UNIVERSITY OF CALIFORNIA,
IRVINE

Investigating the extent and function of RNA editing in animal transcriptomes

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biological Sciences

by

Eddie Park

Dissertation Committee:
Assistant Professor S. Ali Mortazavi, Chair
Professor Ken W. Cho
Professor R. Michael Mulligan
Assistant Professor Susanne M. Rafelski
Associate Professor Kevin R. Thornton

2015

DEDICATION

To my mother for her love and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my committee chair, Ali Mortazavi, who has taken me under his wing and given me the opportunity to work on many interesting projects. In A Mathematician's Survival Guide, Stephen Krantz wrote that when your advisor gives you a problem to work on "he is doing you a tremendous favor. Your advisor is, in effect saying 'Here is something worth doing and is at your level. It is a doable problem and you will get a publishable paper out of it. People are interested in this topic and you will begin your reputation by solving this problem. Moreover, working on this problem will lead you to other worthwhile things later on.'" I am very grateful that I was given such an opportunity and it has paved the way to other promising endeavors.

I would like to thank each of my committee members for their valuable contributions toward my graduate training. Ken Cho has played a key role in graduate career from the beginning by allowing me to do my first rotation in his lab, helping me find a position in the Mortazavi lab, and keeping me involved with ongoing collaborations. I am grateful for the insightful advice and guidance that Michael Mulligan provided during times of tribulation as well as times of celebration. For the thought-provoking discussions during journal club, I am very grateful towards Susanne Rafelski, whose positive attitude has often been like a ray of sunshine on a gloomy day. Lastly, I would like to acknowledge Kevin Thornton for the critical role he played in the ICE-seq project

On a final note, I would like to thank my friends and family who have made this journey worthwhile, my collaborators who have made the pursuit of knowledge enjoyable, and my labmates for their camaraderie.

The text of Chapter 2 is a reprint of the material as it appears in “RNA editing in the human ENCODE RNA-seq data,” which was published in Genome Research. The text of Chapter 3 is a reprint of the material as it appears in “Genome-wide identification and functional analysis of Apobec-1-mediated C-to-U RNA editing in mouse small intestine and liver,” which was published in Genome Biology. The co-authors listed in these publications collaborated, directed and supervised research. Full citations of these articles are listed in the Curriculum Vitae.

CURRICULUM VITAE

Eddie Park

Education

2007 BS in Biochemistry, University of California, Riverside

2007 BS in Mathematics, University of California, Riverside

2009 MS in Biochemistry and Molecular Biology, University of California, Riverside

2010 MA in Mathematics, University of California, Riverside

2015 PhD in Biological Sciences, University of California, Irvine

Research Experience

Graduate Student Researcher
University of California Irvine: 2010-2015
Department of Developmental and Cell Biology
Advisor: Ali Mortazavi

Laboratory Volunteer
University of California Riverside: 2005-2010
Department of Biochemistry
Advisor: Ernest Martinez

Publications

1. RNA editing in the human ENCODE RNA-seq data.
Park E, Williams B, Wold BJ, Mortazavi A.
Genome Res. 2012 Sep;22(9):1626-33. doi: 10.1101/gr.134957.111.

2. Landscape of transcription in human cells.
Djebali S, Davis CA, … Mortazavi A, … Park E, … Guigó R, Gingeras TR.
Nature. 2012 Sep 6;489(7414):101-8. doi: 10.1038/nature11233.

3. An integrated encyclopedia of DNA elements in the human genome.

ENCODE Project Consortium.
Nature. 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247.

4. Genome-wide identification and functional analysis of Apobec-1-mediated C-to-U RNA editing in mouse small intestine and liver.
Blanc V, Park E, Schaefer S, Miller M, Lin Y, Kennedy S, Billing AM, Ben Hamidane H, Graumann J, Mortazavi A, Nadeau JH, Davidson NO.
Genome Biol. 2014 Jun 19;15(6):R79. doi: 10.1186/gb-2014-15-6-r79.

5. Comparative Analysis of RNA editing in two Drosophila species.
Park E, Rogers R, Andolfatto P, Thornton K, Mortazavi A.
In Review. Genome Biology and Evolution.

6. Mutations of NARS2 encoding mitochondrial asparaginyl-tRNA synthetase cause nonsyndromic deafness or Leigh Syndrome.
Simon M, Richard EM, … Park E, … Huang T, Riazuddin S.
PLoS Genet. 2015 Mar 25;11(3):e1005097.

Teaching

University of California Irvine – Biology Teaching Assistant

BioSciD111L	Dev&Cell Bio Lab	Winter12, Spring12, Winter13
BioSci93	DNA to Organisms	Fall13
BioSci97	Genetics	Fall12

University of California Riverside - Mathematics Teaching Assistant

Math4	Intro to College Math	Fall08
Math8A	College Math for Sci. (A)	Fall08
Math8B	College Math for Sci. (B)	Spring09, Winter10
Math9B	First-year Calculus (B)	Winter09
Math10A	Calculus: Several Variables	Winter10
Math22	Calculus for Business	Winter09
Math46	Ordinary Diff. Equations	Spring09

ABSTRACT OF THE DISSERTATION

Investigating the extent and function of RNA editing in animal transcriptomes

By

Eddie Park

Doctor of Philosophy in Biological Sciences

University of California, Irvine, 2015

Assistant Professor Ali Mortazavi, Chair

RNA editing is a post-transcriptional process in which certain nucleotides in RNA become modified to change the sequence of the RNA. The two types of RNA editing involve the deamination of adenosine to inosine (A-to-I) and the deamination of cytosine to uracil (C-to-U) that are mediated respectively by the ADAR family of enzymes and APOBEC1.

Although there are many examples of RNA editing altering the coding sequence of specific genes, the extent of RNA editing within metazoans has remained an open problem. Here, I investigate the extent and function of RNA editing in human cell lines, mice, and flies using RNA-seq as well as a new method called ICE-seq. I first assessed the extent of RNA editing as part of the ENCODE project. RNA-seq datasets from fourteen human cell lines were mined for RNA editing events using a novel method that relies on ENCODE ChIP-seq datasets to filter out genomic SNPs. I separately identified new C-to-U RNA editing targets within mouse livers and intestinal enterocytes by comparing wild-type and Apobec1 knockout mice RNA-seq data. I identified 56 Apobec1 RNA editing targets and found that a subset of the targets have altered mRNA and protein levels. Finally, I used ICE-seq to study RNA editing within two species of *Drosophila*. ICE-seq is a method that includes an acrylonitrile treatment of RNA prior to reverse-transcription. The acrylonitrile specifically reacts with inosine to block reverse transcription. By comparing ICE-seq and RNA-seq samples, RNA editing events were called from *Drosophila melanogaster* and *Drosophila yakuba*. I found that there is a core set of RNA editing targets that are shared between the two species and a set of targets with distinct functional enrichments that are species-specific, which suggests that the biological function of RNA editing in these two *Drosophila* species is still evolving.

