripts where editing is occurring.

To observe if the neighbor preferences characterized in an editome, here using the HEK 293T editome published by Chung et al. (2018), are similar to the neighbor preferences measured *in vitro*, the 5' and 3' neighbor preferences as well as the triplet preferences were calculated and compared to the *in vitro* values measured in Chapter IV. The editing sites in the 293T editome were also corrected for the relative abundance of each triplet in the transcriptome, to observe if editing frequency of each triplet was skewed by the number of possible editing sites available.

One factor that could be involved in observed differences in editing specificity between a simple 50bp dsRNA and the complex editome of a cell is that recognition of RNA by the dsRBDs may not be fully elucidated using a short, perfectly double-stranded segment.

To further explore the role of the dsRBDs in substrate selectivity, more complex pools of RNA were used as substrate for an *in vitro* incubation with ADAR proteins. A pilot experiment, using low concentrations of total RNA from HEK 293T ADAR1/2 KO cells (293T) or reovirus T1L RNA (reoT1L), was performed to identify editing in these longer substrates. HEK 293T total RNA contains a complex pool of RNAs of different lengths and secondary structures, and identification of 5' and 3' neighbour preference differences between ADAR isoforms and deaminase domains may be able to determine if the dsRBDs are having a greater effect on these substrates than on the 50bp dsRNA. Additional characterization of the secondary structures where editing events occur

can provide further insight into the type of structures that ADAR proteins are targeting – adding to work previously done by Thomas and Beal (2017) and Wang et al. (2018), and the degree to which this recognition depends on the dsRBDs and ZBDs.

The use of reoT1L RNA as substrate was also piloted, to develop a method to not only identify editing sites, but to count the relative level of editing at each site. reoT1L has its genome organized into 10 segments of perfectly double-stranded RNA, of lengths varying from 1200 – 4000bp. It has not been shown to be a target of RNA editing by ADAR proteins or affected by upregulation of ADAR1-p150 following interferon induction (Hood et al. 2014). However, reoT1L RNA is a good substrate for characterizing editing specificity due to 5' and 3' neighbors, due to its perfectly double-stranded nature and lack of complex RNA secondary structures. The use of reoT1L RNA as substrate is primarily to develop a version of the *in vitro* editing assay for a longer substrate, as use of the 50bp substrate may not be able to distinguish different editing specificity of the dsRBDs.

Comparing *in vitro* editing frequencies to the 293T editome

Using the list of all editing sites identified in the HEK 293T editome (Chung et al. 2018), editing sites were separated by the identity of the 5' or 3' neighbor of each editing site. **Figure 5.1** shows the relative frequencies of each neighbor across the whole editome, with 5' neighbors shown at left and 3' neighbors shown at right. The 293T editome was generated for ADAR1, identifying sites for p150 and p110 together, as well as identifying ADAR1-p150 edit sites specifically. Each editome was then split into editing sites in Alu elements and the rest of the transcriptome, as Chung et al. found that the majority of editing was occurring in the Alu elements. The 5' and 3' neighbor preferences were also counted for each of the ADAR constructs characterized using the *in vitro* editing assay in Chapter IV.

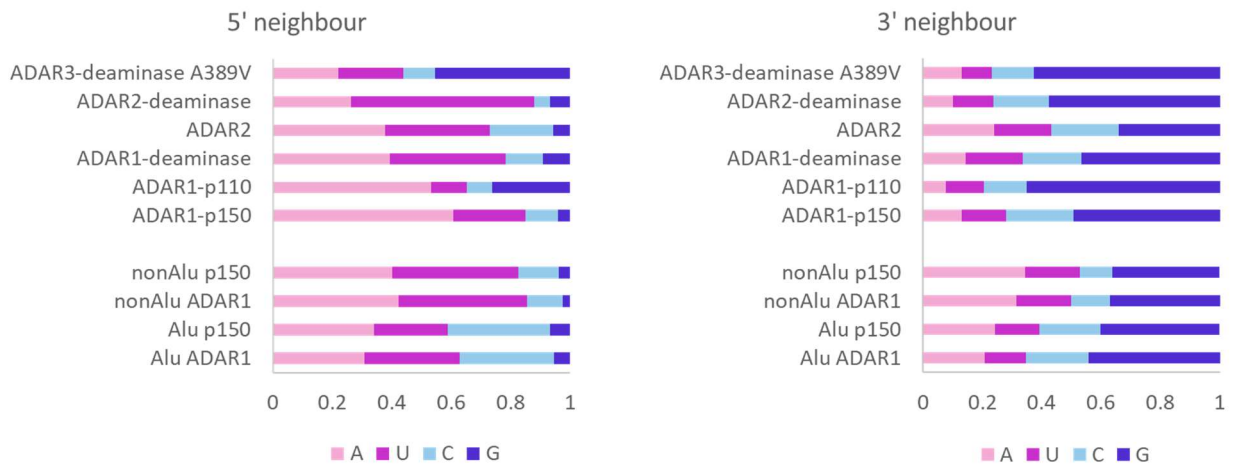

Figure 5.1. 5' (left) and 3' (right) neighbor frequencies for *in vitro* samples ADAR1-p150, ADAR1-p110, ADAR1-deaminase, ADAR2, ADAR2-deaminase, and ADAR3-deaminase A389V, and neighbor frequencies counted from HEK 293T editomes for 'p150' and 'ADAR1' (p110+p150) separated into edits in Alu and nonAlu sequences (Chung et al. 2018).

In **figure 5.1**, the first observation is that the editome samples were all similar to each other, with measurable editing frequency of all neighbors, except 5'G. *in vitro* samples, in comparison, had more variation between the proteins showing distinct characteristics for each protein tested. The editomes were all from ADAR1 editing however, so differences in editing patterns would not be as extreme as those seen for the *in vitro* samples. Indeed, the differences in 5' neighbor were split between the nonAlu and Alu editing sites, with nonAlu having a higher proportion of 5'A, rather than observing any difference in pattern between the ADAR1 (p110+p150) and ADAR1-p150 editomes. The more even spread of neighbor frequency in the editome data sets is likely due to a combination of recognition by dsRBDs and the sheer number of available editing sites in a transcriptome, compared to a 50bp dsRNA with one site for each neighbor combination.

In **figure 5.1**, it can be seen that the majority of samples, *in vitro* and editome, had a low frequency of 5'G, and a high frequency of 5'A/U. ADAR3-deaminase A389V and ADAR1-p110 had a higher frequency of 5'G due to the high background signal at the GAG site. The editome samples had a higher relative amount of 5'C, with Alu elements having higher frequency than nonAlu transcripts, and the editomes all have similar frequency of 5'A and more variable frequency of 5'U. In comparison, the *in vitro* samples had highly variable 5'A frequencies, from ADAR1-p150 with >50% to ADAR3-deaminase A389V with ~20%.

For the 3' neighbors, all samples had high frequency of 3'G. All four of the editome samples had similar frequencies, with the Alu frequencies skewing slightly to 3'C and nonAlu frequencies skewing the 3'A, with 3'U remaining somewhat constant. The *in vitro* samples followed a similar pattern, but with lower 3'A frequencies, especially for ADAR1-p110. The most similar *in vitro* sample to the editomes was ADAR2, which showed significant editing activity at the most sites in the 50bp dsRNA substrate, most likely due to ADAR2 being the most active purified protein.

The 5' and 3' neighbor counts were then combined to calculate the relative frequency of each adenosine triplet. The *in vitro* triplet frequencies had been previously calculated, shown in figure **4.2b**, and those values are also shown below in **table 5.1**. The triplet frequencies for each of the editome data sets were counted and are also shown in **table 5.1**.

The frequency of each triplet was highlighted for relative abundance, and it can be seen in **table 5.1**, that all of the highly edited adenosine-containing triplets – in red – grouped together at the 5'A and 5'U triplets, with low frequency sites – in blue – including the 5'G and most of the 5'C sites. The noisy GAG signal could be seen for ADAR1-p110 and ADAR3-deaminase A389V, with high frequency apparent at that editing site. Overall, there was a similar pattern of 5' and 3' neighbour preference for editing for all samples, with the nonAlu editomes following a similar

pattern to the *in vitro* ADAR1-p150, ADAR2 and deaminase patterns. The AAG editing site in ADAR1-p150 and -p110, and the UAG editing site in ADAR1-and ADAR2-deaminases were highly edited, as observed previously, and none of the editome frequencies were as extreme as the *in vitro* frequencies. The substrate selectivity of the proteins in isolation is not the only determinant in choosing an editing target, as the frequencies of editing do change between the *in vitro* and editome data.

