RNA Editing at Baseline and Following Endoplasmic Reticulum Stress

By

Allison Leigh Richards

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Human Genetics)
in The University of Michigan
2015

Doctoral Committee:

Professor Vivian G. Cheung, Chair
Assistant Professor Santhi K. Ganesh
Professor David Ginsburg
Professor Daniel J. Klionsky

Dedication

To my father, mother, and Matt without whom I would never have made it

Acknowledgements

Thank you first and foremost to my dissertation mentor, Dr. Vivian Cheung. I have learned so much from you over the past several years including presentation skills such as never sighing and never saying “as you can see…” You have taught me how to think outside the box and how to create and explain my story to others. I would not be where I am today without your help and guidance.

Thank you to the members of my dissertation committee (Drs. Santhi Ganesh, David Ginsburg and Daniel Klionsky) for all of your advice and support. I would also like to thank the entire Human Genetics Program, and especially JoAnn Sekiguchi and Karen Grahl, for welcoming me to the University of Michigan and making my transition so much easier. Thank you to Michael Boehnke and the Genome Science Training Program for supporting my work.

A very special thank you to all of the members of the Cheung lab, past and present. Thank you to Xiaorong Wang for all of your help from the bench to advice on my career. Thank you to Zhengwei Zhu who has helped me immensely throughout my thesis even through my panic. Thank you to Jonathan Toung who taught me how to code and was so patient when I had trouble. Thank you to Alan Bruzel who helped me with protocols and made my whole experience more enjoyable. And thank you to everyone else in the Cheung lab for all of your fun and support.

I would especially like to thank my friends who not only helped me with my lab work but also kept me sane: Michelle Nguyen-McCarty, Lauren Brady, Jillian Boden and Becky Glineburg. Michelle, thank you for the support and the distractions. Becky, thank you for introducing me to fun; it has served me well. Lauren, thank you for the understanding and the gossip. Jill, thank you for the lunches and steadying me when I get off-balance.

And finally, I would like to thank my family. To my father for always believing in me and comforting me when all seemed lost. To my mother who always listened even when I repeated myself 20 times. Thank you to Deborah and Ted for your support. And thank you to Matt, who never let me give up. Without even one of you, I could never have accomplished anything.

Table of Contents

List of Figures

List of Tables

Abstract

RNA sequences are expected to be identical to the DNA template. However, some RNA processing steps, such as RNA editing, can lead to differences in the RNA sequence that affect the fate of the RNA transcripts or the resultant proteins. My thesis focuses on the regulation of the canonical A-to-G editing and non-canonical RNA-DNA sequence Differences (RDD). My work contributed to the identification of RDDs throughout the human transcriptome. We identified all 12 types of single base differences across multiple individuals and various tissue types. We also detected peptides matching the RDD-encoded sequences suggesting that RDDs are translated into proteins. In subsequent work, we found that the non-canonical RDDs are found in nascent RNA. Through the use of nuclear run-on assays, we found that RDD occurs within seconds of exiting the RNA polymerase complex.

Chapters 4 and 5 discuss my study of canonical A-to-G editing mediated by Adenosine Deaminase Acting on RNA (ADAR). First, we found that A-to-G editing levels differ across individuals. I searched for and identified genetic variants whose alleles are associated with editing levels of sites in the same gene. These data demonstrate that ADAR editing is *cis* regulated and can lead to individual variability in editing levels. Furthermore, by utilizing individual variability in editing and studying the relationship between editing sites, I learned how ADAR edits multiple sites in a given transcript. My data support a model where ADAR edits multiple sites along one side of a double-stranded RNA structure. To learn about the biological

significance of RNA editing, I focused on endoplasmic reticulum (ER) stress. I found that editing levels change following ER stress suggesting that these RNA processing steps play a role in the ER stress response. Together, this work sheds light on the regulation of RNA editing and RDDs in the human transcriptome and aids in the understanding of how these processes may play a role in cellular response to stress.

Chapter 1: Introduction

RNA is generally assumed to be a copy of its DNA template, but there are many processes that modify RNA transcripts. These modifications give rise to a number of unique proteins from just one genomic locus.

One example of increased variability in proteins comes from RNA splicing. For many years, it was assumed that transcript splicing was constant and always occurred at the same position. However, over the last decade, it has become clear that many genes show alternative splicing. In fact, it is estimated that over 90% of human genes with multiple exons show alternative splicing (Pan et al., 2008, Wang et al., 2008). There are many examples where each isoform has been shown to perform different tasks. For example, there are eight isoforms of *CASP8*, a gene involved in apoptosis, where some of the isoforms produce proteins that have an anti-apoptotic function and the others produce proteins that are pro-apoptotic (Himeji et al., 2002). This example, as well as many others, demonstrates the importance of protein diversity and how changes to RNA transcripts can impact the cell.

In addition to splicing, the RNA transcript goes through numerous processing steps before it is translated into protein or degraded. Some of these steps can affect the RNA sequence such that it no longer matches the corresponding DNA sequence. For example, RNA

editing modifies the sequence content of the RNA transcripts and can change the protein coding sequences or regulatory sequences.

Identification of RNA editing

The first evidence of RNA editing was seen in the mitochondria of trypanosomes in the *coxII* gene (Benne et al., 1986). The *coxII* gene is found in many species. It was found that DNA of the sequence of *coxII* gene in T. brucei was very different than the sequences found in other organisms and was not predicted to form a functional protein (Hensgens et al., 1984, Payne et al., 1985). However, when Benne et al studied the mRNA sequence of *coxII* it was found that uracils were being added into the RNA transcript that could not be found in the corresponding DNA sequence. These uracils restored the functional capacity of the transcript to make the *coxII* protein.

Following this characterization of *coxII,* other examples of uracil additions and deletions were found in trypanosomes demonstrating that this is a widespread phenomenon. Similar to the coxII gene, insertion and deletion of uracils in *COIII* (Feagin et al., 1988a) and *MURF3* (Shaw et al., 1989) restored the sequence that in other organisms would result in a functional protein. This suggested that RNA editing may be a way to bypass any detrimental mutations in the DNA sequence. With further study, it was found that some transcripts are highly edited and are almost unrecognizable in their DNA form (Feagin et al., 1988a, Maslov et al., 1992). In addition to restoring conserved sequences, in some cases, multiple species will use RNA editing at various positions in the RNA sequence to result in the same functional amino acid sequence

(Feagin et al., 1988b, Shaw et al., 1988). For example, two kinetoplastid species have been shown to edit *COIII* in different patterns to result in a protein sequence that only differs by one residue (Shaw et al. 1988). These results demonstrate that RNA editing is a way to maintain a functional protein.

The mechanism of RNA editing in trypanosomes was elucidated by Blum et al (Blum et al., 1990) by the identification of guide RNAs (gRNA). Guide RNAs are encoded on minicircles in the mitochondria of trypanosomes. The gRNAs contain an anchor at the 5' end that is complementary to the mRNA transcript. The middle region of the gRNA contains the sequence complementary to the processed mRNA such that the uracil will be added opposite any adenosine in the gRNA (Blum et al., 1990, Seiwert et al., 1996, Leung and Koslowsky, 1999). In this way, the sequence of the guide RNA directs the insertions or deletions of uracil throughout the target mRNA transcript.

The gRNAs are templates for the editing. The necessary enzymes are found in a complex termed the RNA Editing Core Complex (RECC). There are different forms of the RECC that can lead to insertion or deletion of uracil residues. The complex must include a terminal uridylyl transferase (TUTase) to add uracil residues or an exonuclease to delete uracil residues, scaffolding proteins and an RNA ligase (Rusché et al., 1997, Carnes et al., 2005, Stuart et al., 2005, Trotter et al., 2005, Aphasizhev and Aphasizheva, 2008). First, an endonuclease cleaves at the mismatched base. Next, a terminal uridylyl transferase (TUTase) adds uracil residues which then need to be trimmed back to match the gRNA sequence by a 3' to 5' exonuclease. Finally an RNA ligase puts the cleaved fragments back together (Piller et al., 1995, Aphasizhev et al.,

2002). Though this form of editing has been demonstrated in several kinetoplastids, different types of editing have been found in other organisms.

RNA Editing in Plants

A few years after the identification of RNA editing in the mitochondria of trypanosomes, another type of RNA editing was identified in the mitochondria and plastids of plants (Covello and Gray, 1989, Gualberto et al., 1989, Hiesel et al., 1989, Hoch et al., 1991). The type of editing found in these organelles was a single-base change leading to a C-to-U or U-to-C difference between the DNA and the corresponding mRNA sequence.

Similar to the editing found in trypanosomes, editing in plants is necessary for expression of functional proteins. This can occur through modification of the amino acid sequence (Bock et al., 1994, Sasaki et al., 2001) or changing a start (Hoch et al., 1991) or stop codon (Wintz and Hanson, 1991). RNA editing in plants is mediated through a family of enzymes containing a pentatricopeptide repeat (PPR) motif (Kotera et al., 2005). These proteins had previously been shown to be involved in other types of RNA processing, such as splicing (Delannoy et al., 2007, Schmitz-Linneweber and Small, 2008). The PPR-containing enzymes recognize specific regions of the target mRNA (Bock et al., 1996, Chaudhuri and Maliga, 1996, Okuda et al., 2006). The PPR region of these family members is variable and many plants contain hundreds of PPR proteins (Lurin et al., 2004). Due to the large number of PPR-containing proteins and their variable RNA binding domains, it has been suggested that each

protein may bind to only a few targets leading to the wide variety of target sites. Questions still remain on the identity of the protein able to create the C-to-U or the rare U-to-C edits.

RNA editing in plants is sometimes found in every transcript (Lu and Hanson, 1994), but many times, editing is variable. For example, *nad3* contains multiple editing sites, none of which are edited in every transcript (Schuster et al., 1990). This suggests that there may be another reason for RNA editing besides reverting a specific mutation back to the ancestral form of the protein. It is possible that RNA editing may be a way to introduce various forms of an mRNA transcript to either affect the resulting proteins or regulate expression of the gene (Lu et al., 1996, Phreaner et al., 1996).

RNA editing in organisms as diverse as trypanosomes and plants suggests that RNA editing plays a key role in maintaining protein function. It may also provide variability in the mRNA transcripts to encode multiple functional gene products.

Identification of RNA Editing in Mammals

While editing was being identified in trypanosomes, other scientists were noting a strange discrepancy in apolipoprotein B (APOB) which leads to two forms of the protein, B-48 and B-100. Apolipoprotein B has been a protein of interest for decades due to its involvement in cholesterol metabolism and uptake. In the early 80s it was found that APOB came in two different forms, B-48 and B-100 (Kane et al., 1980, Wu and Windmueller, 1981). Though cholesterol was very well studied, many groups struggled to isolate the two unique forms of the APOB protein due to their hydrophobic nature. Once the two forms were identified, it was

necessary to determine whether they were from the same gene and RNA transcript. This proved to be a difficult question to answer due to the limitations of mass spectrometry and sequencing. Some argued that they must be proteins resulting from different gene loci because there were diseases that caused only one form to be expressed (Malloy et al., 1981). However, there was a single mutation that could be traced to both forms (Young et al., 1986). With the use of specific antibodies, researchers were able to identify large homologies between the two forms of APOB supporting the idea that both forms resulted from a single gene (Marcel et al., 1982, Hospattankar et al., 1986).

The next hurdle was to determine how the two forms could be made from a single gene and why expression of the two forms was tissue specific. In humans, B-48 was found in the small intestine while B-100 was found in liver. Multiple groups found that these two proteins were not only produced from a single gene but that there was a single, post-transcriptional editing process that lead to a premature stop codon (CAA to UAA) (Chen et al., 1987, Powell et al., 1987, Lau et al., 1991). Then in 1993, Apolipoprotein B Editing Catalytic subunit 1 (APOBEC) was found to be responsible for deaminating the cytosine to uracil at position 6666 (Teng et al., 1993). APOBEC, in complex with APOBEC-1 complementation factor (ACF), deaminates the specific cytosine using the surrounding sequence as an anchor (Dance et al., 2002).

APOBEC and *APOB* are the first enzyme and target pair of RNA editing identified in humans. Soon, another form of RNA editing would be found that affects thousands of sites throughout the human transcriptome.

ADAR RNA Editing

The identification of trypanosome editing and APOBEC RNA editing opened the doors to searching for still further examples of RNA editing. In the late 1980s, a protein family was identified in *xenopus* embryos that would be called adenosine deaminases acting on RNA (ADAR) (Bass and Weintraub 1988, Kim et al., 1994). At the time the function of ADAR was not clear. It was found that ADAR bound and unwound double-stranded RNA. Further investigation found that inosine was present in the RNA following ADAR binding (Bass and Weintraub 1988). Finally, it was understood that ADAR deaminates the adenosine to inosine which resulted in disruption of the double-stranded structure and the unwinding phenomenon that was originally found (Wagner et al., 1989).

There are four different ADAR proteins in humans: ADAR1, ADAR2, ADAR3 and ADAR4. ADAR4 is expressed in testes, suggesting that it may play a role in spermatogenesis (Meng et al., 1997, Connolly et al., 2005). ADAR3 is expressed in the brain and is not known to be catalytically active; however, it may play a role in regulating the activity of ADAR1 and ADAR2 (Chen et al., 2000). ADAR1 and ADAR2 are expressed in most tissues and can edit many sites throughout the transcriptome (Wang et al., 2013).

ADAR1 and ADAR2 catalyze deamination of adenosines to inosines in double-stranded RNA (dsRNA) regions. Inosines base-pair similar to guanosine and so inosines are read by translational machinery as a guanosine (Reuter et al., 1999, Bass 2002, Jepson and Reenan, 2008, Nishikura 2010). Therefore, ADAR enzymes effectively induce an A-to-G difference

between the DNA and the corresponding RNA sequence. ADAR enzymes specifically target large stretches of dsRNA regions. Studies have shown that RNA structures less than 100 bases long show far fewer instances of editing than longer ones (Nishikura et al., 1991). In addition to targeting dsRNA regions, ADAR enzymes also use the surrounding sequences to identify their target sites. Both ADAR1 and ADAR2 targets show a depletion of guanosine just upstream of the editing site, while an enrichment of guanosine just downstream of the editing site (Polson and Bass, 1994, Lehmann and Bass, 2000, Eggington et al., 2011).

One of the best characterized examples of ADAR editing is in the AMPA receptor in mammals (Rueter et al., 1995, Yang et al., 1995). The editing of a subunit of the AMPA receptor, GluR2, changes the amino acid sequence at two sites, 607 (Q-to-R) and 764 (R-to-G) (Seeburg et al., 1998). Editing at site 764 changes the recovery from inactivation of the channel (Lomeli et al., 1994). The editing at site 607 (Sommer et al., 1991) influences the conductance of the AMPA receptor and is edited to 100% in human adults. The arginine at site 607 decreases calcium influx and affects the targeting of the receptors to the neuronal synapses (Sommer et al., 1991, Seeburg et al., 1998, Greger et al., 2002). When editing is prevented in mice, seizures or death can occur. This further emphasizes the importance of RNA editing in these neuronal ion channels (Brusa et al., 1995). A number of examples of ADAR editing affecting ion channel function in neurons have been found (Sommer et al., 1991, Köhler et al., 1993, Nutt et al., 1994, Paschen and Djuricic, 1994, Paschen et al., 1994, Burns et al., 1997, Bhalla et al., 2004, Ohlson et al., 2007).

In addition to the editing of ion channels in the brain, recent studies have shown that ADAR editing is widespread throughout the transcriptome (Athanasiadis et al., 2004, Levanon et al., 2004, Li et al., 2009, Wang et al., 2013, Bazak et al., 2014a, Ramaswami and Li, 2014, Sakurai et al., 2014, Ulbricht and Emeson, 2014). The identification of these additional sites has led to the characterization of two types of editing: site-selective and promiscuous (Wahlstedt and Ohman, 2011). These terms are not well defined and generally describe one or a few sites in a given region versus many sites in a region, respectively. The sites commonly found in ion channels would be termed site-selective editing sites. On the other hand, numerous transcripts show evidence of promiscuous editing.

Promiscuous editing, or hyper-editing, does not mean that a site is not consistently edited but rather that multiple editing sites occur in the same region. This has been found to occur in Alu elements in the primate genomes (Ramaswami et al., 2012, Levanon et al., 2004, Athanasiadis et al., 2004). Alu elements are abundant, comprising about 10% of the human genome (Batzer and Deininger, 2002). Editing sites are often found in genes with multiple Alu elements oriented in opposite directions (Athanasiadis et al., 2004). The two opposite Alu elements in the transcript can base-pair to make a long dsRNA region. This is then a target of ADAR editing (Bazak et al., 2014b).

The role of editing in non-coding regions is not as clear as the editing that occurs in the coding regions of the genes. However, several studies have demonstrated that editing in the non-coding regions of a transcript may play an important role in regulating gene expression.

One example is the role of editing on mRNA splicing. ADAR2 editing of its own mRNA transcript leads to differential splicing by creating a 3' acceptor site leading to lower expression of ADAR2 (Rueter et al., 1999, Feng et al., 2006). Additionally, ADAR-mediated editing was shown to affect splicing of other genes, suggesting that this may be a common mechanism of controlling gene expression (Lev-Maor et al., 2007).

RNA editing has also been shown to occur within microRNAs (miRNAs) (Luciano et al., 2004) and affect their miRNA processing. For example, editing of the precursor of miRNA-142 leads to a decrease in the expression of the mature miRNA by affecting cleavage of the transcript by Drosha or Dicer (Chawla and Sokol, 2014, Yang et al., 2006). Some studies have shown that ADAR-mediated editing can result in mature miRNAs with different targets due to changes to the seed sequence (Kawahara et al., 2008, Alon et al., 2012, Kume et al., 2014). Related studies have also shown that ADAR editing can interfere with the related siRNA pathway (Wu et al., 2011).

While the previous two examples demonstrate how an adenosine to guanosine change in sequence can influence protein binding, inosine, itself, can be a target for RNA binding proteins. First, inosine can allow for sequestration of the mRNA transcript in the nucleus by a complex containing an inosine-specific binding protein NURSA, a splicing factor PSF and a nuclear matrix protein (Zhang and Carmichael 2001). Additionally, inosine can lead to recruitment of a factor involved in the RNA-induced silencing complex leading to degradation of the transcript (Scadden and Smith 2001, Scadden 2005, Scadden and O'Connell, 2005). Through these, and likely additional, pathways, RNA editing can affect the regulation of gene expression.

Non-canonical editing

While APOBEC and ADAR can induce C-to-U and A-to-I differences between DNA and corresponding RNA sequence, respectively, some studies have described genes that have other types of single-base differences in sequence. For example, Sharma et al describe a U-to-C difference in the Wilms' tumor gene (WT1) (Sharma et al., 1994). WT1 acts as a transcriptional regulator and loss of WT1 is associated with Wilms' tumor and childhood malignancies in the kidney (Call et al., 1990, Gessler et al., 1990). This group found that a U-to-C difference within the coding region of the gene that changes a leucine to a proline residue. They found that the proline form of the protein is less efficient at repressing a target promoter (EGR1) than the DNA-encoded leucine form. This example demonstrates that other base differences, in addition to the canonical A-to-G and C-to-U, can occur between the DNA and RNA sequence and that they may play a role in protein function.

A recent study identified more single base differences in WT1 including a G-to-A site (Niavarani et al., 2015). Through knock-down and over-expression, they found that APOBEC3A is responsible for the G-to-A difference. APOBEC3A is a part of the APOBEC family but has mostly been known to target viral DNA as a part of the immune response. This finding demonstrates that the other members of the APOBEC family may also play a role in the other types of non-canonical RNA editing in human cells.

Other examples of non-canonical editing can be found in hnRNP K which contains a G-to-A difference, TPH2 that contains multiple differences including A-to-G, C-to-U, T-to-C and G-

to-A, and β amyloid precursor protein and ubiquitin-B protein which contain frameshift RDDs (van Leeuwen et al., 1998, Klimek-Tomczak et al., 2006, Grohmann et al., 2010). The mechanisms that mediate these differences have not yet been elucidated. A few years ago, we showed (described in Chapter 2) that RNA-DNA sequence Differences (RDD) can be found transcriptome-wide.

Endoplasmic Reticulum Stress

During my thesis work, I wondered how RNA editing can be influenced by environmental stress. To answer this question, I studied RNA editing in response to endoplasmic reticulum (ER) stress (chapter 6). The endoplasmic reticulum is a membrane-bound organelle that is split into two parts: smooth ER that plays a role in lipid metabolism and the rough ER that modifies and transports proteins. ER stress occurs when there are excess misfolded proteins in the cell.

The ER stress response is complex and can determine whether a cell will return to homeostatic conditions or commit to apoptosis. When misfolded proteins accumulate in the cell they recruit protein chaperones such as BiP through hydrophobic interactions. When BiP is recruited to the misfolded proteins, it is released from three membrane-bound proteins responsible for triggering the ER stress response: IRE1, PERK and ATF6 (Bertolotti et al., 2000).

IRE1, Inositol-Requiring Enzyme 1, dimerizes and transphosphorylates after release of BiP (Bertolotti et al., 2000). This activation then allows IRE1 to splice the mRNA of XBP1, X-box Binding Protein 1, to XBP1s which is translated into an active transcription factor (Yoshida et al.,

2001, Calfon et al., 2002). XBP1s promotes the expression of many proteins involved in the ER stress response, such as BiP, the protein chaperone, and EDEM which is involved in ER-associated decay to remove misfolded proteins (Kaneko and Nomura, 2003, Lee et al., 2003).

PERK, Pancreatic EIF2α Kinase, like IRE1 oligomerizes and transphosphorylates following release of BiP (Bertolotti et al., 2000). PERK then inactivates eIF2α by phosphorylation. eIF2α is a eukaryotic translational initiation factor. Through phosphorylation of eIF2α, PERK attenuates translation in the cell to attenuate the accumulation of misfolded proteins. Though most proteins cannot be translated under these conditions, certain proteins such as ATF4 are upregulated. ATF4 is a transcription factor that upregulates genes involved in the induction of apoptosis, such as CHOP, C/EBP Homology Protein (Harding et al., 2003).

Finally, ATF6 is activated following release of BiP by transport to the Golgi apparatus where it is proteolytically cleaved. The cytoplasmic portion of ATF6 can then act as a transcription factor to regulate ER stress-related genes (Ye et al., 2000, Chen et al., 2002, Shen and Prywes, 2004, Shen et al., 2005). ATF6 upregulates protein chaperones, such as BiP and GRP94 (Yoshida et al., 1998).

Together these three pathways lead to a complex ER stress response including induction of proteins involved in reestablishing homeostasis, such as the protein chaperones, or induction of apoptosis mediators, such as CHOP. IRE1 and ATF6 are inactivated by feedback loops before PERK. When ER stress continues, only PERK remains active, thereby leading to apoptosis (Lin et al., 2007).

ER stress occurs under many conditions such as rapid growth in cancer cells, in the β-cells of the pancreas that are making large amounts of insulin, or in B-cells that make large amounts of antibody proteins (Ma and Hendershot, 2004, Lee et al., 2005, Zhang et al., 2005). ER stress can lead to diseases such as diabetes. When a β-cell undergoes excess ER stress, it dies and loss of these β-cells can lead to diabetes (Oyadomari et al., 2002). We study editing under ER stress conditions because it may help us to understand these various diseases.

My Thesis Summary

My work in canonical RNA editing and RNA-DNA sequence Differences, broadly defined as single-base differences between the DNA and RNA sequences that cannot be explained by canonical editing mechanisms, describes the importance of these RNA processing steps in RNA regulation and cellular response to stress.

References

Alon S, Mor E, Vigneault F, Church GM, Locatelli F, Galeano F, Gallo A, Shomron N, Eisenberg E. Systematic identification of edited microRNAs in the human brain. Genome Res. 2012 Aug;22(8):1533-40. doi: 10.1101/gr.131573.111. Epub 2012 Apr 12.

Aphasizhev R, Sbicego S, Peris M, Jang SH, Aphasizheva I, Simpson AM, Rivlin A, Simpson L. Trypanosome mitochondrial 3' terminal uridylyl transferase (TUTase): the key enzyme in U-insertion/deletion RNA editing. Cell. 2002 Mar 8;108(5):637-48.

Aphasizhev R, Aphasizheva I. Terminal RNA uridylyltransferases of trypanosomes. Biochim Biophys Acta. 2008 Apr;1779(4):270-80. doi: 10.1016/j.bbagrm.2007.12.007. Epub 2007 Dec 23.

Athanasiadis A, Rich A, Maas S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biol. 2004 Dec;2(12):e391. Epub 2004 Nov 9.

Bass BL, Weintraub H. An unwinding activity that covalently modifies its double-stranded RNA substrate. Cell. 1988 Dec 23;55(6):1089-98.

Bass BL. RNA editing by adenosine deaminases that act on RNA. Annu Rev Biochem. 2002;71:817-46. Epub 2001 Nov 9.

Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nat Rev Genet. 2002 May;3(5):370-9.

a) Bazak L, Haviv A, Barak M, Jacob-Hirsch J, Deng P, Zhang R, Isaacs FJ, Rechavi G, Li JB, Eisenberg E, Levanon EY. A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. Genome Res. 2014 Mar;24(3):365-76. doi: 10.1101/gr.164749.113. Epub 2013 Dec 17.

b) Bazak L, Levanon EY[1], Eisenberg E[2]. Genome-wide analysis of Alu editability. Nucleic Acids Res. 2014 Jun;42(11):6876-84. doi: 10.1093/nar/gku414. Epub 2014 May 14.

Benne R, Van den Burg J, Brakenhoff JP, Sloof P, Van Boom JH, Tromp MC. Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell. 1986 Sep 12;46(6):819-26.

Bertolotti A, Zhang Y, Hendershot LM, Harding HP, Ron D. Dynamic interaction of BiP and ER stress transducers in the unfolded-protein response. Nat Cell Biol. 2000 Jun;2(6):326-32.

Bhalla T, Rosenthal JJ, Holmgren M, Reenan R. Control of human potassium channel inactivation by editing of a small mRNA hairpin. Nat Struct Mol Biol. 2004 Oct;11(10):950-6. Epub 2004 Sep 7.

Blum B, Bakalara N, Simpson L. A model for RNA editing in kinetoplastid mitochondria: "guide" RNA molecules transcribed from maxicircle DNA provide the edited information. Cell. 1990 Jan 26;60(2):189-98.

Bock R, Kössel H, Maliga P. Introduction of a heterologous editing site into the tobacco plastid genome: the lack of RNA editing leads to a mutant phenotype. EMBO J. 1994 Oct 3;13(19):4623-8.

Bock R, Hermann M, Kössel H. In vivo dissection of cis-acting determinants for plastid RNA editing. EMBO J. 1996 Sep 16;15(18):5052-9.

Brusa R, Zimmermann F, Koh DS, Feldmeyer D, Gass P, Seeburg PH, Sprengel R. Early-onset epilepsy and postnatal lethality associated with an editing-deficient GluR-B allele in mice. Science. 1995 Dec 8;270(5242):1677-80.

Burns CM, Chu H, Rueter SM, Hutchinson LK, Canton H, Sanders-Bush E, Emeson RB. Regulation of serotonin-2C receptor G-protein coupling by RNA editing. Nature. 1997 May 15;387(6630):303-8.

Calfon M, Zeng H, Urano F, Till JH, Hubbard SR, Harding HP, Clark SG, Ron D. IRE1 couples endoplasmic reticulum load to secretory capacity by processing the XBP-1 mRNA. Nature. 2002 Jan 3;415(6867):92-6.

Call KM, Glaser T, Ito CY, Buckler AJ, Pelletier J, Haber DA, Rose EA, Kral A, Yeger H, Lewis WH, Jones C, Housman DE. Isolation and characterization of a zinc finger polypeptide gene at the human chromosome 11 Wilms' tumor locus. Cell. 1990 Feb 9;60(3):509-20.

Carnes J, Trotter JR, Ernst NL, Steinberg A, Stuart K. An essential RNase III insertion editing endonuclease in Trypanosoma brucei. Proc Natl Acad Sci U S A. 2005 Nov 15;102(46):16614-9. Epub 2005 Nov 3.

Chaudhuri S, Maliga P. Sequences directing C to U editing of the plastid psbL mRNA are located within a 22 nucleotide segment spanning the editing site. EMBO J. 1996 Nov 1;15(21):5958-64.

Chawla G, Sokol NS. ADAR mediates differential expression of polycistronic microRNAs. Nucleic Acids Res. 2014 Apr;42(8):5245-55. doi: 10.1093/nar/gku145. Epub 2014 Feb 20.

Chen SH, Habib G,Yang CY, Gu ZW, Lee BR, Weng SA, Silberman SR, Cai SJ, Deslypere JP, Rosseneu M, Gotto AM, Li WH, Chan L. Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. Science. 1987 Oct 16;238(4825):363-6.

Chen CX, Cho DS, Wang Q, Lai F, Carter KC, Nishikura K. A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains. RNA. 2000 May;6(5):755-67.

Chen X, Shen J, Prywes R. The luminal domain of ATF6 senses endoplasmic reticulum (ER) stress and causes translocation of ATF6 from the ER to the Golgi. J Biol Chem. 2002 Apr 12;277(15):13045-52. Epub 2002 Jan 30.

Connolly CM, Dearth AT, Braun RE. Disruption of murine Tenr results in teratospermia and male infertility. Dev Biol. 2005 Feb 1;278(1):13-21.

Covello PS, Gray MW. RNA editing in plant mitochondria. Nature. 1989 Oct 19;341(6243):662-6.

Dance GS, Sowden MP, Cartegni L, Cooper E, Krainer AR, Smith HC. Two proteins essential for apolipoprotein B mRNA editing are expressed from a single gene through alternative splicing. J Biol Chem. 2002 Apr 12;277(15):12703-9. Epub 2002 Jan 28.

Delannoy E, Stanley WA, Bond CS, Small ID. Pentatricopeptide repeat (PPR) proteins as sequence-specificity factors in post-transcriptional processes in organelles. Biochem Soc Trans. 2007 Dec;35(Pt 6):1643-7.

Eggington JM, Greene T, Bass BL. Predicting sites of ADAR editing in double-stranded RNA. Nat Commun. 2011;2:319. doi: 10.1038/ncomms1324.

a) Feagin JE, Abraham JM, Stuart K. Extensive editing of the cytochrome c oxidase III transcript in Trypanosoma brucei. Cell. 1988 May 6;53(3):413-22.

b) Feagin JE, Shaw JM, Simpson L, Stuart K. Creation of AUG initiation codons by addition of uridines within cytochrome b transcripts of kinetoplastids. Proc Natl Acad Sci U S A. 1988 Jan;85(2):539-43.

Feng Y, Sansam CL, Singh M, Emeson RB. Altered RNA editing in mice lacking ADAR2 autoregulation. Mol Cell Biol. 2006 Jan;26(2):480-8.

Gessler M, Poustka A, Cavenee W, Neve RL, Orkin SH, Bruns GA. Homozygous deletion in Wilms tumours of a zinc-finger gene identified by chromosome jumping. Nature. 1990 Feb 22;343(6260):774-8.

Greger IH, Khatri L, Ziff EB. RNA editing at arg607 controls AMPA receptor exit from the endoplasmic reticulum. Neuron. 2002 May 30;34(5):759-72.

Grohmann M, Hammer P, Walther M, Paulmann N, Büttner A, Eisenmenger W, Baghai TC, Schüle C, Rupprecht R, Bader M, Bondy B, Zill P, Priller J, Walther DJ. Alternative splicing and extensive RNA editing of human TPH2 transcripts. PLoS One. 2010 Jan 29;5(1):e8956. doi: 10.1371/journal.pone.0008956.

Gualberto JM, Lamattina L, Bonnard G, Weil JH, Grienenberger JM. RNA editing in wheat mitochondria results in the conservation of protein sequences. Nature. 1989 Oct 19;341(6243):660-2.

Harding HP, Zhang Y, Zeng H, Novoa I, Lu PD, Calfon M, Sadri N, Yun C, Popko B, Paules R, Stojdl DF, Bell JC, Hettmann T, Leiden JM, Ron D. An integrated stress response regulates amino acid metabolism and resistance to oxidative stress. Mol Cell. 2003 Mar;11(3):619-33.

Hensgens LA, Brakenhoff J, De Vries BF, Sloof P, Tromp MC, Van Boom JH, Benne R. The sequence of the gene for cytochrome c oxidase subunit I, a frameshift containing gene for cytochrome c oxidase subunit II and seven unassigned reading frames in Trypanosoma brucei mitochondrial maxi-circle DNA. Nucleic Acids Res. 1984 Oct 11;12(19):7327-44.

Hiesel R, Wissinger B, Schuster W, Brennicke A. RNA editing in plant mitochondria. Science. 1989 Dec 22;246(4937):1632-4.

Himeji D, Horiuchi T, Tsukamoto H, Hayashi K, Watanabe T, Harada M. Characterization of caspase-8L: a novel isoform of caspase-8 that behaves as an inhibitor of the caspase cascade. Blood. 2002 Jun 1;99(11):4070-8.

Hoch B, Maier RM, Appel K, Igloi GL, Kössel H. Editing of a chloroplast mRNA by creation of an initiation codon. Nature. 1991 Sep 12;353(6340):178-80.

Hospattankar AV, Fairwell T, Meng M, Ronan R, Brewer HB Jr. Identification of sequence homology between human plasma apolipoprotein B-100 and apolipoprotein B-48. J Biol Chem. 1986 Jul 15;261(20):9102-4.

Jepson JE, Reenan RA. RNA editing in regulating gene expression in the brain. Biochim Biophys Acta. 2008 Aug;1779(8):459-70. Epub 2007 Dec 3.

Kane JP, Hardman DA, Paulus HE. Heterogeneity of apolipoprotein B: isolation of a new species from human chylomicrons. Proc Natl Acad Sci U S A. 1980 May;77(5):2465-9.

Kaneko M, Nomura Y. ER signaling in unfolded protein response. Life Sci. 2003 Dec 5;74(2-3):199-205.

Kawahara Y, Megraw M, Kreider E, Iizasa H, Valente L, Hatzigeorgiou AG, Nishikura K. Frequency and fate of microRNA editing in human brain. Nucleic Acids Res. 2008 Sep;36(16):5270-80. doi: 10.1093/nar/gkn479. Epub 2008 Aug 6.

Kim U, Wang Y, Sanford T, Zeng Y, Nishikura K. Molecular cloning of cDNA for double-stranded RNA adenosine deaminase, a candidate enzyme for nuclear RNA editing. Proc Natl Acad Sci U S A. 1994 Nov 22;91(24):11457-61.

Klimek-Tomczak K, Mikula M, Dzwonek A, Paziewska A, Karczmarski J, Hennig E, Bujnicki JM, Bragoszewski P, Denisenko O, Bomsztyk K, Ostrowski J. Editing of hnRNP K protein mRNA in colorectal adenocarcinoma and surrounding mucosa. Br J Cancer. 2006 Feb 27;94(4):586-92.

Köhler M, Burnashev N, Sakmann B, Seeburg PH. Determinants of Ca2+ permeability in both TM1 and TM2 of high affinity kainate receptor channels: diversity by RNA editing. Neuron. 1993 Mar;10(3):491-500.

Kotera E, Tasaka M, Shikanai T. A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. Nature. 2005 Jan 20;433(7023):326-30.

Kume H, Hino K[2], Galipon J[2], Ui-Tei K[3]. A-to-I editing in the miRNA seed region regulates target mRNA selection and silencing efficiency. Nucleic Acids Res. 2014 Sep;42(15):10050-60. doi: 10.1093/nar/gku662. Epub 2014 Jul 23.

Lau PP, Xiong WJ, Zhu HJ, Chen SH, Chan L. Apolipoprotein B mRNA editing is an intranuclear event that occurs posttranscriptionally coincident with splicing and polyadenylation. J Biol Chem. 1991 Oct 25;266(30):20550-4.

Lee AH, Iwakoshi NN, Glimcher LH. XBP-1 regulates a subset of endoplasmic reticulum resident chaperone genes in the unfolded protein response. Mol Cell Biol. 2003 Nov;23(21):7448-59.

Lee AH, Chu GC, Iwakoshi NN, Glimcher LH. XBP-1 is required for biogenesis of cellular secretory machinery of exocrine glands. EMBO J. 2005 Dec 21;24(24):4368-80. Epub 2005 Dec 15.

van Leeuwen FW, de Kleijn DP, van den Hurk HH, Neubauer A, Sonnemans MA, Sluijs JA, Köycü S, Ramdjielal RD, Salehi A, Martens GJ, Grosveld FG, Peter J, Burbach H, Hol EM. Frameshift mutants of beta amyloid precursor protein and ubiquitin-B in Alzheimer's and Down patients. Science. 1998 Jan 9;279(5348):242-7.

Lehmann KA, Bass BL. Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. Biochemistry. 2000 Oct 24;39(42):12875-84.

Leung SS, Koslowsky DJ. Mapping contacts between gRNA and mRNA in trypanosome RNA editing. Nucleic Acids Res. 1999 Feb 1;27(3):778-87.

Lev-Maor G, Sorek R, Levanon EY, Paz N, Eisenberg E, Ast G. RNA-editing-mediated exon evolution. Genome Biol. 2007;8(2):R29.

Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallegger M, Shemesh R, Fligelman ZY, Shoshan A, Pollock SR, Sztybel D, Olshansky M, Rechavi G, Jantsch MF. Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat Biotechnol. 2004 Aug;22(8):1001-5. Epub 2004 Jul 18.

Li JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E, Zhang K, Gao Y, Church GM. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science. 2009 May 29;324(5931):1210-3. doi: 10.1126/science.1170995.

Lin JH, Li H, Yasumura D, Cohen HR, Zhang C, Panning B, Shokat KM, Lavail MM, Walter P. IRE1 signaling affects cell fate during the unfolded protein response. Science. 2007 Nov 9;318(5852):944-9.

Lomeli H, Mosbacher J, Melcher T, Höger T, Geiger JR, Kuner T, Monyer H, Higuchi M, Bach A, Seeburg PH. Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. Science. 1994 Dec 9;266(5191):1709-13.

Lu B, Hanson MR. A single homogeneous form of ATP6 protein accumulates in petunia mitochondria despite the presence of differentially edited atp6 transcripts. Plant Cell. 1994 Dec;6(12):1955-68.

Lu B, Wilson RK, Phreaner CG, Mulligan RM, Hanson MR. Protein polymorphism generated by differential RNA editing of a plant mitochondrial rps12 gene. Mol Cell Biol. 1996 Apr;16(4):1543-9.