Chapter 1

RNA modifications in the era of high-throughput sequencing

Abstract

Messenger RNA is the primary intermediate between the genome and the proteome and just like DNA and proteins it is modified with post-transcriptional modifications, which have clear and distinct biological functions. High-throughput sequencing has enabled transcriptome-wide study of several of these RNA modifications by allowing researchers to identify and measure these modifications. Here we describe recent advances in the study of RNA modifications such as A-to-I editing, C-to-U editing, RNA methylation, and pseudouridylation and their potential function that represent the basis for the emerging field of epitranscriptomics.

Introduction

There are more than 100 documented post-transcriptional RNA modifications, some of which have been known to occur for more than half a century (Cantara 2011, Machnicka 2013). The vast majority of known modifications occur in tRNAs while 13 occur in mRNA. In the past few years, the field of RNA modifications has seen a huge burst of progress, primarily due to recent high-throughput sequencing assays, such as RNA-seq. Prior to the development of ubiquitous sequencing assays, researchers were limited to labor-intensive low-throughput experimental techniques such as poisoned-primer extension assays and Sanger sequencing. RNA modifications were originally identified by dissociating RNA into individual nucleotides and using chromatography or mass-spectrometry based techniques to estimate the fraction of bases modified. In this review, we will discuss some of the recent advances in the field of post-transcriptional RNA modifications and the role sequencing has played in making these advances.

RNA modifications can be categorized into three broad, overlapping groups: RNA editing, epitranscriptomics, and RNA processing. RNA editing refers to the biological process in which the sequence of RNA is changed by a base substitution or indel. Epitranscriptomics, in analogy with epigenomics, refers to enzymatic modification to a base without changing the sequence. RNA processing encompasses the post-transcriptional events of splicing, 5' capping, and poly-adenylation, which we will not cover in detail in this review. These designations are not rigid and are often used interchangeably. Most modifications are on a subset of target molecules, creating multiple potential functional isoforms within the same cells. Hence, it is necessary to know both the sites of the modification and the rate of modification at a given location on RNA molecules.

General strategies for identifying RNA modifications

There are three types of general strategies that are used to identify RNA modifications using sequencing. The first strategy is to mine sequence variation in RNA-seq where the sequence of the RNA does not match the sequence in the underlying genome. This method can be used to identify RNA editing events and to measure editing levels for A-to-I and C-to-U modifications by mining evidence in cDNA. Initial attempts at mining RNA-seq for new edited sites suggested that there were a spectrum additional types of RNA editing events in human (Li 2011) in addition to the canonical A-to-I and C-to-U RNA editing. However subsequent analysis demonstrated that multiple sources of errors such as sequencing errors, unannotated SNPs, and other noise that could account for these non-canonical edits (Schrider 2011, Pickrell 2012, Lin 2012, Kleinman 2012), highlighting the difficulty of discovering true novel RNA editing sites from RNA-seq data alone. Another RNA modification that can be mined with RNA-seq involves bisulfite treatment of the RNA (Squires 2012) that alters the sequence of unmethylated cytosine but not methylated cytosine to create mismatches in the cDNA to the genome, which in this case are at the unmethylated cytosines. Unfortunately, other types of RNA modifications are not known to alter the cDNA sequence and are not readily detected by this approach.

A second general strategy is to sterically prevent reverse transcription or to cleave the RNA molecule at the position of the modification so that the end of the cDNA will indicate the position of the modification (Carlile 2014, Schwartz 2014). A chemical that is known to specifically react with the modification is required. This enables nucleotide level resolution of the modification site and the level of modification can be inferred indirectly through comparison with controls. The third type of strategy is based on RNA immunoprecipitation (RIP) where an antibody is generated against the RNA modification of interest and used to pull-down RNAs with that modification (Dominissini 2012, Meyer 2012). Some limitations of this approach are that the modification rate is not directly measurable and the resolution is not at the nucleotide level. We discuss below current approaches (Table 1-1) to studying several types of RNA modifications (Figure 1-1) as well as the functional consequences of these RNA modifications.

A-to-I RNA editing

The deamination of adenosine to inosine by the ADAR family of enzymes is the most common editing event in mammalian mRNA and is conserved within metazoans (Keegan 2011). ADARs were originally discovered in Xenopus oocytes for their ability to unwind dsRNA (Bass 1987) and shortly after were shown to catalyze the deamination reaction of adenosine to inosine (Bass 1988). Double stranded RNA is a required substrate for ADARs (Lehmann 1999) and one hypothesis holds that the ancestral function of ADARs may have been to disable viral dsRNAs (Patterson 1995); however, a number of groups have observed a pro-viral effect of ADARs (Samuel 2011), which may indicate a commandeering of cellular machinery that was originally anti-viral. A-to-I RNA editing has been more extensively studied than other modifications because inosines base-pair with cytosine that are reverse-transcribed to guanine. Importantly, inosines are also read by the translation machinery as guanine, which can lead to changes in the resulting protein sequence. The reverse transcription into guanine makes inosine detectable by RNA-seq or by poisoned-primer extension assays. Given the problems with mining RNA editing from RNA-seq reads alone, sequencing the genome of a sample in addition to RNA-seq is one way to accurately identify RNA editing events (Ju 2011) because the genome would carry the same set of SNPs that are present in the transcriptome. However, this is an expensive strategy for studying RNA editing in large genomes. I-seq (Cattenoz 2013) is an inosine-specific cleavage assay that has been developed to identify A-to-I RNA editing. It is based on chemically protecting guanines in the RNA while selectively cleaving inosine sites. Another method is ICE-seq (Sakurai 2014), which involves specific chemical modifications of inosines with acrylonitrile to block reverse transcription. For some species in which ADAR

KO animals are viable, RNA-seq from wild-type and ADAR knockout animals have also been used to identify RNA editing events (St Laurent 2013). Members of the ADAR family include ADAR1, ADAR2, and ADAR3 (also refered to as ADAR, ADARB1, ADARB2 respectively). ADAR1 is expressed ubiquitiously while ADAR2 is only expressed within select cell types. ADAR3 has not been shown to edit RNA.

Within mammals, there are two types of A-to-I RNA editing: the site-specific type is cell type specific and tends to alter the coding sequence while the promiscuous type occurs ubiquitously and predominantly within inverted pairs of repeats such as ALUs in human (Bazak 2014). It is estimated that there are over one hundred million RNA editing sites in the genome based on ALU elements representing the vast majority of RNA editing events and ALUs represent 10% of the human genome, which is three billion bases long. The field of RNA editing identification has expanded to different species (Danecek 2012, Zhao 2015, Alon 2015), different cell types (Park 2012), and across time (Graveley 2011).

There are many biological functions associated with A-to-I RNA editing. The classical examples of A-to-I RNA editing are coding changes within the central nervous system to modulate ion channel activity (Seeburg 1998). In addition, A-to-I RNA editing is known to influence alternative splicing (Reuter 1999) and nuclear retention (Chen 2009). Some of the other functional aspects of A-to-I RNA editing that is being investigated are its role in viral infections (Gelinas 2011) and in innate immunity (Mannion 2014) as well as areas of post-transcriptional regulation of gene expression (Maas 2010).