Table 5.1. Relative triplet frequencies for *in vitro* samples ADAR1-p150, ADAR1-p110, ADAR1-deaminase, ADAR2, ADAR2-deaminase, and ADAR3-deaminase A389V (values from figure 4.2b), and triplet frequencies from HEK 293T editomes for 'p150' and 'ADAR1' (p110+p150) separated into edits in Alu and nonAlu sequences (Chung et al. 2018). Triplets with high editing frequency are labelled in red, and triplets with low editing frequency are labelled in blue.

	ADAR1 P150	ADAR1 P110	ADAR1 deam	ADAR2	ADAR2 deam	ADAR3 deam A389V	ADAR1 Alu	p150 Alu	ADAR1 nonAlu	p150 nonAlu
AAA	0.02	0.01	0.03	0.05	0.01	0.04	0.07	0.09	0.13	0.14
AAU	0.09	0.08	0.12	0.11	0.03	0.04	0.08	0.08	0.08	0.08
AAC	0.16	0.10	0.11	0.13	0.10	0.07	0.04	0.04	0.06	0.04
AAG	0.35	0.34	0.14	0.09	0.13	0.07	0.13	0.13	0.15	0.14
UAA	0.07	0.03	0.08	0.10	0.07	0.05	0.07	0.07	0.15	0.14
UAU	0.05	0.03	0.06	0.07	0.10	0.04	0.03	0.03	0.09	0.10
UAC	0.06	0.02	0.06	0.06	0.08	0.03	0.09	0.08	0.04	0.04
UAG	0.07	0.03	0.18	0.12	0.37	0.10	0.14	0.08	0.15	0.15
CAA	0.04	0.03	0.03	0.06	0.01	0.03	0.07	0.08	0.03	0.05
CAU	0.01	0.01	0.01	0.01	0.01	0.01	0.03	0.03	0.01	0.01
CAC	0.01	0.01	0.01	0.03	0.01	0.03	0.08	0.09	0.02	0.02
CAG	0.06	0.05	0.08	0.10	0.03	0.04	0.14	0.15	0.05	0.05
GAA	0.01	0.01	0.01	0.03	0.01	0.01	0.01	0.01	0.01	0.01
GAU	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
GAC	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01
GAG	0.02	0.23	0.06	0.03	0.05	0.42	0.04	0.04	0.02	0.02

Editing frequencies in Alu elements are interesting as, when compared to the nonAlu editing sites as well as the *in vitro* samples, there was more editing occurring at 5'C sites. The low frequency at CAU however, combined with the low frequency of UAU, implies that 3'U was not favorable for editing in Alu elements, possibly due to the triplets being in RNA structures not amenable for

editing, or just having those triplets at low abundance. The p150-specific editome also had the lowest frequency of UAG editing, compared to the three other editome data sets, and most of the *in vitro* proteins, except for ADAR1-p150 and p110. The absence of UAG editing in the Alu editome could be due to low abundance of the UAG triplet in Alu sequences, but the low frequency in the *in vitro* ADAR1-p150 was due to inherent behavior of the protein, which implies that ADAR1-p150 is not targeting UAG for editing.

To determine if the triplet editing frequencies in the editome data was strongly affected by abundance of each triplet, the abundance of triplets was calculated across the entire transcriptome, as well as abundance in active Alu elements. The frequencies are shown in **table 5.2a**, and it can be seen that UAG is the least common adenosine triplet across the transcriptome and is also rare in Alu elements. 5'U is the least common neighbor in the transcriptome, with AAA and CAG the most common triplets. AAA is also the most common triplet in Alu sequences, due to the poly-A tail, and GAG is also highly frequent. Alu elements, being a specific subset of the transcriptome, have a more extreme spread of relative triplet abundance, and 5'U/C and 3'U/C sites are all rare in these sequences.

For each triplet in the editome data sets, the expected editing level was set based on the frequency of the triplet. Measured frequencies were then converted to fold change of observed editing over the expected value, with positive values indicating that editing was occurring more than expected for the number of editing sites available, and negative values indicating lower than expected. **Table 5.2b** shows the fold change calculated for each editome data set, with positive values in red and negative values in blue.

Table 5.2. Relative frequency of each adenosine triplet in Alu elements (Konkel et al. 2015) and in the whole transcriptome (a). Triplets with high abundance are labelled in red, and triplets with low abundance are labelled in blue. Fold change of observed editing over expected, relative to the frequency of each triplet in either Alu elements or the whole transcriptome (b). Positive fold-change is labelled in red, negative fold-change is labelled in blue.

a) Frequency of triplets

	Frequency in Alu	Frequency in transcriptome
AAA	0.136	0.096
AAU	0.082	0.059
AAC	0.051	0.051
AAG	0.056	0.082
UAA	0.069	0.044
UAU	0.042	0.045
UAC	0.041	0.037
UAG	0.043	0.030
CAA	0.054	0.066
CAU	0.046	0.060
CAC	0.050	0.062
CAG	0.077	0.102
GAA	0.067	0.082
GAU	0.043	0.052
GAC	0.035	0.052
GAG	0.110	0.083

b) Fold change relative to expected frequency

ADAR1 Alu	p150 Alu	ADAR1 nonAlu	p150 nonAlu
-1.04	-0.61	0.42	0.58
-0.12	0.01	0.41	0.41
-0.29	-0.36	0.33	-0.24
1.18	1.22	0.91	0.75
-0.05	-0.05	1.73	1.67
-0.39	-0.63	1.03	1.01
1.05	0.89	0.29	0.23
1.70	0.91	2.32	2.34
0.33	0.58	-0.91	-0.43
-0.84	-0.68	-2.59	-2.32
0.68	0.79	-1.70	-1.36
0.90	0.95	-0.91	-0.96
-3.06	-3.25	-3.77	-2.77
-2.63	-1.63	-4.69	-4.10
-3.56	-3.14	-5.70	n/a
-1.66	-1.36	-2.46	-1.98

It can be seen in **table 5.2b** that even though the Alu editomes had high frequency of editing at most sites except UAU, CAU and GAN (**table 5.1**), the 5'A editing sites were not edited above expected, due to the high frequency of 5'A sites in Alu elements. On the contrary, the AAG, UA(C/G) and CA(C/G) sites were being edited above the expected frequency. ADAR1 appeared to be targeting these sites for editing in Alu elements, whereas editing of 5'A sites was mostly due to the massive number of sites available, meaning some of them were edited even if the protein was not as specific for those sites. The over-abundance of editing at UAG was also seen in the nonAlu editing sites, with UAA and UAG especially overrepresented. CAG, despite having a similar frequency of editing to UAG, was not edited above the expected level, due to CAG being four times more common in the transcriptome than UAG.

In vitro editing of complex substrates

To bridge the gap between ADAR editing against a simple 50bp RNA and the editing seen in editomes, a number of more complex RNAs were trialed as *in vitro* editing substrates.

Counting ADAR editing events in reovirus T1L dsRNA

After characterization of ADAR activity against a perfectly double-stranded 50bp RNA, reoT1L RNA was chosen as a potential editing substrate, as it has a perfectly double-stranded genome. The reoT1L genome is organized into 10 segments of dsRNA that range from ~1500bp for small segments S1, S2, S3 and S4, to ~2500bp for medium segments M1, M2 and M3, and ~4000bp for large segments L1, L2 and L3 (Hood et al. 2014).

The initial pilot for measuring editing in reoT1L RNA was performed to not only identify editing sites, but to be able to count the number of editing events at each specific site so that editing efficiency can be compared across sites. This was achieved by preparing RNAseq libraries using a unique molecular identifier (UMI) linked to the random hexamers used for reverse transcription. When sequenced and aligned, each RT event can be identified by a different UMI sequence, with identical sequences indicative of PCR amplification, and collapsed into single reads.

reoT1L RNA was isolated at a low concentration – 10ng/µl – much lower than the concentration used of the synthetic 50bp dsRNA in the *in vitro* assay. For both the synthetic 50bp dsRNA and the reoT1L RNA, protein and RNA were mixed at a 1:1 molar ratio, but as the reoT1L was a much lower concentration than the 50bp RNA, the concentration of the protein for the reoT1L reaction was also lower than for the synthetic RNA. The reoT1L was incubated at a 1 protein : 50bp RNA molar ratio of 0.25nM : 0.25nM. This is in contrast to the 50bp dsRNA, which was incubated with ADAR proteins at a 1:1 molar ratio of 100nM : 100nM. Each purified protein – ADAR1-p150,

ADAR1-p110, ADAR1-deaminase, ADAR2, and ADAR2-deaminase – was incubated with reoT1L RNA for 1hr at 37°C before libraries were generated to identify editing sites.