Luciano DJ, Mirsky H, Vendetti NJ, Maas S. RNA editing of a miRNA precursor. RNA. 2004 Aug;10(8):1174-7.

Lurin C, Andrés C, Aubourg S, Bellaoui M, Bitton F, Bruyère C, Caboche M, Debast C, Gualberto J, Hoffmann B, Lecharny A, Le Ret M, Martin-Magniette ML, Mireau H, Peeters N, Renou JP, Szurek B, Taconnat L, Small I. Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. Plant Cell. 2004 Aug;16(8):2089-103. Epub 2004 Jul 21.

Ma Y, Hendershot LM. ER chaperone functions during normal and stress conditions. J Chem Neuroanat. 2004 Sep;28(1-2):51-65.

Malloy MJ, Kane JP, Hardman DA, Hamilton RL, Dalal KB. Normotriglyceridemic abetalipoproteinemia. absence of the B-100 apolipoprotein. J Clin Invest. 1981 May;67(5):1441-50.

Marcel YL, Hogue M, Theolis R Jr, Milne RW. Mapping of antigenic determinants of human apolipoprotein B using monoclonal antibodies against low density lipoproteins. J Biol Chem. 1982 Nov 25;257(22):13165-8.

Maslov DA, Sturm NR, Niner BM, Gruszynski ES, Peris M, Simpson L. An intergenic G-rich region in Leishmania tarentolae kinetoplast maxicircle DNA is a pan-edited cryptogene encoding ribosomal protein S12. Mol Cell Biol. 1992 Jan;12(1):56-67.

Meng JP, Zhang FP, Huhtaniemi I, Pakarinen P. Characterization and developmental expression of a testis-specific adenosine deaminase mRNA in the mouse. J Androl. 1997 Jan-Feb;18(1):88-95.

Niavarani A, Currie E, Reyal Y, Anjos-Afonso F, Horswell S, Griessinger E, Luis Sardina J, Bonnet D. APOBEC3A is implicated in a novel class of G-to-A mRNA editing in WT1 transcripts. PLoS One. 2015 Mar 25;10(3):e0120089. doi: 10.1371/journal.pone.0120089. eCollection 2015.

Nishikura K, Yoo C, Kim U, Murray JM, Estes PA, Cash FE, Liebhaber SA. Substrate specificity of the dsRNA unwinding/modifying activity. EMBO J. 1991 Nov;10(11):3523-32.

Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem. 2010;79:321-49. doi: 10.1146/annurev-biochem-060208-105251.

Nutt SL, Hoo KH, Rampersad V, Deverill RM, Elliott CE, Fletcher EJ, Adams SL, Korczak B, Foldes RL, Kamboj RK. Molecular characterization of the human EAA5 (GluR7) receptor: a high-affinity kainate receptor with novel potential RNA editing sites. Receptors Channels. 1994;2(4):315-26.

Ohlson J, Pedersen JS, Haussler D, Ohman M. Editing modifies the GABA(A) receptor subunit alpha3. RNA. 2007 May;13(5):698-703. Epub 2007 Mar 16.

Okuda K, Nakamura T, Sugita M, Shimizu T, Shikanai T. A pentatricopeptide repeat protein is a site recognition factor in chloroplast RNA editing. J Biol Chem. 2006 Dec 8;281(49):37661-7. Epub 2006 Oct 2.

Oyadomari S, Koizumi A, Takeda K, Gotoh T, Akira S, Araki E, Mori M. Targeted disruption of the Chop gene delays endoplasmic reticulum stress-mediated diabetes. J Clin Invest. 2002 Feb;109(4):525-32.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008 Dec;40(12):1413-5. doi: 10.1038/ng.259. Epub 2008 Nov 2.

Paschen W, Djuricic B. Extent of RNA editing of glutamate receptor subunit GluR5 in different brain regions of the rat. Cell Mol Neurobiol. 1994 Jun;14(3):259-70.

Paschen W, Hedreen JC, Ross CA. RNA editing of the glutamate receptor subunits GluR2 and GluR6 in human brain tissue. J Neurochem. 1994 Nov;63(5):1596-602.

Payne M, Rothwell V, Jasmer DP, Feagin JE, Stuart K. Identification of mitochondrial genes in Trypanosoma brucei and homology to cytochrome c oxidase II in two different reading frames. Mol Biochem Parasitol. 1985 May;15(2):159-70.

Phreaner CG, Williams MA, Mulligan RM. Incomplete editing of rps12 transcripts results in the synthesis of polymorphic polypeptides in plant mitochondria. Plant Cell. 1996 Jan;8(1):107-17.

Piller KJ, Decker CJ, Rusché LN, Sollner-Webb B. Trypanosoma brucei mitochondrial guide RNA-mRNA chimera-forming activity cofractionates with an editing-domain-specific endonuclease and RNA ligase and is mimicked by heterologous nuclease and RNA ligase. Mol Cell Biol. 1995 Jun;15(6):2925-32.

Polson AG, Bass BL. Preferential selection of adenosines for modification by double-stranded RNA adenosine deaminase. EMBO J. 1994 Dec 1;13(23):5701-11.

Powell LM, Wallis SC, Pease RJ, Edwards YH, Knott TJ, Scott J. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. Cell. 1987 Sep 11;50(6):831-40.

Ramaswami G, Lin W, Piskol R, Tan MH, Davis C, Li JB. Accurate identification of human Alu and non-Alu RNA editing sites. Nat Methods. 2012 Jun;9(6):579-81. doi: 10.1038/nmeth.1982. Epub 2012 Apr 4.

Ramaswami G, Li JB. RADAR: a rigorously annotated database of A-to-I RNA editing. Nucleic Acids Res. 2014 Jan;42(Database issue):D109-13. doi: 10.1093/nar/gkt996. Epub 2013 Oct 25.

Rueter SM, Burns CM, Coode SA, Mookherjee P, Emeson RB. Glutamate receptor RNA editing in vitro by enzymatic conversion of adenosine to inosine. Science. 1995 Mar 10;267(5203):1491-4.

Rueter SM, Dawson TR, Emeson RB. Regulation of alternative splicing by RNA editing. Nature. 1999 May 6;399(6731):75-80.

Rusché LN, Cruz-Reyes J, Piller KJ, Sollner-Webb B. Purification of a functional enzymatic editing complex from Trypanosoma brucei mitochondria. EMBO J. 1997 Jul 1;16(13):4069-81.

Sakurai M, Ueda H, Yano T, Okada S, Terajima H, Mitsuyama T, Toyoda A, Fujiyama A, Kawabata H, Suzuki T. A biochemical landscape of A-to-I RNA editing in the human brain transcriptome. Genome Res. 2014 Mar;24(3):522-34. doi: 10.1101/gr.162537.113. Epub 2014 Jan 9.

Sasaki Y, Kozaki A, Ohmori A, Iguchi H, Nagano Y. Chloroplast RNA editing required for functional acetyl-CoA carboxylase in plants. J Biol Chem. 2001 Feb 9;276(6):3937-40. Epub 2000 Nov 14.

Scadden AD, Smith CW. Specific cleavage of hyper-edited dsRNAs. EMBO J. 2001 Aug1;20(15):4243-52.

Scadden AD, O'Connell MA. Cleavage of dsRNAs hyper-edited by ADARs occurs at preferred editing sites. Nucleic Acids Res. 2005 Oct 27;33(18):5954-64. Print 2005.

Scadden AD. The RISC subunit Tudor-SN binds to hyper-edited double-stranded RNA and promotes its cleavage. Nat Struct Mol Biol. 2005 Jun;12(6):489-96. Epub 2005 May 15.

Schmitz-Linneweber C, Small I. Pentatricopeptide repeat proteins: a socket set for organelle gene expression. Trends Plant Sci. 2008 Dec;13(12):663-70. doi: 10.1016/j.tplants.2008.10.001. Epub 2008 Nov 12.

Schuster W, Wissinger B, Unseld M, Brennicke A. Transcripts of the NADH-dehydrogenase subunit 3 gene are differentially edited in Oenothera mitochondria. EMBO J. 1990 Jan;9(1):263-9.

Seeburg PH, Higuchi M, Sprengel R. RNA editing of brain glutamate receptor channels: mechanism and physiology. Brain Res Brain Res Rev. 1998 May;26(2-3):217-29.

Seiwert SD, Heidmann S, Stuart K. Direct visualization of uridylate deletion in vitro suggests a mechanism for kinetoplastid RNA editing. Cell. 1996 Mar 22;84(6):831-41.

Sharma PM, Bowman M, Madden SL, Rauscher FJ 3rd, Sukumar S. RNA editing in the Wilms' tumor susceptibility gene, WT1. Genes Dev. 1994 Mar 15;8(6):720-31.

Shaw JM, Feagin JE, Stuart K, Simpson L. Editing of kinetoplastid mitochondrial mRNAs by uridine addition and deletion generates conserved amino acid sequences and AUG initiation codons. Cell. 1988 May 6;53(3):401-11.

Shaw JM, Campbell D, Simpson L. Internal frameshifts within the mitochondrial genes for cytochrome oxidase subunit II and maxicircle unidentified reading frame 3 of Leishmania tarentolae are corrected by RNA editing: evidence for translation of the edited cytochrome oxidase subunit II mRNA. Proc Natl Acad Sci U S A. 1989 Aug;86(16):6220-4.

Shen J, Prywes R. Dependence of site-2 protease cleavage of ATF6 on prior site-1 protease digestion is determined by the size of the luminal domain of ATF6. J Biol Chem. 2004 Oct 8;279(41):43046-51. Epub 2004 Aug 6.

Shen J, Snapp EL, Lippincott-Schwartz J, Prywes R. Stable binding of ATF6 to BiP in the endoplasmic reticulum stress response. Mol Cell Biol. 2005 Feb;25(3):921-32.

Sommer B, Köhler M, Sprengel R, Seeburg PH. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. Cell. 1991 Oct 4;67(1):11-9.

Stuart KD, Schnaufer A, Ernst NL, Panigrahi AK. Complex management: RNA editing in trypanosomes. Trends Biochem Sci. 2005 Feb;30(2):97-105.

Teng B, Burant CF, Davidson NO. Molecular cloning of an apolipoprotein B messenger RNA editing protein. Science. 1993 Jun 18;260(5115):1816-9.

Trotter JR, Ernst NL, Carnes J, Panicucci B, Stuart K. A deletion site editing endonuclease in Trypanosoma brucei. Mol Cell. 2005 Nov 11;20(3):403-12.

Ulbricht RJ, Emeson RB. One hundred million adenosine-to-inosine RNA editing sites: hearing through the noise. Bioessays. 2014 Aug;36(8):730-5. doi: 10.1002/bies.201400055. Epub 2014 May 30.

Wagner RW, Smith JE, Cooperman BS, Nishikura K. A double-stranded RNA unwinding activity introduces structural alterations by means of adenosine to inosine conversions in mammalian cells and Xenopus eggs. Proc Natl Acad Sci U S A. 1989 Apr;86(8):2647-51.

Wahlstedt H, Ohman M. Site-selective versus promiscuous A-to-I editing. Wiley Interdiscip Rev RNA. 2011 Nov-Dec;2(6):761-71. doi: 10.1002/wrna.89. Epub 2011 Apr 21.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008 Nov 27;456(7221):470-6. doi: 10.1038/nature07509.

Wang IX, So E, Devlin JL, Zhao Y, Wu M, Cheung VG. ADAR regulates RNA editing, transcript stability, and gene expression. Cell Rep. 2013 Nov 14;5(3):849-60. doi: 10.1016/j.celrep.2013.10.002. Epub 2013 Oct 31.

Wintz H, Hanson MR. A termination codon is created by RNA editing in the petunia atp9 transcript. Curr Genet. 1991 Jan;19(1):61-4.

Wu AL, Windmueller HG. Variant forms of plasma apolipoprotein B. Hepatic and intestinal biosynthesis and heterogeneous metabolism in the rat. J Biol Chem. 1981 Apr 25;256(8):3615-8.

Wu D, Lamm AT, Fire AZ. Competition between ADAR and RNAi pathways for an extensive class of RNA targets. Nat Struct Mol Biol. 2011 Sep 11;18(10):1094-101. doi: 10.1038/nsmb.2129.

Yang JH, Sklar P, Axel R, Maniatis T. Editing of glutamate receptor subunit B pre-mRNA in vitro by site-specific deamination of adenosine. Nature. 1995 Mar 2;374(6517):77-81.

Yang W, Chendrimada TP, Wang Q, Higuchi M, Seeburg PH, Shiekhattar R, Nishikura K. Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. Nat Struct Mol Biol. 2006 Jan;13(1):13-21. Epub 2005 Dec 20.

Ye J, Rawson RB, Komuro R, Chen X, Davé UP, Prywes R, Brown MS, Goldstein JL. ER stress induces cleavage of membrane-bound ATF6 by the same proteases that process SREBPs. Mol Cell. 2000 Dec;6(6):1355-64.

Yoshida H, Haze K, Yanagi H, Yura T, Mori K. Identification of the cis-acting endoplasmic reticulum stress response element responsible for transcriptional induction of mammalian glucose-regulated proteins. Involvement of basic leucine zipper transcription factors. J Biol Chem. 1998 Dec 11;273(50):33741-9.

Yoshida H, Matsui T, Yamamoto A, Okada T, Mori K. XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. Cell. 2001 Dec 28;107(7):881-91.

Young SG, Bertics SJ, Curtiss LK, Casal DC, Witztum JL. Monoclonal antibody MB19 detects genetic polymorphism in human apolipoprotein B. Proc Natl Acad Sci U S A. 1986 Feb;83(4):1101-5.

Zhang Z, Carmichael GG. The fate of dsRNA in the nucleus: a p54(nrb)-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. Cell. 2001 Aug 24;106(4):465-75.

Zhang K, Wong HN, Song B, Miller CN, Scheuner D, Kaufman RJ. The unfolded protein response sensor IRE1alpha is required at 2 distinct steps in B cell lymphopoiesis. J Clin Invest. 2005 Feb;115(2):268-81.

Chapter 2: Widespread RNA and DNA Sequence Differences in the Human Transcriptome

Mingyao Li[1†], Isabel X. Wang[8†], Yun Li[6,7], Alan Bruzel[8], Allison L. Richards[4], Jonathan M. Toung[5], Vivian G. Cheung[2,3,8]

The research presented in Chapter 2 has been published as an article:

Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG. Widespread RNA and DNA sequence differences in the human transcriptome. Science. 2011 Jul 1;333(6038):53-8. doi: 10.1126/science.1207018. Epub 2011 May 19.

The work is presented herein with modifications to meet thesis formatting requirements.

Richards AL contributed to the optimization of RDD calling and sequence validation of RDDs (Table 2.2 and Figure S2.3)

Abstract

The transmission of information from DNA to RNA is a critical process. We compared RNA sequences from human B cells of 27 individuals to the corresponding DNA sequences from the same individuals and uncovered more than 10,000 exonic sites where the RNA sequences do not match that of the DNA. All 12 possible categories of discordances were observed. These differences were nonrandom as many sites were found in multiple individuals and in different cell types including primary skin cells and brain tissues. Using mass spectrometry, we detected peptides that are translated from the discordant RNA sequences and thus do not correspond

Departments of Biostatistics and Epidemiology[1], Genetics[2] and Pediatrics[3], Cell and Molecular Biology Graduate Program[4], Genomics and Computational Biology Graduate Program[5], University of Pennsylvania School of Medicine, Philadelphia, PA 19104; Departments of Genetics[6] and Biostatistics[7], University of North Carolina School of Medicine, Chapel Hill, NC 27599; Howard Hughes Medical Institute[8].

exactly to the DNA sequences. These widespread RNA-DNA differences in the human transcriptome provide a yet unexplored aspect of genome variation.

Introduction

DNA carries genetic information that is passed onto messenger RNA (mRNA) and proteins that perform cellular functions and it is assumed that the sequence of mRNA reflects that of the DNA. This assumed precision is important since mRNA serves as the template for protein synthesis. Hence, genetic studies have mostly focused on DNA sequence polymorphism as the basis of individual differences in disease susceptibility. Studies of mRNA and proteins analyze their expression and not sequence differences among individuals.

There are, however, known exceptions to the one-to-one relationship between DNA and mRNA sequences. These include errors in transcription(Libby and Gallant, 1992, Sydow and Cramer, 2009) and RNA-DNA differences that result from RNA editing(Chen et al., 1987, Rowell et al., 1987, Bass and Weintraub, 1988, Li et al., 2009, Athanasiadis et al., 2004). Errors are rare since proof-reading and repair mechanisms ensure the fidelity of transcription(Thomas et al., 1998, Wang et al., 2009, Zenkin et al., 2006). RNA editing is carried out by enzymes that target mRNA post-transcriptionally: ADARs that deaminate adenosine to inosine which is then recognized by the translation machineries as a guanosine (A-to-G), and APOBECs which edit cytidine to uridine (C-to-U). Previously, sequence comparisons and computational predictions have identified many A-to-G editing sites(Li et al., 2009, Athanasiadis et al., 2004, Sakurai et al., 2010, Nishikura, 2010, Levanon, et al., 2004). In contrast, C-to-U changes are rare;

apolipoprotein B is one of a small number of known target genes of human APOBEC1(Conticello, 2008, Chester et al., 2000).

Samples

We obtained sequences of DNA and RNA samples from immortalized B-cells of 27 unrelated Centre d'Etude du Polymorphisme Humain (CEPH)(Dausset et al., 1990) individuals, who are part of the International HapMap(International HapMap Consortium, 2003, International HapMap Consortium, 2005) and the 1000 Genomes(1000 Genomes Project Consortium, 2010) Projects. When we compared the DNA and RNA sequences of the same individuals, we found 28,766 events at over 10,000 exonic sites that differ between the RNA and the corresponding DNA sequences. Each of these differences were observed in at least two individuals; many of these were seen in B-cells, as well as in primary skin cells and brain tissues from a separate set of individuals and in expressed sequence tags from cDNA libraries of various cell types. About 43% of the differences are transversions and therefore cannot be the result of typical deaminase-mediated RNA editing. By mass spectrometry, we also found peptide sequences that correspond to the RNA variant sequences, but not the DNA sequences, suggesting that the RNA forms are translated into proteins.

We compared the DNA and RNA sequences from B-cells of 27 unrelated CEPH individuals (Table S2.1). We chose these samples because much information is available on them including dense DNA genotypes obtained using different technologies(Cann, 1992, Matise et al., 2003). The genomes of B-cells from the CEPH collection are stable as evidenced by

Mendelian inheritance of genetic loci that allowed the construction of microsatellite- to SNP-based human genetic maps(Cann, 1992, Matise et al., 2003). More recently the International HapMap Consortium(International HapMap Consortium, 2003, International HapMap Consortium, 2005) obtained millions of SNP genotypes and the 1000 Genomes Project(1000 Genomes Project Consortium, 2010) sequenced the DNA of these individuals. Comparison of sequence data from these two projects showed high concordance (~99%). Here, we used the DNA genotypes and sequences from the two projects for our analyses. First, we considered sites that are monomorphic in the human genome. A monomorphic site is one where there is no evidence for sequence variation at that locus in dbSNP, the HapMap and the 1000 Genomes Projects. Different studies have analyzed these 27 and hundreds of additional individuals for DNA variants; thus if a site has not been identified as polymorphic, most likely all individuals have the same sequences at these sites. But to be certain, for these sites in the 27 individuals, we compared their DNA sequences from the 1000 Genomes Project with the sequences of the human reference genome and carried out traditional Sanger sequencing(Sanger et al., 1977). To be included in our analysis, we required that each site to be covered by at least four reads in the 1000 Genomes Project and that the sequences from 1000 Genomes are the same as the reference genome. To ensure the integrity of the aliquots of B-cells we used for analyses, we carried out Sanger sequencing of their DNA and found perfect concordance of sequences with data from 1000 Genomes (thus also the reference genome sequences) (Table S2.2). Second, we considered SNPs. For each individual, a SNP locus is included only if it is homozygous and the HapMap as well as the 1000 Genomes projects reported the same sequence. We have high confidence in those sequences since despite using different technologies (microarray-based

genotyping in HapMap and high-throughput sequencing in 1000 Genomes), identical sequences were obtained in the two projects.

We sequenced the RNA of B-cells from the same 27 individuals using high-throughput sequencing technology from Illumina (Bentley et al., 2008). The resulting RNA sequence reads were mapped to the Gencode genes(Harrow et al., 2006) in the reference human genome. In total, we generated ~1.1 billion reads of 50 base pairs (bp) (~41 million reads and 2 Gb sequence per individual), of which ~69% of the reads mapped uniquely to the transcriptome (see Methods in Supporting Material). To be confident of the base calls, for each individual, we focused our analysis on high quality reads (quality score ≥ 25) and sites that are covered by at least 10 uniquely mapped reads. Another study (Montgomery et al., 2010) had carried out RNA-Sequencing of the same individuals but at a lower coverage; at these sites we compared our sequences with those from their study, the concordance rate of the sequences is >99.5%. This is reassuring given that the samples were prepared and sequenced in different laboratories.

Differences between RNA and corresponding DNA sequences

For each of 27 individuals, we compared the mRNA sequences from B-cells with the corresponding DNA sequences (Fig. S2.1). The comparison revealed many sites where the mRNA sequences differ from the corresponding DNA sequences of the same individual. To ensure that these are actual differences and to minimize the chance of sequencing errors, we required that at least 10% of the reads covering a site to be different from the DNA sequence and at least two individuals show the same RNA-DNA difference at the site. We call each

occurrence of a difference between RNA and DNA sequences, an event, and the chromosomal location where such a difference occurs, a site. Each person can contribute an event to the site, thus there could be multiple events at a site.

Among our 27 subjects, we identified 28,766 events where the RNA sequences do not match those of the corresponding DNA sequences. These events are found in 10,210 exonic sites (Table S2.10) in the human genome and reside in 4,741 known genes (*36% of 13,214* genes that are covered by 10 or more RNA-Seq reads in at least one part of the gene, in two or more individuals). With gene orientation information in Gencode, we observed all 12 possible categories of base differences between RNA and its corresponding DNA (Fig. 2.1A). All 12 types of differences were found in each of the 27 samples; the relative proportion of each type is similar across individuals. There are 6,698 A-to-G events which can be the result of deamination by ADAR. There are 1,220 C-to-T differences which can also be mediated by a deaminase. However it is important to note that APOBEC1 and its complementation factor A1CF that deaminate cytidine are not expressed in our B-cells (FPKM(Trapnell et al., 2010)~0 for both genes) thus it is likely that an unknown deaminase or other mechanism is involved. Even for relatively well characterized proteins such as APOBEC1, a recent RNA-Seq study of Apobec1-/- mice uncovered many previously unknown targets(Rosenberg et al., 2011). In addition, we found 12,507 transversions (43%); which cannot result from classic deaminase-mediated editing. Since we do not know the mechanism by which these differences between RNA and DNA sequences arise, we refer to them as RNA-DNA Differences or RDD. An example of RDD is a C-to-A difference on chromosome 12 (at position 54,841,626 bp) in the myosin light chain gene

MYL6 where 16 of our subjects have C/C in their DNA but A/C in their RNA sequences. Another example is an A-to-C difference on chromosome 6 (at position 44,328,823 bp) in the gene, *HSP90AB1,* that encodes a heat shock protein, where 8 individuals who have homozygous A/A DNA genotype but have A/C in their RNA. Additional examples are shown in Table 2.1. These sites where RNA sequences differ from the corresponding DNA sequences appear to be non-random since the identical differences were found in multiple individuals: 8,163 (80.0%) of the sites were found in at least 50% of the informative individuals (i.e. with RNA-Seq coverage ≥10 and DNA-Seq coverage ≥4 at the site). Some sites were found in all or nearly all informative individuals. For example, the DNA sequences of all 19 informative individuals at position 49,369,615 bp of chromosome 3 in the *GPX1* gene are G/G whereas their RNA sequences are G/A. (The remaining individuals were not included because available data did not meet our inclusion criteria; although the data suggest the same RDD in all remaining individuals: G/G in DNA, and G/A in RNA).

RDD in expressed sequence tags

Computational and experimental validations also upheld these observed RNA-DNA differences. First, for 120 sites (10 sites per RDD type; randomly selected and all examples cited in this paper; see Tables 2.1 and S2.3), we looked for evidence of RDD in the human EST database by BLAST alignment(Altschul et al., 1990) and manual inspection of each result. For 81 of the 120 sites, we found EST clones that contain the RDD alleles. The numbers of sites found in human ESTs are similar across different RDD types (average 67.5%, range: 60 to 90%). Second, we examined previously identified A-to-G editing sites (Li et al., 2009). Fourteen of the

A-to-G sites that we identified were found in their data despite the fact that different cell types were studied. Even the levels of editing at these sites are similar between the two studies (see Fig. S2.2). Twelve additional sites were found in both studies but were filtered because they did not meet our selection criteria.

Sanger sequencing of B cells, skin, and brain

Next, we validated our findings experimentally by Sanger sequencing of both DNA and RNA at 12 randomly selected sites in B-cells (2 to 9 individuals/ site), primary skin (foreskin; 8 to 10 individuals/ site) and brain cortex (6 to 10 individuals/ site). We regrew the B-cells from our subjects and extracted DNA and mRNA from the same aliquots of cells. From sequencing the paired DNA and RNA samples and analysis of each chromatogram by two individuals independently, we confirmed 57 events in 11 sites (see Table 2.2, Fig. S2.3). In *EIF2AK2*, in all the 8 individuals whose samples were sequenced, three sites were found within 10 nucleotides (see below). RDD was not found in one site in *NDUFC2*. Sanger sequencing is not very sensitive or quantitative thus we do not expect to validate all sites, especially those with low levels of RDD.

To assess if RDD shows cell type specificity, we looked for evidence of RNA-DNA sequence differences using primary human cells. We studied the same sites as above by Sanger sequencing of DNA and RNA samples from primary skin fibroblasts and brain (cortex) of a separate set of normal individuals (for each site, we examined the DNA and RNA of 6 to 10 samples per cell type). We identified 55 RDD events in primary skin cells and 62 events in brain

cortex (Table 2.2). The results suggest that most sites are shared across cell types (Table 2.2); although there are exceptions, for example, an A-to-G difference in *EIF2AK2* (chr2: 37,181,512) which was only found in B-cells and brain cortex but not in primary skin cells. We also queried the EST database for evidence of RDD (Tables 2.1, S2.3). The RNA alleles are seen in a wide range of tissues from embryonic stem cells to brain and testis; they are also found in tumors such as lung carcinoma and neuroblastoma.

Proteomic evidence for RDD

Validation at the sequence level is important but does not address all concerns such as the difficulty in aligning sequences that are highly similar and errors introduced by enzymes in reverse transcription steps. We believe that such artifacts are unlikely considering the consistent patterns across sequencing methods and the fact that we observed all 12 types of nucleotide differences. An alternate and independent validation would be to ask whether the RNA variants in RDD sites are translated to proteins. To do so, first we searched mass spectrometry data from human ovarian cancer cells(Sodek et al., 2008) and leukemic cells for putative RDD sites. Since the levels of most RDDs are less than 100%, both DNA and the RDD-forms of the mRNAs should be available to be translated (from here-on, we refer to mRNAs that correspond identically to the DNA sequences as DNA-forms and those that contain a RDD as RNA-forms). In the ovarian cancer and leukemic cells, we indeed found examples of proteins with peptides encoded by both DNA and RNA forms of mRNA (Table S2.4). Encouraged by the search results and cognizant of possible genome instability and thus DNA mutations in cancer cells, we carried out mass spectrometry analysis of our B-cells.

We analyzed the proteome of our B-cells using liquid chromatography-tandem mass spectrometry and detected peptides for 3,217 proteins. Despite advances in mass spectrometry, far less than 50% of peptides can be detected in most studies(Michalski et al., 2011, de Godoy et al., 2006). We identified 327 peptides that cover RDD sites: 299 of them are encoded by the DNA-forms and 28 by RNA-forms of RDD containing mRNAs (FDR<1%; Table S2.5 and S2.9). For 17 RDD sites, peptides that correspond to both DNA and RNA forms were identified (Table 2.3). By BLAST alignment, we ensured that these 28 peptides are unique to the genes that contain the RDD sites. In addition, we sequenced the DNA of the B-cells used for mass spectrometry and validated that the DNA sequences are the same as the reference genome but differ from the RNA sequences and thus do not encode the RNA-forms of the peptides (Table S2.2). It is easier to detect more abundant proteins by mass spectrometry; for most RDD sites, the unaltered DNA forms are more abundant than variant RNA forms of mRNA (see below) thus it is not surprising to find more peptides that corresponds to the DNA rather than the RNA sequences. However, the counts of peptides corresponding to the DNA and RNA forms of RDD sites should not be taken as a measure of the proportions of DNA versus RNA forms of mRNA that are translated because differences in the amino acid sequences of the DNA and RNA forms of the peptides affect the ability of mass spectrometry to detect them. In addition, when a peptide is not detected, it does not mean that it is absent from the proteome, it could be result of sampling.

The proteomic data provide an independent validation that mRNA sequences are not always identical to DNA sequences, and demonstrate that RNA-forms of genes are translated to

proteins. They also show that there are peptides in human cells that are not exactly encoded by the DNA sequences. An example of a protein variant that results from RDD is RPL28 (T-to-A, chr19: 60,590,467). The RDD led to a loss of a STOP codon. We identified peptides corresponding to the 55 amino acid extension of RPL28 protein in the ovarian cancer cells and in our B-cells (Fig. 2.2). Previously identified cases of RNA editing leading to proteins not encoded by genomic DNA, such as the apolipoprotein B(Chen et al., 1987, Powell et al., 1987), serotonin and glutamate receptors(Burns et al., 1997, Lomeli et al., 1994, Maas et al., 2001) in humans and plant ribosomal protein S12(Phreaner et al., 1996), also support our hypothesis that RDD leads to protein isoforms that do not correspond to the DNA sequences of the encoding genes.

Individual variation in abundance of RDD

Using our selection criteria, we found that in each person among the Gencode genes, there are on average 1,065 exonic events that differ in the RNA and DNA sequences. But the number of events varied among individuals (range: 282 to 1,863) by up to 6-fold across our 27 subjects (Fig. 2.1B). The degree of sequence coverage and sequencing errors in DNA or RNA samples do not explain these individual differences*(*Supporting Material). Thus there is likely a biological basis for the individual variation in the number of editing and RDD events. We found no significant correlation between *ADAR* expression with the number of RDDs or the numbers of A-to-G events (P>0.5). Thus, either ADAR expression does not affect the number of editing or RDD events, or our sample size is not sufficient to detect the correlation.

Characteristics of RDD sites

The 10,210 sites which showed RNA and DNA sequence differences are not evenly distributed across the genome: chromosome 19 has the most whereas chromosome 13 has the fewest number of sites. This pattern is observed after correction for differences in size and gene density among chromosomes. RDD sites are significantly ($P<10^{-10}$) enriched in genes that play a role in helicase activity, protein and nucleotide binding (Table S2.6).

We also noted that the 10,210 sites which showed RNA and DNA sequence differences are not evenly distributed within genes. About 44% (4,453 sites) of them are located in coding exons (10% were found in the last exons), 4% (386 sites) are in the 5' UTRs, and 39% (3,977 sites) are in the 3' UTRs (see Table S2.7, those remaining cannot be classified because of differences in gene structures across isoforms). The results suggest that there are more sites in the 3' ends than the 5' ends of genes; a pattern that was also observed in deamination-mediated RNA editing(Rosenberg et al., 2011, Hundley et al., 2008). Seventy-one percent of the coding sites result in non-synonymous amino acid changes, including 2.1% that lead to the gain or loss of a stop codon if translated into proteins. Relative to other structural features in genes, we found 4% of RDD sites are within 2 nucleotides of exon borders and 5% are within 30 nucleotides of poly(A) signals (Table S2.7). Among RDD types, the numbers of sites near splice junctions are quite similar but the numbers near poly(A) sites are more different. C-to-A and G-to-A differences are found more often near poly(A) sites.

Sites also tended to cluster; for example, 2,613 sites (26%) are within 25 bp, and 1,059 sites (10%) are adjacent to each other. Statistical analysis using a runs test supports that the locations of the sites are not random (median P = 0.22). We did not find obvious patterns or associations with motifs shared across the sites, except for the A-to-G and A-to-C differences that show a preference for a cytidine 5' to the adenosine; as previously observed in ADAR mediated A-to-G changes(Athanasiadis et al., 2004, Maas et al., 2001).

RDD levels

We examined the percentage of mRNAs that differs in sequence from the corresponding DNA. For each site to determine the RDD level, we counted the number of reads with a different nucleotide from that in the corresponding DNA sequence. The distribution of the level is bimodal (Fig. 2.1C); the average level is 20% (median = 13%). However, for some sites, RDD was detected in nearly 100% of the RNA sequences such as the A-to-C difference in the gene that encodes an mRNA decapping enzyme, DCP1A (chr3:53297343). This level is correlated with the frequency and types of RNA-DNA differences. Sites found in more than 50% of the informative individuals tend to have higher levels of RNA editing or RDD than other sites ($P < 10^{-5}$; Fig. S2.5). The levels also differ across individuals. For example at a G-to-A site in the gene *RHOT1*, which encodes a RAS protein that plays a role in mitochondrial trafficking (chr17:27526465), in one person, the level was 90% while in another person, it was only 18%. We identified 437 sites with 10 or more informative individuals where the individuals with the highest levels and the lowest levels differ by 2 fold or more (range: 2 to 8.6 fold).

Conclusions

We have uncovered thousands of exonic sites where the RNA sequences do not match those of the DNA sequences; including transitions and transversions. These findings challenge the long-standing belief that in the same individuals, DNA and RNA sequences are nearly identical. To increase the confidence in our results, we obtained the DNA, RNA and protein sequences from different individuals and cell types using a range of technologies (Fig. S2.1b). The samples included cell lines and primary cells from healthy individuals and tumors. We used data from public resources such as EST databases, The HapMap and 1000 Genomes Projects as well as those that we generated with traditional Sanger sequencing, high-throughput sequencing technologies and mass spectrometry. Table 2.4 showed the DNA, RNA and peptide sequences at 15 confirmed sites which illustrate that the RNA and peptide sequences are the same but differ from the corresponding DNA sequences. The results support our observation that in an individual, DNA and RNA sequences from the same cells are not always identical and some of the variant RNA sequences are translated into proteins. The consistent pattern of the observations suggests that the RDDs have biological significance and are not just "noise." At nearly all RDD sites, we observed only one RDD type across cell types and in different individuals. If the DNA sequence is A/A, and the RNA is A/C in one sample, in other samples, we see the same A-to-C difference, but not other types of differences. These results suggest that there are unknown aspects of transcription and/or post-transcriptional processing of RNA. These differences may now be studied along with those in other genomes and organisms such

as the mitochondrial genomes of trypanosomes and chloroplasts of plants, where RNA editing and modifications are relatively common(Phreaner et al., 1996, Hundley et al., 2008).

The underlying mechanisms for these events are largely unknown. For most of the cases, we do not know yet whether a different base was incorporated into the RNA during transcription or if these events occur post-transcriptionally. About 23% of the sites are A-to-G differences; some of these are likely mediated by ADAR, but other, currently unknown, mechanisms can be involved. If it is a co-transcriptional process, then the signal can be in the DNA or the RNA such as secondary structures or modified nucleotides. In addition, as some of the RDDs are found near splice and poly(A) sites; it is possible that this may be a facet of systematic RNA processing steps such as splicing and cleavage(Rueter et al., 1995, Rueter et al., 1999).

Our findings supplement previous studies demonstrating RNA-DNA differences in the human genome, and show that these differences go beyond A-to-G transition. These findings impact our understanding of genetic variation; in addition to DNA sequence variation, we identify individual variation in RNA sequences. For monomorphic DNA sequences that show RDD there is an overall increase in genetic variation. Thus, this variation contributes not only to individual variation in gene expression but also diversifies the proteome since some identified sites lead to nonsynonymous amino acid changes. We speculate that this RNA sequence variation likely affects disease susceptibility and manifestations. To date, mapping studies have focused on identifying DNA variants as disease susceptibility alleles. Our results suggest that the search may need to include RNA sequence variants that are not in the DNA sequences.

References

1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct 28;467(7319):1061-73. doi: 10.1038/nature09534.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10.

Athanasiadis A, Rich A, Maas S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biol. 2004 Dec;2(12):e391. Epub 2004 Nov 9.

Bass BL, Weintraub H. An unwinding activity that covalently modifies its double-stranded RNA substrate. Cell. 1988 Dec 23;55(6):1089-98.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. Accurate whole human genome sequencing

using reversible terminator chemistry. Nature. 2008 Nov 6;456(7218):53-9. doi: 10.1038/nature07517.

Burns CM[1], Chu H, Rueter SM, Hutchinson LK, Canton H, Sanders-Bush E, Emeson RB. Regulation of serotonin-2C receptor G-protein coupling by RNA editing. Nature. 1997 May 15;387(6630):303-8.

Cann HM. CEPH maps. Curr Opin Genet Dev. 1992 Jun;2(3):393-9.

Chen SH, Habib G,Yang CY, Gu ZW, Lee BR, Weng SA, Silberman SR, Cai SJ, Deslypere JP, Rosseneu M, Gotto AM, Li WH, Chan L. Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. Science. 1987 Oct 16;238(4825):363-6.

Chester A, Scott J, Anant S, Navaratnam N. RNA editing: cytidine to uridine conversion in apolipoprotein B mRNA. Biochim Biophys Acta. 2000 Nov 15;1494(1-2):1-13.

Conticello SG. The AID/APOBEC family of nucleic acid mutators. Genome Biol. 2008;9(6):229. doi: 10.1186/gb-2008-9-6-229. Epub 2008 Jun 17.

Dausset J[1], Cann H, Cohen D, Lathrop M, Lalouel JM, White R. Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. Genomics. 1990 Mar;6(3):575-7.

de Godoy LM, Olsen JV, de Souza GA, Li G, Mortensen P, Mann M. Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. Genome Biol. 2006;7(6):R50.

Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R. GENCODE: producing a reference annotation for ENCODE. Genome Biol. 2006;7 Suppl 1:S4.1-9. Epub 2006 Aug 7.

Hundley HA, Krauchuk AA, Bass BL. C. elegans and H. sapiens mRNAs with edited 3' UTRs are present on polysomes. RNA. 2008 Oct;14(10):2050-60. doi: 10.1261/rna.1165008. Epub 2008 Aug 21.

International HapMap Consortium. The International HapMap Project. Nature. 2003 Dec 18;426(6968):789-96.

International HapMap Consortium. A haplotype map of the human genome. Nature. 2005 Oct 27;437(7063):1299-320.

Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallegger M, Shemesh R, Fligelman ZY, Shoshan A, Pollock SR, Sztybel D, Olshansky M, Rechavi G, Jantsch MF. Systematic identification of

abundant A-to-I editing sites in the human transcriptome. Nat Biotechnol. 2004 Aug;22(8):1001-5. Epub 2004 Jul 18.

Li JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E, Zhang K, Gao Y, Church GM. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science. 2009 May 29;324(5931):1210-3. doi: 10.1126/science.1170995.

Libby RT, Gallant JA. The role of RNA polymerase in transcriptional fidelity. Mol Microbiol. 1991 May;5(5):999-1004.

Lomeli H, Mosbacher J, Melcher T, Höger T, Geiger JR, Kuner T, Monyer H, Higuchi M, Bach A, Seeburg PH. Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. Science. 1994 Dec 9;266(5191):1709-13.

Maas S, Patt S, Schrey M, Rich A. Underediting of glutamate receptor GluR-B mRNA in malignant gliomas. Proc Natl Acad Sci U S A. 2001 Dec 4;98(25):14687-92. Epub 2001 Nov 20.

Matise TC, Sachidanandam R, Clark AG, Kruglyak L, Wijsman E, Kakol J, Buyske S, Chui B, Cohen P, de Toma C, Ehm M, Glanowski S, He C, Heil J, Markianos K, McMullen I, Pericak-Vance MA, Silbergleit A, Stein L, Wagner M, Wilson AF, Winick JD, Winn-Deen ES, Yamashiro CT, Cann HM, Lai E, Holden AL. A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. Am J Hum Genet. 2003 Aug;73(2):271-84. Epub 2003 Jul 3.

Michalski A, Cox J, Mann M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. J Proteome Res. 2011 Apr 1;10(4):1785-93. doi: 10.1021/pr101060v. Epub 2011 Feb 28.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010 Apr 1;464(7289):773-7. doi: 10.1038/nature08903. Epub 2010 Mar 10.

Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem. 2010;79:321-49. doi: 10.1146/annurev-biochem-060208-105251.

Phreaner CG, Williams MA, Mulligan RM. Incomplete editing of rps12 transcripts results in the synthesis of polymorphic polypeptides in plant mitochondria. Plant Cell. 1996 Jan;8(1):107-17.

Powell LM, Wallis SC, Pease RJ, Edwards YH, Knott TJ, Scott J. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. Cell. 1987 Sep 11;50(6):831-40.

Rosenberg BR, Hamilton CE, Mwangi MM, Dewell S, Papavasiliou FN. Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs. Nat Struct Mol Biol. 2011 Feb;18(2):230-6. doi: 10.1038/nsmb.1975. Epub 2011 Jan 23.

Rueter SM, Burns CM, Coode SA, Mookherjee P, Emeson RB. Glutamate receptor RNA editing in vitro by enzymatic conversion of adenosine to inosine. Science. 1995 Mar 10;267(5203):1491-4.

Rueter SM[1], Dawson TR, Emeson RB. Regulation of alternative splicing by RNA editing. Nature. 1999 May 6;399(6731):75-80.

Sakurai M, Yano T, Kawabata H, Ueda H, Suzuki T. Inosine cyanoethylation identifies A-to-I RNA editing sites in the human transcriptome. Nat Chem Biol. 2010 Oct;6(10):733-40. doi: 10.1038/nchembio.434. Epub 2010 Sep 12.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977 Dec;74(12):5463-7.

Sodek KL, Evangelou AI, Ignatchenko A, Agochiya M, Brown TJ, Ringuette MJ, Jurisica I, Kislinger T. Identification of pathways associated with invasive behavior by ovarian cancer cells using multidimensional protein identification technology (MudPIT). Mol Biosyst. 2008 Jul;4(7):762-73. doi: 10.1039/b717542f. Epub 2008 Apr 17.

Sydow JF, Cramer P. RNA polymerase fidelity and transcriptional proofreading. Curr Opin Struct Biol. 2009 Dec;19(6):732-9. doi: 10.1016/j.sbi.2009.10.009. Epub 2009 Nov 13.

Thomas MJ, Platas AA, Hawley DK. Transcriptional fidelity and proofreading by RNA polymerase II. Cell. 1998 May 15;93(4):627-37.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 May;28(5):511-5. doi: 10.1038/nbt.1621. Epub 2010 May 2.

Wang D, Bushnell DA, Huang X, Westover KD, Levitt M, Kornberg RD. Structural basis of transcription: backtracked RNA polymerase II at 3.4 angstrom resolution. Science. 2009 May 29;324(5931):1203-6. doi: 10.1126/science.1168729.

Zenkin N, Yuzenkova Y, Severinov K. Transcript-assisted transcriptional proofreading. Science. 2006 Jul 28;313(5786):518-20.

Dedicated to the memory of Dr. Tom Kadesch who gave us important suggestions, taught us salient and subtle points on gene expression, and inspired us with his enthusiasm. Dr. Kadesch died during the preparation of this manuscript. We thank Drs. D. Epstein, H. Kazazian, D.

Puppione and L. Simpson for suggestions and discussions. We thank Drs. C. Gunter, R. Nussbaum and J. Puck for comments on the manuscript, M. Morley for help with data analysis, W. Ankener for sample processing and J. Devlin for results on Sanger sequencing. The mass spectrometry analysis was carried out at the Wistar Proteomic Facility; we thank K. Speicher for help and suggestions. Funded by grants from the National Institutes of Health (to VGC and ML) and support from the Howard Hughes Medical Institute. The RNA-Seq data have been deposited to NCBI GEO under the accession number GSE25840.

Supporting Material available at: www.sciencemag.org/cgi/content/full/science.1207018/DC1.

Table 2.1 Selected examples of sites that show RNA-DNA Differences in B-cells and EST clones.

Gene	Chr	Position (bp)*	Type	No. of informative individuals†^	No. of individuals with RDD^	Average level‡^ [range]	EST
HSP90AB1	6	44,328,823	A-to-C	11	8	0.39 [0.15, 0.79]	BQ355193 (head neck), BX413896 (B cell)
AZIN1	8	103,910,812	A-to-G	17	10	0.22 [0.12, 0.37]	CD359333 (testis), BF475970 (prostate)
CNBP	3	130,372,812	A-to-T	18	16	0.13 [0.10, 0.21]	EL955109 (eye), BJ995106 (hepatoblastoma)
MYL6	12	54,841,626	C-to-A	16	16	0.35 [0.12, 0.60]	EC496428 (prostate), BG030232 (breast adenocarcinoma)
RBM23	14	22,440,217	C-to-G	11	5	0.18 [0.11, 0.35]	BQ232763 (testis, embryonic)
RPL23	17	34,263,515	C-to-T	12	8	0.16 [0.10, 0.22]	BP206252 (smooth muscle), CK128791 (embryonic stem cell)
BLNK	10	97,957,645	G-to-A	14	7	0.14 [0.11, 0.17]	BF972964 (leiomyosarcoma), BE881159 (lung carcinoma)
C17orf70	17	77,117,583	G-to-C	2	2	0.26 [0.24, 0.28]	AA625546 (melanocyte), AA564879 (prostate)
HMGN2	1	26,674,349	G-to-T	7	4	0.22 [0.14, 0.43]	BX388386 (neuroblastoma), BE091398 (breast)
CANX	5	179,090,533	T-to-A	9	8	0.20 [0.13, 0,30]	EL950052, DB558106
EIF3K	19	43,819,430	T-to-C	19	14	0.16 [0.10, 0.27]	AI250201 (ovarian carcinoma), AI345393 (lung carcinoma)
RPL37	5	40,871,072	T-to-G	6	6	0.27 [0.16, 0.45]	CF124792 (T cell), DW459229 (liver)

* hg18 build of the human genome
^ B-cells
† RNA-Seq ≥ 10 reads, DNA-Seq ≥ 4 reads
‡ Calculated by tallying RNA-Seq reads that contain RDD and those that do not.

Table 2.2 Sanger sequencing of RDD sites.

Gene	Chr	Position (bp)[#]	Type	Location	Amino acid change	B-Cells*		Primary Skin Fibroblast*		Brain (cortex)*	
						# informative individuals	# individuals showing RDD	# informative individuals	# individuals showing RDD	# informative individuals	# individuals showing RDD
EIF2AK2	2	37,181,512	A-to-G	3' UTR	Not applicable	8	8	8	0	10	10
	2	37,181,517	A-to-G	3' UTR	Not applicable	8	8	8	3	10	10
	2	37,181,520	A-to-G	3' UTR	Not applicable	8	8	8	3	10	10
	2	37,181,538	A-to-G	3' UTR	Not applicable	8	8	8	6	10	10
AZIN1[†]	8	103,910,812	A-to-G	coding, exonic	S to G	2	2	10	0	9	8
DPP7	9	139,128,755	C-to-T	coding, exonic	Synonymous (P)	9	2	8	1	10	0
PPWD1	5	64894960	G-to-A	coding, exonic	E to K	2	2	8	8	8	8
HLA-DQB2	6	32,833,537	G-to-A	coding exonic	G to S	2	2	10	10	ne	ne
	6	32,833,545	G-to-A	coding, exonic	R to H	2	2	10	10	ne	ne
	6	32,833,550	C-to-T	coding, exonic	Synonymous (I)	2	2	10	10	ne	ne
BLCAP[‡]	20	35,580,977	A-to-G	coding, exonic	Q to R	6	4	10	4	6	6
NDUFC2	11	77,468,303	C-to-G	coding, exonic	L to V	10	0	10	0	10	0

* In all cases, matched DNA and RNA samples from the same individuals were sequenced
[†] Also reported by Li, Levanon et al, 2009(*6*). [‡] Known site that we used as positive control. ne = not expressed

Table 2.3 Peptides encoded by both DNA and RNA-forms of mRNA at RDD sites.

Protein	Position (bp)[#]	RDD	AA change	DNA-form[†]	RNA-form[†]
AP2A2	chr11:976858	T-to-G	Y-to-D	YLALESMCTLASSEFSHEAVK	DLALESMCTLASSEFSHEAVK
DFNA5*	chr7:24705225	T-to-A	L-to-Q	VFPLLLCITLNGLCALGR	VFPQLLCITLNGLCALGR
ENO1	chr1:8848125	T-to-C	L-to-P	EGLELLK	EGPELLK
ENO3	chr17:4800624	T-to-G	V-to-G	LAQSNGWGVMVSHR	LAQSNGWGGMVSHR
FABP3	chr1:31618424	T-to-A	W-to-R	MVDAFLGTWK	MVDAFLGTR
FH*	chr1:239747217	T-to-A	I-to-K	IEYDTFGELK	KEYDTFGELK
HMGB1	chr13:29935772	T-to-A	Y-to-N	MSSYAFFVQTCR	MSSNAFFVQTCR
NACA	chr12:55392932	G-to-A	D-to-N	DIELVMSQANVSR	NIELVMSQANVSR
NSF	chr17:42161411	T-to-C	V-to-A	LLDYVPIGPR	LLDYAPIGPR
POLR2B	chr4:57567852	T-to-A	L-to-Q	IISDGLK	IISDGQK
RAD50*	chr5:131979610	T-to-G	L-to-R	WLQDNLTLR	WRQDNLTLR
RPL12	chr9:129250509	A-to-G	N-to-D	HSGNITFDEIVNIAR	HSGDITFDEIVNIAR
RPL32*	chr3:12852658	G-to-T	A-to-S	AAQLAIR	SAQLAIR
RPS3AP47*	chr4:152243651	C-to-A	T-to-K	EVQTNDLK	EVQKNDLK
SLC25A17	chr22:39520485	A-to-G	E-to-G	TTHMVLLEIIK	TTHMVLLGIIK
TUBA1*	chr2:219823379	A-to-G	E-to-G	EDMAALEK	EDMAALGK
TUBB2C	chr9:139257297	G-to-A	G-to-D	LHFFMPGFAPLTSR	LHFFMPDFAPLTSR

* DNA sequences of these and others were verified by Sanger sequencing (see Table S2).
^RDD in *RPL28* leads to the loss of a stop codon, the resulting additional peptides are found only in the RNA-form of the mRNA.
hg 18 build of the human genome
† For each peptide, the amino acid that differs between the DNA and RNA forms are underlined.

Table 2.4. Corresponding DNA, RNA and peptide sequences at selected sites.

RDD	Gene	Location	DNA*†	RNA†	Peptide (DNA-form,LC-MS/MS)	Peptide (RNA-form, LC-MS/MS)
TtoG	*CD22*	chr19:40514815	C**T**G	C**G**G	ND	MHLLGPWLLL**R**
TtoA	*DFNA5*	chr7:24705225	C**T**G	C**A**G	VFP**L**LLCITLNGLCALGR	VFP**Q**LLCITLNGLCALGR
TtoC	*ENO1*	chr1:8848125	C**T**G	C**C**G	EG**L**ELLK	EG**P**ELLK
TtoA	*FH*	chr1:239747217	A**T**A	A**A**A	**I**EYDTFGELK	**K**EYDTFGELK
TtoA	*HMGB1*	chr13:29935772	**T**AT	**A**AT	MSS**Y**AFFVQTCR	MSS**N**AFFVQTCR
AtoC	*HMGB1*	chr13:29935469	AA**A**	AA**C**	ND	TMSAKE**N**
AtoC	*ITPR3*	chr6:33755773	G**A**C	G**C**C	ND	DGVEDHSPLMYHISLV**A**LLAACAEGK
TtoG	*RAD50*	chr5:131979610	C**T**A	C**G**A	W**L**QDNLTLR	W**R**QDNLTLR
GtoT	*ROD1*	chr9:114026264	G**G**A	G**T**A	ND	NLFIEA**V**CSVK
GtoT	*RPL32*	chr3:12852658	**G**CT	**T**CT	**A**AQLAIR	**S**AQLAIR
AtoG	*RPS25P8*	chr11:118393375	**A**AC	**G**AC	ND	EVP**D**YK
CtoA	*RPS3AP47*	chr4:152243651	A**C**A	A**A**A	EVQ**T**NDLK	EVQ**K**NDLK
GtoT	*SUPT5H*	chr19:44655806	CA**G**	CA**T**	ND	TPMYGSQTPL**H**DGSR
TtoC	*TOR1AIP1*	chr1:178144365	**T**CA	**C**CA	ND	QPSVLS**P**GYQK
AtoG	*TUBA1*	chr2:219823379	G**A**G	G**G**G	EDMAAL**E**K	EDMAAL**G**K

* DNA sequences are monomorphic according to dbSNP, 1000 Genomes and HapMap Projects; all individuals should have the reference allele. We verified this by Sanger sequencing of the B-cells used for mass spectrometry.
† RDD sites are underlined.
LC-MS/MS = liquid chromatography and tandem mass spectrometry
ND = not detected by mass spectrometry; however this does not mean that the peptides are absent in the B-cell proteome. It is likely a result of sampling.

Figure 2.1

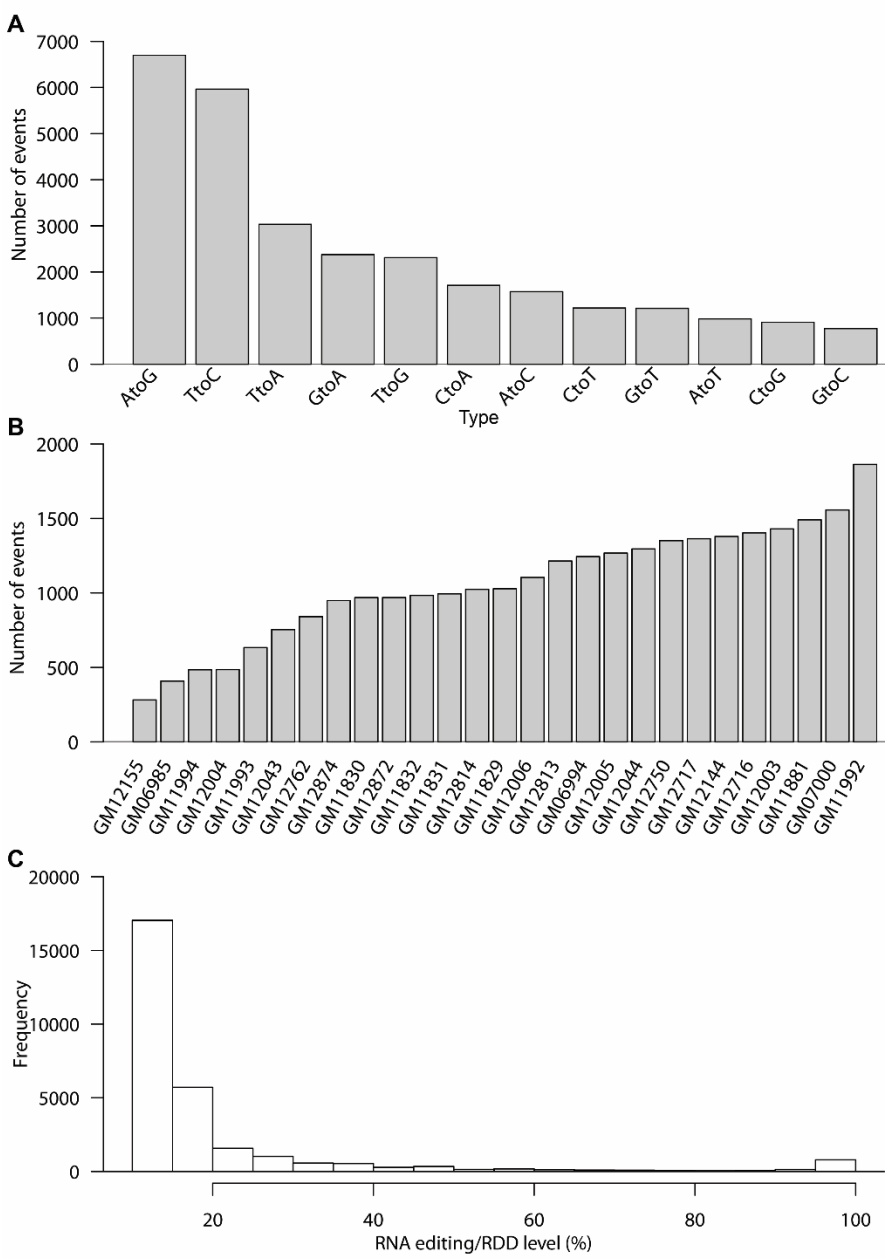

Fig. 2.1. Characteristics of RDD sites. (A) Frequency of the 12 categories of RNA-DNA differences identified in B-cells of 27 normal individuals. (B) Number of RNA editing/ RDD events in 27 normal individuals. (C) Bimodal distribution of the levels of 28,766 RNA editing/ RDD events.

Figure 2.2

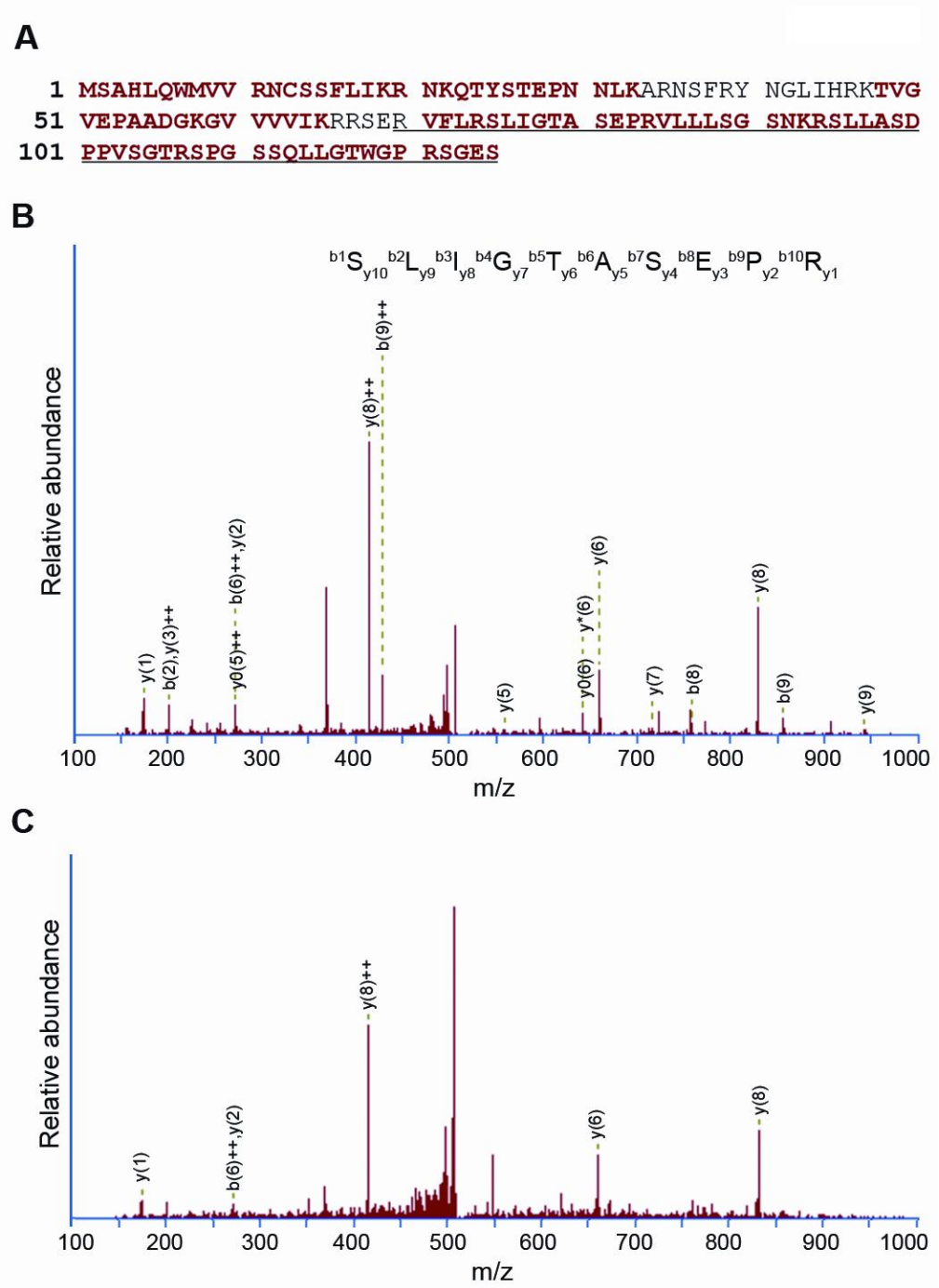

Fig. 2.2. Identification of peptides coded by both RNA and DNA sequences. A) The RNA-form of a RDD leads to loss of a stop codon in RPL28 and extension of 55 amino acids. Peptides detected by mass spectrometry are shown in red. Extended protein sequence due to RDD is underlined. B & C) MS-MS data confirm the detection of peptides encoded by the RDD containing *RPL28* mRNA. The representative spectra of one peptide (SLIGTASEPR) from ovarian cancer cells (B) and cultured B-cells (C) are shown.

Chapter 3: RNA-DNA Differences Are Generated within Seconds After RNA Exits Pol II

Isabel X. Wang[1]*, Leighton J. Core[2]*, Hojoong Kwak[2,4], Lauren Brady[3], Alan Bruzel[1,4], Lee McDaniel[5], Allison L. Richards[6], Ming Wu[4], Grunseich C[7], John T. Lis[2†], Vivian G. Cheung[1,4, 8†]

The research presented in Chapter 3 has been published as an article:

Wang IX, Core LJ, Kwak H, Brady L, Bruzel A, McDaniel L, Richards AL, Wu M, Grunseich C, Lis JT, Cheung VG. RNA-DNA differences are generated in human cells within seconds after RNA exits polymerase II. Cell Rep. 2014 Mar 13;6(5):906-15. doi: 10.1016/j.celrep.2014.01.037. Epub 2014 Feb 20.

The work is presented herein with modifications to meet thesis formatting requirements.

Richards AL contributed to the sequence validation of RDDs and the motif analysis (Table 3.2, Figure S3.2B, Figure S3.4)

Abstract

RNA sequences are expected to be identical to their corresponding DNA sequences. Here, we found all 12 types of RNA-DNA sequence differences (RDDs) in nascent RNA. Our results show that RDDs begin to occur in RNA chains about 55 nucleotides from the RNA polymerase II (Pol II) active site. These RDDs occur so soon after transcription that they are

[1] Life Sciences Institute, University of Michigan, Ann Arbor, MI 48109, USA.
[2] Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA.
[3] Cell and Molecular Biology Graduate Program, University of Pennsylvania, Philadelphia, PA 19104, USA
[4] Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
[5] Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA
[6] Human Genetics Graduate Program, University of Michigan, Ann Arbor, MI 48109, USA.
[7] Neurogenetics Branch, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892, USA.
[8] Departments of Pediatrics and Genetics, University of Michigan, Ann Arbor, MI 48109, USA.
* These authors contributed equally to this work.
†Correspondence to: Vivian G. Cheung (vgcheung@umich.edu), John T. Lis (JohnLis@cornell.edu)

incompatible with known deaminase-mediated RNA editing mechanisms. Moreover, the 55-nucleotide delay in appearance indicates they do not arise during RNA synthesis by Pol II or as a direct consequence of modified base incorporation. Preliminary data suggest a possible couple of RDD formation with R-loops. These findings identify sequence substitution as an early step in co-transcriptional RNA processing.

Introduction

DNA carries instructions for cellular proteins by providing the code that is transcribed into mRNA that in turn is translated into proteins. It is generally assumed that DNA sequences are copied faithfully into RNA. However, there are exceptions to this one-to-one relationship between RNA and its corresponding DNA sequences. The first example of a transcript sequence not encoded by the DNA was reported in 1986 by Benne and colleagues who showed that the *coxII* mRNA in trypanosome has 4 nucleotides not encoded in the DNA. They then coined the term RNA editing for this “novel mechanism of gene expression” (Benne et al., 1986). Other examples of RNA editing were soon discovered in organisms from plants to metazoans (Cattaneo et al., 1989; Driscoll et al., 1989; Gott et al., 1993; Gualberto et al., 1989). In humans, RNA editing occurs in processes mediated by ADAR (adenosine deaminases that act on RNA) (Bass and Weintraub, 1988) and APOBEC (apolipoprotein B mRNA editing enzymes) (Chen et al., 1987; Powell et al., 1987) families of proteins which lead to A-to-G (adenosine to inosine which is then recognized as guanosine) and C-to-U (cytidine to uridine) changes. Recently, advances in sequencing technologies have enabled deep sequencing of DNA and RNA which allowed us (Li et al., 2011) and others (Alon et al., 2012; Bar-Yaacov et al., 2013; Chen, 2013; Chen et al.,

2012; Ju et al., 2011; Lagarrigue et al., 2013; Peng et al., 2012; Silberberg et al., 2012; Vesely et al., 2012) to uncover more RNA-DNA sequence differences (RDDs) than canonical RNA editing events. In different human cells and by using various sequencing and analytical methods, we and others have found all 12 types of RDDs.

While the mechanisms that mediate A-to-G and C-to-U editing in humans are known, we do not know how the other types of RDDs arise. For instance, A-to-C transversions are not likely to be mediated by ADAR and APOBEC families of deaminases. In this project, we ask when RDDs arise in order to distinguish the different types of underlying mechanisms. To address this, we compared nascent RNA sequences with their corresponding DNA sequences. The results show that all 12 types of RDDs occurred early during transcription. We found RDDs in transcripts beginning at approximately 55 bases from the active site or approximately 35 bases beyond the exit channel of RNA polymerase II (Pol II). This demonstrates that the RDD events occur by a mechanism distinct from altered base selectivity during catalysis of chain elongation by Pol II; nonetheless, the RNA processing events that mediate RDDs are closely coupled temporally and spatially to transcription in human cells. Given that RDDs emerge so soon after transcription, we studied cells from a patient with autosomal dominant form of juvenile ALS due to mutation in senataxin gene and found suggestive evidence that RDD formation may be coupled to R-loops.

Nascent RNA from GRO-seq and PRO-seq

To determine whether RDDs occur during or after transcription, we sequenced nascent RNA using two global run-on sequencing methods, GRO-seq (Core et al., 2008) and precision run-on sequencing, PRO-seq (Figure 3.1A) (Kwak et al., 2013). We obtained ~ 100 million 100-nucleotide uniquely mapped GRO-seq reads from B-cells of two individuals. For one subject, we carried out two independent PRO-seq experiments and obtained ~60 million uniquely mapped reads in each. Additionally, we isolated and sequenced nascent RNA with an alternate method (Wuarin and Schibler, 1994) for comparison (chromatin-bound RNA-seq) (~190 million uniquely mapped reads). Finally, we carried out mRNA sequencing (mRNA-seq) and obtained ~135 million uniquely mapped RNA-seq reads, and sequenced the corresponding genomic DNA of the two individuals to 30X and 60X coverage.

We began by assessing the distributions of mapped reads from the libraries obtained by these four independent methods. As expected (Core et al., 2008; Kwak et al., 2013), GRO-seq and PRO-seq enriched for sequences near transcription start sites (Figure 3.1B, TSS). This enrichment in mammalian cells is due to promoter proximal pausing (sense strand) and upstream divergent transcription (antisense strand) (Core et al., 2008; Seila et al., 2008). Additionally, GRO-seq and PRO-seq data provide sensitive detection of active transcription units and identify over 9,000 transcriptionally-active genes. To ensure that we are looking at very nascent RNAs, we assessed the extent of splicing in GRO-seq and PRO-seq relative to chromatin-bound nascent RNA and mRNA. While about 20% of the mRNA-seq reads and 5% of the chromatin-bound nascent transcripts cover exon-exon junctions, less than 1% of the GRO-

seq and PRO-seq reads span junctions. These nascent transcripts map throughout transcription units including introns (Core and Lis, 2008; Core et al., 2008; Core et al., 2012), while mRNA-seq libraries are dramatically depleted of introns but enriched in the 3' untranslated regions due to sample preparation for polyadenylated transcripts. These findings support that GRO-seq and PRO-seq correspond to greatly enriched short nascent RNA that is newly synthesized (also referred to as "very nascent RNAs" below), while chromatin-bound RNA represents longer transcripts on average from a later stage (referred to as "nascent RNA"). Figure 3.1C shows representative results for *UVRAG* and *CAPZB* from sequencing nascent and mature RNAs.

We also compared the expression levels of genes in the very nascent and mature mRNAs. The very nascent RNA differs from mature RNA in that the very nascent RNA levels depend on density of transcribing Pol II, while the mRNA levels depend on the rate of both transcription and mRNA decay. However, levels of transcripts in the two are significantly correlated ($r=0.45$, $P<<0.0001$) (Figure 3.1D) with outliers representing very stable or unstable mRNAs.

RDDs in very nascent RNA

Next, we turned to study RDDs in nascent RNA. Defining when RDDs arise during nascent transcription should help rule out or support particular mechanisms by which they are generated. Therefore, we analyzed the RNA sequences and their corresponding DNA sequences to assess how early during transcription the RNA-DNA differences arise. The steps to identify RDDs are shown in Figure 3.2. At sites that are covered by at least 10 uniquely mapped GRO-

seq or PRO-seq reads and 10 monomorphic DNA reads (that contain only one nucleotide type: A, C, G or T), we compared the nascent RNA and corresponding DNA sequences and identified sites where RNA and DNA sequences are discordant. For a site to be identified as a candidate RDD, at least 10% of the GRO-seq or PRO-seq reads at that site (and a minimum of 2 unique reads) has to contain a sequence that differs from the underlying DNA sequences. All the resulting potential RDD sites were further processed in multiple steps to confirm their unique genomic locations. Several additional experiments and analyses were carried out and they confirmed that the RDDs are not due to sequencing errors or mistakes introduced by reverse transcription (see supplemental experimental procedures for details).

The results uncovered 2,806 RDDs in one subject (GM12004), and 2,881 RDDs in the other individual (GM12750) (Tables S3.1 & S3.2). The orientation-specific sequencing allows us to distinguish all 12 possible types of mismatches between DNA and corresponding RNA sequences. In this analysis, we excluded C-to-T RDDs because the use of 5-bromouridine 5'-triphosphate (BrUTP) in GRO-seq may favor this type of misincorporation (Yu et al., 1993). All the 11 remaining types of RDDs were found; C-to-G was the most common in both samples (Figure 3.3A). We analyzed the PRO-seq data in the same way. Except for the 3' most nucleotide, the sequenced RNAs from the PRO-seq sample are made in the cell (as opposed to about half in GRO-seq) thus it gives us longer segments of *in vivo* synthesized RNAs for analysis. We found 23,093 RDD sites out of about 115 million nucleotides screened, corresponding to one to two RDD per 10,000 bases screened and a frequency of ~ 2 X 10^{-4} RDD in the PRO-seq sample (Table S3.3) which is comparable to the frequency of RDD in mRNAs (also 10^{-4}) (Li et al.,

2011). All 12 types of RDDs were identified (Figure 3.3B). Even though both GRO-seq and PRO-seq are global run-on assays coupled with deep sequencing, they are not identical, therefore different numbers of RDDs were detected in the two assays. Unlike GRO-seq, PRO-seq does not use BrUTP; thus miscorporation that favors C-to-T discordance is not a concern; therefore, we included all 12 types of RDDs in our analysis. This added more than 1,700 RDD sites (1,793 C-to-T). In addition, nearly the entire (except one or at most a few bases) PRO-seq transcripts as compared to ~15 to 20% of the GRO-seq transcripts are made in vivo. Together the addition of the C-to-T sites and the longer in vivo synthesized transcripts, allowed us to identify about 8X more RDD sites in PRO-seq than GRO-seq. Despite the differences in the number, the distributions of RDD types are similar between GRO-seq and PRO-seq samples and across different thresholds of coverage and RDD levels (Figure S3.1). This reflects the robustness of our analysis.

Next, we examined RDDs from different experiments for overlaps. As one expects, the overlaps of RDD sites between the run-on experiments are low, since the ability to resample an RDD site in independent run-on assays depends on several parameters, including the density of transcribing Pol II, sequence depth, and RDD levels. GRO-seq and PRO-seq identify RDD sites in nascent RNA sequences that are closely associated (<100 nucleotides) with actively transcribing polymerases. Finding the same RDD event in two independent samples relies on sampling an RDD-bearing transcript bound to actively transcribing polymerases in both experiments; the chance of such occurrence is very low. The RDD identification also depends on sequence depths and the RDD levels (= number of RDD-containing reads/ total number of reads at the site). The

median RDD level among the sites detected in GRO-seq and PRO-seq is 0.24, therefore high coverage (~40x) is needed to obtain 80% of them in replicate samples (Chen, 2013). Nonetheless, 108 RDD sites were found in more than one sample (among the two GRO-seq and one PRO-seq datasets). The RDD sites we found in nascent RNAs were also present at a later stage of transcription. In chromatin-bound transcripts where we have longer transcripts and deeper coverage, we found over 1,000 RDD sites from one of the GRO-seq and/ or PRO-seq libraries. The distributions of these RDD sites are similar to those in GRO-seq and PRO-seq: T-to-G is one of the more abundant types and A-to-T is less frequent. These results show that the RDDs in nascent RNAs can be identified by different assays.

RDD formation occurs within seconds after transcription

To address how early during transcription do RDD events emerge, we first examined the GRO-seq results. As shown in Figure 3.1A, the GRO-seq reads comprise very nascent RNAs transcribed *in vivo* before nuclei isolation and a portion transcribed *in vitro* during the run-on. Since our very nascent RNAs are triple selected for BrU incorporation and we selectively analyzed reads with an identifiable 3'-end of the nascent RNA, the 3'-portion must contain the *in vitro* transcribed RNA and the 5'-portion contains some *in vivo* synthesized RNA. For both B cell samples, the majority of the RDDs are found in the 5' portion of the GRO-seq samples, which is enriched for the *in vivo* made nascent RNA (Figure 3.4A). These represent newly synthesized transcripts that have just exited the actively transcribing polymerase. These findings suggest that RDDs result from transcription-coupled RNA processing steps.

To further refine the time frame for these RDD events, we used PRO-seq to localize more precisely the RDD sites relative to actively transcribing RNA Pol II. In PRO-seq, the *in vitro* run-on assay was allowed only to proceed for one or at most a few nucleotides, thus the 3' ends of the PRO-seq reads mark precisely the locations of the transcriptionally active RNA polymerases in our B-cells. This offers an opportunity to examine nascent RNAs that have just exited the active site of Pol II. We examined where the RDDs were found relative to the actively transcribing Pol II, and as seen in the GRO-seq data, the RDD events occur after the RNA has exited the polymerase (Figure 3.4B). Moreover, the increased precision and accuracy afforded by PRO-seq allowed us to observe the abrupt increase at ~55 nucleotides from the active site of Pol II, corresponding to the sharp increase in RDD events around position 40 of the PRO-seq reads. As depicted in Figure 3.4B, the first ~20 bases from the 3' ends of the reads are nascent RNAs covered by RNA polymerase II, thus, RDD sites begin to appear about 35 bases after the RNA exits the polymerase. To confirm this observation, we repeated a PRO-seq experiment. The results confirmed our finding of an increase in RDD at ~55 nucleotides from the active site of Pol II (Figure S3.3A). In contrast, the RDDs found in mature mRNAs are more uniformly distributed as expected (Figure 3.4C). These results are consistent with those from GRO-seq and demonstrate that RDD events appear to occur very rapidly (within seconds) after the nascent RNA is exposed, and are not occurring in the Pol II active site during the catalytic step of synthesizing RNA.

RDD frequency is lower in cells from a patient with Senataxin mutation

Our findings that RDDs emerge soon after nascent transcripts exit from transcription bubbles suggest the coupling of RDDs with R-loops (White and Hogness, 1977) which also initiate behind RNA polymerase. We therefore examined and found that RDDs are enriched significantly ($P<0.001$) in regions with R-loop forming sequences (Figure 3.4D & S3.3B) (Ginno et al., 2012; Wongsurawat et al., 2012). To study the co-occurrence of RDDs and R-loops, we carried out PRO-seq using cells from a patient with autosomal dominant form of juvenile Amyotrophic Lateral Sclerosis (ALS4) due to a mutation (L389S) in the Senataxin (SETX) gene that encodes a DNA/RNA helicase (Chen et al., 2004). The senataxin protein, SETX, interacts with RNA polymerase II (Chen et al., 2006; Ursic et al., 2004; Yuce and West, 2013) and plays a role in resolving R-loops particularly in transcription pause sites (Mischo et al., 2011; Skourti-Stathaki et al., 2011; Suraweera et al., 2009; Yuce and West, 2013). The mutation at position 389 corresponds to the N-terminus of SETX that interacts with other nuclear proteins including RNA polymerase II (Yuce and West, 2013). We found that there are 50% fewer RDDs in the very nascent RNA of the ALS4 sample; a frequency of 9×10^{-5} compared to 2×10^{-4}. Compared to controls, the RDD sites in the ALS4 sample skewed away from G-bearing transcripts; there are significantly more ($P=0.03$ (t-test)) RDD events that convert G in the DNA to other bases in the RNA (32% vs. 12% G-to-X, where X = A, C, or T (U)). Since R-loops preferentially form around nascent RNA that is G-rich (Roy and Lieber, 2009), this observation suggests the fewer RDDs in the ALS4 sample may be due to less efficient resolution of R-loops. These results encourage further studies to uncover the mechanistic connection of R-loops and RDDs.