A-to-I RNA editing has been tied to microRNAs and RNAi. Since dsRNA is formed during microRNA biogenesis and during RNAi, it has been proposed that A-to-I RNA editing might edit the intermediate dsRNA and that this might have downstream effects on transcript targeting. A-to-I RNA editing has been shown to have a negative effect on microRNA processing by inhibiting various steps of microRNA processing (Nishikura 2010). It has also been proposed that A-to-I RNA editing within the seed sequence of the mature microRNA can alter targets of the microRNA (Kawahara 2007); however, this is still controversial (de Hoon 2010). Due to the short sequences of microRNAs and the fact the mature microRNAs can be further processed with 5' and 3' modifications, it is difficult to computationally determine whether a sequence originated from A-to-I RNA editing of a transcript from one gene, or another modification from a different gene. In order to determine if A-to-I RNA editing exists within mature microRNA sequences and whether they do indeed target different transcripts it would be interesting to integrate ICE (Sakurai 2010) with small RNA-seq and methods for mapping mRNA and microRNA interactions such as CLASH (Helwak 2013).

C-to-U RNA editing

Another, less-prevalent form of RNA editing found in mammals is the deamination of cytosine to uracil. This reaction is catalyzed by Apobec1, which is the only known C-to-U RNA editing enzyme (Blanc 2003). Apobec1 was found based on its ability to edit the ApoB transcript, which is the best known example of C-to-U RNA editing. Editing of ApoB results in a stop codon that generates a truncated protein that is 48% (ApoB-48) of the length of the full-length transcript (ApoB-100). The ApoB-48 protein is associated with lower LDL levels (Yao 1994).

Apobec-1 mediated C-to-U RNA editing has evolved in the mammalian lineage (Conticello 2007). Knockout of the enzyme is non-lethal in mice. By comparing RNA-seq from wild-type and Apobec1 KO C-to-U RNA, dozens of editing sites were discovered in mouse livers and small intestines (Rosenberg 2011, Blanc 2014). C-to-U RNA editing is unique in that the modified base is converted to an unmodified type base. Thus, the sequence variant approach is the only way to identify this type of RNA modification; the chemical treatment or IP approaches will not be able to distinguish the modified uracils with the naturally occurring uracils. Typical experimental strategies for identifying C-to-U RNA editing events are to knockout the enzyme and/or co-factors and sequence the wild-type and knockout animal transcriptomes.

Although the introduction of a stop codon is the most famous example of C-to-U RNA editing, most other sites are in the 3'UTR and affect translation (Blanc 2014). Since C-to-U RNA editing is very tissue-specific and rare, it has been difficult to study some of the biological consequences. However, overexpression of Apobec1 has been shown to lead to tumorigenesis (Yamanaka 1995) and certain human cancers have been observed to have an abundance of C-to-T substitutions (Alexandrov 2013), which could suggest a carcinogenic role of Apobec1 or possibly a related cytidine deaminase.

5-methylcytosine

5-methylcytosine (m5C) is well known as an epigenetic mark in DNA that is involved with genomic imprinting and silencing of gene expression. m5C methylation has been studied in DNA with bisulfite sequencing (Frommer 1994); briefly, cytosines can be converted to uracil through bisulfite mediated sulphonation, hydrolytic deamination, and desulphonation. Methylated cytosines are not changed by this reaction and are therefore still sequenced as cytosines, whereas unmethylated cytosines are converted to uracil and are sequenced as thymine. A comparison between bisulfite treated and untreated samples can identify sites of m5C methylation and quantitate the level of methylation. Bisulfite treatment has been combined with RNA-seq to identify m5C methylation in transcriptomes (Squires 2012). Using this technique, more than ten thousand m5C sites were identified and it was determined that m5C is enriched in 3'UTRs near Argonaute binding sites. One caveat to this approach is that it might not be specific to m5C; as with bisulfite sequencing of DNA, other Cytosine modifications besides m5C are resistant to bisulfite treatment.

The m5C modified base can be targeted with an antibody and immunoprecipitated to detect the modification (m5C-RIP) (Edelheit 2013); this approach is specific to m5C, but suffers from the same limitations of m6A-seq and meRIP-seq in low resolution and limited quantification ability. Additionally, two other methods have been developed to probe m5C methylation in RNA: Aza-IP (Khoddami 2013) and miCLIP (Hussain 2013). Both of these assays take advantage of the covalent bond formed between the RNA and enzyme as an intermediate step of m5C methylation. Aza-IP uses 5-azacytidine to prevent the release of the RNA from the methylase and the RNA-protein complex is immunoprecipitated with an antibody targeting the enzyme. Similarly, miCLIP immunoprecipitates an RNA-protein complex that has been stalled at the covalently bound intermediate, but miCLIP relies on a cysteine to alanine mutation to prevent dissociation of the RNA-protein complex.

The exact biological functions of m5C methylation in mRNA still remains a mystery; however, it would not be surprising if m5C methylation is as dynamic as other RNA modification or has its own set of readers, writers, and erasers in analogy to m6A methylation and DNA m5C methylation.

N6-methyladenosine

N6-methyladenosine (m6A) is the most abundant internal (non-cap) RNA modification within mRNA (Wei 1975) and recently has been shown to be present in DNA (Greer 2015, Zhang 2015). Two similar assays were developed to detect m6A methylation: m6a-seq (Dominissini 2012) and meRIP-seq (Meyer 2012). Both assays are RIP-based where an antibody against the m6A is used to pull-down RNA containing the modified base. These initial studies found that there were tens of thousands of m6A sites within thousands of genes and that these sites were enriched within long internal exons, near stop codons, around transcription start sites, and within 3’UTRs.

Unfortunately, m6A-seq and meRIP-seq are unable to detect the level of m6A methylation (i.e. what fraction of adenosines are methylated) and unable to detect the actual base of the modification (low resolution). SCARLET (Liu 2013), on the other hand, is a method to measure the level of methylation that is based on site-specific cleavage, radiolabeling, site-specific ligation, and nuclease digestion. PA-m6A-seq (Chen 2015) is a modified version of m6A-seq that provides higher resolution using the incorporation of 4-thiouridine into cells to facilitate cross-linking. The procedure involves poly-A selection, m6A immunoprecipitation, cross-linking, and RNA digestion to generate sequencing libraries. The higher resolution is achieved by cross-linking the 4-thiouridine to the antibody. Due to the requirement of 4-thiouridine, one limitation with this assay is that it can only be used in systems where 4-thiouridine can be incorporated (i.e. cell lines). Although m6A-seq and meRIP-seq does not allow for direct methylation quantification or base-level resolution, the read density can provide a rough estimate of methylation levels and the consensus sequence can aid in base-level resolution. The consensus sequence for m6A methylation is RRACH (R=A/G, H=A/C/U) (Harper 1990) and m6A binding proteins are able to recognize the modified base through the YTH domain (Xu 2014).