Due to the low concentrations tested, a small number of editing events were identified, primarily edited by ADAR1-p150. At each of the editing sites identified, the presence of a UMI tag allowed individual editing events to be counted, to be able to measure efficiency of editing per site. Identification of editing sites was stringent, with mismatches needing to be present for at least two samples and not present in any of the negative controls. Due to the low level of editing in the pilot experiment, all editing sites identified were combined into a single list, incorporating sites edited by ADAR1-p150, ADAR1-p110 and the two deaminase domains from ADAR1 and ADAR2. The full-length ADAR2 failed to generate any identifiable editing events in this data set. **Table 5.3** shows the editing sites identified, across the 10 reovirus segments, detailing the site of the identified edit, the sequence context and the number of edits counted for that site.

Table 5.3. A-to-I editing events identified in reovirus T1L RNA. The reovirus segment and position is listed for each site, as well as whether the editing event was found in the positive or negative strand of the RNA segment. The triplet sequence surrounding the adenosine is also noted, as is the number of unique edits seen at each site.

Reovirus segment	Position in segment [bp]	Positive or negative strand	Editing site triplet	No. editing events counted
L3	1391	-	GAG	26
M2	429	-	GAC	24
M3	1982	-	CAC	29
S1	604	+	UAG	19
S3	640	+	AAG	22
S3	974	+	AAG	19
S4	948	-	CAC	69

The small number of editing events identified, due to low concentration of RNA and protein during the editing reaction, prevents any conclusion from this data. The main takeaway from this pilot experiment is that ADAR proteins are capable of editing reoT1L dsRNA, and that the use of UMI-linked hexamers to label allowed for identification of identification of individual editing events.

Measuring *in vitro* editing frequencies on HEK 293T substrate RNA

To better characterize the role of the dsRBDs in substrate selection, 293T total RNA was used as substrate for editing *in vitro* by ADAR constructs. Each purified protein – ADAR1-p150, ADAR1-p110, ADAR1-deaminase, ADAR2, and ADAR2-deaminase – was incubated at a 1 protein : 50bp RNA molar ratio for 1hr at 37°C. This is the same 1:1 ratio as the *in vitro* 50bp assay, but the 293T RNA was isolated to a concentration of 1μg/ul, so while the 50bp dsRNA was incubated at a ratio of 100nM : 100nM, the 293T RNA was incubated with ADARs at a ratio of 30nM : 30nM.

After generating libraries for each 293T sample after incubation with an ADAR protein, analysis did show that the total number of editing events was much lower than the *in vitro* samples, as expected due to the lower concentrations. Along with the low total editing measured, the majority of editing sites were clustered in Alu elements, with much lower frequency of editing in non-Alu sequences. This is shown in **figure 5.2a**, that has counts of all nucleotide conversion and illustrates that A/G and T/C conversions are more prevalent than other mismatches, indicating A-to-I edits, and those mismatches are mainly in Alu elements.

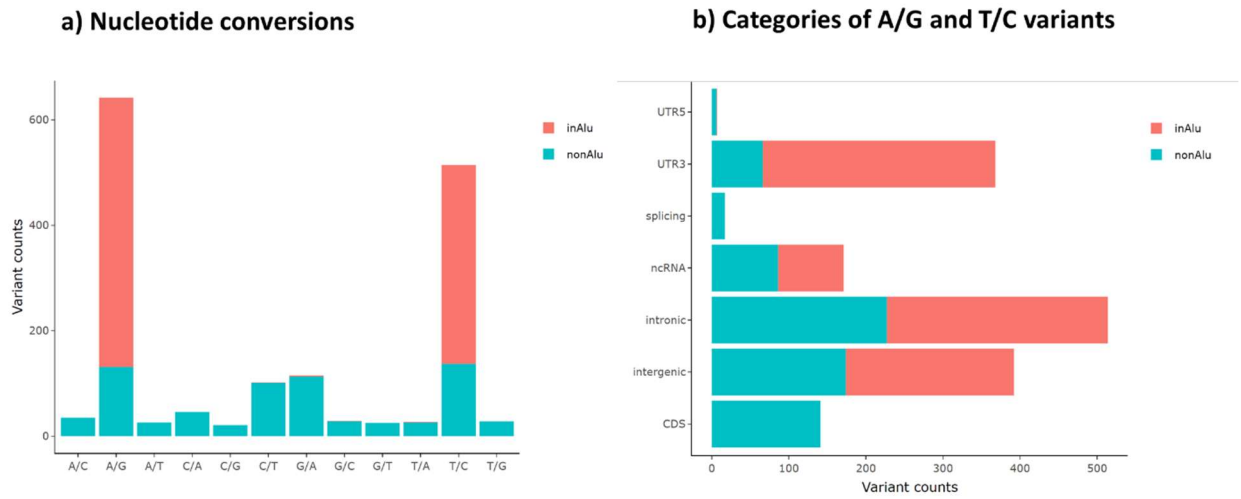

Figure 5.2. All nucleotide variants identified in 293T transcripts, following incubation of 293T RNA *in vitro* with purified protein ADAR1-p150, ADAR1-p110, ADAR1-deaminase, ADAR2, ADAR2-deaminase, or ADAR3-deaminase A389V (a), with A/G and T/C variants indicating editing sites located in Alu or nonAlu sequences. Distribution of A/G and T/C variants across transcript categories (b), with majority of editing events in 3'UTR, introns and intergenic regions.

For the A/G and T/C mismatches identified, **figure 5.2b** shows the regions of the transcriptome where these sites are located, with intronic and intergenic having the highest number of identified editing sites, both in Alu and nonAlu sequences, and a high number of editing sites in Alu elements contained in 3'UTRs. For all of the editing sites identified in each protein condition, the 5' and 3' neighbors were counted and a plot of relative frequency of each neighbor was generated, shown in **figure 5.3**, with frequencies counted for editing of 293T *in vitro* by proteins ADAR1-p150, ADAR1-p110, ADAR1-deaminase, ADAR2, and ADAR2-deaminase, as well as the frequencies found for endogenous edits in WT 293T.

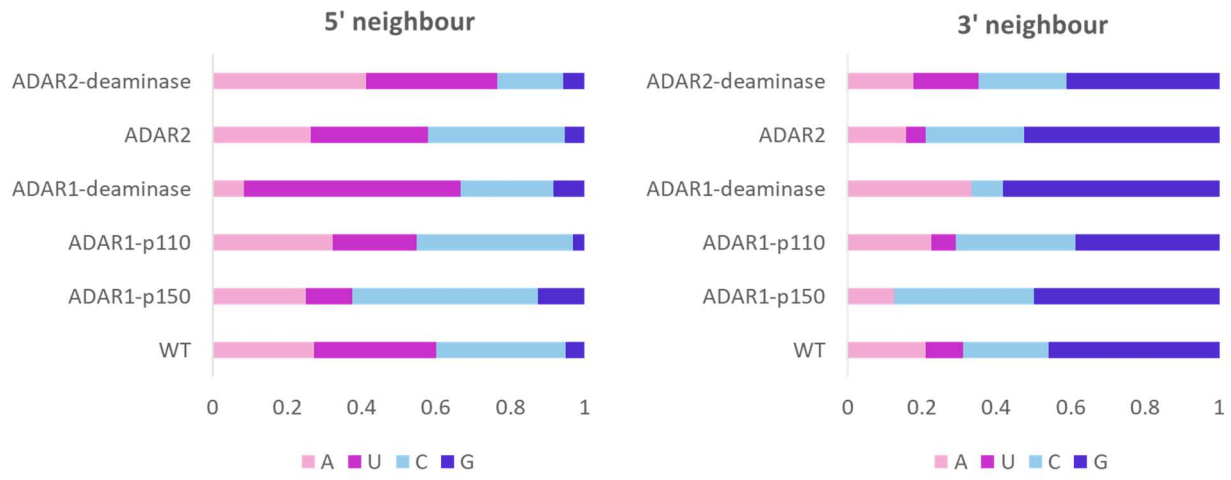

Figure 5.3. 5' (left) and 3' (right) neighbor frequencies for editing sites in 293T RNA, following incubation of ADAR1/2$^{-/-}$ HEK 293T RNA *in vitro* with purified proteins ADAR1-p150, ADAR1-p110, ADAR1-deaminase, ADAR2, ADAR2-deaminase, and ADAR3-deaminase A389V. The typical editing frequencies for endogenous ADAR in HEK 293T WT is shown as 'WT'.

Figure 5.3 shows that for editing of 293T total RNA by ADAR proteins *in vitro*, there is more variation in the 5' neighbor between proteins than there is for the 3' neighbor and, like the *in vitro* assay using the 50bp substrate, the different ADAR constructs do show distinct patterns of editing. The 3' neighbor consistently has ~40-50% 3'G, low frequency of 3'U and variable 3'A and 3'C. The ADAR2-deaminase has the most similar profile to endogenous editing in WT 293T cells, while ADAR1-p150 and ADAR1-deaminase are the least similar, with no observed 3'U at any editing site.