A-to-G RDDs in very nascent RNA are not mediated by ADAR

In our B-cells, the only known editing mechanism is ADAR-mediated A-to-G editing (APOBEC1 is not expressed), so we asked if the A-to-G discrepancies in the nascent RNAs can be explained by ADAR proteins. Previously, ADAR-mediated editing was found in nascent RNA of Drosophila (Rodriguez et al., 2012) where nascent RNA was defined as chromatin-bound transcripts. We examined our chromatin-bound transcripts and mature poly-adenylated RNA, and found A-to-G editing events in both fractions, consistent with results in Drosophila. However, we did not find these A-to-G sites in GRO-seq or PRO-seq. For example, from mRNA-seq, we identified 65 A-to-G sites in *POLH*, and 48 of the adenosines were also edited in chromatin-bound RNA; but, none of these A-to-G sites were detected in the nascent RNA from GRO-seq or PRO-seq despite good sequence coverage (Figure S3.3C). For a more comprehensive analysis, we turned to results from several recent studies that have identified over 10,000 A>G editing sites (Bahn et al., 2011; Carmi et al., 2011; Kiran and Baranov, 2010; Li et al., 2009; Peng et al., 2012). None of the RDD sites in GRO-seq overlap with the editing sites reported in those studies. However, there are some A-to-G events in nascent RNA from GRO-seq and PRO-seq, so we compared the features of these A-to-G sites in nascent RNA with those known to be edited by ADAR-mediated deamination. We found that the sequence characteristics of the A-to-G sites in nascent and mature RNAs appear to be different. Most (>95%) of the ADAR-mediated A-to-G sites in polyadenylated mRNAs are found in Alu repeats (Athanasiadis et al., 2004; Chen, 2013), but in contrast, the A-to-G sites in very nascent (GRO- or PRO-seq) RNAs are not in Alu containing regions. In addition, the A-to-G sites in very nascent

RNAs do not have the sequence motif (5' depletion of G (Lehmann and Bass, 2000)) that flanks ADAR-edited adenosines (see Fig S4A) (Wang et al., 2013). The data suggest that there are two distinct classes of A-to-G mismatches; those that are mediated by ADAR, and others that use a separate mechanism occurring on very nascent RNA during transcription.

Other characteristics of RDDs in very nascent RNA

Previous studies of RDDs focused on polyadenylated mRNAs (Bahn et al., 2011; Ju et al., 2011; Li et al., 2011; Peng et al., 2012); the very nascent RNAs in the present study allowed us to assess RDDs in regions such as introns that were spliced out in mature transcripts. Many of the RDDs in very nascent RNAs are found in intronic regions (28%), which could potentially affect downstream RNA processing steps. In addition, nearly half (44%) of the RDDs are intergenic (many of these correspond to gene isoforms with longer 5' and 3' UTRs relative to the REFseq forms). The remaining (28%) are found in exonic regions and evenly divided among coding exons and UTRs (48% and 52%, respectively). As we found previously (Li et al, 2011), unlike SNPs, there is no bias against nonsynonymous changes as ~70% of the coding RDD sites lead to alternate amino acids as predicted by the codon table. We studied the genes that contain RDD sites in nascent RNA and found that they are significantly ($P<10^{-30}$) enriched for roles in regulation and metabolism of nucleic acids and other macromolecules (see Table 3.3).

We also examined the sequences (10 bases) surrounding the RDD sites and showed that sequence context may be important. RDDs with the same DNA base share similar sequence characteristics. In particular, C-to-A and C-to-G, and the G-to-A, G-to-C and G-to-T RDDs share

similar surrounding sequences. The RDDs whose DNA base is C reside in regions that are significantly more C-rich, while RDDs whose DNA base is G reside in regions that are significantly more G-rich than negative controls (Figure S3.4B & C) (t-test, $P< 0.05$). The enrichments of these nucleotides extend in both the 5' and 3' directions. These regions are more C-rich and G-rich, but they are not homopolymer tracts of Cs or Gs (Figure S3.4D). Thus, these are different from the co-transcriptional editing of homopolymer tracts in Ebola (Volchkov et al., 1995) and paramyxoviruses (Cattaneo et al., 1989; Paterson and Lamb, 1990). Additionally, RDDs whose DNA base is C show depletion of G at the base 3' of the RDD, and those whose reference base is G show depletion of C at the base 5' of the RDD. These features may affect the DNA and/or RNA structures, or possibly an RNA/DNA hybrid, which in turn signals for an RDD event as mentioned above.

Conclusions

We presented data from studying where RDDs occur and put them in context of known RNA editing mechanisms. We showed all 12 types of RDDs are found in RNAs that have recently extruded from the RNA Pol II exit channel. The RDD events occurred *in vivo* on transcripts about 35 nucleotides from the exit channel of Pol II. Pol II elongates in mammalian cells at 20 to 60 bases per second (Ardehali and Lis, 2009). Therefore, the RDD events found ~35 bases from the exit channel must occur very shortly after nascent RNA synthesis. Thus, our results indicate that RDDs are likely to occur within a few seconds of RNA synthesis and before classic RNA editing events. RNAs synthesized by RNA polymerase II are quickly modified: 5' caps are added as the RNA end exits the Pol II RNA channel (Rasmussen and Lis, 1993), introns are often spliced co-

transcriptionally (Carrillo Oesterreich et al., 2010; Vargas et al., 2011) and 3'-ends are cleaved and polyadenylated before Pol II terminates transcription (Osheim et al., 2002). Based on knowledge of co-transcriptional processing events and results from the present study, we suggest that RDD occurs soon after the capping of the transcripts and before splicing.

The reason that we looked for timing of RDD is to help us to narrow the search for the underlying mechanisms that mediate its formation. A co-transcriptional event that coincides temporally with RDD formation is the emergence of R-loop (Broccoli et al., 2000; Drolet et al., 1995; Masse and Drolet, 1999). As a preliminary examination of whether there is association between RDD and R-loop, we studied RDDs in very nascent RNA of cells from a juvenile ALS patient with a mutation in the senataxin gene(Chen et al., 2004). The RNA/DNA helicase, senataxin, interacts with RNA polymerase and mediates the resolution of R-loops. We found that the patient has about 50% fewer RDDs in her nascent RNAs. The RDDs seem to be associated with R-loop since there is enrichment in R-loop forming sequences (Ginno et al., 2012) around RDD sites and depletion of G-bearing RDD transcripts in patient. These findings points to possible coupling of RDD and R-loop formations, and encourage further studies to uncover the molecular basis.

GRO-seq and PRO-seq assays allowed us to study very nascent RNA for RDD formation. But these methods also limit us to study sequences that are covered by or immediately adjacent (<100 bases) to actively transcribing polymerases. It is possible that there are other mechanisms, like ADAR-mediated editing, that modify RNA transcripts at a later stage of RNA processing. While our results show that RDD formation occurs very soon after RNA synthesis,

they do not imply that all RDD formations have to occur as early co-transcriptional steps. Additional methods may be needed to identify or rule out existence of other processing steps that modify RNA sequences. Comparison of RNA sequences at different stages of maturity alone will not provide a comprehensive view because the levels of many RDD sites are low (below 30%), therefore the depth of sequencing necessary to conclude that a RDD site is absent in one stage of transcript synthesis but present in subsequent stages is difficult to achieve with current sequencing technologies given the constraints of error rate and cost. However, technologies to isolate RNA from different subcellular compartments and advances in sequence analysis are improving quickly; they soon will allow the tracking of individual transcripts through various processing steps and thus facilitate the determination of whether there are additional events that modify RNA sequences.

In summary, we have identified sequence modification as an early RNA-processing step thus adding to the already complex set of events that add diversity to transcriptomes.

Experimental Procedures

DNA sequencing. Cultured B-cells from two normal individuals in the Centre d'Étude du Polymorphisme Humain database, GM12004 and GM12750, were obtained from Coriell Cell Repositories (NJ, USA). DNA-seq libraries were prepared and sequenced on HiSeq instrument to obtain 60X and 30X coverage, respectively (Illumina).

mRNA-seq and chromatin-bound nascent RNA-seq. For mRNA sequencing, RNA-seq libraries were prepared following Illumina TruSeq RNA sample preparation protocol. Chromatin-bound

nascent RNA was extracted as previously described (Wuarin and Schibler, 1994). The mRNA and chromatin RNA were sequenced on HiSeq instrument.

GRO-seq and PRO-seq. Nuclei were isolated from cultured B cells and GRO-seq libraries were prepared with $5X10^6$ nuclei as described previously (Core et al., 2008, 2012). PRO-seq libraries were prepared as described previously (Kwak et al., 2013). Briefly, 5×10^6 nuclei were added to 2 X Nuclear Run-On (NRO) reaction mixture (10 mM Tris-HCl pH 8.0, 300 mM KCl, 1% Sarkosyl, 5 mM $MgCl_2$, 1 mM DTT, 0.375 mM each of biotin-11-A/C/G/UTP (Perkin-Elmer), 0.8 u/µl RNase inhibitor) and incubated for 3 min at 30˚C. Nascent RNA was extracted and fragmented by base hydrolysis in 0.2 N NaOH on ice for 10~12 min, and neutralized by adding 1 X volume of 1 M Tris-HCl pH 6.8. Fragmented nascent RNA was purified using streptavidin beads, ligated with reverse 3’ RNA adapter (5'p-GAUCGUCGGACUG-UAGAACUCUGAAC-/3’InvdT/), and biotin-labeled products were enriched by another round of streptavidin bead binding and extraction. For 5’ end repair, the RNA products were successively treated with tobacco acid pyrophosphatase (TAP, Epicentre) and polynucleotide kinase (PNK, NEB). 5’ repaired RNA was ligated to reverse 5’ RNA adaptor (5'-CUGAACAAGCAGAAGACGGCAUACGA-3') before being further purified by the third round of streptavidin bead binding and extraction. RNA was reverse transcribed using 25 pmol RT primer (5’AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA-3’). The product was amplified 15±3 cycles and products greater than 150 bp (insert > 70 bp) were PAGE purified before being analyzed by Illumina HiSeq 2500 instrument. Two PRO-seq experiment was carried out, one at

the Lis lab at Cornell University (Figure 3.3&3.4), and one at the Cheung lab at University of Pennsylvania (Figure S3.3A).

Sequence analysis. DNA-seq and RNA-seq reads were aligned to human reference genome (hg18) using GSNAP (Wu and Nacu, 2010) (version 2012-04-10). A list of SNP sites in the CEU population from Hapmap (release #28) and 1000 Genomes (pilot project) was used for SNP-tolerant alignments. Alignments with (read length + 2)/12 – 2 or fewer mismatches were obtained for each read. PRO-seq sequences were converted to the reverse-complements before alignment. For RNA sequence analysis, known exon-exon junctions (defined by RefSeq (downloaded March 7, 2011) and Gencode (version 3c)) and novel junctions (defined by GSNAP) were accepted. Read coverage was analyzed using RSeQC and RPKM (read per kilobase per million reads) for each gene were calculated (Wang et al., 2012). For GRO-seq and PRO-seq, we include all the reads covering exon or intron region in computing RPKM, while excluding 1kb-region downstream of TSS which is overrepresented by short transcripts associated with proximally paused Pol II.

RNA-DNA differences. To identify RDDs, we compared RNA sequence to its corresponding DNA sequence. Low-quality bases (Phred quality score < 20) in both the RNA and DNA were removed. To be included as RDD sites in the final lists, the following criteria had to be met: 1) a minimum of 10 total DNA-seq reads covering that site; 2) DNA sequence at this site is 100% concordant, without any DNA-seq reads containing alternative alleles; 3) a minimum of 10 total RNA-seq reads covering that site; 4) level of RDD (# of RNA-seq reads containing non-DNA allele/# all RNA-seq reads covering a given site) is ≥10% (a minimum of two RNA-seq reads

containing RDD). To ensure the accuracy of the RDD sites, additional filtering steps were performed using two additional mapping algorithms. See supplemental experimental procedures for further details.

References

Alon, S., Mor, E., Vigneault, F., Church, G., Locatelli, F., Galeano, F., Gallo, A., Shomron, N., and Eisenberg, E. (2012). Systematic identification of edited microRNAs in the human brain. Genome Res *22*, 1533-1540.

Ardehali, M.B., and Lis, J.T. (2009). Tracking rates of transcription and splicing in vivo. Nature structural & molecular biology *16*, 1123-1124.

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

Bahn, J., Lee, J., Li, G., Greer, C., Peng, G., and Xiao, X. (2011). Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. Genome Res *29*, 29.

Bar-Yaacov, D., Avital, G., Levin, L., Richards, A., Hachen, N., Rebolledo Jaramillo, B., Nekrutenko, A., Zarivach, R., and Mishmar, D. (2013). RNA-DNA differences in human mitochondria restore ancestral form of 16S ribosomal RNA. Genome Res.

Bass, B.L., and Weintraub, H. (1988). An unwinding activity that covalently modifies its double-stranded RNA substrate. Cell *55*, 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*, 819-826.

Broccoli, S., Phoenix, P., and Drolet, M. (2000). Isolation of the topB gene encoding DNA topoisomerase III as a multicopy suppressor of topA null mutations in Escherichia coli. Mol Microbiol *35*, 58-68.

Carmi, S., Borukhov, I., and Levanon, E.Y. (2011). Identification of widespread ultra-edited human RNAs. PLoS Genet *7*, e1002317.

Carrillo Oesterreich, F., Preibisch, S., and Neugebauer, K.M. (2010). Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. Mol Cell *40*, 571-581.

Cattaneo, R., Kaelin, K., Baczko, K., and Billeter, M.A. (1989). Measles virus editing provides an additional cysteine-rich protein. Cell *56*, 759-764.

Chen, L. (2013). Characterization and comparison of human nuclear and cytosolic editomes. Proc Natl Acad Sci U S A *110*, E2741-2747.

Chen, R., Mias, G.I., Li-Pook-Than, J., Jiang, L., Lam, H.Y., Miriami, E., Karczewski, K.J., Hariharan, M., Dewey, F.E., Cheng, Y.*, et al.* (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell *148*, 1293-1307.

Chen, S.H., Habib, G., Yang, C.Y., Gu, Z.W., Lee, B.R., Weng, S.A., Silberman, S.R., Cai, S.J., Deslypere, J.P., Rosseneu, M.*, et al.* (1987). Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. Science *238*, 363-366.

Chen, Y.Z., Bennett, C.L., Huynh, H.M., Blair, I.P., Puls, I., Irobi, J., Dierick, I., Abel, A., Kennerson, M.L., Rabin, B.A.*, et al.* (2004). DNA/RNA helicase gene mutations in a form of juvenile amyotrophic lateral sclerosis (ALS4). Am J Hum Genet *74*, 1128-1135.

Chen, Y.Z., Hashemi, S.H., Anderson, S.K., Huang, Y., Moreira, M.C., Lynch, D.R., Glass, I.A., Chance, P.F., and Bennett, C.L. (2006). Senataxin, the yeast Sen1p orthologue: characterization of a unique protein in which recessive mutations cause ataxia and dominant mutations cause motor neuron disease. Neurobiology of disease *23*, 97-108.

Core, L., and Lis, J. (2008). Transcription regulation through promoter-proximal pausing of RNA polymerase II. Science *319*, 1791-1792.

Core, L., Waterfall, J., and Lis, J. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845-1848. Epub 2008 Dec 1844.

Core, L.J., Waterfall, J.J., Gilchrist, D.A., Fargo, D.C., Kwak, H., Adelman, K., and Lis, J.T. (2012). Defining the status of RNA polymerase at promoters. Cell reports *2*, 1025-1035.

Driscoll, D.M., Wynne, J.K., Wallis, S.C., and Scott, J. (1989). An in vitro system for the editing of apolipoprotein B mRNA. Cell *58*, 519-525.

Drolet, M., Phoenix, P., Menzel, R., Masse, E., Liu, L.F., and Crouch, R.J. (1995). Overexpression of RNase H partially complements the growth defect of an Escherichia coli delta topA mutant: R-loop formation is a major problem in the absence of DNA topoisomerase I. Proc Natl Acad Sci U S A *92*, 3526-3530.

Ginno, P.A., Lott, P.L., Christensen, H.C., Korf, I., and Chedin, F. (2012). R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol Cell *45*, 814-825.

Gott, J.M., Visomirski, L.M., and Hunter, J.L. (1993). Substitutional and insertional RNA editing of the cytochrome c oxidase subunit 1 mRNA of Physarum polycephalum. J Biol Chem *268*, 25483-25486.

Gualberto, J.M., Lamattina, L., Bonnard, G., Weil, J.H., and Grienenberger, J.M. (1989). RNA editing in wheat mitochondria results in the conservation of protein sequences. Nature *341*, 660-662.

Ju, Y.S., Kim, J.I., Kim, S., Hong, D., Park, H., Shin, J.Y., Lee, S., Lee, W.C., Yu, S.B., Park, S.S.*, et al.* (2011). Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. Nat Genet *43*, 745-752.

Kiran, A., and Baranov, P.V. (2010). DARNED: a DAtabase of RNa EDiting in humans. Bioinformatics *26*, 1772-1776.

Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science *339*, 950-953.

Lagarrigue, S., Hormozdiari, F., Martin, L.J., Lecerf, F., Hasin, Y., Rau, C., Hagopian, R., Xiao, Y., Yan, J., Drake, T.A.*, et al.* (2013). Limited RNA editing in exons of mouse liver and adipose. Genetics *193*, 1107-1115.

Lehmann, K.A., and Bass, B.L. (2000). Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. Biochemistry *39*, 12875-12884.

Li, J.B., Levanon, E.Y., Yoon, J.K., Aach, J., Xie, B., Leproust, E., Zhang, K., Gao, Y., and Church, G.M. (2009). Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science *324*, 1210-1213.

Li, M., Wang, I.X., Li, Y., Bruzel, A., Richards, A.L., Toung, J.M., and Cheung, V.G. (2011). Widespread RNA and DNA sequence differences in the human transcriptome. Science *333*, 53-58.

Masse, E., and Drolet, M. (1999). Escherichia coli DNA topoisomerase I inhibits R-loop formation by relaxing transcription-induced negative supercoiling. J Biol Chem *274*, 16659-16664.

Mischo, H.E., Gomez-Gonzalez, B., Grzechnik, P., Rondon, A.G., Wei, W., Steinmetz, L., Aguilera, A., and Proudfoot, N.J. (2011). Yeast Sen1 helicase protects the genome from transcription-associated instability. Mol Cell *41*, 21-32.

Osheim, Y.N., Sikes, M.L., and Beyer, A.L. (2002). EM visualization of Pol II genes in Drosophila: most genes terminate without prior 3' end cleavage of nascent transcripts. Chromosoma *111*, 1-12.

Paterson, R.G., and Lamb, R.A. (1990). RNA editing by G-nucleotide insertion in mumps virus P-gene mRNA transcripts. Journal of virology *64*, 4137-4145.

Peng, Z., Cheng, Y., Tan, B.C., Kang, L., Tian, Z., Zhu, Y., Zhang, W., Liang, Y., Hu, X., Tan, X.*, et al.* (2012). Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nat Biotechnol *30*, 253-260.

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*, 831-840.

Rasmussen, E.B., and Lis, J.T. (1993). In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. Proc Natl Acad Sci U S A *90*, 7923-7927.

Rodriguez, J., Menet, J.S., and Rosbash, M. (2012). Nascent-seq indicates widespread cotranscriptional RNA editing in Drosophila. Mol Cell *47*, 27-37.

Roy, D., and Lieber, M.R. (2009). G clustering is important for the initiation of transcription-induced R-loops in vitro, whereas high G density without clustering is sufficient thereafter. Mol Cell Biol *29*, 3124-3133.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent transcription from active promoters. Science *322*, 1849-1851.

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

Skourti-Stathaki, K., Proudfoot, N.J., and Gromak, N. (2011). Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. Mol Cell *42*, 794-805.

Suraweera, A., Lim, Y., Woods, R., Birrell, G.W., Nasim, T., Becherel, O.J., and Lavin, M.F. (2009). Functional role for senataxin, defective in ataxia oculomotor apraxia type 2, in transcriptional regulation. Hum Mol Genet *18*, 3384-3396.

Ursic, D., Chinchilla, K., Finkel, J.S., and Culbertson, M.R. (2004). Multiple protein/protein and protein/RNA interactions suggest roles for yeast DNA/RNA helicase Sen1p in transcription, transcription-coupled DNA repair and RNA processing. Nucleic Acids Res *32*, 2441-2452.

Vargas, D.Y., Shah, K., Batish, M., Levandoski, M., Sinha, S., Marras, S.A., Schedl, P., and Tyagi, S. (2011). Single-molecule imaging of transcriptionally coupled and uncoupled splicing. Cell *147*, 1054-1065.

Vesely, C., Tauber, S., Sedlazeck, F.J., von Haeseler, A., and Jantsch, M.F. (2012). Adenosine deaminases that act on RNA induce reproducible changes in abundance and sequence of embryonic miRNAs. Genome Res *22*, 1468-1476.

Volchkov, V.E., Becker, S., Volchkova, V.A., Ternovoj, V.A., Kotov, A.N., Netesov, S.V., and Klenk, H.D. (1995). GP mRNA of Ebola virus is edited by the Ebola virus polymerase and by T7 and vaccinia virus polymerases. Virology *214*, 421-430.

White, R.L., and Hogness, D.S. (1977). R loop mapping of the 18S and 28S sequences in the long and short repeating units of Drosophila melanogaster rDNA. Cell *10*, 177-192.

Wongsurawat, T., Jenjaroenpun, P., Kwoh, C.K., and Kuznetsov, V. (2012). Quantitative model of R-loop forming structures reveals a novel level of RNA-DNA interactome complexity. Nucleic Acids Res *40*, e16.

Wuarin, J., and Schibler, U. (1994). Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. Mol Cell Biol *14*, 7219-7225.

Yu, H., Eritja, R., Bloom, L.B., and Goodman, M.F. (1993). Ionization of bromouracil and fluorouracil stimulates base mispairing frequencies with guanine. J Biol Chem *268*, 15935-15943.

Yuce, O., and West, S.C. (2013). Senataxin, defective in the neurodegenerative disorder ataxia with oculomotor apraxia 2, lies at the interface of transcription and the DNA damage response. Mol Cell Biol *33*, 406-417.

Acknowledgments: We dedicate this paper to the memory of Dr. Tom Kadesch who suggested the collaboration between the Lis and Cheung labs to study mechanisms that underlie RDDs.

We thank Dr. Nancy Zhang for discussions of estimation of sequencing errors, Jonathan Toung for analysis of the sequencing data, and Zhengwei Zhu for data analysis. We thank Dr. Kenneth Fischbeck for discussion of juvenile ALS and patient samples. Part of this work was carried out at the University of Pennsylvania prior to the Cheung lab's move to the University of Michigan. This work is supported by grants from the National Institutes of Health (to VGC MH087384, ES015733, JTL GM25232) and funds from the Howard Hughes Medical Institute (to VGC).

Data: The sequence data have been deposited to the National Center for Biotechnology Information under the accession number GSE38233 (mRNA-seq), GSE39878 (chromatin-bound RNA-seq, GRO-seq and PRO-seq) and ERP001478 (DNA-seq).

Supplemental Material available at: http://www.sciencedirect.com.proxy.lib.umich.edu/science/article/pii/S2211124714000710#app3

Table 3.1. Results of genome-walking confirm that RDDs are in unique regions of the genome.

Genomic location	RDD type	Plus strand		Minus strand	
		Sequence	# clones	Sequence	# clones
chr1:152175284	G-to-A	G	31	G	9
chr6: 107088915	T-to-A	T	19	T	21
chr9:34336911	G-to-A	G	14	G	10
chr11:72079055	T-to-C	T	10	T	10
chr12:100980077	A-to-C	A	11	A	18
chr14:20221257	G-to-T	G	14	G	16
chr16: 2140620	A-to-G	A	2	A	14
chr19: 2427122	C-to-A	C	15	C	1
chr22:42867269	T-to-C	T	13	T	14
chrX:7004437	G-to-T	G	11	G	10

Table 3.2. ddPCR validation of GRO-seq RDDs.

Genomic Location	Gene Name	RDD Type	Feature	Individual	Level in nascent RNA	
					GRO-seq (%)[ı]	ddPCR (%)
1:152175284	*DENND4B**	G>A	Coding exon	GM12004	20	15
				GM12750	0	0
3:197100758	*TNK2*	G>T	Intron	GM12004	0	0
				GM12750	9	9
6:161450890	*MAP3K4**	G>T	Coding exon	GM12004	17	1
				GM12750	0	2
6:37987903	*ZFAND3*	G>C	Intron	GM12004	0	0
				GM12750	7	3
9:34336911	----	G>A	Intergenic	GM12004	8	19
				GM12750	14	27
11:58103493	*ZFP91*	G>C	Coding exon	GM12004	0	0
				GM12750	18	10
11:72079055	*ARAP1**	T>C	Intron	GM12004	0	0
				GM12750	11	7
12:100980077	----	A>C	Intergenic	GM12004	18	17
				GM12750	0	0
16:69880869	*FTSJD1*	C>G	5' UTR	GM12004	9	3
				GM12750	0	0
17:30949447	*AP2B1*	G>C	Coding exon	GM12004	0	0
				GM12750	10	16
18:8628755	*RAB12**	T>C	3' UTR	GM12004	0	0
				GM12750	11	9
17:34815068	*MED1*	G>T	3' UTR	GM12004	0	0
				GM12750	13	10
19:2197783	*SF3A2*	G>C	Coding exon	GM12004	0	0
				GM12750	33	10
X:7004437	*HDHD1A**	G>T	Intron	GM12004	43	50
				GM12750	0	0

* Also found in nuclear RNA fractions of both individuals (Figure S4) except the site in HDHD1A was found in GM12004 but not GM12750.

[ı] We included a few RDD sites with levels <10% in the validations even though the analyses focused on sites whose levels are >10%. As shown, even the sites with lower levels were validated by ddPCR analysis of these same libraries.

Table 3.3. Genes with RDDs in their nascent RNAs are enriched for roles in regulation and metabolism of macromolecules.

GO Term	Examples	P-value
gene expression	*RNF10, ZNF791, KDM2B; DHX9; ELF4*	1.8×10^{-60}
nucleic acid metabolic process	*SP3, MAX, RPS6KA4; PSMD11; UTP23*	6.2×10^{-60}
RNA metabolic process	*RPS24; ELF1; CPEB2; DHX9; NFX1*	2.6×10^{-58}
cellular macromolecule biosynthetic process	*DPF1; SEC14L2; RPL18A; UPF1; HARS*	4.5×10^{-53}
macromolecule biosynthetic process	*ARFRP1;CTBP2; TSG101; GTF3C2; PARP10*	4.3×10^{-51}
regulation of macromolecule metabolic process	*AXIN1; FYN; VCP; SMARCA5; ZNF7*	3.9×10^{-50}
regulation of cellular metabolic process	*BCOR; ELL; MTF1; STAT5A; VPS36*	2.4×10^{-49}
cellular protein metabolic process	*CCT8; TCF3; RNF115; UBE4B; LNX1*	4.9×10^{-49}
regulation of primary metabolic process	*ATG7; CLIP3; YLPM1; CD44; POGK*	8.6×10^{-47}
regulation of nitrogen compound metabolic process	*AGRN; SMARCC1; MOV10; SUMO1; HSPA8*	4.6×10^{-36}

Figure 3.1

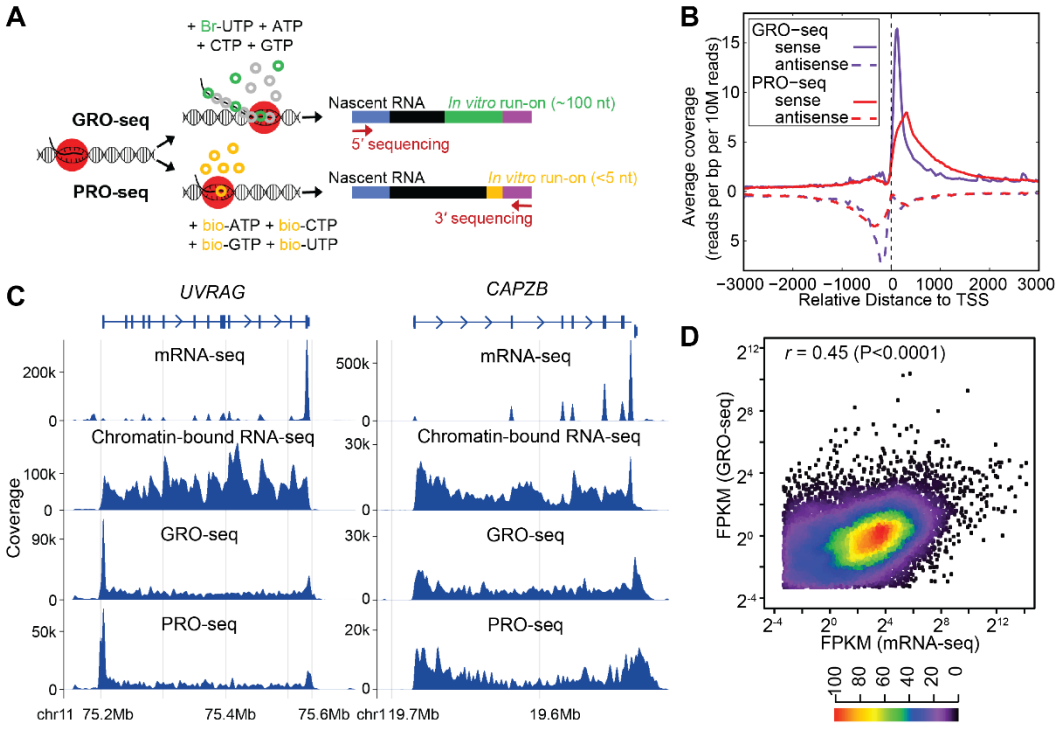

Figure 3.1. GRO-seq and PRO-seq analysis. (**A**) Schematic of GRO-seq and PROseq. (**B**) Comparison between GRO-seq and PRO-seq. Sense and antisense transcripts associated with transcription start sites (TSS) are shown for GRO-seq and PRO-seq samples. The slight shift of the PRO-seq promoter-proximal peak downstream relative to the GRO-seq peak is because the PRO-seq reads that were less than 35 nucleotides were not mapped in the analysis, and because GRO-seq maps 5' ends and PRO-seq maps 3' ends of nascent RNAs. (**C**) mRNA-seq, chromatin-bound nascent RNA-seq, GRO-seq and PRO-seq results for two representative genes, *UVRAG* and *CAPZB*. For genes with proximal Pol II pausing such as *UVRAG*, there are more reads mapping to the 5' ends of genes in both GRO-seq and PRO-seq samples. Schematic gene structure is aligned to mRNA-seq results, with boxes representing exons, lines representing introns and arrowheads showing direction of transcription. Coverage is calculated using bin size of ~ 1500 bp and 600 bp, respectively. (**D**) Scatter plot of gene expression levels from GRO-seq and mRNA-seq (FPKM>0.1). Results from GM12750 (shown) and GM12004 are similar (r=0.45 for both samples). Heatmap indicates frequency of different expression levels.

Figure 3.2

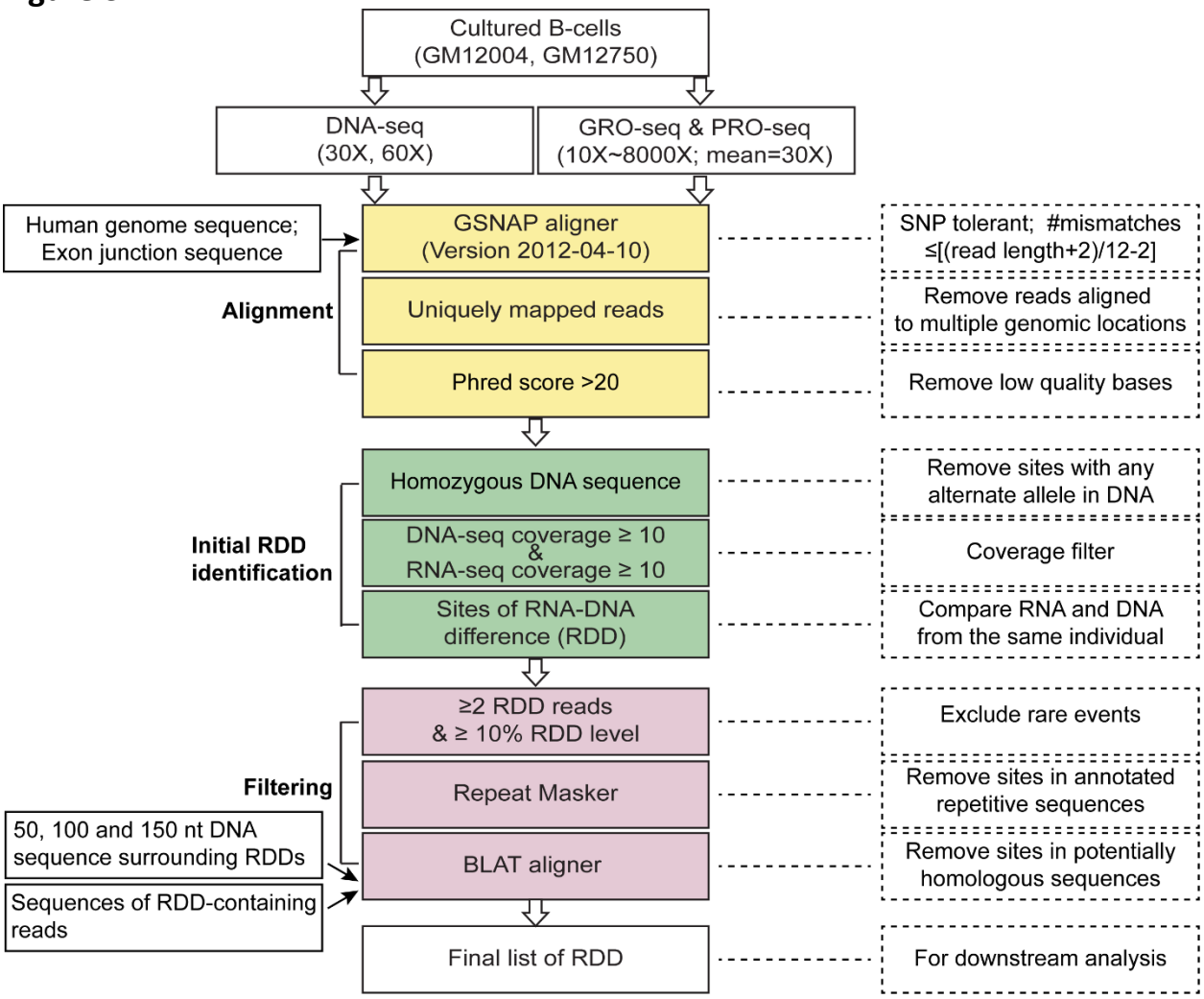

Figure 3.2. Analysis steps to identify RNA-DNA sequence differences. See also Table S1-S3.

Figure 3.3

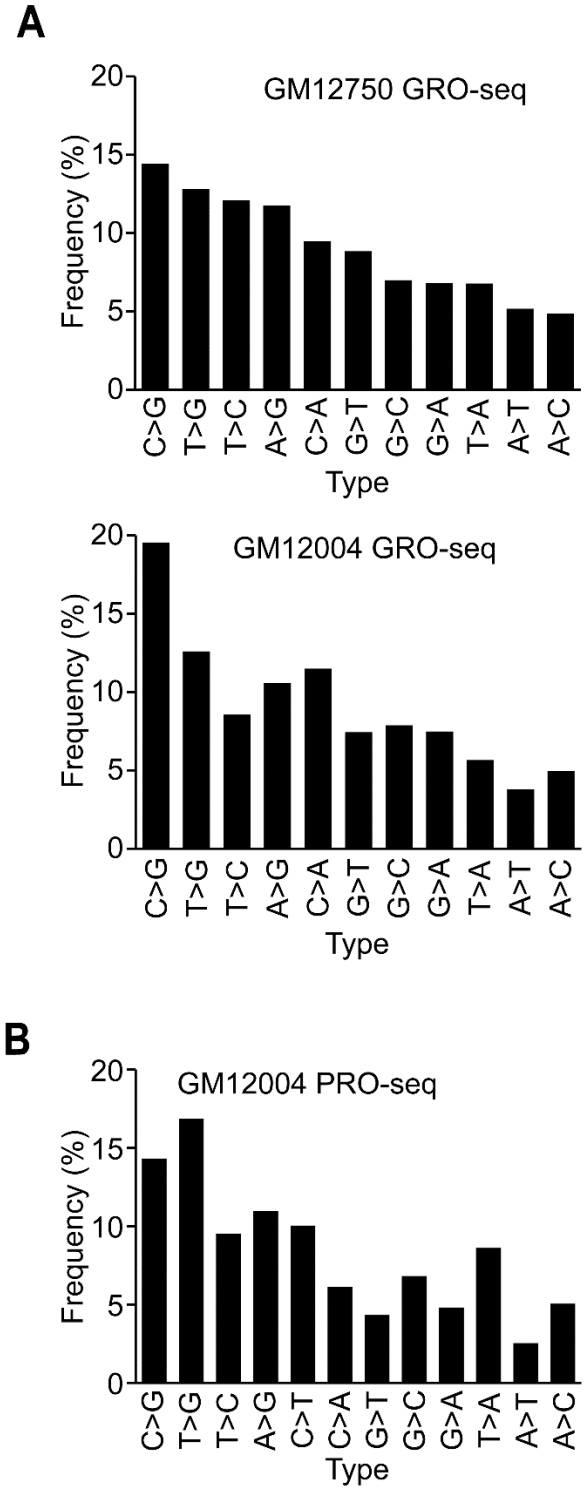

Figure 3.3. RNA-DNA differences in very nascent transcripts. Distributions of RDD types (**A**) in GRO-seq samples of two individuals, (**B**) in PRO-seq. RDD types were ordered as in (**A**) and C-to-T RDDs for the PRO-seq sample.