There have been many biological functions that are associated with m6A methylation. m6A methylation reduces mRNA half-lives and translatability through sequestration away from the translation machinery (Wang 2014). Key transcripts involved in pluripotency are methylated and loss of m6A methylation in ESC promotes self-renewal and inhibits differentiation (Batista 2014, Geula 2015). m6A is key player in establishing circadian rhythm by regulating RNA stability to help maintain periodicity (Fustin 2013). m6A methylation is known to be reversible and this reversibility is important in mRNA export and fertility (Zheng 2013). RNA methylation can suppress RNA-mediated immune stimulation (Kariko 2005) and act as a switch to regulate RNA-protein interactions and modulate alternative splicing (Liu 2015). Lastly, microRNAs influence m6A methylation (Chen 2015) and conversely m6A methylation is enriched at microRNA target sequences in the 3'UTR (Meyer 2012).

While most RNA modifications only have known enzymes that add the modification to RNA (writers), m6A modifications are known to be dynamic with known sets of writers that introduce the modification to RNA, readers that detect the modification, and erasers that remove the modification. The known components of the m6A methyltransferase (writer) consist of two catalytic components (METTL3 and METTL14) and a splicing factor (WTAP) (Liu 2014). The first identified reader of m6A is YTHDF2 (Wang 2014) which sequesters m6A mRNA to inhibit translation and promote degradation. Additionally, any member of the YTH domain family could be a potential reader of m6A, since the YTH domain recognizes the m6A base. Alternatively, there are readers that can read the modification indirectly by detecting changes in secondary structure that is introduced by the m6A (Liu 2015). FTO and ALKBH5 are the known erasers of m6A (Jia 2011, Zheng 2013). Thus, m6A modifications are analogous to an RNA version of protein phosphorylation in that there are players that add, detect, and remove the modification. This versatility allows m6A methylation to be spatiotemporally dynamic and allows for rapid response to signal.

Pseudouridine

Pseudouridine, an isomer of uridine, is the most abundant RNA modification in noncoding RNA (Davis 1957). The post-transcriptional process of pseudouridylation is catalyzed by pseudouridine synthases, of which there are 23 putative genes in human (Hunter 2012). Defects in pseudouridine synthases have been implicated in disease, such as dyskeratosis congenital (DKC) and mitochondrial myopathy and sideroblastic anemia (MLASA) (Heiss 1998, Bykhovskaya 2004). Pseudouridine is found in tRNAs and rRNAs (Charette 2000); however, little is known about the extent and function of pseudouridine in mRNA. Recently two groups have independently developed high-throughput techniques (Pseudo-seq and Psi-seq) to identify the locations of pseudouridine transcriptome-wide (Carlile 2014, Schwartz 2014). Pseudo-seq and Psi-seq are based on selective pseudouridine modification with N-cyclohexyl-N′-(2-morpholinoethyl)carbodiimide metho-p-toluenesulphonate (CMC) (Bakin 1993) which is able to sterically inhibit reverse transcription. Hundreds of sites were identified and found to be regulated by environmental changes. In coding RNAs, pseudouridines are enriched in the coding sequence. The functional knowledge of pseudouridine in mRNA is limited; however, it is tempting to hypothesize that pseudouridines affect RNA secondary structure similarly to their role in tRNAs and rRNAs (Arnez 1994). Alternatively, pseudouridines in stop codons could allow for ribosomal read-through (Karijolich 2011) and altered tRNA base-pairing (Fernández 2013). It would be interesting to investigate whether pseudouridylation of the identified sites correlate with altered amino acid sequences.

The Road Ahead

The study of RNA modifications and epitranscriptomics is a rapidly growing discipline thanks to the growth of high-throughput sequencing based methods and advances in advances in computational analysis, which have greatly expanded the breadth of knowledge of well-studied modifications as well as reigniting interest in forgotten modifications in mRNA. They have greatly increased the abilities of biologist by adding the equivalent of power-tools to the biologist's toolbox, high-throughput versions of old molecular biology tools. Although these new methods allow users to obtain large amounts of data, one needs to be cautious because they also allow users to obtain large amounts of data that could be misinterpreted if not analyzed carefully using appropriate controls, which was the case for (Li 2011).

As sequencing technologies improve, we will be able to sequence all the full-length short and long RNA transcripts in a given sample in their entirety at a high level of confidence and know the locations of all the RNA modifications. However, it could take many years before this level of sequencing becomes available. In the mean time, we need to develop clever molecular biology strategies and computational algorithms to study RNA modifications using current sequencing technologies that either have high-quality short reads or error-prone long-reads. It should be no surprise that the development of computational methods to analyze epitranscriptomic data can be just as difficult as the experimental method, if not more so.

The crucial next steps in the study of RNA modifications are to associate functional and mechanistic effects of the individual sites of modifications. In particular, it will be interesting to investigate structural changes that are induced by RNA modifications (Spitale 2015) and to integratively analyze the relationship between the different modifications during the lifetime of the modified RNA.

Figure 1-1

RNA modifications.

The modifications and nitrogenous base for A-to-I RNA editing, C-to-U RNA editing, N6-methyladenosine methylation, 5-methylcytosine methylation, and pseudouridylation are shown.