For the 5' neighbor, all samples have low frequency of 5'G, as seen previously for the *in vitro* editing specificity of the proteins, and somewhat equal frequencies of 5'A, 5'U and 5'C. Again, the least similar samples are ADAR1-p150, with higher 5'C frequency and lower 5'U, and

ADAR1-deaminase, with lower 5'C and higher 5'A than the other ADAR constructs. It still needs to be seen if these patterns will be consistent when a higher level of editing is achieved, using higher concentrations of ADAR proteins and RNA.

Comparing *in vitro* editing frequencies from 293T to synthetic 50bp dsRNA

For each ADAR construct tested, we now have well characterized 5' and 3' neighbor preferences against a 50bp dsRNA and initial preferences against total RNA from 293T cells. An example of how the editing preferences of each ADAR construct can be compared is shown in **figure 5.4**, with comparisons of both 5' neighbor and 3' neighbor between samples 'in vitro' (the 50bp dsRNA) and '293T' (total RNA from HEK 293T). As the 293T samples are preliminary, no in-depth analysis of differences can be performed until further samples are generated with higher levels of editing achieved. Even with a much lower overall number of editing events identified, and much more editing in Alus versus nonAlus compared to WT, the neighbor frequencies observed are not that different to the *in vitro* determined frequencies.

Future developments for complex RNA editing assay

Having observed that editing events can be identified in RNAseq libraries generated from 293T or reoT1L RNA, the next step is to concentrate the reoT1L and 293T RNA 20-fold, to be able to incubate protein and RNA using the same conditions used for the 50bp substrate: 100nM : 100nM. For the reoT1L substrate, a sequencing analysis pipeline has been developed to identify individual editing events marked by UMIs, enabling editing efficiency at each adenosine to be measured. Generating libraries with higher levels of editing, the total count of editing for each triplet can be calculated and the relative editing frequency for substrates 1200-4000bp can be compared to the previously determined frequencies for a substrate of 50bp.

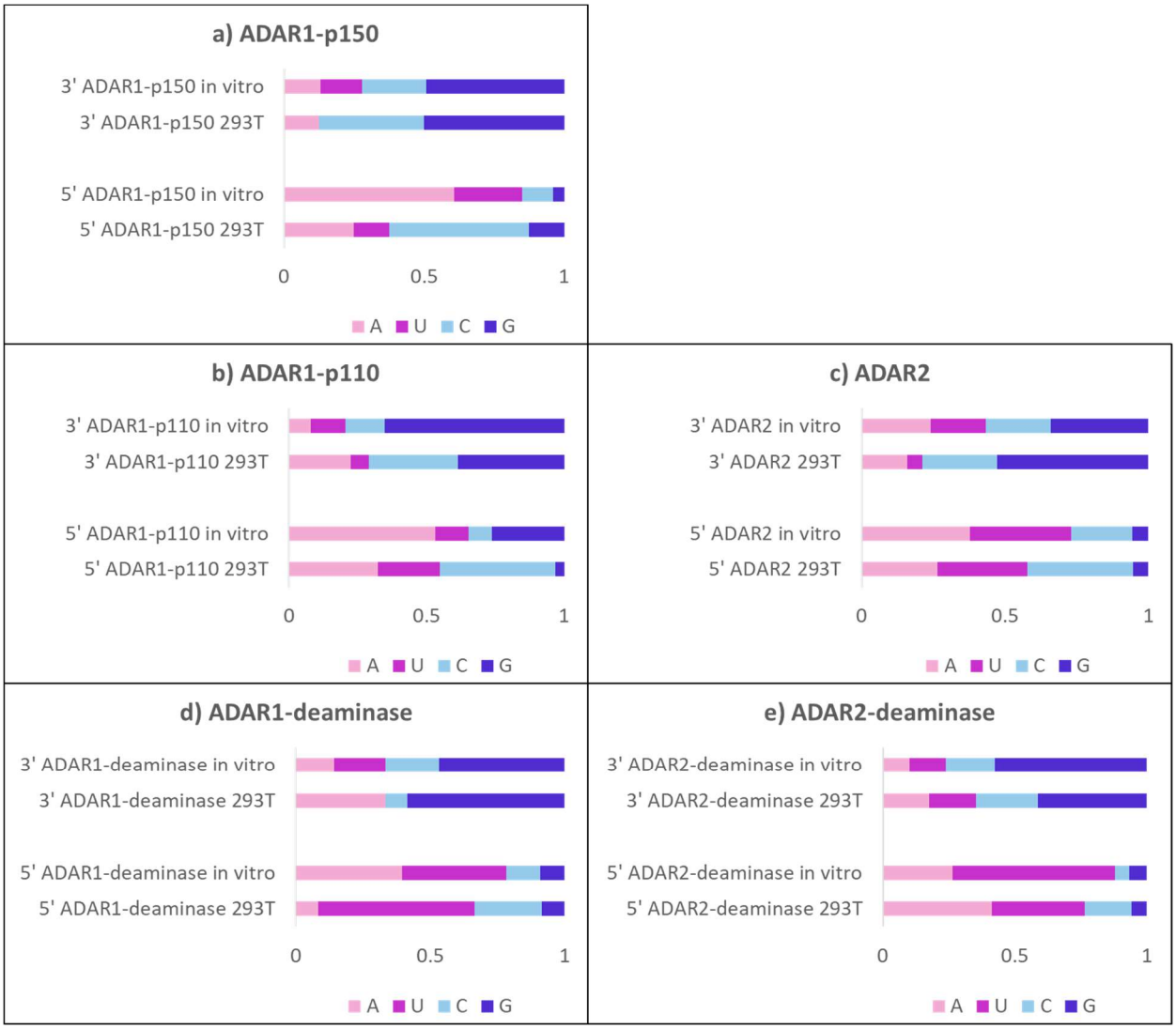

Figure 5.4. 5' and 3' neighbor frequencies for editing sites in 293T RNA ('293T') and in synthetic 50bp dsRNA ('in vitro'), following incubation of each RNA substrate *in vitro* with purified proteins ADAR1-p150 (a), ADAR1-p110 (b), ADAR2 (c), ADAR1-deaminase (d), and ADAR2-deaminase (e).

Specifically, editing of reoT1L can be compared to the short substrate to observe if the frequency of editing for the full-length ADAR1-p110 and ADAR2 change in relation to the deaminase constructs, to determine if dsRBDs are playing more of a role in substrate selection with a longer substrate. Following the observation using the 50bp substrate that ADAR1-p110 increases AAG editing and ADAR2 increases CAG editing, the increase or decrease of editing of these sites can be observed over the many adenosines present in the reoT1L genome segments.

For the HEK 293T substrate RNA, generating samples with increased levels of editing can similarly be used to observe relative editing frequency of the adenosine-containing triplets, and compare the patterns to the shorter 50bp substrate. Further to this, once editing sites have been identified in 293T transcripts, the RNA secondary structures can be surveyed to determine if patterns emerge for site selection by ADAR1-p110, ADAR2 and the two deaminases. Comparisons of editing of specific RNA secondary structures can then be made to work by Thomas and Beal (2017) and Wang et al. (2018), which found that ADAR2-deaminase favors perfect dsRNA 5' of the site and/or a loop immediately downstream, ADAR1-p110 and ADAR1-deaminase favor a hairpin or loop 3-6bp 5' of the editing site, and ADAR2 favors sites with 2-3 mismatches at least 10bp downstream of the site.

Concluding remarks

Characterising the editing specificity of ADAR proteins using simple or complex substrates, or using editome data, will produce different patterns of behavior.

Identification of editing sites in editomes takes into account the sites that ADARs target in vivo, with complex RNA structures and adenosine-containing triplets of different abundances. This provides information on how ADARs are functioning in cells and, combined with work identifying the specific RNA structures that ADARs target, the endogenous behavior of ADAR proteins can be characterized.

The inherent editing specificity of the ADAR proteins in the absence of complicated structures or sequences is characterized in this thesis. The activity of each ADAR protein, and selection of substrates at the protein:RNA interacting level, is one component of the selection described above for complex pools of RNA.

Characterising ADAR deaminases in vitro found that ADAR1-deaminase skews slightly to favoring 5'A while ADAR2-deaminase skews to 5'U, and this difference could be due to the interaction between the orphan-base binding loop of each protein with the RNA substrate. The ADAR1-deaminase loop does not appear to interact as strongly with the substrate, and so would require the duplex RNA to be less stable so that the targeted adenosine can be flipped out of the structure. A 5'A would not form cross-stacking interactions, while a 5'U would, and so ADAR1-deaminase may favor 5'A as it forms a less stable editing substrate.