Figure 3.4

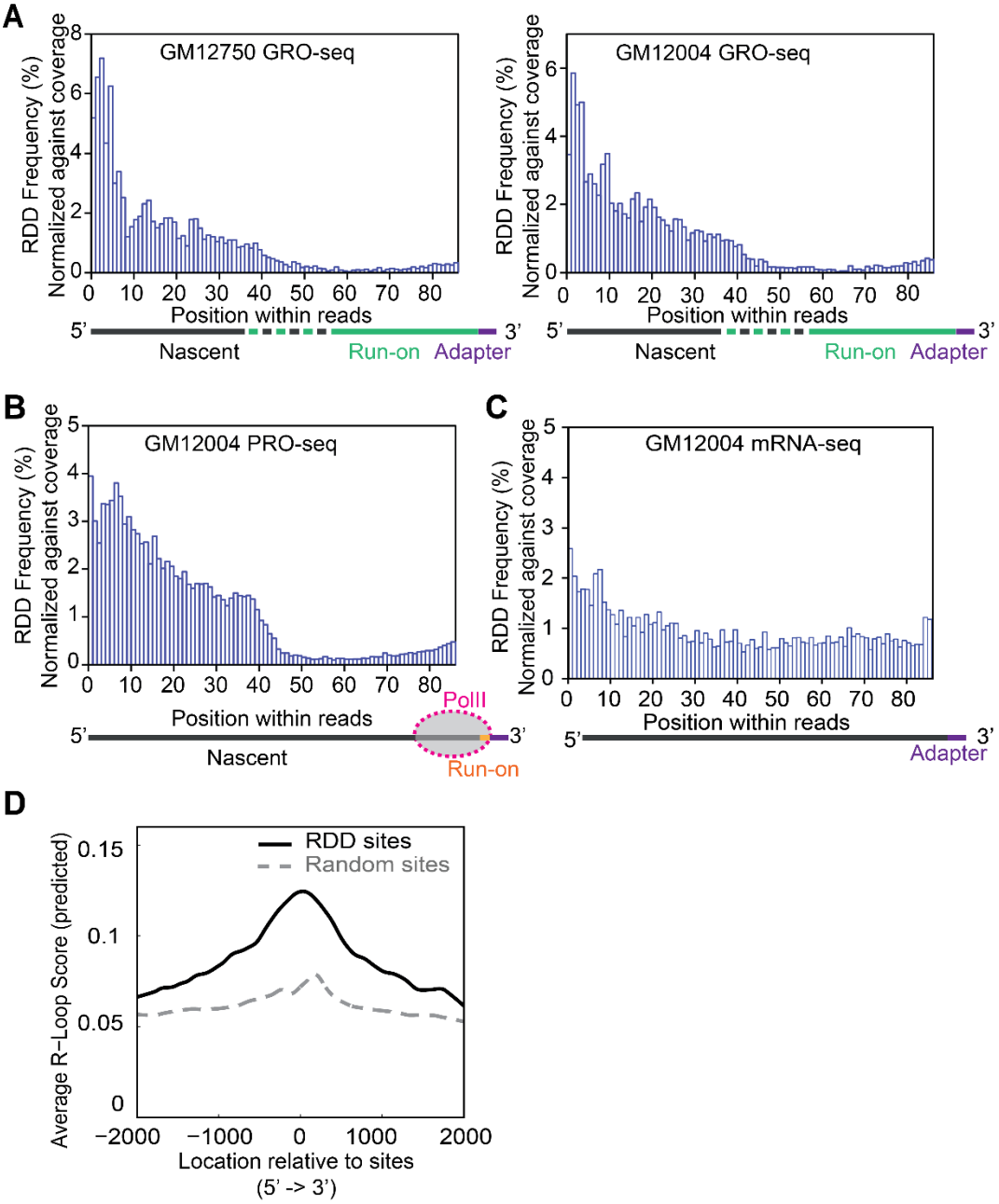

Figure 3.4. Locations of RDD sites within sequencing reads. (**A**) Locations of RDD sites along GRO-seq reads. Only reads that have defined 3' ends (reads that contain 3' adapter sequences) were included in our analysis. (**B**) Locations of RDD sites along PRO-seq reads. Schematic diagrams indicate the locations of the different segments of GRO-seq (**A**) and PRO-seq (**B**) transcripts along the sequence reads. (**C**) Locations of RDD sites along mRNA-seq reads. (**D**) R-loop forming sequences are enriched in regions immediately adjacent to RDD sites. Average R-loop scores for 2 kb of regions up and downstream of RDD sites are shown. RDDs on positive and negative strands are calculated separately. RDD sites have significantly higher R-loop scores ($P<0.001$, t-test) than random control sites shown in Figure S3B.

Chapter 4: *Cis* Regulation of RNA Editing in Human

Abstract

RNA sequences are expected to be identical to the corresponding DNA sequences. Recent studies have found that the RNA sequences do not match their DNA templates at thousands of sites throughout the human transcriptome. Many of these mismatches can be attributed to RNA editing mediated by Adenosine Deaminase Acting on RNA (ADAR). ADAR enzymes deaminate adenosine to inosine which is then read by translational machinery as guanosine. In our analysis, we focus on 8,837 A-to-G editing sites found in B-cells of ten individuals. We found that RNA editing levels vary among individuals. We searched for *cis* variation near these editing sites and found evidence of allelic association with editing levels. We further studied an editing site found in the 3' untranslated region of *SEC16A*, a gene involved in protein transport. Our data suggest that editing of *SEC16A* may influence the binding of proteins involved in RNA transcript stability. This work sheds light on the regulation of editing and the effect of editing on RNA processing.

Introduction

Recent studies have shown that RNA editing and other types of RNA-DNA sequence Differences (RDDs) are widespread throughout the human transcriptome (Li et al., 2011, Ulbricht and Emeson 2014). An abundant type of RNA editing in humans is mediated by Adenosine Deaminase Acting on RNA (ADAR). ADAR enzymes deaminate adenosine to inosine, which is read as guanosine by the protein translation and RNA splicing machineries (Savva et al., 2012). ADAR recognizes double-stranded RNA structures. This preference causes many ADAR editing sites to occur in Alu elements as multiple Alu insertions in a given transcript can lead to a long double-stranded RNA structure (Athanasiadis et al., 2004, Levanon et al., 2004).

The majority of Alu elements, and in turn editing sites, occur in non-coding regions. Specifically, about 90% of ADAR editing sites are found in non-coding regions, such as intronic or untranslated regions (Wang et al., 2013). Studies have demonstrated that RNA editing in non-coding regions can affect RNA processing, such as splicing (Athanasiadis et al., 2004, Lev-Maor et al., 2007) and transport (Zhang and Carmichael 2001).

ADAR edits thousands of sites in the human transcriptome and little is understood about the regulation of editing. Previous studies have demonstrated that ADAR preferentially edits adenosines without a guanosine at the upstream base but with a guanosine at the downstream base (Polson and Bass 1994, Lehmann and Bass 2000, Eggington et al., 2011). While these preferences have been well-established, it is still unclear how sequences located further from the target site may affect the efficiency of ADAR editing. ADAR editing levels vary not only

across sites but also across individuals (Li et al., 2011). This caused us to ask whether the variability in editing levels may be affected by genomic variations acting in *cis*.

In this work, we leveraged individual variation to study the heritability and regulation of RNA editing. We find allelic association of RNA editing with nearby DNA sequence polymorphisms thus suggesting that RNA editing level is genetically regulated. Further, we characterize an editing site in the 3'UTR of *SEC16A*, a gene whose protein product is involved in vesicle transport from the endoplasmic reticulum to the Golgi apparatus (Watson et al., 2006). Our data suggest that editing of this site in *SEC16A* may influence RNA transcript stability through affecting an RNA binding protein. Together, this study demonstrates that RNA editing levels can be heritable and may contribute to individual differences in cellular processes, such as RNA decay.

Identification of RNA editing and RDDs in Human B-cells

In order to identify ADAR A-to-G RNA editing sites and other single-base differences, we utilized RNA-sequencing data from ten B-cell lines derived from unrelated individuals (Figure S4.1). The RNA sequencing data were compared to the reference genome (hg18) to identify positions where the RNA did not match the corresponding DNA sequence. We did not consider sites that were found to be SNPs by dbSNP (build 138), 1000 genomes, or the Exome Sequencing Project. We removed repetitive elements (except for Alu elements due to their high probability of containing ADAR editing sites) and pseudogenes. To call an RNA editing site or RNA-DNA sequence Difference (RDD), we required that there are at least 10 total reads at the position and that 5% of the reads contain the alternative base. If this requirement was met in at

least two of the ten individuals, the site was included in the analysis. We define an RDD to be any single-base difference that cannot be explained by canonical editing in our cells. This means that an RDD in this dataset is any difference except A-to-G. The known cytidine deaminases, APOBEC1 and A1CF, are not expressed in our cells (fragments per kilobase of exon per million fragments mapped (FPKM) ~0 for both genes across B-cells from 10 individuals) so we include C-to-U in our RDD list.

These sites were further queried against the 1000 genome database of 99 CEU individuals and by DNA-sequencing of 4 of our 10 individuals to ensure homozygous DNA sequences. In this way, we identified 8,837 A-to-G RNA editing sites and 3,965 RDD sites (Figure 4.1a).

Characteristics of RNA editing and RDD

To verify that these 8,837 A-to-G editing sites are mediated by ADAR, we looked for well-known characteristics of ADAR editing. First, we found that 90% of our A-to-G editing sites are found in Alu elements, similar to the proportion seen in other studies (Wang et al., 2013). ADAR editing sites are often in non-coding regions. Seventy-seven percent (6,846) of the editing sites are found in 1,387 annotated genes. Of these 6,846 editing sites, 50% are found in introns, 24% are found in 3' UTRs, and less than 1% are found in coding regions. In addition, 66% of the genes that contain editing sites contain more than one editing site. There are an average of 5 A-to-G editing sites per gene. In our dataset, the gene with the most editing sites is *USP4* with 59 editing sites. Furthermore, our A-to-G RNA editing sites show the same motif that has been

found by others to be targeted by ADAR (depletion of 5' G and enrichment of 3' G) (Figure 4.1b).

RNA editing sites have different characteristics compared to RDDs identified from the same RNA sequencing data. Using the same filtering steps, we identified 3,965 RDDs of the other 11 types of single-base differences (not including A-to-G). Only 51% of RDDs are found in Alu elements, compared to the 90% of A-to-G sites. 51% of RDDs are found in 1,339 annotated genes, of which 23% are in introns, 14% in 3' UTRs, and 11% in coding regions. Many more RDDs are found in coding regions compared to A-to-G editing sites. Only 21% of genes containing an RDD show multiple RDDs, with an average of 2 RDDs per gene.

Cis-regulation of RNA editing levels

To study the relationship between editing sites, we focused on sites within genes that had at least five A-to-G editing sites. There are 2,373 A-to-G editing sites in 223 genes that met this criterion. We studied the correlation of editing levels across the ten individuals. For each of the 2,373 sites, we computed pairwise correlations for all sites in the same gene and averaged those correlation coefficients. We then permuted the data (10 permutations), preserving the number of sites per gene, and calculated correlation coefficients for the permuted data. We compared the correlations in the observed data with those from the permutations and found higher correlations among sites in the observed data (Figure 4.2a). Specifically, of the 223 genes studied, 131 genes showed higher correlations within the genes than compared to the randomized set. We narrowed our search to genes with sites that showed correlation coefficients of 0.1 and found that of the 66 genes to meet these criteria, 63 had greater

correlation of sites in the gene than to a random set. These data demonstrate that some genes show a correlation and suggest that sites in the same gene are similarly regulated.

We next wanted to delve further into the regulation of editing level by studying the correlation between sites in the same Alu element versus sites outside of the Alu element but all within the same gene. We focused our analysis on 231 editing sites in 25 genes that contained 2 edited Alu elements. We focused on these genes because of our knowledge of RNA secondary structure. It has been shown that Alu elements are common targets of ADAR editing due to their likelihood of creating double-stranded structures. Multiple Alu elements occurring in close proximity can create double-stranded structures due to high similarity of sequence. Therefore, we focused on the 25 genes with two Alu elements because we can assume that these two elements base-pair with each other to make an ADAR editing target. When we studied the 231 sites in 25 genes we found that there was a higher correlation of editing levels when comparing sites in the same Alu element as compared to other sites in the same gene but not found in the same Alu element (Figure 4.2b). Specifically, we found that 14 out of the 25 genes had sites that showed higher correlation to other sites in the same Alu. While this group of 25 genes has only 2 edited Alu elements, it may have more Alu elements that are unedited and influence the secondary structure. We focused on 8 genes that had only 2 or 3 Alu elements in total in the region and found that 6 out of the 8 genes had higher correlation of sites in the same Alu than between. These findings suggest that sites in the same Alu elements have more similar editing levels. Furthermore, these data allow us to deduce that ADAR editing occurs along a given Alu and thereby along a single side of a double-stranded RNA region rather than on both sides simultaneously.

As an example of this relationship between sites, we focused on one gene, *ATG14*, which has five editing sites spread across two Alu elements in its 3'UTR. Two of its five editing sites are correlated while the other three are correlated to each other, creating two groups of correlated sites. The first two appear in one Alu element and the rest in the other Alu element located in the 3'UTR of *ATG14*. The two Alu elements are predicted to make a long double-stranded RNA secondary structure (Lorenz et al., 2011) with the first Alu on one side of the structure and the second Alu on the opposite side of the hairpin (Figure S4.2). This structure, with the correlations, suggest that sites in the same Alu element have more similar editing levels, and that ADAR edits one side of a hairpin structure at a time. Further study of editing level correlations may aid in the prediction of RNA secondary structure.

Individual variation in RNA editing levels

We found that many A-to-G editing sites show evidence of variability in editing level across individuals. 40% of editing sites show a greater than 5-fold range in editing level across the ten individuals (Figure 4.3a). We therefore utilized this variability to understand more about the regulation of RNA editing.

Due to the large amount of variability in editing level across our ten unrelated individuals, we studied the possible genetic effects that could regulate ADAR editing. We began by studying ten sets of monozygotic twins. RNA was collected from B-cells derived from these 20 individuals. The RNA was then sequenced. Using these data, we studied 4,498 A-to-G editing sites that were covered by at least 5 RNA reads in each of the 20 individuals. We then compared that variation within each twin pair to the variation in editing level between each

twin pair using the intraclass correlation analysis. Using this test, we identified 372 editing sites that show less variation within a twin pair than between pairs (ANOVA p-value < 0.01, ICC > 0.66) (Figure 4.3b-c, Table S4.2). For the sites that have highly similar levels among monozygotic twins, we asked if they are genetically regulated by carrying out association analyses.

Effect of genomic variation on RNA editing

We focused on 372 editing sites that showed significantly (p-value < 0.01) similar editing levels within twin pairs and performed association analyses. Of the 372 editing sites, there are 54 sites that were found to have a previously identified SNP in the same gene that was covered by RNA-sequencing in all 20 individuals. These 54 sites were compared to a total of 51 SNPs in 25 genes. We compared each SNP to any A-to-G editing site in the same gene leading to 116 comparisons.

We identified 11 A-to-G editing sites that showed allelic association of SNP genotype to editing level across the 20 individuals (p-value < 0.05)(Figure 4.3d-e, Table S4.3). These sites show a trend where editing may be affected by nearby sequencing variation. However, of these 11 sites, none reach significance following Bonferroni correction for multiple testing. Additional individuals would give more power to detect these associations.

SEC16A

We next determined if editing in a non-coding region plays a role in RNA processing and gene expression. To this end, we focused on an editing site in the gene *SEC16A.*

SEC16A encodes a protein involved in the marking of endoplasmic reticulum exit sites (ERES). These sites define where a vesicle will bud from the ER in order to transfer its contents to the Golgi apparatus (Watson et al., 2006). The 3'UTR of *SEC16A* contains an editing site (Figure 4.4a). Our first step was to verify that this editing site is mediated through ADAR RNA editing. To test this, we used data from previous work in our lab (Wang et al., 2013) where an ADAR knock-down was performed followed by RNA-sequencing. In these data, editing of this site is reduced from 3% to undetectable following knock-down of ADAR1 but not of ADAR2. This suggests that the editing site in *SEC16A* is indeed a target of ADAR1 RNA editing.

Next we asked if the editing level of this site varies across tissues. We sequenced DNA and RNA from three different tissues (skeletal muscle, brain cortex, and liver) of three unrelated individuals. We verified that the DNA contained no evidence of guanosine at the *SEC16A* editing site in any of the three individuals. By studying the RNA from each of the tissues in the individuals, we found that RNA editing can be found in each of the nine samples (three tissues in three individuals) (Figure S4.3). We found that the editing levels at this site vary across individuals and across tissues. Across all three individuals, skeletal muscle shows the lowest editing levels of editing of the three tissues.

Because the editing site in *SEC16A* is found in the 3'UTR, a region known to regulate RNA processing and decay, we asked if editing affects transcript stability. Previously, data from our lab showed that there is an interaction between ADAR and HuR (Wang et al., 2013), a protein that promotes the stability of mRNA transcripts (Peng et al., 1998). Kishore et al.

performed a CLIP-seq with HuR and identified several HuR binding sites in the vicinity of the RNA editing site in *SEC16A* (Kishore et al., 2011).

To study if editing affects stability of *SEC16A* transcripts, we treated cells with actinomycin D and then collected RNA before treatment and 2, 4, 8, 12, and 24 hours after treatment. Sanger sequencing of these samples showed an increase in editing level over the time course (Figure 4.4b). These data suggest two potential scenarios. First, ADAR could be continuing to edit the pool of RNA and as the pool of RNA is gets smaller (as there is no active transcription), the editing level increases. The second possible reason for these results would be that the edited form of *SEC16A* is more stable and so as RNA is continually degraded, only the most stable transcripts remain, leading to an increased editing level. Since previous work has shown that HuR binds to *SEC16A* transcripts, the second scenario can be explained by HuR binding and stabilizing the edited transcripts.

To investigate this possibility, we performed an RNA immunoprecipitation of HuR to determine whether it bound to the edited region of *SEC16A* in our B-cells. We verified binding of HuR to *SEC16A* (Figure S4.4). Using droplet digital PCR (Bio-Rad), we compare the number of droplets with edited transcripts to the total number of droplets with any *SEC16A* transcript to calculate the editing level of each sample. Using these data, we found that the edited transcript was enriched in the HuR-bound RNA (Figure 4.4c, Figure S4.5). While these data suggest that HuR may interact with edited transcripts of *SEC16A,* further study will help to elucidate whether editing plays a role in transcript stability or if the results following actinomycin treatment are caused by continual ADAR editing.

Conclusions

By studying RNA sequences from ten unrelated individuals, we identified 12,802 locations in the transcriptome where the RNA did not match the corresponding DNA sequence. 8,837 of these sites showed an adenosine in the DNA and a guanosine in the RNA suggesting a canonical ADAR-mediated RNA editing site. The A-to-G editing sites show features typical of ADAR editing, including 90% being located in Alu elements. These sites also have the motif associated with ADAR editing, including enrichment of guanosine downstream of the edited adenosine and depletion of guanosine one base upstream. Once we had identified the ADAR targets, we then characterized further the canonical A-to-G editing sites.

We found that over 30% of A-to-G editing sites show at least a 5-fold difference between the highest and lowest editing levels among the ten individuals. We decided to utilize this variability to learn about the regulation of RNA editing. We found that sites within the same Alu element and the same gene show higher correlation of editing levels.

To delve further into the regulation of editing, we studied ten twin pairs to determine heritability of editing levels. We identified 372 editing sites that had a significant intraclass correlation coefficient (p-value < 0.01), thereby suggesting a heritable factor that regulates gene expression. To search for this heritable factor, we studied single-nucleotide polymorphisms (SNPs) to determine association with editing level. We identified 11 sites that had editing levels associated with SNPs in the same gene across 20 individuals. Our data leads us to conclude that variability in RNA editing level can be affected by genomic variation in cis.

Finally, we decided to examine one editing site more closely. We focused on an editing site in the 3'UTR of *SEC16A*. Editing of this site may help to regulate transcript stability through influencing binding of HuR. This demonstrates a way by which a variably edited site may lead to differences in the cell, such as transcript stability, among individuals.

Further work should be done to expand the number of individuals as this will allow for an increase in power to detect associations and correlations among SNP genotypes and editing levels. Studying the twin sequencing data and identifying sites that show greater similarity in editing level within a twin pair than between suggests that our editing levels are accurately measured. However, replicating the sequencing of each sample would further verify accuracy of editing levels.

Together, this work sheds light on the regulation of ADAR editing transcriptome-wide, demonstrating that regulation of ADAR editing can be due to cis-acting elements, and helps elucidate the interaction of editing with RNA-binding proteins, such as HuR.

Methods

Cell culture: B-cells

Cultured B-cells from monozygotic twins and individuals in the CEPH-Utah collection (Dausset et al., 1990) were used (Table S4.1). B-cells were cultured at a density of 5 x 10^5 cells/mL in RPMI 1640 with 15% fetal bovine serum, 100 units/mL penicillin-streptomycin and 2mM L-glutamine. Each batch was cultured separately.

RNA-sequencing and Analysis

RNA-seq from B-cells was performed as recommended by the manufacturer (Illumina). RNA was extracted from the cultured B-cells using Qiagen RNeasy purification kits. The quality of the RNA was measured using the Agilent BioAnalyzer, and only those with RIN > 9.0 were used in the study. cDNA was made using Illumina TruSeq library construction kits. Briefly, RNA was first denatured, then fragmented and first and second strand synthesis was performed using oligo-dT primers. Adapters were added to the cDNA and samples were PCR amplified. The libraries were sequenced on a Hi-Seq instrument to synthesize 100bp reads (for original ten individuals and the individuals from the ten twin pairs) and 50bp (for 28 additional samples) (Illumina, San Diego, USA).

The resulting sequences were analyzed as follows. Low quality bases (Phred-score ≤2) at the 3' ends of reads were trimmed prior to alignment of the reads to the human reference genome (hg18) using GSNAP (version 2012.01.11, Wu and Nacu, 2010), and the following parameters: number of mismatches ≤ [(read length+2)/12-2], mapping score ≥ 20, soft-clipping on (trim mismatch score = -3), known exon-exon junctions as defined by RefSeq, Gencode (version 3c) and novel junctions as defined by GSNAP. SNP tolerant alignment was used with SNPs from the CEU population of HapMap (release #28, International HapMap Consortium 2003) and 1000 genomes (pilot project, The 1000 Genomes Project Consortium 2010). Only reads that aligned to only one genomic locations were used. Gene expression was quantified using Cufflinks v2.1.1.

DNA-sequencing and Analysis

DNA-seq of 4 of the original 10 individuals (GM10838, GM10839, GM12877, and GM12878) was performed as recommended by the manufacturer (Illumina). DNA was extracted from cultured B-cells followed by preparation of DNA-sequencing libraries using the Illumina TruSeq Paired-End construction kits. Briefly, DNA is fragmented, adapters are ligated onto each fragment and fragments are amplified via PCR. The libraries were sequenced on a HiSeq instrument to generate paired-end 100 bp reads.

Low quality bases at the 3' ends of reads were trimmed. The resulting reads were aligned to the human reference genome (hg18) using GSNAP. The following parameters were used for mapping: number of mismatches ≤ [(read length+2)/12-2], mapping score ≥ 20, and soft-clipping on (trim mismatch score = -3). Only reads that aligned to only one genomic locations were used.

Identifying RDD and RNA editing sites

To identify RNA-DNA Differences and RNA editing sites we compared the RNA-sequencing reads to the reference human genome (hg18). To be included as an RNA editing site or RDD for a given sample, a site had to pass multiple criteria: 1) 10 total reads from RNA-seq, 2) 2 reads (5% of the total reads at position) contained the a base that differs from the corresponding DNA sequence, 3) if within an intron, the site must be greater than 10 base pairs from the intron-exon junction, 5) site is not considered a SNP by dbSNP (build 138), 1000 genomes or Exome Sequencing Project, and 6) no RDD read is found in any of the 4 DNA

samples. Additionally, the final list comprises of sites that are found in at least 10 out of the 30 samples whose RNA was sequenced.

In order to ensure uniqueness of RDD and RNA editing sites, further filtering steps were implemented. All sites found in pseudogenes (as annotated by RefSeq) and repetitive regions (RepeatMasker) except Alu elements were removed. Next, the genomic regions around each site (in 25bp, 50bp and 75bp each direction) must be unique as determined by mapping with local sequence alignment (BLAT). Sites were removed if BLAT revealed an alternative alignment with 3 or fewer mismatches. To ensure accurate mapping of RNA-sequencing reads, each read containing the RDD/edited form of the transcript was aligned using a local sequence aligner (BLAT). For a site to be included, over 50% of the RDD/edited reads must align to the same position with BLAT as with GSNAP.

To ensure that the genomic sequence is homozygous across individuals, we used sequencing data from 99 CEU individuals (the same population as the 10 in our dataset) in the 1000 Genomes collection. We required 1) that there be at least 10 reads covering the site among the 99 individuals (only 3 sites were removed by this requirement) and 2) that less than 1% of reads contain the alternative allele. 1% of reads were allowed to contain the alternative allele due to the fact that many sites had over 500 reads and so we must allow for sequencing error. Furthermore, the DNA was sequenced to 30x coverage for 4 of our 10 individuals (as described above). If any RNA editing site showed even a single read containing the alternative allele from these four DNA-sequencing datasets, it was removed.

Validation of SEC16A by Sanger Sequencing

cDNA was prepared using the Taqman Reverse Transcription kit according to manufacturer's protocol (Applied Biosystems). To validate the editing site in SEC16A (chr9: 138455418) and the SNP two bases away (chr9:138455420), the region was amplified using PCR (Forward: 5'-ACCTGGCTGAATGAGTGGAG, Reverse: 5'-AAAATCACCCATGGTCCTCA). The samples were denatured at 94°C for 3 min then put through a cycle of 45 seconds at 94°C, 45 seconds at 55°C and 1 minute at 72°C for 40 cycles followed by 10 minutes at 72°C. PCR product was then purified using Qiaquick PCR Purification Kit (Qiagen 28106) and sequenced using the primers listed above.

Human Tissue Samples

Human skeletal muscle, liver tissue and brain cortex were obtained from the National Disease Research Interchange from three individuals (64998, 65080, and 65288). Tissues were collected between seven and twelve hours post-mortem during autopsies of donors that suffered from cardiac failure or respiratory failure. Individuals were between 62 and 79 years of age. Samples were snap-frozen and kept at -80°C. DNA was extracted from all tissue types using Puregene Tissue Kit (Qiagen). RNA was extracted from the skeletal muscle using RNeasy Maxi Kit (Qiagen), from liver using RNeasy Lipid Tissue Mini Kit (Qiagen) and from brain cortex using MaXtract High Density Kit (Qiagen). RNA was reverse transcribed, PCR amplified and sequenced as described in the previous section.

Treatment with Actinomycin D

B-cells were seeded at 500,000 cells/mL for eight to twelve hours, then they were treated with 5 µg/mL Actinomycin D (Sigma A-1410) that was dissolved in DMSO. Cells were collected before treatment and at the following time points 1, 2, 4, 8, 12, and 24 hours post treatment with Actinomycin D. DNA and RNA was extracted using AllPrep DNA/RNA Mini Kit (Qiagen 80204). RNA was reverse transcribed as described above. Genomic DNA and cDNA were PCR amplified using primers specific to SEC16A and sequenced as described above.

ddPCR of SEC16A editing

Taqman assays were prepared (Applied Biosystems) to amplify the selected region of SEC16A (Forward: 5'-CCGAGGAGCCGTGGG, Reverse: 5'-TGCAACAGGAAAGAAATTCACTG) and probes were prepared correspond to the DNA and edited form of the transcript. The probe specific to the edited form of the transcript (5'-AGGCCCTGGTTACTG) was labeled with FAM while the probe specific to the unedited (DNA) form of the transcript (5'-AGAGGCCCTAGTTACTG) was labeled with VIC. The PCR master mix was made by using cDNA or genomic DNA from the selected samples plus the two primers and two probes above and Taqman reagents. Emulsion PCR was then carried out according to manufacturer's protocol (Bio-Rad Laboratories). PCR amplification conditions were 10 minutes at 95°C, followed by 40 cycles of 30 seconds at 94°C and 1 minute at 57.6°C, followed by termination for 10 minutes at 98°C. The fluorescent signal representing transcripts with either the unedited or edited base was then quantified using QuantaLife Droplet Reader (Bio-Rad Laboratories).

RNA Immunoprecipitation

RNA immunoprecipitation was done using Magna RNA-Binding Protein Immunoprecipitation Kit (Millipore) per manufacturer's protocol. Briefly, $40x10^6$ B-cells were harvested and lysed with 200 µL of lysis buffer with protease and RNase inhibitors (cells were later split in half for HuR or IgG antibodies). 5 µg of anti-HuR (Millipore CS203212) or negative control rabbit IgG (Millipore PP64B) was conjugated to protein A/G beads. 100 µL of the original 200 µL cell lysate was added to 900 µL immunoprecipitation buffer with RNase inhibitor and incubated with 50 µL of beads and appropriate antibody overnight at 4°C. Then, the beads and the bound immunoprecipitate were washed six times with wash buffer and RNase inhibitor and then incubated for 30 minutes at 55°C in protease K and 1% SDS. RNA was extracted from the supernatants using TriPure reagent (Roche 11667157001) and chloroform. The resulting RNA was reverse transcribed into cDNA as above and used to detect binding of SEC16A by HuR using real-time PCR (Forward: 5'-CGCTGTGTTCTCAATCAGC, Reverse: 5'-CAACAGGAAAGAAATTCACTGC).

References

1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct 28;467(7319):1061-73. doi: 10.1038/nature09534.

Athanasiadis A, Rich A, Maas S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biol. 2004 Dec;2(12):e391. Epub 2004 Nov 9.

Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM, White R. Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. Genomics. 1990 Mar;6(3):575-7.

Eggington JM, Greene T, Bass BL. Predicting sites of ADAR editing in double-stranded RNA. Nat Commun. 2011;2:319. doi: 10.1038/ncomms1324.

International HapMap Consortium. The International HapMap Project. Nature. 2003 Dec 18;426(6968):789-96.

Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, Zavolan M. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nat Methods. 2011 May 15;8(7):559-64. doi: 10.1038/nmeth.1608.

Lehmann KA, Bass BL. Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. Biochemistry. 2000 Oct 24;39(42):12875-84.

Lev-Maor G, Sorek R, Levanon EY, Paz N, Eisenberg E, Ast G. RNA-editing-mediated exon evolution. Genome Biol. 2007;8(2):R29.

Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallegger M, Shemesh R, Fligelman ZY, Shoshan A, Pollock SR, Sztybel D, Olshansky M, Rechavi G, Jantsch MF. Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat Biotechnol. 2004 Aug;22(8):1001-5. Epub 2004 Jul 18.

Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG. Widespread RNA and DNA sequence differences in the human transcriptome. Science. 2011 Jul 1;333(6038):53-8. doi: 10.1126/science.1207018. Epub 2011 May 19.

Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011 Nov 24;6:26. doi: 10.1186/1748-7188-6-26.

Peng SS, Chen CY, Xu N, Shyu AB. RNA stabilization by the AU-rich element binding protein, HuR, an ELAV protein. EMBO J. 1998 Jun 15;17(12):3461-70.

Polson AG, Bass BL. Preferential selection of adenosines for modification by double-stranded RNA adenosine deaminase. EMBO J. 1994 Dec 1;13(23):5701-11.

Savva YA, Rieder LE, Reenan RA. The ADAR protein family. Genome Biol. 2012 Dec 28;13(12):252. doi: 10.1186/gb-2012-13-12-252.

Sommer B, Köhler M, Sprengel R, Seeburg PH. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. Cell. 1991 Oct 4;67(1):11-9.

Ulbricht RJ, Emeson RB. One hundred million adenosine-to-inosine RNA editing sites: hearing through the noise. Bioessays. 2014 Aug;36(8):730-5. doi: 10.1002/bies.201400055. Epub 2014 May 30.

Wang IX, So E, Devlin JL, Zhao Y, Wu M, Cheung VG. ADAR regulates RNA editing, transcript stability, and gene expression. Cell Rep. 2013 Nov 14;5(3):849-60. doi: 10.1016/j.celrep.2013.10.002. Epub 2013 Oct 31.

Watson P, Townley AK, Koka P, Palmer KJ, Stephens DJ. Sec16 defines endoplasmic reticulum exit sites and is required for secretory cargo export in mammalian cells. Traffic. 2006 Dec;7(12):1678-87. Epub 2006 Sep 27.

Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010 Apr 1;26(7):873-81. doi: 10.1093/bioinformatics/btq057. Epub 2010 Feb 10.

Zhang Z, Carmichael GG. The fate of dsRNA in the nucleus: a p54(nrb)-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. Cell. 2001 Aug 24;106(4):465-75.

Figure 4.1

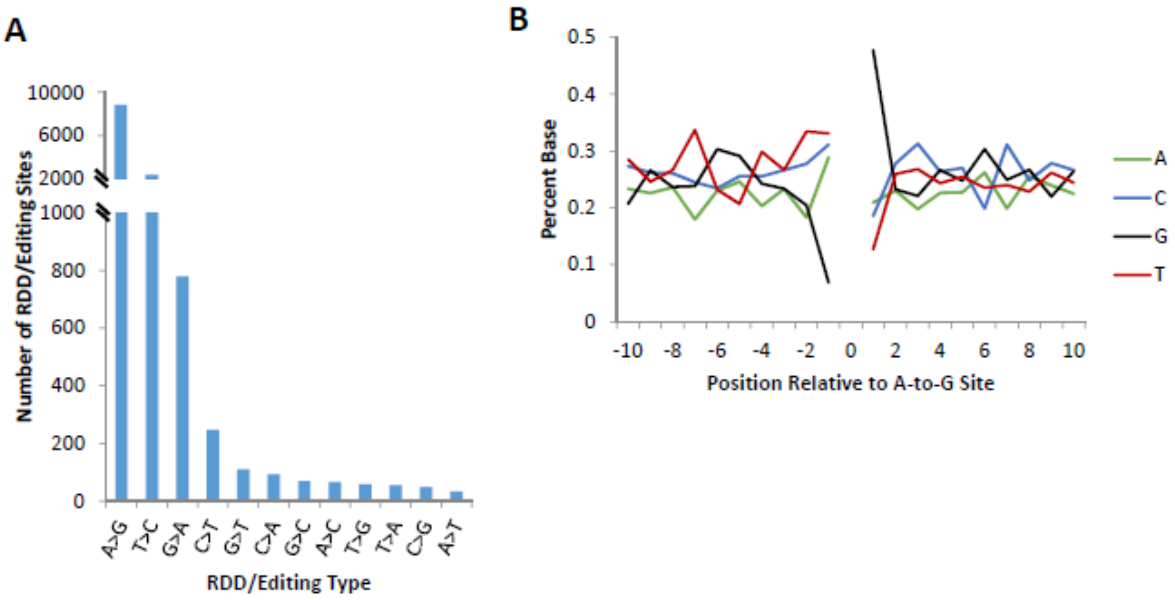

Figure 4.1: Characteristics of RNA editing sites and RDDs. A) Graph showing the number of RDD or editing sites of each type. B) Percentage of each base 10 bases upstream and downstream from A-to-G sites.

Figure 4.2

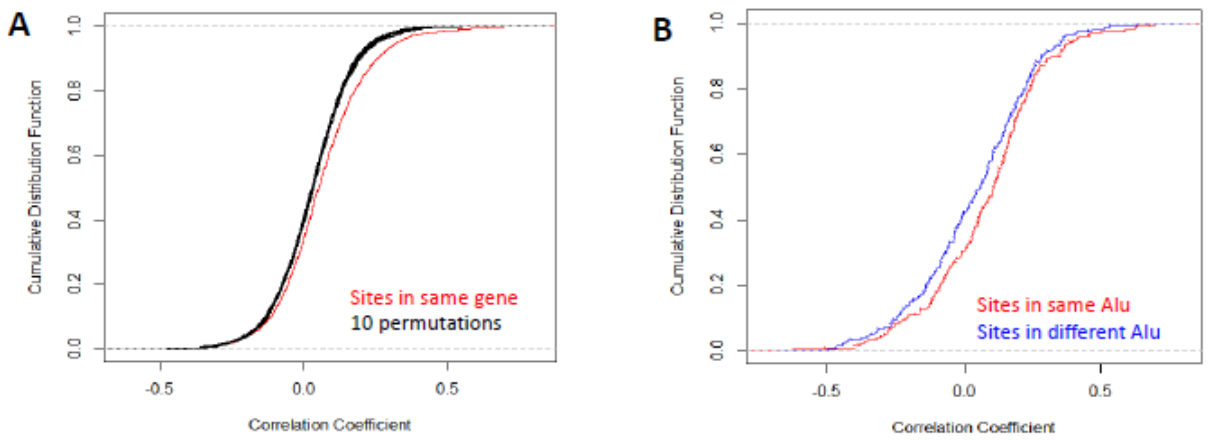

Figure 4.2: Correlation of editing levels of sites in the same transcript. A) Cumulative distribution function comparing the correlation of editing levels across 2,373 sites in the ten individuals. Red line indicates average correlation of sites in the same gene. Black lines indicate average correlation of the same number of random sites. Ten black lines denote ten permutations of random sites. B) Cumulative distribution function comparing the correlation of editing levels for 231 sites. Red indicates average correlations to other sites in the same Alu element while blue denotes average correlations of sites in the same gene but not in the same Alu element.