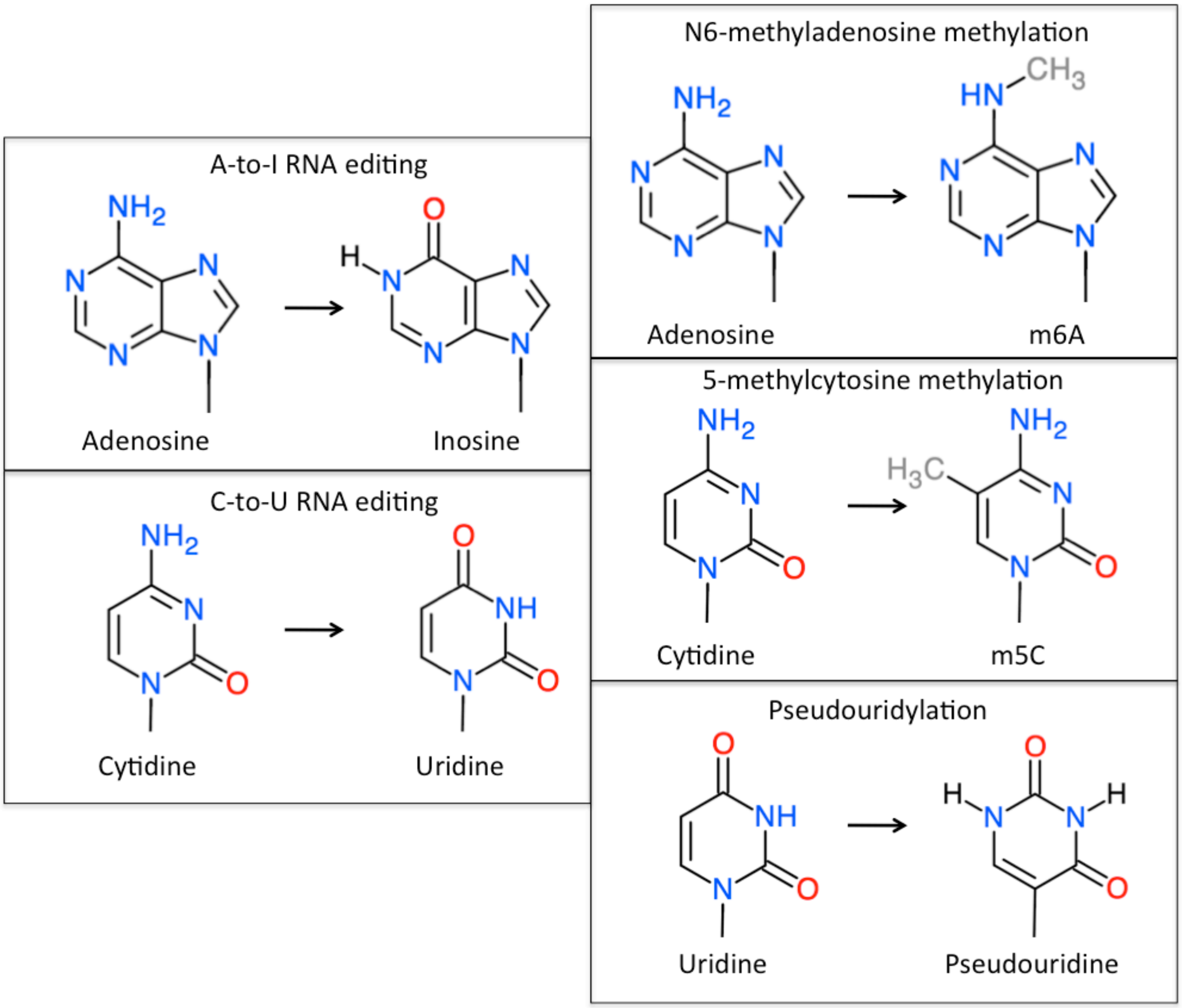
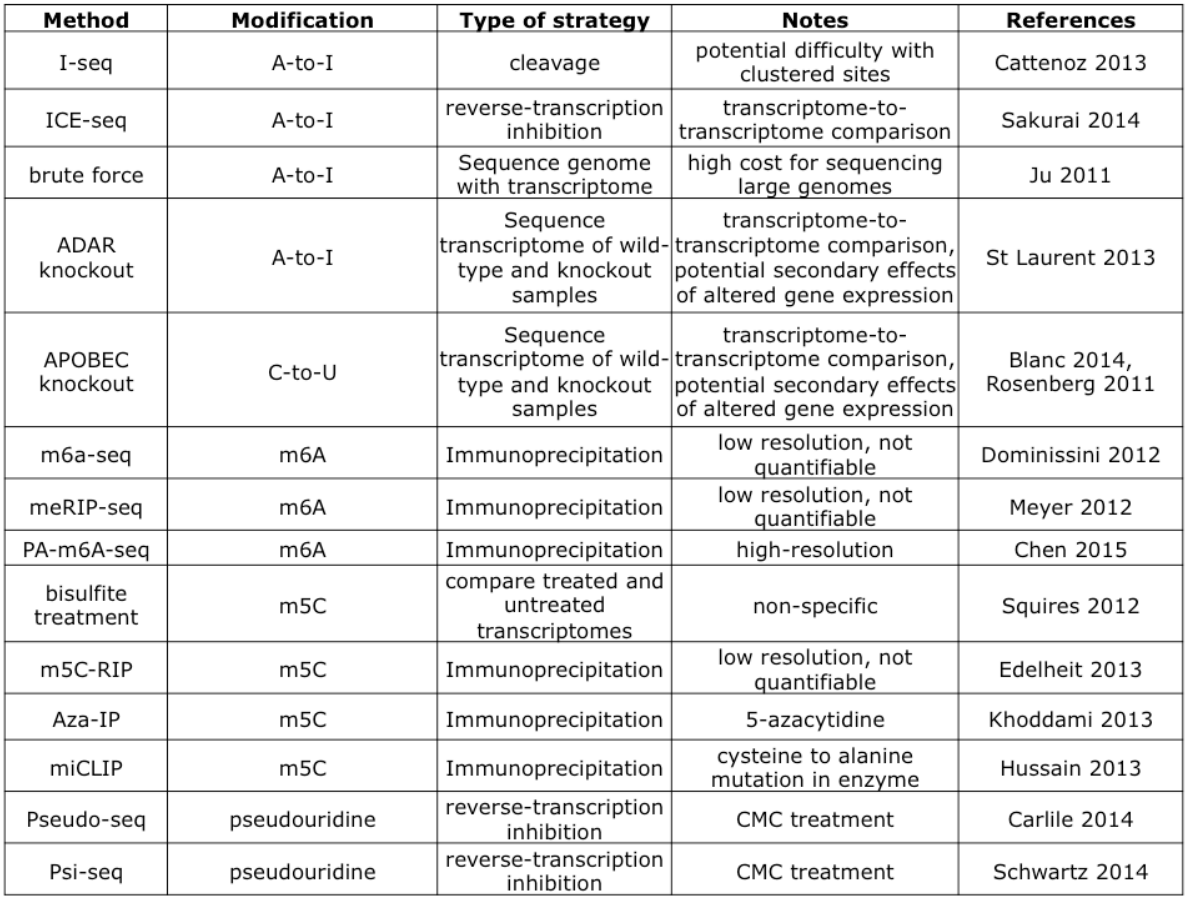


Table 1-1

Published high-throughput methods for RNA modification determination



Method	Modification	Type of strategy	Notes	References
I-seq	A-to-I	cleavage	potential difficulty with clustered sites	Cattenoz 2013
ICE-seq	A-to-I	reverse-transcription inhibition	transcriptome-to-transcriptome comparison	Sakurai 2014
brute force	A-to-I	Sequence genome with transcriptome	high cost for sequencing large genomes	Ju 2011
ADAR knockout	A-to-I	Sequence transcriptome of wild-type and knockout samples	transcriptome-to-transcriptome comparison, potential secondary effects of altered gene expression	St Laurent 2013
APOBEC knockout	C-to-U	Sequence transcriptome of wild-type and knockout samples	transcriptome-to-transcriptome comparison, potential secondary effects of altered gene expression	Blanc 2014, Rosenberg 2011
m6a-seq	m6A	Immunoprecipitation	low resolution, not quantifiable	Dominissini 2012
meRIP-seq	m6A	Immunoprecipitation	low resolution, not quantifiable	Meyer 2012
PA-m6A-seq	m6A	Immunoprecipitation	high-resolution	Chen 2015
bisulfite treatment	m5C	compare treated and untreated transcriptomes	non-specific	Squires 2012
m5C-RIP	m5C	Immunoprecipitation	low resolution, not quantifiable	Edelheit 2013
Aza-IP	m5C	Immunoprecipitation	5-azacytidine	Khoddami 2013
miCLIP	m5C	Immunoprecipitation	cysteine to alanine mutation in enzyme	Hussain 2013
Pseudo-seq	pseudouridine	reverse-transcription inhibition	CMC treatment	Carlile 2014
Psi-seq	pseudouridine	reverse-transcription inhibition	CMC treatment	Schwartz 2014