The characterization of the full-length ADAR proteins ADAR-p110 and ADAR2 found that the when compared to the patterns of editing by the deaminases, UAG was favored by the deaminases over the full-length proteins, AAG editing increased for ADAR1-p110 relative to its deaminase and decreased for ADAR2, and CAG showed the opposite phenotype: increased editing by ADAR2 and decreased editing by ADAR1-p110. This indicates that the dsRBDs are playing a role in increasing selectivity of specific sites, with ADAR1-p110 skewing to AAG and ADAR2 skewing to CAG.

Characterization of ADAR1-p150 found that, on a substrate of this length, there was no significant difference in editing to ADAR1-p110. To be able to better characterize the role of the dsRBDs in substrate selectivity, longer and more complex substrates are required.

Initial tests of 293T and reoT1L RNA demonstrated that editing events can be identified and, in the case of reoT1L, individual editing events can be counted by incorporation of UMIs during library generation. Once the assay conditions have been improved – increasing protein and substrate concentrations to generate more editing events – the reoT1L can be used to characterize relative editing frequencies, to be compared to those frequencies determined for the 50bp substrate. Identification of editing sites in 293T RNA will hopefully identify RNA secondary structures where editing is occurring, and if the ADAR1 and ADAR2 proteins are targeting different structures. Comparisons to the deaminase-domain constructs will further investigate the role of dsRBDs in selectivity of editing sites.

The inactivity of the ADAR3-deaminase was also probed, with the deaminase confirmed to lack catalytic activity and the point mutant A389V found to rescue activity. Interestingly, although ADAR3 is more similar by sequence to ADAR2, the editing pattern was more similar to ADAR1.

Characterising the editing specificity of deaminase domains can help inform protein engineering for site-directed RNA editing, which is currently progressing in two styles: editing by endogenous protein targeted to the chosen site with guide RNAs and use of exogenous ADAR proteins with guide RNAs. Use of exogenous ADAR has the potential to involve engineering of ADAR deaminases with improved specificity of editing against specific sites, such as the difficult-to-edit 5'C, 5'G and 3'U. Observing that the rescued ADAR3 mutant behaves more like ADAR1 than ADAR2 implies that something in the structure of that domain is determining editing specificity, which could be the 5' RNA binding loop. Further investigations into residues important for selecting adenosines for editing is required to be able to generate ADAR proteins with more variation in editing specificity.

Bibliography

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A. and Blankenberg, D. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research* 46(W1), pp. W537–W544.

Alon, S., Garrett, S.C., Levanon, E.Y., Olson, S., Graveley, B.R., Rosenthal, J.J.C. and Eisenberg, E. 2015. The majority of transcripts in the squid nervous system are extensively recoded by A-to-I RNA editing. *eLife* 4.

Athanasiadis, A., Placido, D., Maas, S., Brown, B.A., Lowenhaupt, K. and Rich, A. 2005. The crystal structure of the Zbeta domain of the RNA-editing enzyme ADAR1 reveals distinct conserved surfaces among Z-domains. *Journal of Molecular Biology* 351(3), pp. 496–507.

Athanasiadis, A., Rich, A. and Maas, S. 2004. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biology* 2(12), p. e391.

Avesson, L. and Barry, G. 2014. The emerging role of RNA and DNA editing in cancer. *Biochimica et Biophysica Acta* 1845(2), pp. 308–316.

Bahn, J.H., Lee, J.-H., Li, G., Greer, C., Peng, G. and Xiao, X. 2012. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Research* 22(1), pp. 142–150.

Bakhtiarizadeh, M.R., Salehi, A. and Rivera, R.M. 2018. Genome-wide identification and analysis of A-to-I RNA editing events in bovine by transcriptome sequencing. *Plos One* 13(2), p. e0193316.

Barraud, P. and Allain, F.H.-T. 2012. ADAR proteins: double-stranded RNA and Z-DNA binding domains. *Current Topics in Microbiology and Immunology* 353, pp. 35–60.

Barraud, P., Banerjee, S., Mohamed, W.I., Jantsch, M.F. and Allain, F.H.-T. 2014. A bimodular nuclear localization signal assembled via an extended double-stranded RNA-binding domain acts as an RNA-sensing signal for transportin 1. *Proceedings of the National Academy of Sciences of the United States of America* 111(18), pp. E1852-61.

Barraud, P., Heale, B.S.E., O'Connell, M.A. and Allain, F.H.-T. 2012. Solution structure of the N-terminal dsRBD of Drosophila ADAR and interaction studies with RNA. *Biochimie* 94(7), pp. 1499–1509.

Bass, B.L. 2002. RNA editing by adenosine deaminases that act on RNA. *Annual Review of Biochemistry* 71, pp. 817–846.

Bass, B.L., K. Nishikura, W. Keller, P.H. Seeburg, R.B. Emeson, M.A. O'Connell, C.E. Samuel, and A. Herbert. 1997. A standardized nomenclature for adenosine deaminases that act on RNA. *RNA* 3(9), pp. 947–949.

Bass, B.L. and Weintraub, H. 1987. A developmentally regulated activity that unwinds RNA duplexes. *Cell* 48(4), pp. 607–613.

Bass, B.L. and Weintraub, H. 1988. An unwinding activity that covalently modifies its double-stranded RNA substrate. *Cell* 55(6), pp. 1089–1098.

Benne, R., Van den Burg, J., Brakenhoff, J.P., Sloof, P., Van Boom, J.H. and Tromp, M.C. 1986. Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* 46(6), pp. 819–826.

Betts, L., Xiang, S., Short, S.A., Wolfenden, R. and Carter, C.W. 1994. Cytidine deaminase. The 2.3 A crystal structure of an enzyme: transition-state analog complex. *Journal of Molecular Biology* 235(2), pp. 635–656.

Brennicke, A., Marchfelder, A. and Binder, S. 1999. RNA editing. *FEMS Microbiology Reviews* 23(3), pp. 297–316.

Burns, C.M., Chu, H., Rueter, S.M., Hutchinson, L.K., Canton, H., Sanders-Bush, E. and Emeson, R.B. 1997. Regulation of serotonin-2C receptor G-protein coupling by RNA editing. *Nature* 387(6630), pp. 303–308.

Casey, J.L. 2006. RNA editing in hepatitis delta virus. *Current Topics in Microbiology and Immunology* 307, pp. 67–89.

Chang, K.-Y. and Ramos, A. 2005. The double-stranded RNA-binding motif, a versatile macromolecular docking platform. *The FEBS Journal* 272(9), pp. 2109–2117.

Chen, C.X., Cho, D.S., Wang, Q., Lai, F., Carter, K.C. and Nishikura, K. 2000. A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains. *RNA (New York)* 6(5), pp. 755–767.

Cheng, X., Kumar, S., Posfai, J., Pflugrath, J.W. and Roberts, R.J. 1993. Crystal structure of the HhaI DNA methyltransferase complexed with S-adenosyl-L-methionine. *Cell* 74(2), pp. 299–307.

Chilibeck, K.A., Wu, T., Liang, C., Schellenberg, M.J., Gesner, E.M., Lynch, J.M. and MacMillan, A.M. 2006. FRET analysis of in vivo dimerization by RNA-editing enzymes. *The Journal of Biological Chemistry* 281(24), pp. 16530–16535.

Cho, D.-S.C., Yang, W., Lee, J.T., Shiekhattar, R., Murray, J.M. and Nishikura, K. 2003. Requirement of dimerization for RNA editing activity of adenosine deaminases acting on RNA. *The Journal of Biological Chemistry* 278(19), pp. 17093–17102.

Chung, H., Calis, J.J.A., Wu, X., Sun, T., Yu, Y., Sarbanes, S.L., Dao Thi, V.L., Shilvock, A.R., Hoffmann, H.-H., Rosenberg, B.R. and Rice, C.M. 2018. Human ADAR1 Prevents Endogenous RNA from Triggering Translational Shutdown. *Cell* 172(4), pp. 811-824.e14.

Colizzi, F., Perez-Gonzalez, C., Fritzen, R., Levy, Y., White, M.F., Penedo, J.C. and Bussi, G. 2019. Asymmetric base-pair opening drives helicase unwinding dynamics. *Proceedings of the National Academy of Sciences of the United States of America*.

Conticello, S.G., Thomas, C.J.F., Petersen-Mahrt, S.K. and Neuberger, M.S. 2005. Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. *Molecular Biology and Evolution* 22(2), pp. 367–377.

Cox, D.B.T., Gootenberg, J.S., Abudayyeh, O.O., Franklin, B., Kellner, M.J., Joung, J. and Zhang, F. 2017. RNA editing with CRISPR-Cas13. *Science* 358(6366), pp. 1019–1027.