Figure 4.3

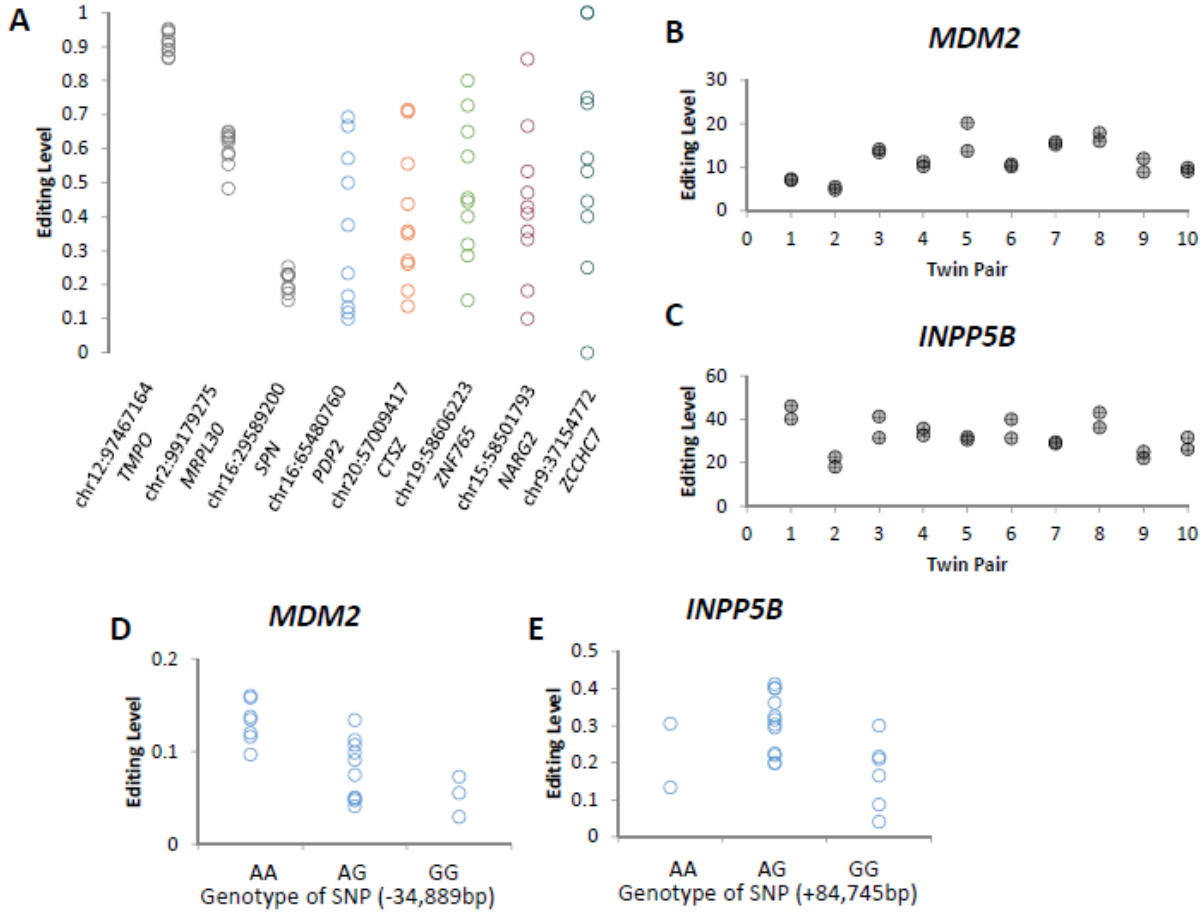

Figure 4.3: *Cis*-association of editing levels. A) Levels of A-to-G editing at ten sites in ten, unrelated individuals. Each point denotes one individual. B-E) Editing levels at two sites (*MDM2* at chr12:67523320 and *INPP5B* at chr1:38099287). B-C) Editing levels across 10 twin pairs with each point denoting the editing level of one individual. ICC of site in *MDM2* is 0.84 and ICC of *INPP5B* is 0.72. D-E) Editing level from the 20 individuals used for association analysis including one individual from each twin pair and the 10 original individuals (ANOVA, for *MDM2* and *INPP5B* p-values are <0.005 and <0.05, respectively). Each point denotes editing level of one individual.

Figure 4.4

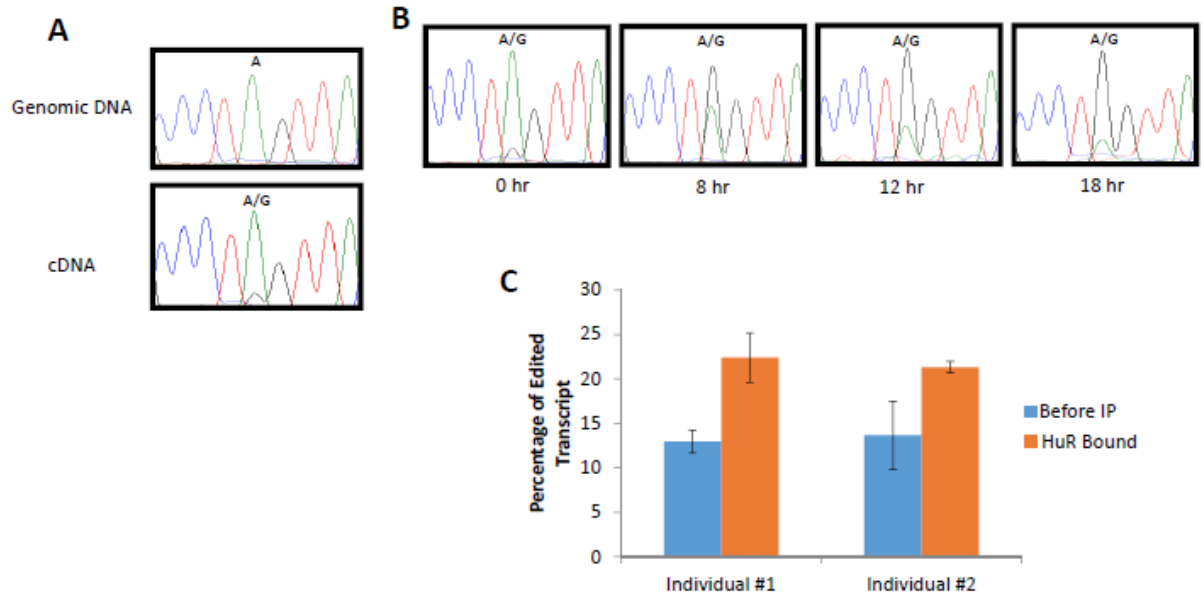

Figure 4.4: RNA editing of *SEC16A.* A) Sanger sequencing trace of editing site at chr9:138,455,418. Genomic and cDNA derived from the same individual, individual 1 (GM11994). B) Sanger sequencing of the same region as A) following time course of Actinomycin D in GM11831. C) HuR RNA-immunoprecipitation results for two individuals (Individual 1 – GM11994, Individual 2 – GM12716) are denoted in the graph. Percentage of edited allele was found for the HuR RNA-IP fraction and the input sample by ddPCR and shown in the graph.

Table S4.1: RNA-sequencing datasets.

Original 10 Individual	Twin Pairs
GM07000	**Pair 1: GM14831**
GM10838	Pair 1: GM14832
GM10839	**Pair 2: GM14408**
GM11992	Pair 2: GM14409
GM11993	**Pair 3: GM14432**
GM11994	Pair 3: GM14433
GM12716	**Pair 4: GM14447**
GM12717	Pair 4: GM14448
GM12877	**Pair 5: GM14452**
GM12878	Pair 5: GM14453
	Pair 6: GM14467
	Pair 6: GM14468
	Pair 7: GM14506
	Pair 7: GM14507
	Pair 8: GM14520
	Pair 8: GM14521
	Pair 9: GM14568
	Pair 9: GM14569
	Pair 10: GM14581
	Pair 10: GM14582

Table S4.1: Two datasets were used in this paper as denoted in this table. For the association analysis, one individual from each twin pair (shown in bold) was used along with the original ten individuals.

Table S4.2: Editing sites with significant ICC.

Chrom	Position	Gene Name	P-values (ANOVA)	ICC
1	1243644	CPSF3L	2.13E-03	0.76
1	1585423	CDK11B	5.15E-03	0.71
1	7900815	TNFRSF9	7.42E-03	0.69
1	17614868	RCC2	1.77E-03	0.77
1	20954404	HP1BP3	7.62E-03	0.68
1	20954546	HP1BP3	1.71E-04	0.86
1	35840512	PSMB2	2.31E-03	0.76
1	38099181	INPP5B	4.91E-03	0.71
1	38099287	INPP5B	4.63E-03	0.72
1	38100287	INPP5B	4.61E-03	0.72
1	45015565	RPS8	4.49E-03	0.72
1	54337885	TCEANC2	8.36E-03	0.68
1	78115317	FAM73A	4.33E-03	0.72
1	78880781	IFI44L	2.93E-14	1.00
1	85170094	MCOLN2	1.79E-03	0.77
1	85170195	MCOLN2	9.56E-04	0.80
1	85170289	MCOLN2	4.63E-04	0.83
1	85170620	MCOLN2	6.21E-03	0.70
1	93573001	LOC100131564	4.25E-03	0.72
1	93576124	LOC100131564	1.96E-03	0.77
1	93576210	LOC100131564	3.00E-05	0.90
1	100262423	N/A	9.66E-03	0.67
1	117426856	TTF2	7.13E-04	0.81
1	153551677	FDPS	3.69E-04	0.84
1	165940932	RCSD1	1.31E-12	1.00
1	177592486	N/A	8.16E-03	0.68
1	177593631	N/A	7.66E-03	0.68
1	177593634	N/A	9.97E-03	0.66
1	202787065	MDM4	8.59E-03	0.68
1	202791169	MDM4	1.95E-03	0.77
1	206008959	CD46	6.32E-03	0.70
1	206008966	CD46	2.69E-03	0.75
2	24076663	UBXN2A	1.74E-04	0.86
2	39555485	LOC728730	7.37E-04	0.81
2	85403736	TGOLN2	2.13E-03	0.76
2	113050712	POLR1B	8.50E-07	0.95
2	171886747	METTL8	1.88E-03	0.77
2	179006651	MIR548N	1.31E-03	0.79
2	198064890	HSPD1	3.52E-03	0.73
2	198064915	HSPD1	5.77E-03	0.70

2	198069877	HSPD1	3.98E-03	0.73
2	201550806	FAM126B	7.42E-03	0.69
2	201550814	FAM126B	1.15E-03	0.79
2	201740214	CFLAR	2.71E-03	0.75
2	216611743	PECR	6.08E-03	0.70
2	218808332	ARPC2	7.57E-03	0.68
2	218808394	ARPC2	7.11E-04	0.81
2	218808418	ARPC2	1.19E-03	0.79
2	218808823	ARPC2	5.34E-06	0.93
2	218808824	ARPC2	3.82E-03	0.73
2	218808911	ARPC2	4.08E-03	0.73
2	218809372	ARPC2	4.44E-03	0.72
2	223132588	N/A	5.82E-03	0.70
2	223132607	N/A	5.99E-03	0.70
2	224329272	AP1S3	1.03E-03	0.80
2	241912658	SEPT2	2.67E-03	0.75
3	10167825	VHL	4.85E-03	0.71
3	10169555	VHL	4.11E-03	0.73
3	10169706	VHL	7.09E-03	0.69
3	10169983	VHL	1.50E-04	0.87
3	10169984	VHL	7.89E-04	0.81
3	10169988	VHL	4.67E-03	0.72
3	15426906	METTL6	4.03E-03	0.73
3	38541028	EXOG	1.32E-03	0.79
3	40553935	ZNF621	9.72E-03	0.67
3	45697104	LIMD1	2.67E-04	0.85
3	50109312	RBM5	8.48E-03	0.68
3	122840500	HCLS1	9.36E-03	0.67
3	137533573	N/A	7.03E-03	0.69
3	143151628	TFDP2	4.76E-04	0.83
3	157741923	SSR3	1.92E-04	0.86
3	188100875	N/A	7.73E-05	0.88
4	116769	ZNF718	4.58E-03	0.72
4	57020883	PAICS	3.30E-03	0.74
4	57020916	PAICS	2.09E-03	0.76
4	57021090	PAICS	1.92E-04	0.86
4	57021632	PAICS	2.67E-04	0.85
4	57021636	PAICS	6.43E-04	0.82
4	71892206	RUFY3	7.96E-06	0.93
4	84036824	LOC100499177	2.18E-03	0.76
4	191120826	FRG1	1.93E-04	0.86
5	1531060	LPCAT1	9.38E-04	0.80
5	34989203	DNAJC21	4.18E-03	0.72

5	40861693	LOC100506548	8.38E-03	0.68
5	43417357	CCL28	3.24E-03	0.74
5	74710543	COL4A3BP	3.40E-04	0.84
5	74710678	COL4A3BP	2.28E-03	0.76
5	125990381	PHAX	8.20E-04	0.81
5	130523546	HINT1	7.77E-03	0.68
5	130523712	HINT1	9.66E-03	0.67
5	130523736	HINT1	7.06E-03	0.69
5	130565567	LYRM7	1.41E-05	0.92
5	130565823	LYRM7	8.16E-04	0.81
5	130565975	LYRM7	1.07E-03	0.80
5	130567745	LYRM7	1.98E-03	0.77
5	131836275	C5orf56	1.66E-03	0.78
5	150627930	GM2A	2.92E-03	0.75
5	179199904	C5orf45	2.65E-03	0.75
5	179199915	C5orf45	4.18E-03	0.72
5	179200010	C5orf45	3.10E-03	0.74
6	32549218	N/A	2.67E-05	0.91
6	33080993	HLA-DOA	7.97E-04	0.81
6	33081005	HLA-DOA	9.52E-03	0.67
6	33081401	HLA-DOA	5.87E-04	0.82
6	33081535	HLA-DOA	7.09E-04	0.81
6	33081553	HLA-DOA	1.71E-03	0.77
6	33484523	KIFC1	8.39E-03	0.68
6	42153541	TAF8	8.08E-03	0.68
6	42155340	TAF8	5.57E-03	0.71
6	43691647	POLH	4.69E-03	0.72
6	43692054	POLH	6.87E-03	0.69
6	43693170	POLH	9.83E-04	0.80
6	53265496	ELOVL5	1.28E-05	0.92
6	57030260	N/A	6.04E-03	0.70
6	109892696	ZBTB24	5.48E-03	0.71
6	149811911	ZC3H12D	5.78E-03	0.70
6	150087691	NUP43	8.82E-03	0.67
6	167271993	RNASET2	6.24E-03	0.70
7	17351405	AHR	7.97E-03	0.68
7	17351412	AHR	3.14E-04	0.84
7	20144144	MACC1	4.30E-03	0.72
7	20144152	MACC1	6.91E-03	0.69
7	20144179	MACC1	2.51E-03	0.75
7	20144193	MACC1	9.98E-04	0.80
7	44807913	PPIA	5.63E-03	0.71
7	44808044	PPIA	1.93E-03	0.77

7	44808624	PPIA	6.79E-03	0.69
7	44839012	H2AFV	1.93E-03	0.77
7	44839264	H2AFV	5.11E-04	0.83
7	44839352	H2AFV	5.16E-03	0.71
7	56088661	CCT6A	1.91E-03	0.77
7	65255714	CRCP	5.75E-03	0.70
7	73275649	LAT2	5.94E-04	0.82
7	73284232	RFC2	8.79E-05	0.88
7	86654508	DMTF1	1.07E-03	0.80
7	89857309	GTPBP10	1.13E-05	0.92
7	100521478	N/A	1.49E-03	0.78
7	128080669	FLJ45340	8.96E-06	0.93
7	128086502	FLJ45340	8.29E-03	0.68
7	130280164	FLJ43663	6.71E-04	0.82
7	130280166	FLJ43663	1.79E-05	0.91
7	130280532	FLJ43663	3.36E-03	0.74
7	149766473	LOC285972	1.66E-03	0.78
7	149766499	LOC285972	3.15E-03	0.74
8	30655694	GSR	8.84E-03	0.67
8	30655952	GSR	6.87E-03	0.69
8	38948894	PLEKHA2	3.76E-03	0.73
8	38948932	PLEKHA2	6.33E-03	0.70
8	38948996	PLEKHA2	1.40E-04	0.87
8	38949002	PLEKHA2	8.98E-04	0.80
8	38949743	PLEKHA2	7.73E-03	0.68
8	66805472	PDE7A	5.39E-04	0.82
8	99125087	RPL30	5.11E-03	0.71
8	99125514	RPL30	1.03E-03	0.80
8	142221069	DENND3	9.70E-03	0.67
8	144808685	N/A	7.12E-03	0.69
9	6708349	N/A	8.44E-05	0.88
9	37493392	POLR1E	1.03E-03	0.80
9	131713235	FNBP1	2.60E-03	0.75
9	131716139	FNBP1	6.43E-03	0.70
9	131717541	FNBP1	5.61E-03	0.71
9	131725162	FNBP1	8.82E-03	0.67
9	131725249	FNBP1	2.04E-04	0.86
9	132570300	N/A	6.70E-03	0.69
9	132571724	N/A	6.39E-03	0.70
9	138455418	SEC16A	4.13E-03	0.73
10	70189289	CCAR1	8.59E-03	0.68
10	73663789	ANAPC16	8.86E-03	0.67
10	74679850	MRPS16	6.23E-03	0.70

10	75206082	N/A	8.50E-04	0.81
10	75206083	N/A	5.27E-03	0.71
10	75206178	N/A	1.75E-03	0.77
10	75206298	N/A	4.08E-03	0.73
10	75206318	N/A	2.02E-04	0.86
10	75206324	N/A	8.55E-03	0.68
10	75206464	N/A	4.82E-03	0.72
10	75207418	N/A	6.19E-03	0.70
10	96360520	N/A	4.15E-03	0.72
10	101982673	CWF19L1	2.14E-03	0.76
10	103537618	MGEA5	2.18E-03	0.76
10	103561417	MGEA5	2.27E-03	0.76
10	134927644	ADAM8	9.54E-03	0.67
11	759779	PDDC1	4.26E-03	0.72
11	31409161	DNAJC24	4.25E-06	0.94
11	33052850	N/A	2.68E-04	0.85
11	33053606	N/A	2.01E-03	0.77
11	62263887	N/A	5.17E-03	0.71
12	9732528	CLEC2D	3.35E-03	0.74
12	9732535	CLEC2D	3.28E-03	0.74
12	14543879	N/A	3.46E-05	0.90
12	27078716	N/A	1.38E-05	0.92
12	27078903	N/A	2.71E-03	0.75
12	27079281	N/A	4.15E-04	0.83
12	27079369	N/A	3.14E-03	0.74
12	27465468	N/A	1.56E-03	0.78
12	27465694	N/A	1.55E-04	0.87
12	27838791	KLHDC5	6.69E-03	0.69
12	47618511	ARF3	2.09E-03	0.76
12	63130434	N/A	6.29E-03	0.70
12	67523271	MDM2	5.03E-03	0.71
12	67523309	MDM2	3.29E-04	0.84
12	67523310	MDM2	2.75E-08	0.98
12	67523320	MDM2	3.92E-04	0.84
12	67523773	MDM2	6.27E-03	0.70
12	67523786	MDM2	4.04E-04	0.83
12	97467164	TMPO	1.38E-07	0.97
12	109420433	VPS29	3.26E-03	0.74
12	112203790	TPCN1	3.76E-05	0.90
12	112313307	N/A	3.58E-04	0.84
12	112313308	N/A	1.92E-03	0.77
12	117059632	PEBP1	5.88E-04	0.82
12	119383458	GATC	6.36E-03	0.70

12	120256202	ANAPC5	3.73E-03	0.73
12	122675039	EIF2B1	6.35E-03	0.70
12	131807977	PGAM5	6.89E-04	0.81
12	131808168	PGAM5	3.24E-05	0.90
12	131808194	PGAM5	3.03E-03	0.74
13	20846502	ZDHHC20	3.85E-06	0.94
13	20847107	ZDHHC20	1.27E-03	0.79
13	44988372	COG3	5.48E-04	0.82
13	110349045	ANKRD10	1.55E-03	0.78
14	19905128	TEP1	4.49E-03	0.72
14	23983669	N/A	2.70E-03	0.75
14	34564938	SRP54	5.25E-03	0.71
14	52312609	GNPNAT1	2.86E-03	0.75
14	54592126	MAPK1IP1L	1.92E-03	0.77
14	54597427	MAPK1IP1L	4.68E-03	0.72
14	54904719	ATG14	2.28E-03	0.76
14	90752682	C14orf159	1.23E-05	0.92
14	103231015	KLC1	5.93E-03	0.70
15	39384939	N/A	4.21E-03	0.72
15	39384949	N/A	5.96E-05	0.89
15	40647240	HAUS2	1.01E-03	0.80
15	48447516	N/A	5.15E-04	0.83
15	62234026	SNX22	1.10E-03	0.79
15	73433139	NEIL1	2.45E-03	0.76
15	83440922	PDE8A	1.26E-03	0.79
16	11836425	RSL1D1	9.23E-04	0.80
16	11836984	RSL1D1	6.18E-03	0.70
16	13949951	ERCC4	5.68E-03	0.70
16	15701474	NDE1	1.03E-03	0.80
16	15701825	NDE1	2.73E-04	0.85
16	15701988	NDE1	7.25E-04	0.81
16	15702345	NDE1	6.25E-03	0.70
16	15702536	NDE1	7.10E-07	0.96
16	22204361	EEF2K	2.58E-04	0.85
16	22204439	EEF2K	9.69E-03	0.67
16	27369486	IL21R	5.02E-03	0.71
16	27370394	IL21R	2.10E-03	0.76
16	28751005	ATXN2L	2.73E-04	0.85
16	29586447	SPN	3.03E-03	0.74
16	29587051	SPN	1.31E-04	0.87
16	29587323	SPN	3.33E-03	0.74
16	29587522	SPN	9.34E-03	0.67
16	29587532	SPN	4.30E-05	0.90

16	29587784	SPN	3.54E-05	0.90
16	29587918	SPN	5.78E-03	0.70
16	29588423	SPN	2.09E-03	0.76
16	29753224	MVP	8.05E-03	0.68
16	46945855	MIR548AE2	7.96E-03	0.68
16	55647735	NLRC5	1.25E-03	0.79
16	55648552	NLRC5	4.30E-03	0.72
16	79620781	CENPN	7.82E-03	0.68
16	83258175	N/A	2.40E-03	0.76
16	88157407	RPL13	6.68E-04	0.82
16	88157527	RPL13	7.23E-04	0.81
17	2268356	METTL16	3.52E-04	0.84
17	3524357	P2RX5	5.84E-03	0.70
17	3525364	P2RX5	1.88E-03	0.77
17	4869087	KIF1C	3.87E-03	0.73
17	31178703	TAF15	7.35E-03	0.69
17	40575770	ACBD4	7.73E-03	0.68
17	44296677	CALCOCO2	2.42E-03	0.76
17	46397215	SPAG9	1.52E-03	0.78
17	54635387	PRR11	8.44E-04	0.81
17	59910467	POLG2	1.70E-03	0.77
17	71397599	TRIM65	1.93E-05	0.91
17	71451828	ACOX1	8.16E-03	0.68
17	73617255	LOC100131096	6.70E-03	0.69
17	73930206	PGS1	8.12E-03	0.68
17	73930313	PGS1	6.74E-03	0.69
18	54569458	N/A	3.85E-03	0.73
19	2031647	MOBKL2A	1.20E-03	0.79
19	3491316	C19orf28	1.60E-03	0.78
19	4605537	TNFAIP8L1	8.18E-04	0.81
19	4605554	TNFAIP8L1	5.72E-03	0.70
19	5660146	LONP1	9.33E-03	0.67
19	7665229	FCER2	6.10E-03	0.70
19	7665859	FCER2	8.32E-03	0.68
19	7666681	FCER2	6.98E-04	0.81
19	7666958	FCER2	4.60E-03	0.72
19	7667260	FCER2	5.14E-03	0.71
19	10889920	CARM1	2.24E-03	0.76
19	13744855	MRI1	2.18E-03	0.76
19	13744923	MRI1	9.32E-03	0.67
19	13745419	MRI1	8.55E-03	0.68
19	14569513	CLEC17A	5.55E-04	0.82
19	14569530	CLEC17A	1.66E-03	0.78

19	14582497	CLEC17A	1.97E-04	0.86
19	14582498	CLEC17A	2.70E-05	0.91
19	14582553	CLEC17A	7.71E-04	0.81
19	14582594	CLEC17A	4.79E-03	0.72
19	18337948	PGPEP1	1.05E-03	0.80
19	18338312	PGPEP1	2.28E-03	0.76
19	18338815	PGPEP1	1.32E-03	0.79
19	18338901	PGPEP1	3.27E-04	0.84
19	18530846	C19orf50	1.46E-05	0.92
19	19653335	N/A	3.70E-03	0.73
19	19653408	N/A	8.45E-04	0.81
19	21266341	ZNF708	5.77E-04	0.82
19	40512956	CD22	1.07E-03	0.80
19	40513928	CD22	4.63E-03	0.72
19	40514342	CD22	2.05E-03	0.76
19	40517474	CD22	1.40E-03	0.78
19	41729014	ZNF529	4.42E-04	0.83
19	44050992	RINL	9.47E-04	0.80
19	44674071	N/A	5.57E-03	0.71
19	57898485	ZNF611	8.68E-03	0.67
19	58689888	N/A	6.07E-04	0.82
19	58689956	N/A	4.49E-03	0.72
19	62508825	N/A	9.46E-03	0.67
19	62960197	ZNF776	6.30E-03	0.70
19	63046861	LOC100293516	4.04E-03	0.73
19	63046881	LOC100293516	8.59E-03	0.68
19	63070149	N/A	6.84E-03	0.69
19	63785572	MGC2752	1.97E-03	0.77
19	63787385	MGC2752	3.37E-04	0.84
20	3796078	MAVS	2.75E-04	0.85
20	3796188	MAVS	5.87E-03	0.70
20	3798191	MAVS	1.54E-03	0.78
20	3801067	MAVS	3.99E-04	0.84
20	33681293	CPNE1	5.34E-03	0.71
20	33681371	CPNE1	9.88E-03	0.66
20	33764606	RBM39	1.69E-03	0.77
20	33765106	RBM39	1.11E-03	0.79
20	33765163	RBM39	2.54E-03	0.75
20	43140284	STK4	5.72E-04	0.82
20	43140402	STK4	7.24E-03	0.69
20	43141123	STK4	1.76E-03	0.77
20	44186067	CD40	3.88E-03	0.73
21	29358242	CCT8	6.24E-03	0.70

21	33558893	N/A	4.45E-03	0.72
21	34199188	ATP5O	5.10E-03	0.71
21	43137033	WDR4	4.12E-03	0.73
22	16951271	PEX26	7.70E-03	0.68
22	17818564	UFD1L	2.36E-03	0.76
22	22644046	DDT	4.02E-04	0.83
22	22644078	DDT	1.31E-03	0.79
22	22644094	DDT	4.13E-04	0.83
22	22644147	DDT	4.40E-05	0.90
22	22644244	DDT	1.90E-04	0.86
22	23300542	N/A	5.35E-04	0.82
22	24451085	ADRBK2	6.03E-03	0.70
22	37744869	N/A	6.18E-03	0.70
22	37745837	N/A	5.39E-03	0.71
22	37752842	APOBEC3D	5.45E-03	0.71
22	37759009	APOBEC3D	9.25E-06	0.93
22	37779445	APOBEC3F	4.62E-03	0.72
22	37779451	APOBEC3F	2.99E-04	0.85
22	43969023	KIAA0930	4.04E-03	0.73
X	156177	PLCXD1	5.32E-04	0.82
X	156473	PLCXD1	6.28E-05	0.89
X	157393	PLCXD1	8.17E-03	0.68
X	157402	PLCXD1	1.10E-03	0.79
X	157844	PLCXD1	1.07E-04	0.88
X	157852	PLCXD1	4.32E-03	0.72
X	157946	PLCXD1	9.24E-04	0.80
X	158215	PLCXD1	9.68E-04	0.80
X	24003875	EIF2S3	7.54E-03	0.68
X	84232559	APOOL	2.97E-03	0.74
X	118642447	SEPT6	1.40E-03	0.78
X	118642993	SEPT6	2.78E-04	0.85

Table S4.2: This table denotes editing sites with ICC > 0.66 across the ten twin pairs.

Table S4.3: *Cis*-associated editing sites across twenty individuals.

Editing Site Position (hg18)	Gene Name	Genic Region of Editing Site	SNP Position (hg18)	SNP rs#	P-values
chr1:38099181	INPP5B	3'UTR	chr1:38184032	rs871524	0.021
chr1:38099287	INPP5B	3'UTR	chr1:38184032	rs871524	0.019
chr1:38100287	INPP5B	3'UTR	chr1:38184032	rs871524	0.019
chr12:67523309	MDM2	3'UTR	chr12:67488431	rs937283	0.020
chr12:67523310	MDM2	3'UTR	chr12:67488431	rs937283	0.001
chr12:67523320	MDM2	3'UTR	chr12:67488431	rs937283	0.001
chr16:88157407	RPL13	3'UTR	chr16:88157812	rs12709089	0.026
chr16:88157527	RPL13	3'UTR	chr16:88157812	rs12709089	0.024
chr20:3796188	MAVS	3'UTR	chr20:3799600	rs34419413	0.015
chr20:3796188	MAVS	3'UTR	chr20:3801093	rs17212649	0.015
chr20:3801067	MAVS	3'UTR	chr20:3798968	rs14161	0.034

Table S4.3: This table denotes editing sites with p-values < 0.05 from ANOVA across the twenty, unrelated individuals.

Figure S4.1

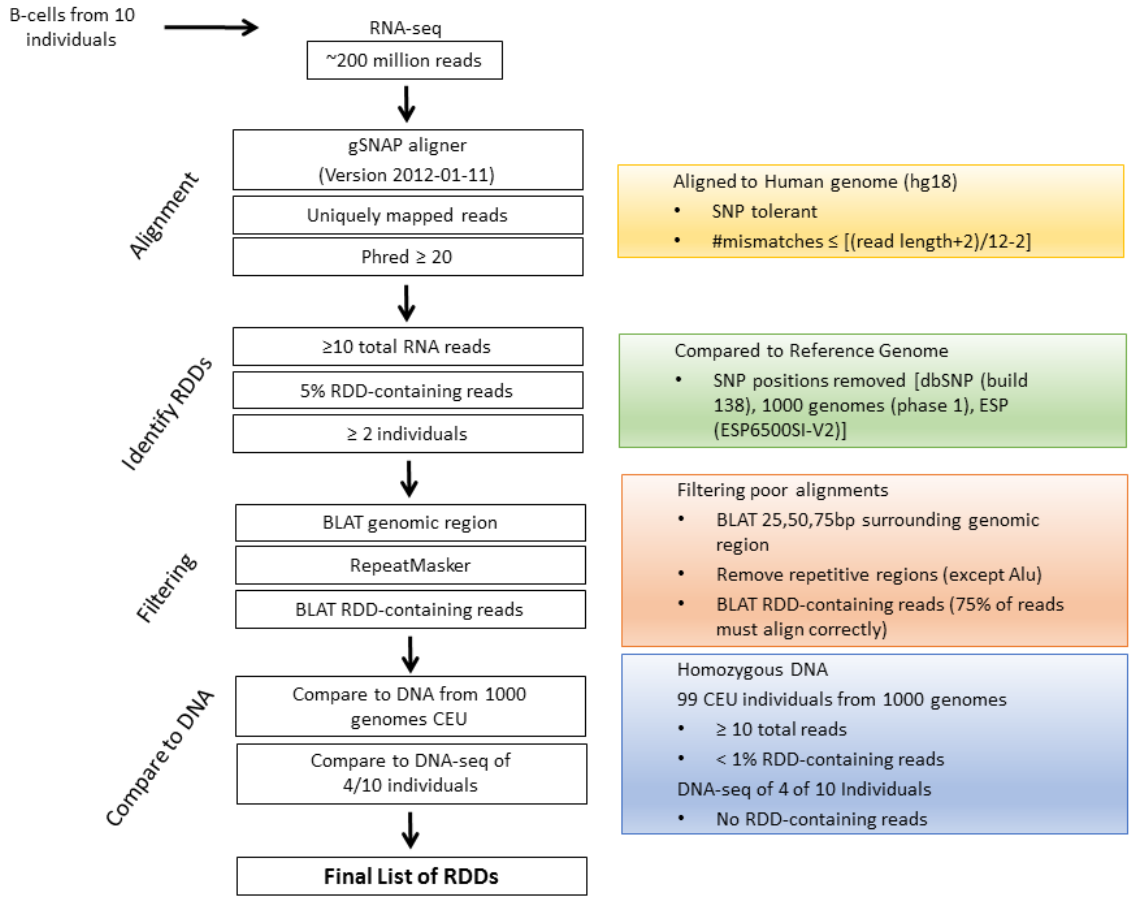

Figure S4.1: Analysis for identification of RDDs among ten individuals. This flow-chart indicates the steps used to identify all RDDs and editing sites from the original ten individuals.

Figure S4.2

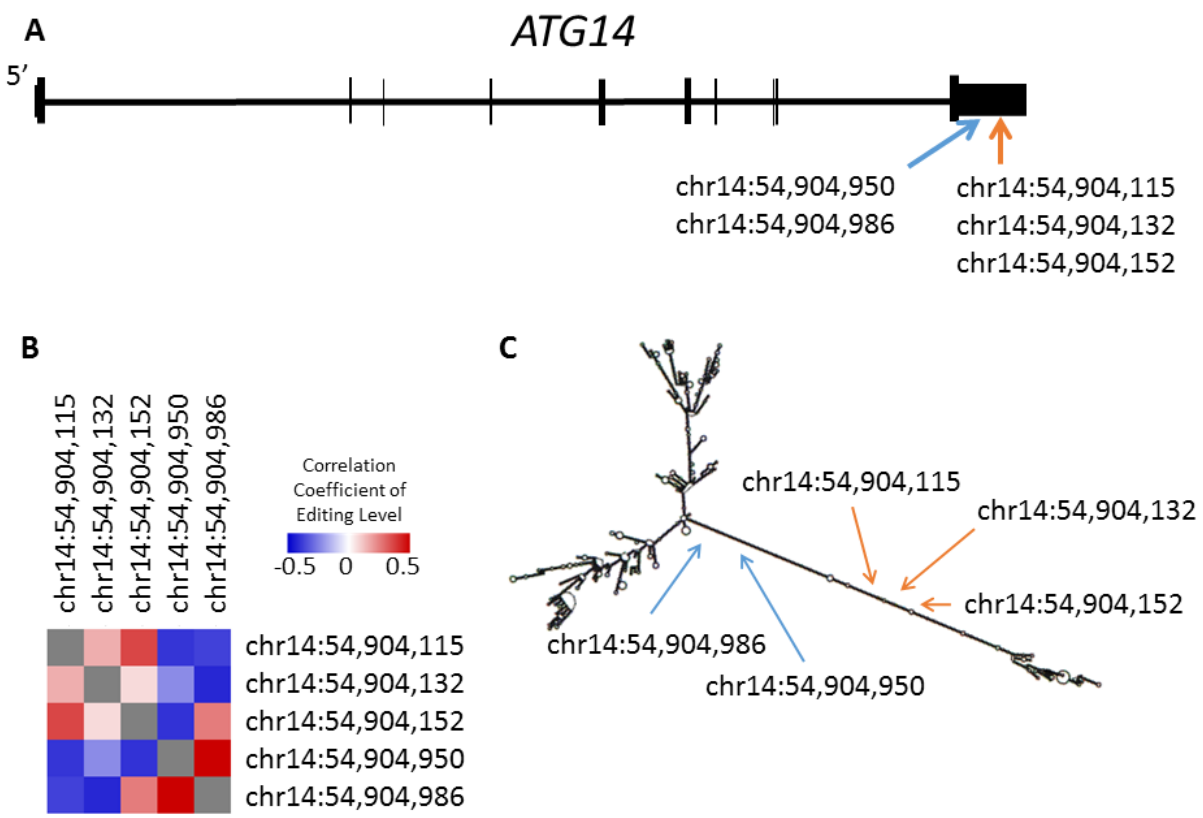

Figure S4.2: *ATG14:* Example of RNA folding and editing. A) Gene structure for *ATG14* with the five editing sites indicated. There are two groups of editing sites: 1) 54,904,950 and 54,904,986 and 2) 54,904,115, 54,904,132 and 54,904,152 as indicated by the arrows (blue shows the first group, orange shows the second group). B) Heatmap indicating the correlation coefficient determined by comparing the editing levels of the indicated sites across ten, unrelated individuals. C) RNA secondary structure (Lorenz et al., 2011) with the editing sites indicated by arrows.

Figure S4.3

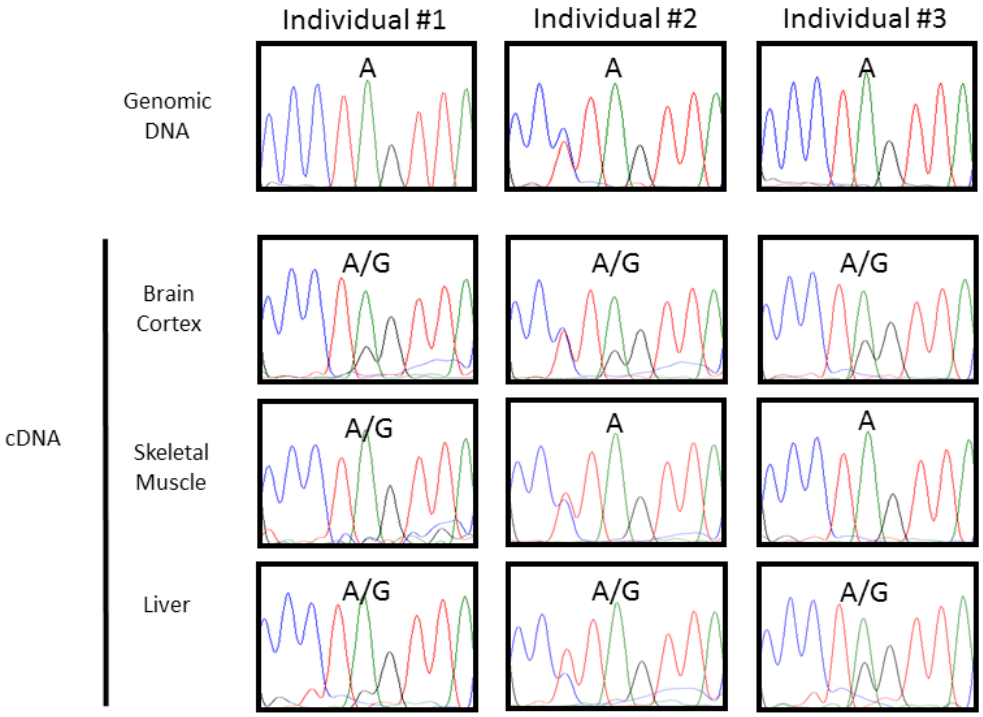

Figure S4.3: Editing of *SEC16A* in three tissues. Sanger sequencing traces of editing site and C/T SNP at chr9:138,455,418 and 13,455,420, respectively. Genomic and cDNA derived from the same individuals (Individual 1 – 64998, Individual 2 – 65080, Individual 3 – 65288 from NDRI).