Chapter 2

RNA editing in the human ENCODE RNA-seq data

Abstract

RNA-seq data can be mined for sequence differences relative to the reference genome to identify both genomic SNPs and RNA editing events. We analyzed the long, polyA-selected, unstranded, deeply sequenced RNA-seq data from the ENCODE Project across 14 human cell lines for candidate RNA editing events. On average, 43% of the RNA sequencing variants that are not in dbSNP and are within gene boundaries are A-to-G(I) RNA editing candidates. The vast majority of A-to-G(I) edits are located in introns and 3′ UTRs, with only 123 located in protein-coding sequence. In contrast, the majority of non–A-to-G variants (60%–80%) map near exon boundaries and have the characteristics of splice-mapping artifacts. After filtering out all candidates with evidence of private genomic variation using genome resequencing or ChIP-seq data, we find that up to 85% of the high-confidence RNA variants are A-to-G(I) editing candidates. Genes with A-to-G(I) edits are enriched in Gene Ontology terms involving cell division, viral defense, and translation. The distribution and character of the remaining non–A-to-G variants closely resemble known SNPs. We find no reproducible A-to-G(I) edits that result in nonsynonymous substitutions in all three lymphoblastoid cell lines in our study, unlike RNA editing in the brain. Given that only a fraction of sites are reproducibly edited in multiple cell lines and that we find a stronger association of editing and specific genes suggests that the editing of the transcript is more important than the editing of any individual site.

Introduction

RNA editing is a post-transcriptional process that modifies the primary RNA and microRNA transcripts. This process can result in nonsynonymous protein coding substitutions, alternative splicing, nuclear retention of mRNA, or alterations of microRNA seed regions (for a review, see Nishikura 2010). The most common form of RNA editing in mammals is A-to-I editing, in which adenosine is deaminated to produce inosine by members of the ADAR (adenosine deaminases acting on RNA) family of enzymes (Wagner et al. 1989; Kim et al. 1994; Kumar and Carmichael et al. 1997). C-to-U editing in mammals by the APOBEC family of enzymes is thought to be much less frequent and much more specific (Gerber and Keller 2001). In mammals, ADAR is found within several tissues, while ADARB1 is known to be active in the brain. Abnormal RNA editing has been reported in epilepsy, amyotrophic lateral sclerosis (ALS), brain ischemia, depression, and brain tumors (Maas et al. 2006; Nishikura 2006; Peng et al. 2006; Paz et al. 2007; Cenci et al. 2008).

ADARs recognize double-stranded RNA as their major substrate, but editing at some sites is very selective for specific A residues, while other sites are edited promiscuously and mainly in clusters (Nishikura et al. 1991; Polson and Bass 1994). Because inosine pairs preferentially with cytidine, I is read as G during protein synthesis or during reverse transcription for RNA-seq. Known functional consequences of this post-transcriptional modification include changes in amino acids in the protein product such as in the glutamate and serotonin receptors, creation or deletion of entire exons by changes in splicing, retention of mRNA in the nucleus, changes in RNA stability, heterochromatin formation, protection against viral RNA, and microRNA modification (Zheng et al. 1992; Burns et al. 1997; Seeburg et al. 1998; Athanasiadis et al. 2004; Luciano et al. 2004; Prasanth et al.

2005; Wang et al. 2005; Agranat et al. 2008). Although the best-studied RNA editing sites are in coding sequences that qualitatively change the protein product, the majority of known RNA edits in human occur within Alu sequences embedded within introns and UTRs (Kim et al. 2004). ADAR mouse knockouts are embryonic lethal at day E11.5 (Wang et al. 2004), where it plays an important role in suppressing interferon signaling to block premature apoptosis in hematopoiesis (Iizasa and Nishikura 2009). While microRNAs are known to be important RNA editing targets, this present study focuses on RNA editing in messenger RNA as measured from polyA-selected RNA.

We surveyed our human polyA+ ENCODE RNA-seq data from 14 cell types for RNA editing events using a rigorous computational pipeline designed to filter out sequencing and read mapping artifacts. We further filtered private genomic single nucleotide variants (SNVs) for 12 of the 14 cell types using either 1000 Genomes resequencing data (The 1000 Genomes Project Consortium 2010) or ENCODE ChIP-seq data sets. We identified between 500 and 3000 reproducible A-to-I RNA editing events per cell type in biological duplicate RNA-seq samples. We then focused on genes that are frequently edited across multiple cell types for further analysis, and found enrichment for genes involved in basic housekeeping processes such as cell division, viral defense, and translation.

Results

Development and refinement of an RNA-editing pipeline tuned based on data from the GM12878 lymphoblastoid cell line

RNA-seq data has been mined for known SNPs in expressed genes to study allele-specific expression (Montgomery et al. 2010). While sequence variants in RNA-seq that are not in the genome are RNA editing candidates, we expect some level of mapping artifact and sequencing error in the data, and these could be mistaken for RNA editing. We reasoned that a pipeline that maximizes the fraction of called known SNPs would then produce the most conservative set of RNA editing candidates. One ENCODE cell line, GM12878, was particularly well-suited for tuning our pipeline, as it was deeply sequenced as part of the 1000 Genomes Project with the results incorporated into dbSNP (Sherry et al. 2001). GM12878 2 × 75 bp RNA-seq reads were mapped onto an expanded genome that includes known splice junctions (Mortazavi et al. 2008) using Bowtie as described in the Methods (Fig. 2-1A, Fig. 2-1B). Reads that mapped onto splice junctions were set aside because they are more prone to mismapping artifacts, and the remainder were used for SNV calling. An additional source of false-positive SNVs are reads with errors that are amplified during the library construction process. We can avoid PCR artifacts by collapsing reads, i.e., counting reads with identical starts only once. We further restricted ourselves to sites that had a minimum 10% frequency in both the entire data set ("uncollapsed set") as well as in the smaller data set of nonredundant reads ("collapsed set"). We kept only candidate RNA editing sites called from two independent RNA-seq replicates (Fig. 2-1C). Of the SNVs present in the union of collapsed and uncollapsed sets, 47% were in the intersection, while 37% were only in the uncollapsed set and 16% were only in the collapsed set (Fig. 2-1D). We found that using the intersection strategy delivered a higher fraction of A-to-G calls (20%), while the sets that were only in the uncollapsed or only in the collapsed sets had A-to-G calls of 14% and 18%, respectively (Fig. 2-1E). We then scored our SNVs for occurrence within dbSNP and found that the intersection had the highest percentage of known SNPs (71%) (Fig. 2-1F).