Daniel, C., Silberberg, G., Behm, M. and Öhman, M. 2014. Alu elements shape the primate transcriptome by cis-regulation of RNA editing. *Genome Biology* 15(2), p. R28.

Eckmann, C.R., Neunteufl, A., Pfaffstetter, L. and Jantsch, M.F. 2001. The human but not the Xenopus RNA-editing enzyme ADAR1 has an atypical nuclear localization signal and displays the characteristics of a shuttling protein. *Molecular Biology of the Cell* 12(7), pp. 1911–1924.

Eggington, J.M., Greene, T. and Bass, B.L. 2011. Predicting sites of ADAR editing in double-stranded RNA. *Nature Communications* 2, p. 319.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J.A., Dutta, A., et al., et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146), pp. 799–816.

Garncarz, W., Tariq, A., Handl, C., Pusch, O. and Jantsch, M.F. 2013. A high-throughput screen to identify enhancers of ADAR-mediated RNA-editing. *RNA Biology* 10(2), pp. 192–204.

George, C.X. and Samuel, C.E. 1999. 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. *Proceedings of the National Academy of Sciences of the United States of America* 96(8), pp. 4621–4626.

George, C.X., Wagner, M.V. and Samuel, C.E. 2005. Expression of interferon-inducible RNA adenosine deaminase ADAR1 during pathogen infection and mouse embryo development involves tissue-selective promoter utilization and alternative splicing. *The Journal of Biological Chemistry* 280(15), pp. 15020–15028.

Gerber, A.P. and Keller, W. 2001. RNA editing by base deamination: more enzymes, more targets, new mysteries. *Trends in Biochemical Sciences* 26(6), pp. 376–384.

Grice, L.F. and Degnan, B.M. 2015. The origin of the ADAR gene family and animal RNA editing. *BMC Evolutionary Biology* 15, p. 4.

Ha, S.C., Choi, J., Hwang, H.-Y., Rich, A., Kim, Y.-G. and Kim, K.K. 2009. The structures of non-CG-repeat Z-DNAs co-crystallized with the Z-DNA-binding domain, hZ alpha(ADAR1). *Nucleic Acids Research* 37(2), pp. 629–637.

Herbert, A., Alfken, J., Kim, Y.G., Mian, I.S., Nishikura, K. and Rich, A. 1997. A Z-DNA binding domain present in the human editing enzyme, double-stranded RNA adenosine deaminase. *Proceedings of the National Academy of Sciences of the United States of America* 94(16), pp. 8421–8426.

Herbert, A. and Rich, A. 2001. The role of binding domains for dsRNA and Z-DNA in the in vivo editing of minimal substrates by ADAR1. *Proceedings of the National Academy of Sciences of the United States of America* 98(21), pp. 12132–12137.

Hogg, M., Paro, S., Keegan, L.P. and O'Connell, M.A. 2011. RNA editing by mammalian ADARs. *Advances in genetics* 73, pp. 87–120.

Hood, J.L., Morabito, M.V., Martinez, C.R., Gilbert, J.A., Ferrick, E.A., Ayers, G.D., Chappell, J.D., Dermody, T.S. and Emeson, R.B. 2014. Reovirus-mediated induction of ADAR1 (p150) minimally alters RNA editing patterns in discrete brain regions. *Molecular and Cellular Neurosciences* 61, pp. 97–109.

Hoopengardner, B., Bhalla, T., Staber, C. and Reenan, R. 2003. Nervous system targets of RNA editing identified by comparative genomics. *Science* 301(5634), pp. 832–836.

Hough, R.F. and Bass, B.L. 1997. Analysis of Xenopus dsRNA adenosine deaminase cDNAs reveals similarities to DNA methyltransferases. *RNA (New York)* 3(4), pp. 356–370.

Jain, M., Jantsch, M.F. and Licht, K. 2019. The editor's I on disease development. *Trends in Genetics*.

Jin, Y., Zhang, W. and Li, Q. 2009. Origins and evolution of ADAR-mediated RNA editing. *IUBMB Life* 61(6), pp. 572–578.

Källman, A.M., Sahlin, M. and Ohman, M. 2003. ADAR2 A-->I editing: site selectivity and editing efficiency are separate events. *Nucleic Acids Research* 31(16), pp. 4874–4881.

Keegan, L.P., Gallo, A. and O'Connell, M.A. 2001. The many roles of an RNA editor. *Nature Reviews. Genetics* 2(11), pp. 869–878.

Keegan, L.P., Leroy, A., Sproul, D. and O'Connell, M.A. 2004. Adenosine deaminases acting on RNA (ADARs): RNA-editing enzymes. *Genome Biology* 5(2), p. 209.

Keegan, L.P., McGurk, L., Palavicini, J.P., Brindle, J., Paro, S., Li, X., Rosenthal, J.J.C. and O'Connell, M.A. 2011. Functional conservation in human and Drosophila of Metazoan ADAR2 involved in RNA editing: loss of ADAR1 in insects. *Nucleic Acids Research* 39(16), pp. 7249–7262.

Kim, D.D.Y., Kim, T.T.Y., Walsh, T., Kobayashi, Y., Matise, T.C., Buyske, S. and Gabriel, A. 2004. Widespread RNA editing of embedded alu elements in the human transcriptome. *Genome Research* 14(9), pp. 1719–1725.

Kim, U., Wang, Y., Sanford, T., Zeng, Y. and Nishikura, K. 1994. Molecular cloning of cDNA for double-stranded RNA adenosine deaminase, a candidate enzyme for nuclear RNA editing. *Proceedings of the National Academy of Sciences of the United States of America* 91(24), pp. 11457–11461.

Koeris, M., Funke, L., Shrestha, J., Rich, A. and Maas, S. 2005. Modulation of ADAR1 editing activity by Z-RNA in vitro. *Nucleic Acids Research* 33(16), pp. 5362–5370.

Kohn, A.B., Sanford, R.S., Yoshida, M. and Moroz, L.L. 2015. Parallel Evolution and Lineage-Specific Expansion of RNA Editing in Ctenophores. *Integrative and Comparative Biology* 55(6), pp. 1111–1120.

Konkel, M.K., Walker, J.A., Hotard, A.B., Ranck, M.C., Fontenot, C.C., Storer, J., Stewart, C., Marth, G.T., 1000 Genomes Consortium and Batzer, M.A. 2015. Sequence analysis and characterization of active human alu subfamilies based on the 1000 genomes pilot project. *Genome Biology and Evolution* 7(9), pp. 2608–2622.

Kuttan, A. and Bass, B.L. 2012. Mechanistic insights into editing-site specificity of ADARs. *Proceedings of the National Academy of Sciences of the United States of America* 109(48), pp. E3295-304.

Lee, Y.-M., Kim, H.-E., Park, C.-J., Lee, A.-R., Ahn, H.-C., Cho, S.J., Choi, K.-H., Choi, B.-S. and Lee, J.-H. 2012. NMR study on the B-Z junction formation of DNA duplexes induced by Z-DNA binding domain of human ADAR1. *Journal of the American Chemical Society* 134(11), pp. 5276–5283.

Lehmann, K.A. and Bass, B.L. 2000. Double-Stranded RNA Adenosine Deaminases ADAR1 and ADAR2 Have Overlapping Specificities†. *Biochemistry* 39(42), pp. 12875–12884.

Lehmann, K.A. and Bass, B.L. 1999. The importance of internal loops within RNA substrates of ADAR1. *Journal of Molecular Biology* 291(1), pp. 1–13.

Levanon, E.Y. and Eisenberg, E. 2015. Does RNA editing compensate for Alu invasion of the primate genome? *Bioessays: News and Reviews in Molecular, Cellular and Developmental Biology* 37(2), pp. 175–181.

Liddicoat, B.J., Piskol, R., Chalk, A.M., Ramaswami, G., Higuchi, M., Hartner, J.C., Li, J.B., Seeburg, P.H. and Walkley, C.R. 2015. RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. *Science* 349(6252), pp. 1115–1120.

Liscovitch-Brauer, N., Alon, S., Porath, H.T., Elstein, B., Unger, R., Ziv, T., Admon, A., Levanon, E.Y., Rosenthal, J.J.C. and Eisenberg, E. 2017. Trade-off between Transcriptome Plasticity and Genome Evolution in Cephalopods. *Cell* 169(2), pp. 191-202.e11.

Liu, Y., Emeson, R.B. and Samuel, C.E. 1999. Serotonin-2C receptor pre-mRNA editing in rat brain and in vitro by splice site variants of the interferon-inducible double-stranded RNA-specific adenosine deaminase ADAR1. *The Journal of Biological Chemistry* 274(26), pp. 18351–18358.