Figure S4.4

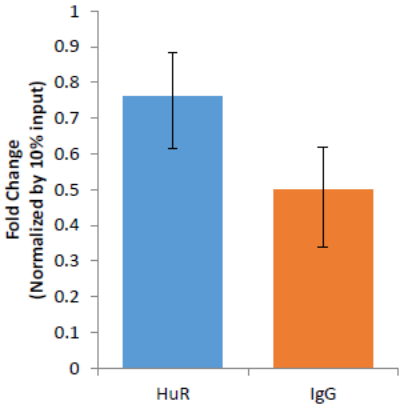

Figure S4.4: HuR RNA-IP of *SEC16A.* HuR RNA-immunoprecipitation performed on two individuals (GM11994 and GM12716). Quantitative real-time PCR was performed to target *SEC16A* and the graph shows enrichment of this region in the HuR RNA-IP as compared to sample immunoprecipitated by IgG.

Figure S4.5

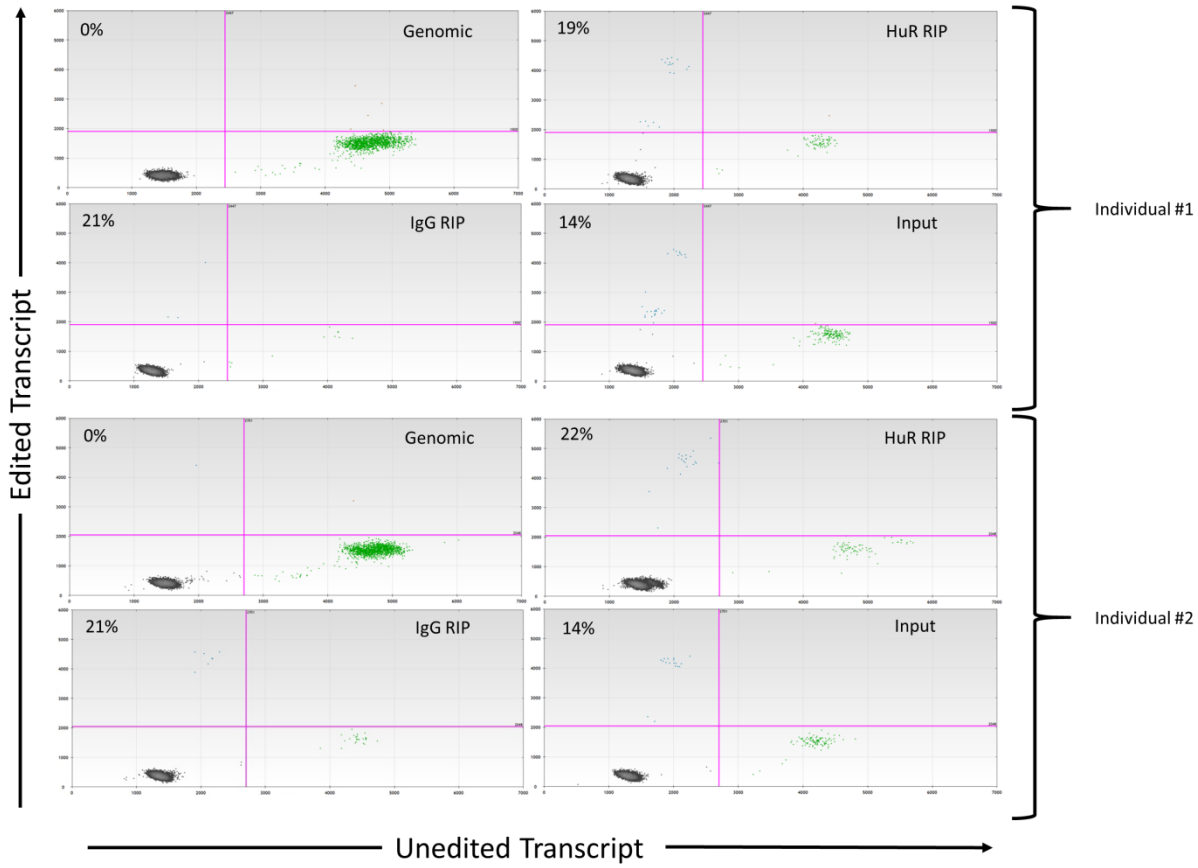

Figure S4.5: HuR RNA-IP by ddPCR. Representative examples of ddPCR results comparing edited versus unedited transcripts in HuR RNA-immunoprecipitation samples from two individuals (Individual 1 – GM11994, Individual 2 – GM12716). Results comparing HuR RIP to input over four or three replicates (for GM11994 and GM12716, respectively) are denoted in the graph in Figure 4c (IgG is not utilized because few droplets contain any *SEC16A* transcript). Editing level was calculated for the HuR RNA-IP fraction, IgG RNA-IP fraction and the input sample by ddPCR by comparing the number of droplets with the edited transcript compared to the total number of droplets containing *SEC16A* transcripts.

Chapter 5: RNA Editing in Response to ER Stress in Human B-cells

Abstract

Endoplasmic reticulum (ER) stress leads to many changes as the cell attempts to re-establish homeostasis. ER stress is caused by an excess of misfolded proteins in the cell. This leads to changes in expression of genes involved in protein folding, transport and cell cycle to help the cell cope with the misfolded proteins. In this study, we show that certain RNA processing steps, including RNA editing, are also influenced by ER stress. We collected and sequenced RNA from B-cells of ten individuals before, and two and eight hours following ER stress induction. By comparing the RNA and corresponding DNA sequences, we identified over 4,000 A-to-G editing sites among the ten individuals. A subset of these editing sites, found in genes important for the ER stress response, show changes in editing level following ER stress. Additionally, some of these sites have editing levels correlated to gene expression following ER stress. These data suggest that the changes in editing level may play a role in the ER stress response. In addition to changes in canonical A-to-G editing, we also find other types of single-base differences that change level following ER stress. Our study demonstrates that RNA editing responds to ER stress and may play a role in the ER stress response.

Introduction

The endoplasmic reticulum (ER) is an organelle that synthesizes and modifies proteins. ER stress occurs when there is an excess of misfolded proteins in the cell. It has been shown in previous studies that ER stress leads to gene expression changes in a wide variety of pathways from protein transport to cell cycle control (Dombroski et al., 2010). These changes in expression will either re-establish homeostasis or initiate apoptosis in cells. Changes in gene expression are influenced by a variety of mechanisms, including RNA processing steps such as RNA editing. We asked whether RNA editing and other RNA-DNA sequence Differences (RDD) were also affected by ER stress and how changes in these modes of RNA processing may influence expression of ER stress-responsive genes.

One form of RNA editing in humans is mediated by Adenosine Deaminase Acting on RNA (ADAR) which results in thousands of A-to-G differences between the DNA and corresponding RNA sequences throughout the transcriptome (Athanasiadis et al., 2004, Levanon et al., 2004, Li et al., 2009, Wang et al., 2013). The best characterized examples of this type of RNA editing are found in coding regions of ion channels. These sites cause changes to the amino acid sequence and affect the channel permeability and function (Sommer et al., 1991, Lomeli et al., 1994, Seeburg et al., 1998). Though there are a handful of editing sites in coding sequences, many editing sites are in non-coding regions and are predicted to affect RNA expression. For example, studies have demonstrated that some RNA editing can affect RNA processing, such as splicing (Athanasiadis et al., 2004, Lev-Maor et al., 2007), transport (Zhang and Carmichael, 2001) and decay (Scadden 2005). In addition to canonical RNA editing, our lab has previously

characterized the other eleven types of single-base differences between the RNA and DNA. RDDs are also found throughout the transcriptome, in coding and non-coding regions of transcripts (Li et al., 2011).

In this study, we explore the role of RNA editing and RDDs in the ER stress response. We identified both canonical ADAR editing and RDD sites in genes involved in the ER stress response. Furthermore, a subset of these sites show changes in level following ER stress and the levels of editing correlate to changes in gene expression. This suggests that changes in editing and RDD level may play a role in the ER stress response.

Induction of ER stress

We induced ER stress in cultured human B-cells with tunicamycin. Tunicamycin induces ER stress by blocking N-glycosylation, which is necessary for proper protein folding (Tkacz and Lampen, 1975). We confirmed induction of ER stress by quantifying two steps in the unfolded protein response, XBP1 splicing and up-regulation of a protein chaperone, BiP (Yoshida et al., 2001, Kozutsumi et al., 1988). ER stress is induced by tunicamycin in our human B-cells as shown by an increase in the spliced form of XBP1 following 2 hours of tunicamycin treatment and an increase in BiP expression following 8 hours of tunicamycin treatment (Figure 5.1a,b).

Gene expression changes following ER stress

To study the response of B-cells to ER stress, we treated cells derived from ten unrelated individuals with tunicamycin. We collected RNA from before, and 2 and 8 hours

following treatment. We used RNA-sequencing of these 30 samples to characterize the gene expression and RNA editing response to ER stress.

We identified 12,426 genes that were expressed (FPKM ≥ 1) in at least five of the 30 samples. Of these genes, 4,182 genes showed significant changes in gene expression following 2 or 8 hours of ER stress across the ten individuals (ANOVA, nominal p-value < 0.01) (Figure 5.1c). 1,244 genes were significant even after Bonferroni correction. In order to learn about the ER stress response more comprehensively, we further analyzed the 4,182 genes that met the lower cut-off. At 2 hours after tunicamycin treatment, 2,012 genes showed changes in expression ranging from 27-fold increase to 4-fold decrease in expression (1,154 increase, 858 decrease). Following 8 hours of ER stress, 3,580 genes changed expression from 31-fold increase to 11-fold decrease in expression (1,798 increase, 1,782 decrease). These groups are not exclusive as we see 1,530 changing at both time points (post hoc t-test, p < 0.01). Over 60 genes had previously been characterized as ER stress-responsive genes (Ashburner et al., 2000, Gene Ontology Consortium 2015, GO:0034976). For example, *EDEM1*, a gene that is involved in ER-associated decay of proteins following ER stress, increases after 8 hours of tunicamycin treatment.To further characterize these expression changes, we performed hierarchical clustering on the 4,182 ER stress-responsive genes. These genes cluster into five groups that separate genes that upregulated or downregulated at 2 versus 8 hours following ER stress (Figure 5.1c). Each of these groups was found to be enriched for certain pathways. One group contains 1,440 genes that show an increase in gene expression at both time points with larger increases at 8 hours following treatment (6- to 31-fold maximum increases at 2 and 8 hours,

respectively). These genes show enrichment for certain pathways, such as protein folding (*HSPA5*) and localization (*SEC61A1)* pathways.

We also identified expression changes in genes involved in RNA regulation. Genes that show increases in expression early (at two hours following tunicamycin treatment) are enriched for transcription factors (*E2F7, ZNF135*). Furthermore, genes found in RNA processing pathways also changed following ER stress. Specifically, splicing factors show decreased expression (*HNRNPD*) following ER stress while factors involved in decay (*DCP1A)* increase in expression. The changes in expression of genes regulating RNA led us to ask whether RNA processing, specifically RNA editing, may also be affect by ER stress.

ADAR RNA editing levels following ER stress

First, we focused on canonical, A-to-G editing by comparing the RNA sequences to the reference genome (hg18) to identify sites where the RNA did not match the corresponding DNA sequences. In order to compare editing across individuals and across time points we required that the site be found in at least 10 out of the total 30 samples sequenced (which include before, 2 and 8 hours following ER stress in ten B-cell lines). In this way, we identified 4,020 A-to-G editing sites, 3,023 of which are located in 739 annotated genes while the rest are found in intergenic regions. Of those found in annotated genes, 2,915 editing sites (96%) are located in non-coding regions (1,565 in introns and 1,350 in untranslated regions). As the majority of these sites are located in non-coding, regulatory regions, these data suggest that editing in these regions may affect gene expression.

We next asked whether ADAR editing changes under ER stress conditions. We used ANOVA to identify 342 sites in 185 genes whose level change following ER stress induction (p-value <0.05) (Figure 5.2a). Though none of these 342 editing sites are significant following Bonferroni correction for multiple testing, we proceeded to study these 342 sites to learn about trends of editing that may occur following ER stress. At 2 hours after tunicamycin treatment, 114 sites showed changes in editing levels ranging from 12-fold increase to 9-fold decrease in level (88 increase, 26 decrease). Following 8 hours of ER stress, 207 sites changed editing levels from 13-fold increase to 9-fold decrease in expression (192 increase, 15 decrease). These groups are not exclusive as we see 51 editing sites that change levels at both time points (post hoc t-test, $p < 0.05$). While the majority of sites show an increase in editing level at 2 hours, over 92% of sites show an increase in editing level following 8 hours of ER stress. This coordinated response suggests that a factor responding to ER stress affects the editing of these 192 sites.

To determine if ADAR expression was affected by ER stress, we first measured ADAR mRNA expression by quantitative PCR in five individuals and did not detect a significant change in ADAR mRNA expression following 8 hours of ER stress (Figure S5.1a). Additionally, we measured ADAR protein expression in two individuals and no significant change in expression was found following 2 or 8 hours of ER stress (Figure S5.1b). Therefore, changes in ADAR expression alone, do not account significantly for the changes in editing level that we see following ER stress.

Next, we analyzed the editing level changes following ER stress using hierarchical clustering of the 342 editing sites. These editing sites cluster into four groups that separate editing level patterns based on increases or decreases in editing level following 2 or 8 hours of ER stress (Figure 5.2a). Many of the editing sites are found in genes involved in processes important for ER stress response, such as protein modification and metabolism. For example, 202 editing sites show an increase in editing level at both time points (up to about 13-fold increases) are found in genes involved in the protein modification (*RFWD3, PRKCSH*) and protein trafficking (*SRP9, GGA1)*. Other sites can be found in genes involved in a variety of processes, such as mRNA processing (*RBM3, APOBEC3D*) and cell cycle control (*KNTC1, DMTF1*). These results suggest a role for ADAR editing in regulating the ER stress response.

Validation of ADAR editing following ER stress

As a way to validate some editing sites and better assess individual variability in editing level, we performed targeted RNA sequencing. We focused on 41 regions with 144 editing sites from the original dataset. 27 of these regions are in 3' untranslated regions, 12 in introns, and 2 in an intron and exon. These regions are in 38 genes involved in ER stress response or protein processing (Table S5.1). We sequenced the 41 target regions in 16 additional individuals before treatment and following 2 and 8 hours of treatment with either tunicamycin or DMSO, as a control.

We obtained sequencing data for 125 A-to-G editing sites identified with the original dataset. Among these 118 sites were covered by at least ten sequencing reads. Of the 125 sites,

114 (91%) had G-containing reads in in the targeted sequencing, thus confirming the presence of editing at these sites.

To study trends of RNA editing in this additional dataset, we identified RNA editing sites using similar criteria. A site was required to be found in 10 of the 80 samples (16 individuals under five conditions). We identified 207 RNA editing sites, including 150 additional A-to-G sites not seen in the original dataset. Though 114 sites from the original dataset showed evidence of editing in the targeted sequencing, only 57 were called as editing sites using our thresholds. Among these 207 editing sites, 11 were found to show significant changes in editing level following ER stress when comparing baseline editing levels to 2 and to 8 hours following ER stress (ANOVA, p-value < 0.05). As seen in the original dataset, 94% of sites that change in level show an increase in editing level following ER stress. In addition, seven sites that were found to be significant in the original 10 individuals were identified as editing sites in the targeted sequencing. Of these seven sites, three had a p-value less than 0.1 by ANOVA determining change in editing level following ER stress in the additional 16 individuals. One site (chr2:223132607) in a region 3' of a gene *SGPP2,* involved in ER stress-induced autophagy (Lépine et al., 2011), shows significant changes in editing level in both datasets (Figure 5.2b). The targeted sequencing validates our initial findings of RNA editing level changes in response to ER stress and expands our list of responsive editing sites.

ADAR editing and gene expression

Editing levels across individuals are highly variable before and following induction of ER stress. Specifically, 37% of editing sites show greater than a 5-fold difference between the

maximum and the minimum editing level across the 10 individuals at baseline (38% and 35% at 2 hours and 8 hours, respectively). In the previous chapter, my data suggest that *cis* variation may play a role in editing variability. Here, we asked whether variability in editing level could lead to variability in gene expression following ER stress.

Next, we asked whether ADAR editing levels are correlated to changes in gene expression following ER stress. Of the 342 editing sites that change levels following ER stress, 101 sites are found in 76 genes that show changes in gene expression as well. We compared the editing level at one time point to the gene expression at the next time point to see if editing level may affect gene expression changes. We found 17 sites in 15 genes where the editing level was highly correlated to gene expression at a later time point (correlation coefficient, $r > 0.632$, p-values < 0.05). Eleven sites have baseline editing levels that are correlated with 2 hour gene expression levels while 7 sites have 2 hour editing levels that are correlated with 8 hour gene expression levels. A few examples of these correlations are shown in Figure 3. When comparing editing levels from baseline to gene expression at 2 hours, we found that 73% of sites (8 out of 11) show a negative correlation. Only 42% of sites (3 out 7) show a negative correlation of editing level at 2 hours to gene expression at 8 hours. Furthermore, the same site may be correlated in opposite directions. For example, such as chr12:119383458 found in *GATC* (Figure 5.3) has baseline editing level that is negatively correlated to gene expression at 2 hours but the editing level at 2 hours is positively correlated to gene expression at 8 hours after tunicamycin. This demonstrates that there can be a change in the relationship between editing level and gene expression following cellular stress. Furthermore, editing sites can be both positively and

negatively correlated to gene expression (Table S5.2) suggesting that each site may play a unique effect on its transcript.

Conclusions

ER stress response utilizes many pathways to cope with excess misfolded proteins. In this study, we describe how RNA editing and gene expression change in response to ER stress. Using high-throughput sequencing we identified over 300 editing sites that change levels following ER stress. Many of the editing sites can be found in genes relating to the ER stress response such as those involved in protein transport and protein modification. Previous reports of editing demonstrate that editing sites can modulate gene expression in a variety of ways, including affecting splicing and RNA degradation. Together, this suggests that a change in RNA editing level following ER stress may affect expression of genes involved in the ER stress response. In this way, RNA editing may play a role in the ER stress response. Another report suggests that lack of ADAR editing can lead to ER stress and apoptosis (Qiu et al., 2013) while our work demonstrates how this may be accomplished, through an effect of editing on gene expression.

Furthermore, almost 90% of editing sites show an increase in editing level following ER stress despite only minor changes to ADAR expression level. This suggests that many of these editing levels may be acted upon by a factor that modifies ADAR editing. Changes in this factor following ER stress then lead to an increase in editing level throughout the transcriptome. This type of mechanism is similar to that seen in transcription factors where change in expression of one transcription factor can lead to changes in the expression of many genes. These data

suggest another way that the cell may be able to induce transcriptome-wide changes in gene expression during cellular stress.

While, this is the first transcriptome-wide description of RNA-DNA Differences during a cellular response and demonstrates the importance of studying RNA processing changes to cellular stress, it is important to stress caveats in the experimental design. First, only ten individuals were analyzed in the main dataset and are used for comparison of gene expression levels and editing levels. This limited sample size reduces the power to detect significant changes in gene expression or editing levels or correlation between these two factors. Furthermore, each of these individuals was only measured once. Future experiments should be aimed at repeating the sequencing of these individuals to ensure that any variability is due to biological individual differences. Although further experiments should verify these results, the data thus far suggest that individual variation in editing level may contribute to differences in ER stress response.

Methods

Cell culture: B-cells

Cultured B-cells from 26 unrelated individuals from the CEPH-Utah collection were used (GM07000, GM10838, GM10839, GM11992, GM11993, GM11994, GM12716, GM12717, GM12877, GM12878, GM10842, GM10843, GM12752, GM12753, GM12864, GM12865, GM10836, GM10837, GM10852, GM10857, GM12766, GM12767, GM10830, GM10831, GM10846, and GM10847). B-cells were cultured at a density of 5 x 10^5 cells/mL in RPMI 1640 with 15% fetal bovine serum, 100 units/mL penicillin-streptomycin and 2mM L-glutamine.

Induction of ER Stress

To induce ER stress, we treated the cells with 4 μg/mL tunicamycin (Sigma-Aldrich) in DMSO. Control cultures were grown in media as described above with 0.5% DMSO. Samples were collected at indicated time points (before treatment, 2, 8, 12, 24, 36, or 48 hours after treatment). To validate ER stress, we assayed for x box-binding protein 1 (XBP1) splicing via PCR resolved by a 2% agarose gel. The following primers were used: forward, 5'-GCTGAAGAGGAGGCGGAAG-3'; reverse, 5'-GTCCAGAATGCCCAACAGG-3'. Protein expression of BiP was assessed via western blot (Cell Signaling #3177). Western blot was also probed with antibody specific for GAPDH as a loading control (Santa Cruz sc-137179).

RNA-sequencing and analysis of ten individuals

RNA-seq of 30 samples was performed as recommended by the manufacturer (Illumina). The 30 samples are comprised of B-cells from 10 individuals (GM07000, GM10838, GM10839, GM11992, GM11993, GM11994, GM12716, GM12717, GM12877, and GM12878) with RNA extracted from three time points each: before treatment and two and eight hours following tunicamycin treatment. RNA was extracted from the cultured B-cells using Qiagen RNeasy purification kits. The quality of the RNA was measured using the BioAnalyzer and only those with RIN > 9.0 were used in this study. cDNA was made using Illumina TruSeq library construction kits. Briefly, RNA was first denatured, then fragmented and first and second strand synthesis was performed using oligo-dT primers. Adapters were added to the cDNA and samples were PCR amplified. The libraries were sequenced on a Hi-Seq instrument (Illumina, San Diego, USA).

The resulting sequences were analyzed as follows. Low quality bases (Phred-scale ≤2) at the 3' ends of reads were trimmed prior to alignment of the reads to the human reference genome (hg18) using GSNAP (version 2012.01.11, Wu and Nacu, 2010). The following parameters were used for mapping: number of mismatches ≤ [(read length+2)/12-2], mapping score ≥ 20, soft-clipping on (trim mismatch score = -3), known exon-exon junctions as defined by RefSeq, Gencode (version 3c) and novel junctions as defined by GSNAP. SNP tolerant alignment was used with SNPs from the CEU population of HapMap (International HapMap Consortium 2003, release #28) and 1000 genomes (1000 Genomes Project Consortium, pilot project). Reads that aligned to only one genomic location were used. Gene expression was quantified using Cufflinks v2.1.1. A gene is considered "expressed" if at least five of the 30 samples had an FPKM ≥ 1.

Targeted RNA-sequencing and Analysis

Targeted RNA-seq of 16 individuals (GM10842, GM10843, GM12752, GM12753, GM12864, GM12865, GM10836, GM10837, GM10852, GM10857, GM12766, GM12767, GM10830, GM10831, GM10846, and GM10847) was performed using the TruSeq Custom Amplicon Library preparation (Illumina) in conjunction with Maxima H Minus Double-Stranded cDNA Synthesis Kit (Thermo Scientific). Briefly, RNA was reverse transcribed to cDNA using oligo-dT primers. A second strand of cDNA was synthesized and the resulting double-stranded cDNA was then purified (QIAGEN). The double-stranded cDNA was then used for the TruSeq Custom Amplicon kit (Illumina). This protocol entailed hybridization of 2 probes unique to a given target (82 total targets), extension and ligation of DNA complementary to the sample,

addition of indices and adapters. This was performed on 96 samples which were then multi-plexed on a MiSeq sequencer (property of Hudson Alpha, Huntsville, Alabama) to produce 250bp, paired-end reads. Probes were designed to 82 regions of genes involved in ER and Golgi function that contained A-to-G editing sites.

The resulting reads were aligned to the targeted sequences from the human reference genome (hg18) using GSNAP. The number of mismatches allowed was less than or equal to [(read length+2)/12-2].

DNA-sequencing and Analysis

DNA-seq of 4 of the 10 individuals (GM10838, GM10839, GM12877, and GM12878) was performed as recommended by the manufacturer (Illumina). DNA was extracted from cultured B-cells followed by preparation of DNA-sequencing libraries using the Illumina TruSeq Paired-End construction kits. Briefly, DNA is fragmented, adapters are ligated onto each fragment and fragments are amplified via PCR. The libraries were sequenced on a HiSeq instrument to generate paired-end 100bp reads.

Low quality bases at the 3' ends of reads were trimmed. The resulting reads were aligned to the human reference genome (hg18) using GSNAP. The following parameters were used for mapping: number of mismatches ≤ [(read length+2)/12-2], mapping score ≥ 20, and soft-clipping on (trim mismatch score = -3). SNP sites found in dbSNP and 1000 genomes were included for SNP-tolerant alignments. Reads that aligned to only one genomic location were used.

Identifying RNA editing sites

To identify RNA editing sites we compared the RNA-sequencing reads to the reference human genome (hg18). To be included as an RNA editing site for a given sample, a site had to pass multiple criteria: 1) 10 total reads from RNA-seq, 2) 2 G-containing reads (5% of the total reads at position), 3) if within an intron, the site must be greater than 10 base pairs from the intron-exon junction, 5) site is not considered a SNP by dbSNP (build 138), 1000 genomes or the Exome Sequencing Project, and 6) no G-containing read is found in any of the 4 DNA samples. Additionally, the final list comprises of sites that are found in at least 10 out of the 30 samples whose RNA was sequenced.

In order to ensure uniqueness of RNA editing sites, further filtering steps were implemented. All sites found in pseudogenes (as annotated by RefSeq) and repetitive regions (RepeatMasker) except Alu elements were removed. Next, the genomic regions around each site (in 25bp, 50bp and 75bp each direction) must be unique as determined by mapping with local sequence alignment (BLAT). Sites were removed if BLAT revealed an alternative alignment with 3 or fewer mismatches. To ensure accurate mapping of RNA-sequencing reads, each read containing the RDD/edited form of the transcript was aligned using a local sequence aligner (BLAT). For a site to be included, over 50% of the RDD/edited reads must align to the same position with BLAT as with GSNAP.

To ensure that the genomic sequence is homozygous across individuals, we used sequencing data from 99 CEU individuals (the same population as the 10 in our dataset) in the 1000 Genomes collection. We identified sites that have at least 10 reads across all individuals

and 99% of the reads containing the reference base (to allow for sequencing error at such great sequencing depths). We also required that all of the DNA sequences contain the reference base if covered by the DNA-sequencing of four individuals (as described above).

ADAR Protein Expression Level

Quantitative PCR was used to detect RNA expression of ADAR (Forward: 5'-GGTAGAGAAGGCTACGTGGTG, Reverse: 5'- CGGGTCTTGCACTTCCTC). A housekeeping gene, NDUF4A, was used as a control (Forward: 5'- GTCAGGCCAAGAAGCATCC, Reverse: 5'-GCTCCAGTAGCTCCAGTTCC). Protein expression of ADAR was assessed via western blot (Sigma HPA003890). Western blot was also probed with antibody specific for GAPDH as a loading control (Santa Cruz sc-137179).

References

1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct 28;467(7319):1061-73. doi: 10.1038/nature09534.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25-9.

Athanasiadis A, Rich A, Maas S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biol. 2004 Dec;2(12):e391. Epub 2004 Nov 9.

Dombroski BA, Nayak RR, Ewens KG, Ankener W, Cheung VG, Spielman RS. Gene expression and genetic variation in response to endoplasmic reticulum stress in human cells. Am J Hum Genet. 2010 May 14;86(5):719-29. doi: 10.1016/j.ajhg.2010.03.017. Epub 2010 Apr 15.

Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2015 Jan;43(Database issue):D1049-56. doi: 10.1093/nar/gku1179. Epub 2014 Nov 26.

International HapMap Consortium. The International HapMap Project. Nature. 2003 Dec 18;426(6968):789-96.

Kozutsumi Y, Segal M, Normington K, Gething MJ, Sambrook J. The presence of malfolded proteins in the endoplasmic reticulum signals the induction of glucose-regulated proteins. Nature. 1988 Mar 31;332(6163):462-4.

Lépine S, Allegood JC, Park M, Dent P, Milstien S, Spiegel S. Sphingosine-1-phosphate phosphohydrolase-1 regulates ER stress-induced autophagy. Cell Death Differ. 2011 Feb;18(2):350-61. doi: 10.1038/cdd.2010.104. Epub 2010 Aug 27.

Lev-Maor G, Sorek R, Levanon EY, Paz N, Eisenberg E, Ast G. RNA-editing-mediated exon evolution. Genome Biol. 2007;8(2):R29.

Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallegger M, Shemesh R, Fligelman ZY, Shoshan A, Pollock SR, Sztybel D, Olshansky M, Rechavi G, Jantsch MF. Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat Biotechnol. 2004 Aug;22(8):1001-5. Epub 2004 Jul 18.

Li JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E, Zhang K, Gao Y, Church GM. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science. 2009 May 29;324(5931):1210-3. doi: 10.1126/science.1170995.

Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG. Widespread RNA and DNA sequence differences in the human transcriptome. Science. 2011 Jul 1;333(6038):53-8. doi: 10.1126/science.1207018. Epub 2011 May 19.

Lomeli H, Mosbacher J, Melcher T, Höger T, Geiger JR, Kuner T, Monyer H, Higuchi M, Bach A, Seeburg PH. Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. Science. 1994 Dec 9;266(5191):1709-13.

Qiu W, Wang X, Buchanan M, He K, Sharma R, Zhang L, Wang Q, Yu J. ADAR1 is essential for intestinal homeostasis and stem cell maintenance. Cell Death Dis. 2013 Apr 18;4:e599. doi: 10.1038/cddis.2013.125.

Scadden AD. The RISC subunit Tudor-SN binds to hyper-edited double-stranded RNA and promotes its cleavage. Nat Struct Mol Biol. 2005 Jun;12(6):489-96. Epub 2005 May 15.

Seeburg PH, Higuchi M, Sprengel R. RNA editing of brain glutamate receptor channels: mechanism and physiology. Brain Res Brain Res Rev. 1998 May;26(2-3):217-29.

Sommer B, Köhler M, Sprengel R, Seeburg PH. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. Cell. 1991 Oct 4;67(1):11-9.

Tkacz JS, Lampen O. Tunicamycin inhibition of polyisoprenyl N-acetylglucosaminyl pyrophosphate formation in calf-liver microsomes. Biochem Biophys Res Commun. 1975 Jul 8;65(1):248-57.

Wang IX, So E, Devlin JL, Zhao Y, Wu M, Cheung VG. ADAR regulates RNA editing, transcript stability, and gene expression. Cell Rep. 2013 Nov 14;5(3):849-60. doi: 10.1016/j.celrep.2013.10.002. Epub 2013 Oct 31.

Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010 Apr 1;26(7):873-81. doi: 10.1093/bioinformatics/btq057. Epub 2010 Feb 10.

Yoshida H, Matsui T, Yamamoto A, Okada T, Mori K. XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. Cell. 2001 Dec 28;107(7):881-91.

Zhang Z, Carmichael GG. The fate of dsRNA in the nucleus: a p54(nrb)-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. Cell. 2001 Aug 24;106(4):465-75.

Figure 5.1

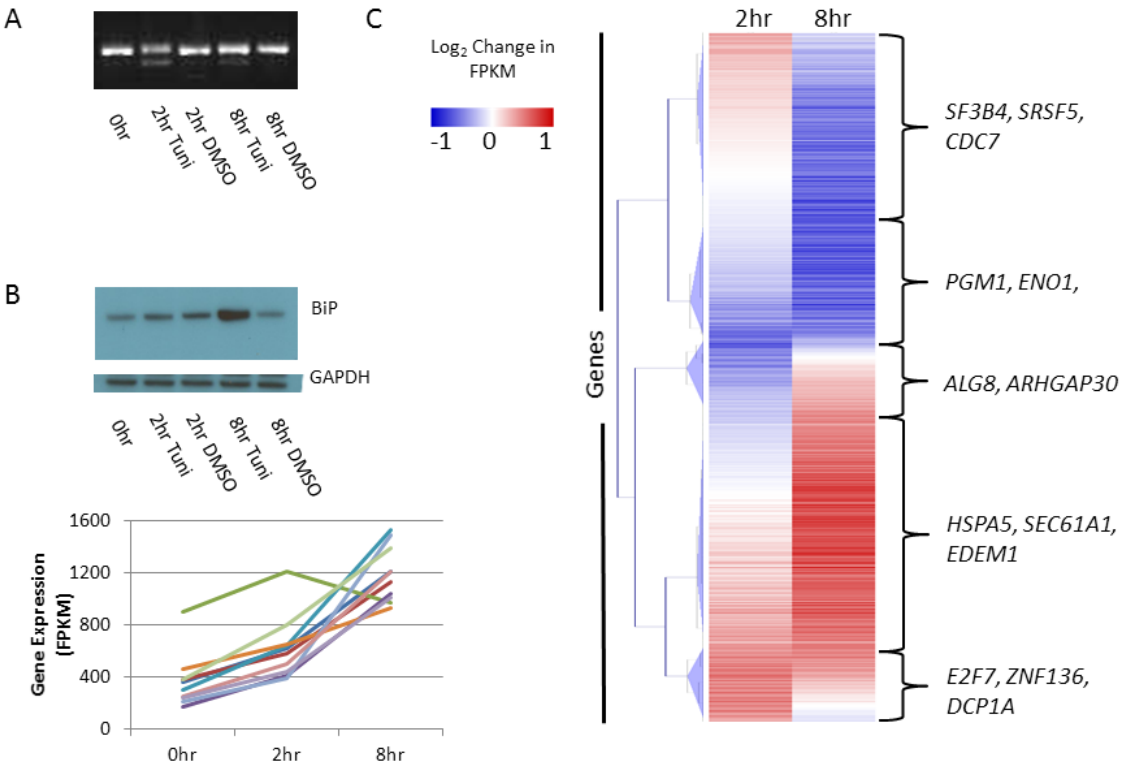

Figure 5.1: Gene expression changes following ER stress. A) B-cells were treated with 4µg/mL tunicamycin. XBP1 splicing was assessed on a 2% agarose gel before treatment (0hr), 2 and 8 hours after treatment with tunicamycin or vehicle, DMSO. The upper band is the unspliced form while the lower band is the spliced form of XBP1. The spliced form of XBP1 is increased at 2 hours following tunicamycin treatment. B) Protein expression of BiP was assessed via western blot using GAPDH as a loading control and is shown for a representative individual (GM11994). Increase in BiP protein expression is apparent at 8 hours following tunicamycin treatment. Gene expression of BiP was determined from ten individuals by RNA-sequencing. FPKM values are shown on the graph for before treatment (0hr), and 2 and 8 hours following tunicamycin treatment. Each line represents one individual. C) Heatmap showing the 4,182 genes that show changes in gene expression following ER stress. Genes were clustered using hierarchical clustering into five groups as shown to the right. Examples of genes are listed for each group. Blue indicates a decrease in gene expression as compared to baseline while red correspond to increases in gene expression.

Figure 5.2

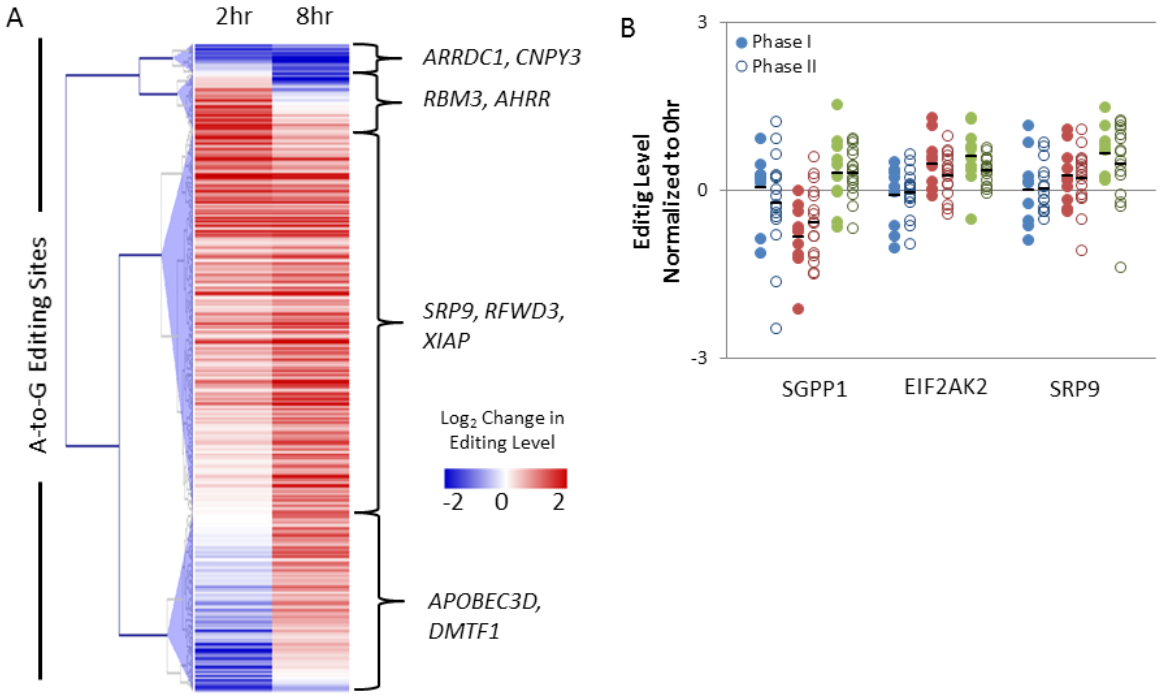

Figure 5.2: Editing level changes following ER stress. A) Heatmap showing the 342 A-to-G editing sites that show changes in level following ER stress. Sites were clustered using hierarchical clustering into four groups as shown to the right. Examples from each cluster are listed. Blue indicates a decrease in level as compared to baseline while red shows an increase in editing level. B) Editing level for three sites are graphed (chr2:223132607 downstream of *SGPP2*, chr2:37181094 downstream of *EIF2AK2*, and chr1:224041268 in *SRP9*). Editing levels are shown for three time points: baseline (0hr) in blue, 2 hours after tunicamycin treatment in red and 8 hours after tunicamycin treatment in green. Solid circles denote editing levels of ten individuals while hollow circles indicate editing levels at the same site from the 16 individuals from the targeted sequencing. Each point shows the editing level of one individual. Black line indicates the average of each group.