We further filtered the set of SNVs absent in dbSNP (candidates for editing because they are not known polymorphisms) for their overlap with known transcript boundaries from GENCODE v7 protein coding genes, which retained 86% of the candidates. SNV calls within known transcripts on the minus strand were reassigned to reflect the appropriate substitution in the sense of transcription. We then used ANNOVAR (Wang et al. 2010) to annotate the part of the transcript in which the variants occurred. We found that A-to-G variants were present primarily in introns and UTRs, whereas 82% of the non–A-to-G calls were annotated as "splicing," which is defined here as intronic and within 5 bp of the splice junction (Fig. 2-2A). Inspection of these calls revealed that they were primarily due to mismapped reads that should have mapped across splice junctions (Fig. 2-2B). After removing all calls within 5 bp of splice sites, we found that >80% of our novel SNVs were A-to-G calls, which is 20-fold higher than the 4% of SNV calls passing all the same filters that are G-to-A calls (Fig. 2-2C). This enrichment is 20% higher in the intersection of the collapsed and uncollapsed sets than it is in either of the outersects (Fig. 2-S1). However, there remains a chance that there are private genomic SNVs that are not yet represented in dbSNP because of low-coverage or read mapping issues. We therefore used the genomic reads for GM12878 (and GM12891 and GM12892 below) from the 1000 Genomes Project to remove transcriptome SNVs with one or more genomic reads also supporting that candidate variant. This further increased our A-to-G SNV to >85% of the remaining calls. If we assume that our G-to-A calls are false positives with respect to editing (whether they are true SNPs not found in dbSNP or are sequencing/mapping artifacts), then our false-discovery rate (FDR) for A-to-G RNA editing would be <2%. No other SNV class accounted for >6% of our most stringently filtered set. T-to-C calls were the second most frequent class of SNV. Thity-three percent of these T-to-C were found in regions with overlapping GENCODE gene models on opposite strands, and another 33% had an unannotated transcript in the opposite sense as the overlapping gene model.

We then tested whether using ChIP-seq reads could be used to filter candidate SNVs in cases for which genomic reads are not available, as is the case for many ENCODE cell lines analyzed in the next section. Surprisingly, we found that we filtered more SNVs using ChIP-seq than using the 1000 Genomes data. In order to better understand filtering the SNV calls using ChIP-seq data versus 1000 Genomes data, we compared the coverage of SNVs in the two data-types. Although there were fewer total reads in the ENCODE ChIP-seq data than in the 1000 Genomes data, we found overall higher coverage in the ChIP data (mean coverage, 57.4) than in the 1000 Genomes data (mean coverage, 33.8). Our ChIP-seq mean coverage was even higher (60.6) over the 1457 SNVs that were filtered when using only the ChIP data, thus allowing us to detect rare variants (Fig 2-S2). However, there were other regions for which the 1000 Genomes data filter out SNVs that were missed using the ChIP data.

SNVs that matched dbSNP have an expected bimodal distribution with one mode at a frequency of 1.0 and the second mode at a frequency of 0.5, which are due to homozygous and heterozygous SNPs, respectively (Fig. 2-2D). In contrast, we find that the distribution of A-to-G RNA editing calls are skewed rightward with a mode at a frequency of 0.2. Nonsplicing, non–A-to-G SNVs show a distribution similar to dbSNP SNVs, with a mode at a frequency 1.0 (Fig. 2-2E). Relatively few candidate RNA editing events were in open reading frames (Fig. 2-S3). As expected, our power to call SNVs within exons exceeded that of calling them in introns for a given expression level (Fig. 2-2F), given the much greater depth of RNA-seq coverage in exons versus introns. We next asked whether genes with A-to-(G)I candidate edits had related functions, and found that they are enriched for Gene Ontology terms that are mainly broad anabolic functions such as translation, translational elongation, ribonucleoprotein complex, chromosome, centromeric region, ribosome, cytosolic ribosome, mitochondrial nucleoid, melanosome, and coated pit.

Survey of 14 ENCODE cell lines

We then applied the above pipeline to polyA RNA-seq from 14 ENCODE cell types with 2 × 75 bp non-strand-specific protocol, all of which express only ADAR (Fig. 2-S4). Since sequenced genomes were not available for any cell line other than the three lymphoblastoid cell lines, we substituted ENCODE ChIP-seq data from HudsonAlpha and histone modifications data from the Broad Institute to filter out private genomic SNVs that were not represented in dbSNP. The fraction of editing candidates that were of the A-to-G class ranged from 50%–85% (Fig. 2-3A; Fig. 2-S5). HepG2 and HUVEC cells had the lowest number of candidate RNA editing sites with about 500 calls each. The filtering of private genomic SNVs using ChIP-seq data increased the percentage of A-to-G calls by 5%–20%, with the most SNVs filtered out for growth-transformed tumor cell types (Fig. 2-3B). The number of candidate sites called by our pipeline did not depend on the sequencing depth of the RNA-seq data set (Fig. 2-S6) over the range represented in our samples (28–135 million reads per replicate), but the amount of filtering did depend on the aggregate depth of coverage in the ChIP-seq data sets. The individual non–A-to-G classes ranged between 0%–10% of the total, with T-to-C again being the second most prevalent modification. Based on the results in GM12878, it is likely that the bulk of the T-to-C edits are A-to-G(I) edits from transcripts on the opposite strand relative to their original annotation. Although members of the APOBEC family are detectably expressed in all cell lines (Fig. 2-S7), we observed a very modest proportion of C-to-T(U) editing events in HepG2, HUVEC, and HCT116. While some of these C-to-T(U) sites might prove to be true RNA edits, only a handful of these C-to-T(U) sites are located in AU-rich regions known to be associated with APOBEC editing. We therefore provisionally conclude that most C-to-T(U) candidates are false positives.

The number of candidate SNVs within coding domains summed over all cell types are below 1000, and these are primarily nonsynonymous for the non–A-to-G SNVs (Fig. 2-3C). We find that 94% (5349 of 5695) of the candidate A-to-G(I) calls are within known repeat families, and 98% (5247 of 5349) of these are in Alu's. Alu families with the most members also had the most edits This is certainly an underestimate, as our conservative mapping strategy would underreport hyper-edited regions with more than three simultaneous edits within our 75-bp reads.

We next focused on individual sites with evidence of editing in multiple cell types. Overall 33.5% (1905 out of 5695 possible) of individual A-to-G(I) candidate editing sites occur independently in two or more different cell types (Fig. 2-4A; for non–A-to-G classes, see Fig. 2-S8). We also found that 24% (1386/5695) of our candidate A-to-G(I) calls intersect with the DARNED database of RNA editing in human (Kiran and Baranov 2010), which was generated from mining human ESTs. This low overlap is expected, since the ENCODE RNA-seq data analyzed here did not include neuronal tissues or cell lines. We found 28 genes that were edited in all of our cell types, and 20 genes of these (71%) are also in the DARNED set. We found that 47.4% (662 out of 1396 possible) of genes edited are called as edited in at least two or more different cell types (Fig. 2-4B). We therefore conclude that gene-level association with editing is more robust than the identity of an individual site edited.

We then focused on the distribution of calls within gene models. For most genes, A-to-G(I) RNA editing candidates were either all in intronic regions or all in UTRs, although 1%–4% of edited genes had both intronic and UTR sites, depending on the cell type (Fig. 2-4C). Because editing events are—overall—rare in the transcriptome and because editing often covers only a fraction of transcripts from a given gene, we wanted to probe our sensitivity for calling events in replicate samples. While most ENCODE data is in duplicates, we had an instance of data from four replicate determinations for Human H1 ES. We compared calls from Human H1 ES cell reps 1 and 2 with calls from reps 3 and 4 (Fig. 2-4D) and found that the number of edits per gene that we had called was quite noisy when there were only a few candidate sites that were called per gene. This is possibly due to the stochastic and promiscuous nature of ADAR's ability to hyper-edit dsRNA (Polson and Bass 1994) or due to the sensitivity of pipeline to coverage at the low end of the RNA expression spectrum.