Liu, Y. and Samuel, C.E. 1999. Editing of glutamate receptor subunit B pre-mRNA by splice-site variants of interferon-inducible double-stranded RNA-specific adenosine deaminase ADAR1. *The Journal of Biological Chemistry* 274(8), pp. 5070–5077.

Liu, Y., Wolff, K.C., Jacobs, B.L. and Samuel, C.E. 2001. Vaccinia virus E3L interferon resistance protein inhibits the interferon-induced adenosine deaminase A-to-I editing activity. *Virology* 289(2), pp. 378–387.

Luckow, V.A., Lee, S.C., Barry, G.F. and Olins, P.O. 1993. Efficient generation of infectious recombinant baculoviruses by site-specific transposon-mediated insertion of foreign genes into a baculovirus genome propagated in Escherichia coli. *Journal of Virology* 67(8), pp. 4566–4579.

Lunde, B.M., Moore, C. and Varani, G. 2007. RNA-binding proteins: modular design for efficient function. *Nature Reviews. Molecular Cell Biology* 8(6), pp. 479–490.

Maas, S. and Gommans, W.M. 2009. Novel exon of mammalian ADAR2 extends open reading frame. *Plos One* 4(1), p. e4225.

Macbeth, M.R. and Bass, B.L. 2007. Large-scale overexpression and purification of ADARs from Saccharomyces cerevisiae for biophysical and biochemical studies. *Methods in Enzymology* 424, pp. 319–331.

Macbeth, M.R., Lingam, A.T. and Bass, B.L. 2004. Evidence for auto-inhibition by the N terminus of hADAR2 and activation by dsRNA binding. *RNA (New York)* 10(10), pp. 1563–1571.

Macbeth, M.R., Schubert, H.L., Vandemark, A.P., Lingam, A.T., Hill, C.P. and Bass, B.L. 2005. Inositol hexakisphosphate is bound in the ADAR2 core and required for RNA editing. *Science* 309(5740), pp. 1534–1539.

Matthews, M.M., Thomas, J.M., Zheng, Y., Tran, K., Phelps, K.J., Scott, A.I., Havel, J., Fisher, A.J. and Beal, P.A. 2016. Structures of human ADAR2 bound to dsRNA reveal base-flipping mechanism and basis for site selectivity. *Nature Structural & Molecular Biology* 23(5), pp. 426–433.

McMahon, A.C., Rahman, R., Jin, H., Shen, J.L., Fieldsend, A., Luo, W. and Rosbash, M. 2016. TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins. *Cell* 165(3), pp. 742–753.

Melcher, T., Maas, S., Herb, A., Sprengel, R., Higuchi, M. and Seeburg, P.H. 1996. RED2, a brain-specific member of the RNA-specific adenosine deaminase family. *The Journal of Biological Chemistry* 271(50), pp. 31795–31798.

Melcher, T., Maas, S., Herb, A., Sprengel, R., Seeburg, P.H. and Higuchi, M. 1996. A mammalian RNA editing enzyme. *Nature* 379(6564), pp. 460–464.

Merkle, T., Merz, S., Reautschnig, P., Blaha, A., Li, Q., Vogel, P., Wettengel, J., Li, J.B. and Stafforst, T. 2019. Precise RNA editing by recruiting endogenous ADARs with antisense oligonucleotides. *Nature Biotechnology* 37(2), pp. 133–138.

Mizrahi, R.A., Phelps, K.J., Ching, A.Y. and Beal, P.A. 2012. Nucleoside analog studies indicate mechanistic differences between RNA-editing adenosine deaminases. *Nucleic Acids Research* 40(19), pp. 9825–9835.

Ng, S.K., Weissbach, R., Ronson, G.E. and Scadden, A.D.J. 2013. Proteins that contain a functional Z-DNA-binding domain localize to cytoplasmic stress granules. *Nucleic Acids Research* 41(21), pp. 9786–9799.

Nishikura, K., Yoo, C., Kim, U., Murray, J.M., Estes, P.A., Cash, F.E. and Liebhaber, S.A. 1991. Substrate specificity of the dsRNA unwinding/modifying activity. *The EMBO Journal* 10(11), pp. 3523–3532.

Oakes, E., Anderson, A., Cohen-Gadol, A. and Hundley, H.A. 2017. Adenosine Deaminase That Acts on RNA 3 (ADAR3) Binding to Glutamate Receptor Subunit B Pre-mRNA Inhibits RNA Editing in Glioblastoma. *The Journal of Biological Chemistry* 292(10), pp. 4326–4335.

Oakes, E., Vadlamani, P. and Hundley, H.A. 2017. Methods for the Detection of Adenosine-to-Inosine Editing Events in Cellular RNA. *Methods in Molecular Biology* 1648, pp. 103–127.

O'Connell, M.A., Gerber, A. and Keller, W. 1997. Purification of human double-stranded RNA-specific editase 1 (hRED1) involved in editing of brain glutamate receptor B pre-mRNA. *The Journal of Biological Chemistry* 272(1), pp. 473–478.

O'Connell, M.A., Krause, S., Higuchi, M., Hsuan, J.J., Totty, N.F., Jenny, A. and Keller, W. 1995. Cloning of cDNAs encoding mammalian double-stranded RNA-specific adenosine deaminase. *Molecular and Cellular Biology* 15(3), pp. 1389–1397.

Ohman, M., Källman, A.M. and Bass, B.L. 2000. In vitro analysis of the binding of ADAR2 to the pre-mRNA encoding the GluR-B R/G site. *RNA (New York)* 6(5), pp. 687–697.

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

Patterson, J.B. and Samuel, C.E. 1995. Expression and regulation by interferon of a double-stranded-RNA-specific adenosine deaminase from human cells: evidence for two forms of the deaminase. *Molecular and Cellular Biology* 15(10), pp. 5376–5388.

Peroutka, R.J., Elshourbagy, N., Piech, T. and Butt, T.R. 2008. Enhanced protein expression in mammalian cells using engineered SUMO fusions: secreted phospholipase A2. *Protein Science* 17(9), pp. 1586–1595.

Phelps, K.J., Tran, K., Eifler, T., Erickson, A.I., Fisher, A.J. and Beal, P.A. 2015. Recognition of duplex RNA by the deaminase domain of the RNA editing enzyme ADAR2. *Nucleic Acids Research* 43(2), pp. 1123–1132.

Pokharel, S., Jayalath, P., Maydanovych, O., Goodman, R.A., Wang, S.C., Tantillo, D.J. and Beal, P.A. 2009. Matching active site and substrate structures for an RNA editing reaction. *Journal of the American Chemical Society* 131(33), pp. 11882–11891.

Polson, A.G. and Bass, B.L. 1994. Preferential selection of adenosines for modification by double-stranded RNA adenosine deaminase. *The EMBO Journal* 13(23), pp. 5701–5711.

Polson, A.G., Crain, P.F., Pomerantz, S.C., McCloskey, J.A. and Bass, B.L. 1991. 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(49), pp. 11507–11514.

Poulsen, H., Jorgensen, R., Heding, A., Nielsen, F.C., Bonven, B. and Egebjerg, J. 2006. Dimerization of ADAR2 is mediated by the double-stranded RNA binding domain. *RNA (New York)* 12(7), pp. 1350–1360.

Poulsen, H., Nilsson, J., Damgaard, C.K., Egebjerg, J. and Kjems, J. 2001. CRM1 mediates the export of ADAR1 through a nuclear export signal within the Z-DNA binding domain. *Molecular and Cellular Biology* 21(22), pp. 7862–7871.

Powell, L.M., Wallis, S.C., Pease, R.J., Edwards, Y.H., Knott, T.J. and Scott, J. 1987. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell* 50(6), pp. 831–840.

Qu, L., Yi, Z., Zhu, S., Wang, C., Cao, Z., Zhou, Z., Yuan, P., Yu, Y., Tian, F., Liu, Z., Bao, Y., Zhao, Y. and Wei, W. 2019. Programmable RNA editing by recruiting endogenous ADAR using engineered RNAs. *Nature Biotechnology* 37(9), pp. 1059–1069.

Ramaswami, G. and Li, J.B. 2014. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Research* 42(Database issue), pp. D109-13.

Ramos, A., Grünert, S., Adams, J., Micklem, D.R., Proctor, M.R., Freund, S., Bycroft, M., St Johnston, D. and Varani, G. 2000. RNA recognition by a staufen double-stranded RNA-binding domain. *The EMBO Journal* 19(5), pp. 997–1009.

Rebagliati, M.R. and Melton, D.A. 1987. Antisense RNA injections in fertilized frog eggs reveal an RNA duplex unwinding activity. *Cell* 48(4), pp. 599–605.