Figure 5.3

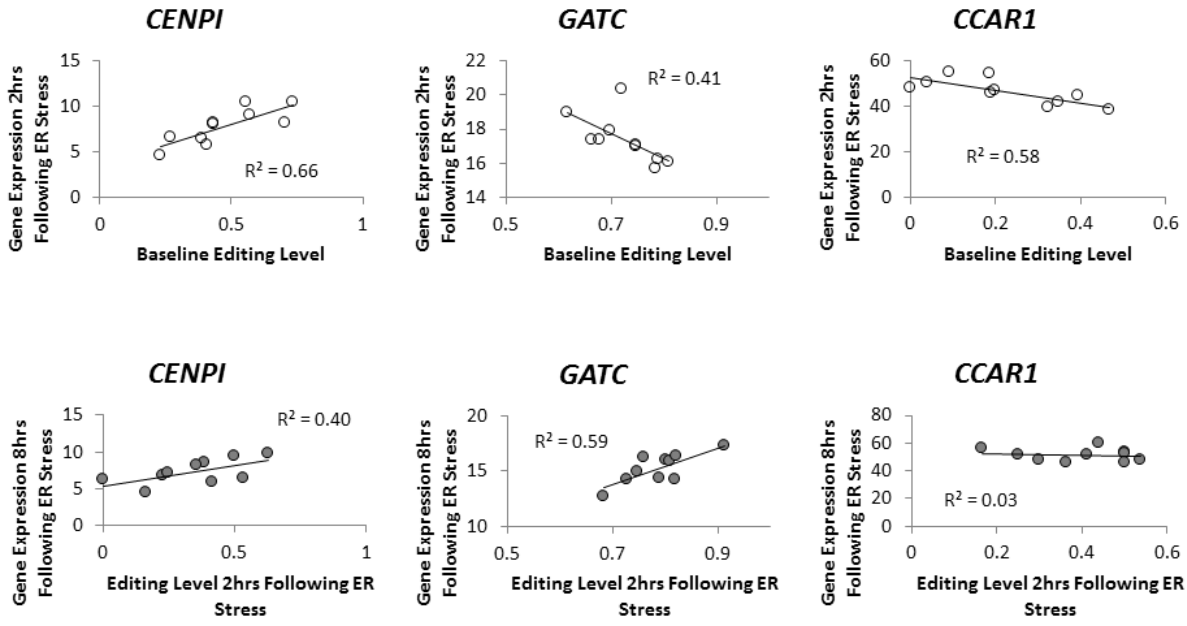

Figure 5.3: Editing level and gene expression following ER stress. Graphs depicting the correlation of editing levels at three sites compared to gene expression at the following time point (chrX:100241566 in *CENPI*, chr12:119383458 in *GATC*, and chr10:70190507 in *CCAR1*). Hollow circles depict baseline editing level compared to gene expression at 2 hours following tunicamycin treatment. Solid circles indicate editing level 2 hours following tunicamycin treatment compared to gene expression 8 hours following tunicamycin treatment. In each graph, data from each individual are represented as dots.

Table S5.1: Regions for targeted sequencing.

Chromosome	Begin Position	End Position	Gene Name
chr1	6205582	6206131	ICMT
chr1	38099203	38099705	INPP5B
chr1	177309321	177309822	FAM20B
chr1	177593282	177593811	SOAT1
chr1	224040924	224041425	SRP9
chr2	24077228	24077729	UBXN2A
chr2	32387609	32389540	YIPF4
chr2	37180818	37181896	EIF2AK2
chr2	37182498	37182998	EIF2AK2
chr2	37184420	37184919	EIF2AK2
chr2	223132510	223133010	SGPP2
chr3	10167622	10168421	VHL
chr3	157741721	157742259	SSR3
chr3	197563406	197563919	UBXN7
chr4	177489112	177489616	SPCS3
chr5	34988582	34989114	DNAJC21
chr5	74709176	74709679	COL4A3BP
chr5	138646161	138646661	MATR3
chr6	43018756	43019275	CNPY3
chr7	44808215	44808743	PPIA
chr7	56113957	56114473	SUMF2
chr8	42995016	42995515	HOOK3
chr8	98809764	98810275	MTDH
chr9	138455144	138456005	SEC16A
chr11	31409531	31410070	DNAJC24
chr11	61373634	61374183	FADS2
chr12	97652638	97653167	APAF1
chr14	52311891	52312399	GNPNAT1
chr15	72106716	72107226	PML
chr16	23383814	23384323	GGA2
chr16	55531587	55532113	HERPUD1
chr16	68847596	68848108	AARS
chr17	4868795	4869294	KIF1C
chr17	4871577	4872100	KIF1C
chr17	40232423	40232923	GJC1
chr17	73930091	73930617	PGS1
chr17	77138910	77139442	NPLOC4
chr19	1409253	1409765	APC2
chr19	12912805	12913332	CALR
chr21	41741431	41741964	MX1
chr22	35213724	35214224	FOXRED2

Table S5.1: Chromosome position is based on hg18.

Table S5.2: Correlation of editing level and gene expression.

Chromosome	Position	Gene Name	r (baseline editing, 2hr expression)	r (2hr editing, 8hr expression)
1	6633172	DNAJC11	0.02	0.77
1	10443352	DFFA	0.64	0.12
2	171886747	METTL8	0.71	0.09
3	40553935	ZNF621	0.77	0.17
5	74709587	COL4A3BP	0.07	0.76
5	74710652	COL4A3BP	0.72	0.03
5	138646437	MATR3	0.38	0.66
7	86653592	DMTF1	0.84	0.39
10	70189946	CCAR1	0.70	0.55
10	70190507	CCAR1	0.76	0.18
12	119383458	GATC	0.64	0.77
12	131807977	PGAM5	0.65	0.12
16	73256026	RFWD3	0.29	0.94
16	88143470	SPG7	0.13	0.65
17	23959638	SGK494	0.71	0.23
19	63046245	LOC100293516	0.59	0.72
X	100241566	CENPI	0.81	0.63

Table S5.2: Correlation coefficient denoted by “r”.

Figure S5.1

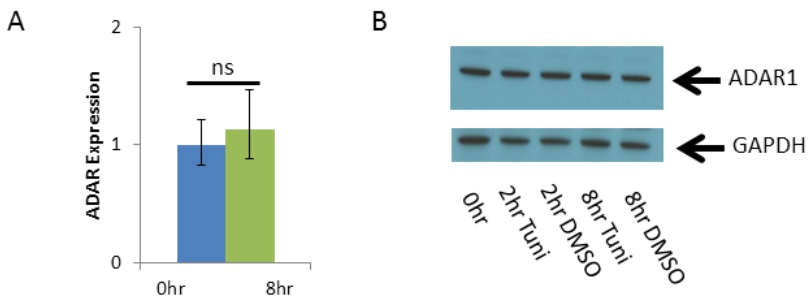

Figure S5.1: ADAR expression under ER stress conditions. A) Quantitative PCR of ADAR expression across five individuals (GM07345, GM11993, GM11994, GM12044, GM12716) before and following 8 hours of tunicamycin treatment. There is not a significant change in expression (p-value > 0.05). B) Protein expression of ADAR was assessed via western blot using GAPDH as a loading control. Western blot shown for a representative example (GM11993).

Chapter 6: Conclusion

Summary of Findings

My study describes RNA editing and RNA-DNA sequence Differences (RDD) in human B-cells and how these processes may play a role within the cell. While there had been a few previous reports demonstrating a non-canonical single-base difference between the RNA and DNA sequence (Sharma et al., 1994, van Leeuwen et al., 1998, Klimek-Tomczak et al., 2006, Grohmann et al., 2010), the Cheung lab was the first to identify these RDDs transcriptome-wide. Since our first paper in 2011, many other groups have also identified RDDs (Ju et al., 2011, Li et al., 2011, Alon et al., 2012, Silberberg et al., 2012, Vesely et al., 2012, Peng et al., 2012, Chen 2013, Lagarrigue et al., 2013). We and others have found that RDDs can be translated into proteins and some are predicted to affect protein function (Sharma et al., 1994, Bar-Yaacov et al., 2013, Wang et al., 2013, Zhu et al., 2014).

The mechanism leading to RDDs was still elusive so next we determined when RDDs are formed during transcription. The Cheung lab found that RDDs occur soon after RNA transcripts exit the RNA polymerase (Wang et al., 2014). This demonstrates that RDDs are not inserted during the process of polymerase elongation but instead, RDDs are mediated through a process that closely follows the polymerase. Furthermore, the Cheung lab found that individuals with a

mutation causing Amyotrophic Lateral Sclerosis 4 resolve R-loops more efficiently and therefore have fewer R-loops and fewer RDDs (Wang et al., 2013). Together, these studies determine the frequencies of RNA editing and RDD occur as well as bring us closer to understanding the mechanism of RDD formation.

In addition to my study of RNA-DNA Differences, I also studied canonical, ADAR-mediated A-to-G editing in human B-cells. Previous studies had reported that ADAR enzymes preferentially edit adenosines in double-stranded RNA structures with no guanosine at the base 5' but with a guanosine at the base 3' to the target site (Polson and Bass, 1994, Lehmann and Bass, 2000). My work has expanded upon this knowledge to demonstrate that genomic variation can affect editing across individuals. Specifically, I identified 11 A-to-G editing sites whose levels were associated with alleles of nearby SNPs. Though we have found these associations, further work will be needed to determine the causative SNP that affects editing level. Thus far, my work suggests that genetic variation may play a role in the individual differences we have seen in editing levels.

Furthermore, I was able to use variability in editing levels across individuals to learn about the mode by which ADAR edits multiple sites within a long double-stranded RNA region. I determined that correlation of editing levels was stronger for sites within the same gene and within the same Alu element. Previous work has demonstrated that Alu elements often pair to form a double-stranded RNA structure that is can be targeted by ADAR enzymes (Athanasiadis et al., 2004, Bazak et al., 2014b). Therefore, my results suggest that ADAR deaminates adenosines on one side of a double-stranded structure at a time.

In addition to studying RNA editing and RDDs at baseline, I also studied the role of RNA editing in the endoplasmic reticulum (ER) stress response. We found over 300 sites that showed changes in editing level following ER stress. Almost 90% of editing sites that changed level following ER stress show an increase in editing level. In addition, over 100 editing sites that change level are found within genes that change expression following ER stress. Of these, 17 sites are correlated with changes in gene expression suggesting that editing may play a role in the gene expression changes occurring in the ER stress response. This work suggests that RNA editing and RDD may play a role in the ER stress response through an effect on gene expression.

Finally, I have studied the RNA editing of a particular site in the 3' UTR of *SEC16A,* a gene involved in protein trafficking from the ER to the Golgi apparatus (Watson et al., 2006). The edited form of *SEC16A* is preferentially bound by HuR suggesting that editing may play a role in transcript stability. If this is found to be the case, then editing of *SEC16A* would be an example of how variability in editing level may affect gene expression. Further studies will help to illuminate the role of editing in this transcript and others.

Significance

There is a large amount of phenotypic variation across many traits, from height to blood pressure. Some of these individual differences can be attributed to genetic variation seen in the DNA (Knoblauch et al., 2002, Rost et al., 2004, Tishkoff et al., 2007, Soranzo et al., 2009, CARDIoGRAMplusC4D Consortium 2013). However, there are many traits that we have not been able to map to a specific location in the genome. My work suggests that there is another form of variation, that seen in the RNA, that may be affecting these various traits. Through

work in the Cheung lab, we have identified ADAR-mediated RNA editing and non-canonical RDDs that show varying levels across individuals (Li et al., 2011). We have also shown that RDDs can be translated into proteins (Li et al., 2011) and that RNA editing in *SEC16A* can influence transcript stability. These data suggest that variability in RNA editing and RDD may affect cellular phenotypes that are not attributed to genetic variation in the DNA.

In addition, RNA editing and RNA-DNA sequence differences can help us to understand various phases of RNA processing. Previous studies have shown that canonical RNA editing can affect splicing, translation and mRNA decay (Rueter et al., 1999, Zhang and Carmichael 2001, Scadden 2005, Lev-Maor et al., 2007). Our work demonstrates that RDDs may be linked to R-loop formation as well (Wang et al., 2013). These works suggest that RNA editing and RDDs are coupled with several RNA processing steps. Therefore, the differences in editing and RDD across individuals and following ER stress could help to explain differences variability in gene expression.

On-going Work

As noted in Chapter 4, I began to study an editing site in the 3'UTR of a gene, *SEC16A.* Previous work from our lab indicated that the editing of this site was mediated by ADAR1 (Wang et al., 2013). Further data that I collected showed that when transcription was halted using actinomycin, the remaining transcripts of *SEC16A* showed an increase in the percentage of editing over time (Figure 4.4).

These data suggest two different scenarios: As RNA is decayed (and not replenished due to the actinomycin treatment), 1) only the most stable transcripts are left at the later time

points suggesting that the edited transcript is more stable or 2) ADAR is continuing to edit the remaining RNA such that by the later time points, most of the *SEC16A* RNA transcripts are edited.

Previous work from our lab demonstrates that ADAR interacts with HuR, an RNA-binding protein[1], to regulate transcript stability[2]. To study this further, I performed an RNA-immunoprecipitation and found that HuR binds preferentially to the edited transcript of *SEC16A* (Figure 4.4). Furthermore, data from the lab demonstrates that editing level is lower in cells depleted of HuR (7% editing versus 12% editing in control samples). Together, these data suggest that scenario 1 may be the cause of the increase in editing level following actinomycin treatment.

In order to follow-up on these results, I treated cells with siRNA designed to knock-down ADAR1 and HuR to study whether the edited transcript is more stable or if the increase in editing level following actinomycin treatment was due to additional ADAR editing. In this experiment, there were 15 samples (5 knock-downs: no siRNA, non-target siRNA, ADAR1 siRNA, HuR siRNA and ADAR1 and HuR siRNAs together) under 3 conditions (no treatment, actinomycin treatment, and DMSO [vehicle] treatment for 12 hours).

First, I validated that ADAR1 expression was reduced in their respective samples (Figure 6.1a,c). I verified that ADAR1 editing activity was also reduced by studying an ADAR1 target, *PPIA* (Figure 6.1d). Next, I studied the changes in editing level of *SEC16A* in ADAR1 knock-down samples following actinomycin treatment. Using ddPCR to quantify editing level of *SEC16A,* I found an initial decrease in editing level of *SEC16A* following ADAR1 knock-down. Under

actinomycin treatment, I still saw an increase in editing level. As ADAR1 protein is reduced by greater than 80% by siRNA treatment, these data suggest that the increase in editing level following actinomycin treatment is not solely due to additional ADAR1 editing.

This led us to the effect of HuR on changes in editing level following actinomycin treatment. I tried 2 different siRNAs designed to knock-down HuR. One siRNA worked more efficiently to knock-down HuR expression and so "HuR 1" was used for the follow-up experiments. Using HuR 1, I found that while HuR was not effectively knocked-down at the RNA level, at the protein level we found a 50% decrease in expression (Figure 6.1b,c). However, while our knock-down of HuR was moderately effective without additional treatment, actinomycin led to an increase in HuR protein expression such that the protein was only knocked-down 25% (Figure 6.1c). This increase in HuR expression with actinomycin treatment led to difficulty in validating loss of HuR activity under these conditions. We found that one target gene, *H1F0,* did show a decrease in transcript stability following HuR knock-down suggesting that the HuR knock-down was sufficient (Figure 6.1e). I, therefore, continued by studying the editing of *SEC16A* under these conditions. Following actinomycin treatment and HuR knock-down I still saw an increase in editing level of *SEC16A* suggesting that HuR is not solely responsible for this change with the caveat being that HuR expression is not completely suppressed.

Future Directions

My thesis describes the role and regulation of RNA editing and RNA-DNA sequence Differences. The first identified examples of RNA editing were in select genes (Sommer 1991,

Köhler 1993, Nutt 1994, Burns 1997). Advances in technology, namely the development of high-throughput sequencing technology, has allowed for the study of RNA editing throughout the entire transcriptome. This has led to the detection of thousands of RNA editing and RDD sites in humans and in other organisms (Athanasiadis et al., 2004, Levanon et al., 2004, Li et al., 2009, Bazak et al., 2014a, Wang et al., 2013). My thesis and the work of others in the Cheung lab has helped us to understand the regulation of these sites and why they may be important.

I have found that editing and RDD levels change for over 300 sites following ER stress. An increase in editing has previously been characterized for a different environmental state, viral infection. When a cell is invaded by an RNA virus, it promotes the interferon response. This induces an isoform of ADAR that then edits the virus in an attempt to destroy the virus's genetic information (Patterson and Samuel, 1995, Liu et al., 1997). While the goal of editing a virus is to make mistakes and create mismatches, the role of changes in editing in the host system is less clear when found under ER stress conditions. My work suggests that the changes in editing levels may affect the cellular response to stress through their influence on gene expression. The sites that change levels following ER stress are located in the introns and 3'UTRs of genes involved in the ER stress response, such as protein chaperones. The intronic and 3' untranslated regions of an mRNA transcript are often targets of proteins that regulate gene expression through multiple mechanisms, including splicing and mRNA decay. This suggests that editing in these regions may regulate gene expression, as in the case of *SEC16A*, of genes involved in the ER stress response. If these editing sites play a role in gene expression then a change in editing and RDD level at over 300 sites following ER stress may affect the ability of the cell to respond to the cellular stress. Future studies will be aimed at understanding

the role of each editing site and how editing may influence the cellular response to stress. My work suggests that RNA editing may play a larger role in cellular response than previously thought. This work also begs the question of whether RNA editing and RDDs affect cellular response to other stimuli.

In addition to differences in editing and RDD level following ER stress, we also see differences in level across individuals. Some editing sites, such as those in the GluR2 gene, have been shown to be edited to almost 100% in every individual (Sommer et al., 1991). However, over 40% of editing sites in my study show variable editing levels. When we consider their possible effect on gene expression, as seen with *SEC16A*, we can see that editing may play a role in the previously described individual variation in gene expression (Cheung et al., 2003). Editing may not only be important in cellular response but may be responsible for individual variation that we see at baseline. While some of my work suggests that *cis*-factors regulate editing level, not all of the variation in editing level that we see can be attributed to these factors. As described above, editing levels can be affected by the environment making them a way by which genotypes and environment may interact to create a phenotype. Future studies of editing and RDDs at baseline and following cellular stress will be able to dissect the interaction of genotype and environment and their effect on editing and RDD level. Our work in the Cheung lab studying the regulation of RNA editing and RDD should aid in this process.

In summary

RNA editing and RNA-DNA sequence Differences can be found transcriptome-wide across tissues and individuals. RNA editing and RDD level can be affected by cellular stress

conditions, genomic variation in *cis* and RNA secondary structure. Variability in levels may lead differences in gene expression and RNA stability as demonstrated by editing in *SEC16A*. My work in RNA editing and RDDs sheds light on the regulation of these processing steps and how they may influence cellular phenotypes.

Methods

Knock-down and sample analysis

Cultured B-cell from GM11994 was cultured at a density of 5 x 10^5 cells/mL in RPMI 1640 with 15% fetal bovine serum, 100 units/mL penicillin-streptomycin and 2mM L-glutamine. Knock-down was performed following Accell protocol (Dharmacon) using four different siRNAs. SiRNAs designed to target HuR were custom designed (HuR 1: 5' – AAGAGGCAAUUACCAGUUUCA, HuR 2: 5' – AAUCUUAAGUUUCGUAAGUUA) while those designed to target ADAR or act as a control were SmartPools (E-008630-00-0020 and D-001910-10-20, respectively). Briefly, cells were seeded at 600,000 cells/mL with 1µM siRNA in Accell media for 36 hours. After 36 hours, the cells were spun and resuspended in complete media for 24 hours. After 24 hours, cells were untreated, or treated with 5µg/mL actinomycin or vehicle, DMSO, and collected after an additional 12 hours. DNA, RNA and protein were extracted from these samples.

Primers to detect changes in RNA expression or perform Sanger sequencing can be found in Table 6.1. To detect changes at the protein level, three antibodies were utilized: ADAR (Sigma HPA003890), HuR (Millipore 03-102) and Tubulin (Millipore 05-661).

References

Alon S, Mor E, Vigneault F, Church GM, Locatelli F, Galeano F, Gallo A, Shomron N, Eisenberg E. Systematic identification of edited microRNAs in the human brain. Genome Res. 2012 Aug;22(8):1533-40. doi: 10.1101/gr.131573.111. Epub 2012 Apr 12.

Athanasiadis A, Rich A, Maas S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biol. 2004 Dec;2(12):e391. Epub 2004 Nov 9.

Bar-Yaacov D, Avital G, Levin L, Richards AL, Hachen N, Rebolledo Jaramillo B, Nekrutenko A, Zarivach R, Mishmar D. RNA-DNA differences in human mitochondria restore ancestral form of 16S ribosomal RNA. Genome Res. 2013 Nov;23(11):1789-96. doi: 10.1101/gr.161265.113. Epub 2013 Aug 2.

a) Bazak L, Haviv A, Barak M, Jacob-Hirsch J, Deng P, Zhang R, Isaacs FJ, Rechavi G, Li JB, Eisenberg E, Levanon EY. A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. Genome Res. 2014 Mar;24(3):365-76. doi: 10.1101/gr.164749.113. Epub 2013 Dec 17.

b) Bazak L, Levanon EY[1], Eisenberg E[2]. Genome-wide analysis of Alu editability. Nucleic Acids Res. 2014 Jun;42(11):6876-84. doi: 10.1093/nar/gku414. Epub 2014 May 14.

Bertrand E, Chartrand P, Schaefer M, Shenoy SM, Singer RH, Long RM. Localization of ASH1 mRNA particles in living yeast. Mol Cell. 1998 Oct;2(4):437-45.

Burns CM, Chu H, Rueter SM, Hutchinson LK, Canton H, Sanders-Bush E, Emeson RB. Regulation of serotonin-2C receptor G-protein coupling by RNA editing. Nature. 1997 May 15;387(6630):303-8.

CARDIoGRAMplusC4D Consortium, Deloukas P, Kanoni S, Willenborg C, Farrall M, Assimes TL, Thompson JR, Ingelsson E, Saleheen D, Erdmann J, Goldstein BA, Stirrups K, König IR, Cazier JB, Johansson A, Hall AS, Lee JY, Willer CJ, Chambers JC, Esko T, Folkersen L, Goel A, Grundberg E, Havulinna AS, Ho WK, Hopewell JC, Eriksson N, Kleber ME, Kristiansson K, Lundmark P, Lyytikäinen LP, Rafelt S, Shungin D, Strawbridge RJ, Thorleifsson G, Tikkanen E, Van Zuydam N, Voight BF, Waite LL, Zhang W, Ziegler A, Absher D, Altshuler D, Balmforth AJ, Barroso I, Braund PS, Burgdorf C, Claudi-Boehm S, Cox D, Dimitriou M, Do R; DIAGRAM Consortium; CARDIOGENICS Consortium, Doney AS, El Mokhtari N, Eriksson P, Fischer K, Fontanillas P, Franco-Cereceda A, Gigante B, Groop L, Gustafsson S, Hager J, Hallmans G, Han BG, Hunt SE, Kang HM, Illig T, Kessler T, Knowles JW, Kolovou G, Kuusisto J, Langenberg C, Langford C, Leander K, Lokki ML, Lundmark A, McCarthy MI, Meisinger C, Melander O, Mihailov E, Maouche S, Morris AD, Müller-Nurasyid M; MuTHER Consortium, Nikus K, Peden JF, Rayner NW, Rasheed A, Rosinger S, Rubin D, Rumpf MP, Schäfer A, Sivananthan M, Song C, Stewart AF, Tan ST, Thorgeirsson G, van der Schoot CE, Wagner PJ; Wellcome Trust Case Control Consortium, Wells GA, Wild PS, Yang TP, Amouyel P, Arveiler D, Basart H, Boehnke M, Boerwinkle E, Brambilla P, Cambien F, Cupples AL, de Faire U, Dehghan A, Diemert P, Epstein SE, Evans A, Ferrario MM, Ferrières J, Gauguier D, Go AS, Goodall AH, Gudnason V, Hazen SL, Holm H, Iribarren C, Jang Y,

Kähönen M, Kee F, Kim HS, Klopp N, Koenig W, Kratzer W, Kuulasmaa K, Laakso M, Laaksonen R, Lee JY, Lind L, Ouwehand WH, Parish S, Park JE, Pedersen NL, Peters A, Quertermous T, Rader DJ, Salomaa V, Schadt E, Shah SH, Sinisalo J, Stark K, Stefansson K, Trégouët DA, Virtamo J, Wallentin L, Wareham N, Zimmermann ME, Nieminen MS, Hengstenberg C, Sandhu MS, Pastinen T, Syvänen AC, Hovingh GK, Dedoussis G, Franks PW, Lehtimäki T, Metspalu A, Zalloua PA, Siegbahn A, Schreiber S, Ripatti S, Blankenberg SS, Perola M, Clarke R, Boehm BO, O'Donnell C, Reilly MP, März W, Collins R, Kathiresan S, Hamsten A, Kooner JS, Thorsteinsdottir U, Danesh J, Palmer CN, Roberts R, Watkins H, Schunkert H, Samani NJ. Large-scale association analysis identifies new risk loci for coronary artery disease. Nat Genet. 2013 Jan;45(1):25-33. doi: 10.1038/ng.2480. Epub 2012 Dec 2.

Chen L. Characterization and comparison of human nuclear and cytosolic editomes. Proc Natl Acad Sci U S A. 2013 Jul 16;110(29):E2741-7. doi: 10.1073/pnas.1218884110. Epub 2013 Jul 1.

Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, Spielman RS. Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet. 2003 Mar;33(3):422-5. Epub 2003 Feb 3.

Gong C, Tang Y, Maquat LE. mRNA-mRNA duplexes that autoelicit Staufen1-mediated mRNA decay. Nat Struct Mol Biol. 2013 Oct;20(10):1214-20. doi: 10.1038/nsmb.2664. Epub 2013 Sep 22.

Grohmann M, Hammer P, Walther M, Paulmann N, Büttner A, Eisenmenger W, Baghai TC, Schüle C, Rupprecht R, Bader M, Bondy B, Zill P, Priller J, Walther DJ. Alternative splicing and extensive RNA editing of human TPH2 transcripts. PLoS One. 2010 Jan 29;5(1):e8956. doi: 10.1371/journal.pone.0008956.

Ju YS, Kim JI, Kim S, Hong D, Park H, Shin JY, Lee S, Lee WC, Kim S, Yu SB, Park SS, Seo SH, Yun JY, Kim HJ, Lee DS, Yavartanoo M, Kang HP, Gokcumen O, Govindaraju DR, Jung JH, Chong H, Yang KS, Kim H, Lee C, Seo JS. Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. Nat Genet. 2011 Jul 3;43(8):745-52. doi: 10.1038/ng.872.

Klimek-Tomczak K, Mikula M, Dzwonek A, Paziewska A, Karczmarski J, Hennig E, Bujnicki JM, Bragoszewski P, Denisenko O, Bomsztyk K, Ostrowski J. Editing of hnRNP K protein mRNA in colorectal adenocarcinoma and surrounding mucosa. Br J Cancer. 2006 Feb 27;94(4):586-92.

Knoblauch H, Bauerfeind A, Krähenbühl C, Daury A, Rohde K, Bejanin S, Essioux L, Schuster H, Luft FC, Reich JG. Common haplotypes in five genes influence genetic variance of LDL and HDL cholesterol in the general population. Hum Mol Genet. 2002 Jun 1;11(12):1477-85.

Köhler M, Burnashev N, Sakmann B, Seeburg PH. Determinants of Ca2+ permeability in both TM1 and TM2 of high affinity kainate receptor channels: diversity by RNA editing. Neuron. 1993 Mar;10(3):491-500.

Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science. 2013 Feb 22;339(6122):950-3. doi: 10.1126/science.1229386.

Lagarrigue S, Hormozdiari F, Martin LJ, Lecerf F, Hasin Y, Rau C, Hagopian R, Xiao Y, Yan J, Drake TA, Ghazalpour A, Eskin E, Lusis AJ. Limited RNA editing in exons of mouse liver and adipose. Genetics. 2013 Apr;193(4):1107-15. doi: 10.1534/genetics.112.149054. Epub 2013 Feb 14.

Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Rådmark O, Kim S, Kim VN. The nuclear RNase III Drosha initiates microRNA processing. Nature. 2003 Sep 25;425(6956):415-9.

van Leeuwen FW, de Kleijn DP, van den Hurk HH, Neubauer A, Sonnemans MA, Sluijs JA, Köycü S, Ramdjielal RD, Salehi A, Martens GJ, Grosveld FG, Peter J, Burbach H, Hol EM. Frameshift mutants of beta amyloid precursor protein and ubiquitin-B in Alzheimer's and Down patients. Science. 1998 Jan 9;279(5348):242-7.

Lehmann KA, Bass BL. Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. Biochemistry. 2000 Oct 24;39(42):12875-84.

Lev-Maor G, Sorek R, Levanon EY, Paz N, Eisenberg E, Ast G. RNA-editing-mediated exon evolution. Genome Biol. 2007;8(2):R29.

Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallegger M, Shemesh R, Fligelman ZY, Shoshan A, Pollock SR, Sztybel D, Olshansky M, Rechavi G, Jantsch MF. Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat Biotechnol. 2004 Aug;22(8):1001-5. Epub 2004 Jul 18.

Li JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E, Zhang K, Gao Y, Church GM. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science. 2009 May 29;324(5931):1210-3. doi: 10.1126/science.1170995.

Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG. Widespread RNA and DNA sequence differences in the human transcriptome. Science. 2011 Jul 1;333(6038):53-8. doi: 10.1126/science.1207018. Epub 2011 May 19.

Liu Y, George CX, Patterson JB, Samuel CE. Functionally distinct double-stranded RNA-binding domains associated with alternative splice site variants of the interferon-inducible double-stranded RNA-specific adenosine deaminase. J Biol Chem. 1997 Feb 14;272(7):4419-28.

Nutt SL, Hoo KH, Rampersad V, Deverill RM, Elliott CE, Fletcher EJ, Adams SL, Korczak B, Foldes RL, Kamboj RK. Molecular characterization of the human EAA5 (GluR7) receptor: a high-affinity kainate receptor with novel potential RNA editing sites. Receptors Channels. 1994;2(4):315-26.

Patterson JB, Samuel CE. Expression and regulation by interferon of a double-stranded-RNA-specific adenosine deaminase from human cells: evidence for two forms of the deaminase. Mol Cell Biol. 1995 Oct;15(10):5376-88.

Peng Z, Cheng Y, Tan BC, Kang L, Tian Z, Zhu Y, Zhang W, Liang Y, Hu X, Tan X, Guo J, Dong Z, Liang Y, Bao L, Wang J. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nat Biotechnol. 2012 Feb 12;30(3):253-60. doi: 10.1038/nbt.2122.

Polson AG, Bass BL. Preferential selection of adenosines for modification by double-stranded RNA adenosine deaminase. EMBO J. 1994 Dec 1;13(23):5701-11.

Rost S, Fregin A, Ivaskevicius V, Conzelmann E, Hörtnagel K, Pelz HJ, Lappegard K, Seifried E, Scharrer I, Tuddenham EG, Müller CR, Strom TM, Oldenburg J. Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2. Nature. 2004 Feb 5;427(6974):537-41.

Rueter SM, Dawson TR, Emeson RB. Regulation of alternative splicing by RNA editing. Nature. 1999 May 6;399(6731):75-80.

Scadden AD. The RISC subunit Tudor-SN binds to hyper-edited double-stranded RNA and promotes its cleavage. Nat Struct Mol Biol. 2005 Jun;12(6):489-96. Epub 2005 May 15.

Sharma PM, Bowman M, Madden SL, Rauscher FJ 3rd, Sukumar S. RNA editing in the Wilms' tumor susceptibility gene, WT1. Genes Dev. 1994 Mar 15;8(6):720-31.

Silberberg G, Lundin D, Navon R, Öhman M. Deregulation of the A-to-I RNA editing mechanism in psychiatric disorders. Hum Mol Genet. 2012 Jan 15;21(2):311-21. doi: 10.1093/hmg/ddr461. Epub 2011 Oct 7.

Sommer B, Köhler M, Sprengel R, Seeburg PH. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. Cell. 1991 Oct 4;67(1):11-9.

Soranzo N, Rivadeneira F, Chinappen-Horsley U, Malkina I, Richards JB, Hammond N, Stolk L, Nica A, Inouye M, Hofman A, Stephens J, Wheeler E, Arp P, Gwilliam R, Jhamai PM, Potter S, Chaney A, Ghori MJ, Ravindrarajah R, Ermakov S, Estrada K, Pols HA, Williams FM, McArdle WL, van Meurs JB, Loos RJ, Dermitzakis ET, Ahmadi KR, Hart DJ, Ouwehand WH, Wareham NJ, Barroso I, Sandhu MS, Strachan DP, Livshits G, Spector TD, Uitterlinden AG, Deloukas P. Meta-analysis of genome-wide scans for human adult stature identifies novel Loci and associations with measures of skeletal frame size. PLoS Genet. 2009 Apr;5(4):e1000445. doi: 10.1371/journal.pgen.1000445. Epub 2009 Apr 3.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, Ibrahim M, Omar SA, Lema G, Nyambo TB, Ghori J, Bumpstead S, Pritchard JK, Wray GA, Deloukas P. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet. 2007 Jan;39(1):31-40. Epub 2006 Dec 10.

Vesely C, Tauber S, Sedlazeck FJ, von Haeseler A, Jantsch MF. Adenosine deaminases that act on RNA induce reproducible changes in abundance and sequence of embryonic miRNAs. Genome Res. 2012 Aug;22(8):1468-76. doi: 10.1101/gr.133025.111. Epub 2012 Feb 6.

Wang IX, So E, Devlin JL, Zhao Y, Wu M, Cheung VG. ADAR regulates RNA editing, transcript stability, and gene expression. Cell Rep. 2013 Nov 14;5(3):849-60. doi: 10.1016/j.celrep.2013.10.002. Epub 2013 Oct 31.

Wang IX, Core LJ, Kwak H, Brady L, Bruzel A, McDaniel L, Richards AL, Wu M, Grunseich C, Lis JT, Cheung VG. RNA-DNA differences are generated in human cells within seconds after RNA exits polymerase II. Cell Rep. 2014 Mar 13;6(5):906-15. doi: 10.1016/j.celrep.2014.01.037. Epub 2014 Feb 20.

Watson P, Townley AK, Koka P, Palmer KJ, Stephens DJ. Sec16 defines endoplasmic reticulum exit sites and is required for secretory cargo export in mammalian cells. Traffic. 2006 Dec;7(12):1678-87. Epub 2006 Sep 27.

Zhang Z, Carmichael GG. The fate of dsRNA in the nucleus: a p54(nrb)-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. Cell. 2001 Aug 24;106(4):465-75.

Zhu Y, Luo H, Zhang X, Song J, Sun C, Ji A, Xu J, Chen S. Abundant and selective RNA-editing events in the medicinal mushroom Ganoderma lucidum. Genetics. 2014 Apr;196(4):1047-57. doi: 10.1534/genetics.114.161414. Epub 2014 Feb 4.

Figure 6.1

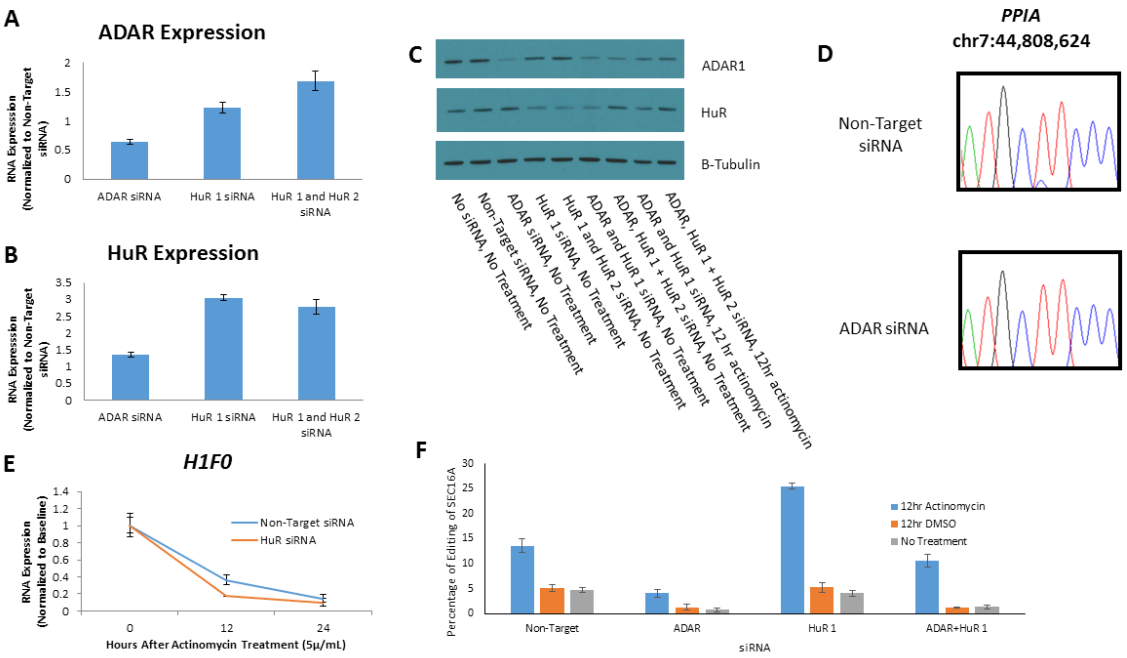

Figure 6.1: Effect of HuR and ADAR knock-down on editing of *SEC16A*. A) ADAR RNA expression normalized to sample treated with non-targeting siRNA. B) HuR RNA expression normalized to sample treated with non-targeting siRNA. C) Protein expression of ADAR and HuR using β-Tubulin as a loading control. D) Decrease in editing level of *PPIA* following ADAR knock-down. E) Decreased stability of *H1F0* transcript in samples treated with HuR siRNA as compared to non-targeting siRNA. F) Changes in editing level in one individual (GM11994) following knock-down of ADAR, HuR or both and treatment with 5µg/mL actinomycin or vehicle, DMSO. Samples without treatment of either actinomycin or DMSO and treated with non-targeting siRNA are shown as a control.

Table 6.1: Primers for ADAR and HuR knock-down.

Target	Use	Forward	Reverse
ADAR	Expression	GGTAGAGAAGGCTACGTGGTG	CGGGTCTTGCACTTCCTC
HuR	Expression	ACCTCCCTCAGAACATGACC	CCAAGCTGTGTCCTGCTACT
NDUF4A	Expression	GTCAGGCCAAGAAGCATCC	GCTCCAGTAGCTCCAGTTCC
H1F0	Expression	TGTCCTCAAGCAGACCAAAG	TGAAGGCCACTGACTTCTTG
PPIA	Sanger Seq	GAACACTGTTGATGTTCTTGAGG	CCTCTGCAGGGAGACTGACT
SEC16A	Sanger Seq	ACCTGGCTGAATGAGTGGAG	AAAATCACCCATGGTCCTCA

Table 6.1: All primers listed from 5’ to 3’.