Rice, G.I., Kasher, P.R., Forte, G.M.A., Mannion, N.M., Crow, Y.J., et al. 2012. Mutations in ADAR1 cause Aicardi-Goutières syndrome associated with a type I interferon signature. *Nature Genetics* 44(11), pp. 1243–1248.

Rueden, C.T., Schindelin, J., Hiner, M.C., DeZonia, B.E., Walter, A.E., Arena, E.T. and Eliceiri, K.W. 2017. ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics* 18(1), p. 529.

Salter, J.D., Bennett, R.P. and Smith, H.C. 2016. The APOBEC protein family: united by structure, divergent in function. *Trends in Biochemical Sciences* 41(7), pp. 578–594.

Salter, J.D. and Smith, H.C. 2018. Modeling the embrace of a mutator: APOBEC selection of nucleic acid ligands. *Trends in Biochemical Sciences* 43(8), pp. 606–622.

Samuel, C.E. 2012. ADARs: viruses and innate immunity. *Current Topics in Microbiology and Immunology* 353, pp. 163–195.

Samuel, C.E. 2001. Antiviral actions of interferons. *Clinical Microbiology Reviews* 14(4), pp. 778–809, table of contents.

Schaub, M. and Keller, W. 2002. RNA editing by adenosine deaminases generates RNA and protein diversity. *Biochimie* 84(8), pp. 791–803.

Schumacher, J.M., Lee, K., Edelhoff, S. and Braun, R.E. 1995. Distribution of Tenr, an RNA-binding protein, in a lattice-like network within the spermatid nucleus in the mouse. *Biology of Reproduction* 52(6), pp. 1274–1283.

Schwartz, T., Rould, M.A., Lowenhaupt, K., Herbert, A. and Rich, A. 1999. Crystal structure of the Zalpha domain of the human editing enzyme ADAR1 bound to left-handed Z-DNA. *Science* 284(5421), pp. 1841–1845.

Serra, M.J., Smolter, P.E. and Westhof, E. 2004. Pronounced instability of tandem IU base pairs in RNA. *Nucleic Acids Research* 32(5), pp. 1824–1828.

Stefl, R., Oberstrass, F.C., Hood, J.L., Jourdan, M., Zimmermann, M., Skrisovska, L., Maris, C., Peng, L., Hofr, C., Emeson, R.B. and Allain, F.H.-T. 2010. The solution structure of the ADAR2 dsRBM-RNA complex reveals a sequence-specific readout of the minor groove. *Cell* 143(2), pp. 225–237.

Stefl, R., Skrisovska, L. and Allain, F.H.-T. 2005. RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. *EMBO Reports* 6(1), pp. 33–38.

Stefl, R., Xu, M., Skrisovska, L., Emeson, R.B. and Allain, F.H.-T. 2006. Structure and specific RNA binding of ADAR2 double-stranded RNA binding motifs. *Structure* 14(2), pp. 345–355.

Strehblow, A., Hallegger, M. and Jantsch, M.F. 2002. 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. *Molecular Biology of the Cell* 13(11), pp. 3822–3835.

Tan, M.H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A.N., Liu, K.I., Zhang, R., Ramaswami, G., Ariyoshi, K., Gupte, A., Keegan, L.P., George, C.X., Ramu, A., Huang, N., Pollina, E.A., Leeman, D.S., Rustighi, A., Goh, Y.P.S., GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration &Visualization—EBI, Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz, Chawla, A., Del Sal, G., Peltz, G., Brunet, A., Conrad, D.F., Samuel, C.E., O'Connell, M.A., Walkley, C.R., Nishikura, K. and Li, J.B. 2017. Dynamic landscape and regulation of RNA editing in mammals. *Nature* 550(7675), pp. 249–254.

Teng, B., Burant, C.F. and Davidson, N.O. 1993. Molecular cloning of an apolipoprotein B messenger RNA editing protein. *Science* 260(5115), pp. 1816–1819.

Thomas, J.M. and Beal, P.A. 2017. How do ADARs bind RNA? New protein-RNA structures illuminate substrate recognition by the RNA editing ADARs. *Bioessays: News and Reviews in Molecular, Cellular and Developmental Biology* 39(4).

Tian, B., Bevilacqua, P.C., Diegelman-Parente, A. and Mathews, M.B. 2004. The double-stranded-RNA-binding motif: interference and much more. *Nature Reviews. Molecular Cell Biology* 5(12), pp. 1013–1023.

Tomaselli, S., Bonamassa, B., Alisi, A., Nobili, V., Locatelli, F. and Gallo, A. 2013. ADAR enzyme and miRNA story: a nucleotide that can make the difference. *International Journal of Molecular Sciences* 14(11), pp. 22796–22816.

Toth, A.M., Zhang, P., Das, S., George, C.X. and Samuel, C.E. 2006. Interferon action and the double-stranded RNA-dependent enzymes ADAR1 adenosine deaminase and PKR protein kinase. *Progress in Nucleic Acid Research and Molecular Biology* 81, pp. 369–434.

Valente, L. and Nishikura, K. 2005. ADAR Gene Family and A-to-I RNA Editing: Diverse Roles in Posttranscriptional Gene Regulation. In: Progress in Nucleic Acid Research and Molecular Biology. Elsevier, pp. 299–338.

Valente, L. and Nishikura, K. 2007. RNA binding-independent dimerization of adenosine deaminases acting on RNA and dominant negative effects of nonfunctional subunits on dimer functions. *The Journal of Biological Chemistry* 282(22), pp. 16054–16061.

Vogel, P., Moschref, M., Li, Q., Merkle, T., Selvasaravanan, K.D., Li, J.B. and Stafforst, T. 2018. Efficient and precise editing of endogenous transcripts with SNAP-tagged ADARs. *Nature Methods* 15(7), pp. 535–538.

Vu, L.T. and Tsukahara, T. 2017. C-to-U editing and site-directed RNA editing for the correction of genetic mutations. *Bioscience trends* 11(3), pp. 243–253.

Wang, Q., Miyakoda, M., Yang, W., Khillan, J., Stachura, D.L., Weiss, M.J. and Nishikura, K. 2004. Stress-induced apoptosis associated with null mutation of ADAR1 RNA editing deaminase gene. *The Journal of Biological Chemistry* 279(6), pp. 4952–4961.

Wang, Y., Chung, D.H., Monteleone, L.R., Li, J., Chiang, Y., Toney, M.D. and Beal, P.A. 2019. RNA binding candidates for human ADAR3 from substrates of a gain of function mutant expressed in neuronal cells. *Nucleic Acids Research*.

Wang, Y., Park, S. and Beal, P.A. 2018. Selective recognition of RNA substrates by ADAR deaminase domains. *Biochemistry* 57(10), pp. 1640–1651.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R. and Schwede, T. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research* 46(W1), pp. W296–W303.

Wu, H., Henras, A., Chanfreau, G. and Feigon, J. 2004. Structural basis for recognition of the AGNN tetraloop RNA fold by the double-stranded RNA-binding domain of Rnt1p RNase III. *Proceedings of the National Academy of Sciences of the United States of America* 101(22), pp. 8307–8312.

Xu, W., Rahman, R. and Rosbash, M. 2018. Mechanistic implications of enhanced editing by a HyperTRIBE RNA-binding protein. *RNA (New York)* 24(2), pp. 173–182.

Yablonovitch, A.L., Fu, J., Li, K., Mahato, S., Kang, L., Rashkovetsky, E., Korol, A.B., Tang, H., Michalak, P., Zelhof, A.C., Nevo, E. and Li, J.B. 2017. Regulation of gene expression and RNA editing in Drosophila adapting to divergent microclimates. *Nature Communications* 8(1), p. 1570.

Yang, Y., Zhu, M., Fan, X., Yao, Y., Yan, J., Tang, Y., Liu, S., Li, K. and Tang, Z. 2019. Developmental atlas of the RNA editome in Sus scrofa skeletal muscle. *DNA Research* 26(3), pp. 261–272.

Zhang, Yuanyuan, Han, D., Dong, X., Wang, J., Chen, J., Yao, Y., Darwish, H.Y.A., Liu, W. and Deng, X. 2019. Genome-wide profiling of RNA editing sites in sheep. *Journal of animal science and biotechnology* 10, p. 31.

Zhang, Yuebo, Zhang, L., Yue, J., Wei, X., Wang, Ligang, Liu, X., Gao, H., Hou, X., Zhao, F., Yan, H. and Wang, Lixian 2019. Genome-wide identification of RNA editing in seven porcine tissues by matched DNA and RNA high-throughput sequencing. *Journal of animal science and biotechnology* 10, p. 24.