We compared the filtered A-to-G(I) SNV calls in GM12878 to those in its parents, GM12891 and GM12892, to help assess the stability of RNA editing within a single cell-type, i.e., EBV-transformed lymphoblastoid cells. GM12891 and GM12892 had 1885 (86%) and

843 (87.7%) candidate A-to-G(I) sites located in 479 and 265 genes, respectively. GM12878 shared 490 sites (337 genes) with GM12891 and 292 sites (218 genes) with GM12892. While 26.2% of the individual editing SNVs were found in at least two individuals of the trio, >49.6% of the genes were in common (Fig. 2-4E). Thus the gene-level association with editing is also more reproducible than the identity of individual sites edited within different cell lines of the same type.

Overall, there were 248 out of 1396 genes that were edited in at least five out of the 10 distinct cell types after we applied the ChIP-seq filter (Fig. 2-4A). Our GO analysis of these genes showed enrichments that included interspecies interaction between organisms, cell division, DNA metabolic process, positive regulation of defense response to virus by host, protein folding, ER-Golgi intermediate compartment, ribosome, and ribonucleoprotein complex. Two genes that are especially highly edited within multiple cell types are the inhibitor of apoptosis XIAP and the caspase-3 target DFFA, suggesting an explicit and direct link to the apoptosis phenotype of the mouse ADAR knockout.

The overwhelming prevalence of A-to-G(I) RNA editing candidates in our lymphoblastoid trio agrees with a recent publication (Peng et al. 2012) but differs qualitatively from a previously reported analysis of RNA editing in a nonoverlapping set of HapMap lymphoblastoid lines (Li et al. 2011). To make a more informed comparison, we applied our pipeline to the data sets from Li et al. (2011) though we note that their data were single replicate RNA-seq measurements with shorter 50-bp reads that were more shallowly sequenced (40 million reads vs. our average of 100 million reads). We found that some of their individual samples produced similar genic, nonsplicing A-to-G enrichments as in our lymphoblastoid lines, while others had an especially high percentage of A-to-C and

T-to-G classes (Fig. 2-S9, Fig. 2-S10). To check whether these could be DNA sequencing artifacts from earlier and different (2008–2009) Illumina chemistry, we checked the output of our pipeline on an independent data set (ENCODE GM12878 DNase-seq) from that same earlier timeframe. We found relatively low numbers of non-dbSNP SNVs in the DNase-seq data and found that A-to-C and T-to-G classes were again comparatively enriched (Fig. 2-S11), suggesting that a similar systematic bias exists in some of the Li et al. (2011) samples as in genomic DNA sequenced with that technology in that timeframe; these are not therefore attributable to RNA editing. Upon further inspection, we found that the particular sites most affected are embedded in G-rich regions (Fig. 2-S12). In addition, recent reports (Schrider et al. 2011; Pickrell et al. 2012) attribute the majority of the non–A-to-G calls in Li et al. (2011) to sequences that are paralogs of the gene that were reported as edited, i.e., a mismapping error. Therefore we only report our ChIP-filtered A-to-G(I) RNA editing candidates for each cell line.

Discussion

We developed an intentionally conservative strategy to identify candidate RNA edits from RNA-seq data for the human ENCODE cell lines. We analyzed the transcriptomes of 14 cell types and compared the RNA sequences with the reference genome, filtered out known SNPs from dbSNP, and filtered out private SNPs detected in ChIP-seq data from each cell line. We found that SNVs in the intersection of our "collapsed" and "uncollapsed" mapping sets yielded the highest fraction of known SNPs; this observation functions as a positive control that our SNV calls are not significantly biased by mapping artifacts. We also found that incorrect read mapping across splice junctions was the source of the majority of non–A-to-G calls. We further developed a strategy to filter out nonediting transcriptome SNVs by using up to 2.0 billion ChIP-seq reads in GM12878 and 0.1–1.7 billion ChIP-seq reads in the other nine cell lines that do not have resequencing data available. Together with filtered calls in two additional lymphoblastoid with 1000 Genomes data, we therefore have reliable A-to-G(I) editing candidates in 12 of our 14 cell lines.

Up to 87% of SNVs that are not SNPs (either in dbSNP or private genomic SNVs) are A-to-G calls; this suggests they are likely to be A-to-I editing candidates. Furthermore, >97% of these candidates are located in introns and 3′ UTRs, which is consistent with what was previously known about RNA editing based on earlier EST surveys and recent reports (Bahn et al. 2012; Peng et al. 2012). Our candidate A-to-G(I) RNA editing sites have a different variant frequency from known SNPs. They tend to cluster predominantly in the 3′ UTR or in introns. In the three cell lines with the least amount of A-to-G(I) editing, there were relatively more C-to-T(U) SNVs, but these were not associated with AU-rich regions, as would have been expected if these are due to APOBEC activity. We also found that individual RNA editing calls are noisy for lowly expressed genes because of depth of coverage requirements for editing calls. If sensitivity to a conservative threshold was the major source of noise from one data set to another, and one replicate to another, then the identity of genes that are edited would be more consistent than the actual edits as long as edits occur in clusters.

Overall, we report 5695 unique candidate A-to-G(I) RNA editing events in 1396 genes, including a subset of 248 genes that were consistently edited across more than five cell types. Ninety-nine percent of the candidate RNA editing calls occurred within known repeats (with 98% of those in Alu elements), and only 24% were annotated in DARNED, which is expected since none of the ENCODE cell lines in this study cover neuronal phenotypes. Comparing our results to previously reported widespread evidence of noncanonical RNA editing besides A-to-G(I) in a different set of HapMap lymphoblastoid cell lines (Li et al. 2011), we found that sequencing and mapping artifacts could account for the vast majority of the unconventional (non-ADAR) variant calls. Moreover, we showed that this is caused, in part, by RNA data generated using older sequencing chemistry.

Genes with at least one edit within all cell types show GO term enrichments in "housekeeping" annotations related to cell division, DNA metabolic process, protein folding, and ribosome, as well as terms relating to viruses and defenses against them. The latter functions are consistent with the view that editing arose first in this context. For example, the protein EIF2AK2 inhibits protein synthesis upon activation by viral RNA and has an average of 13 edits across all cell-types. This meshes well with reports that RNA editing participates in host-viral interactions, such as the editing of the hepatitis C virus (HCV) genome by ADAR (Taylor et al. 2005). Interestingly, some viruses have been able to take advantage of ADAR for their own purposes; endogenous ADAR has been shown to stimulate HIV-1 replication (Doria et al. 2009).

When searching for RNA editing events to create a global map of high-quality candidates, there is a difficult tradeoff between sensitivity (identifying a highly inclusive set of possible edits) and specificity (being more confident that a call is in fact a true RNA edit). We judged it better to have a smaller number of candidate RNA editing events that are highly likely to be true than to have a larger number with an increased percentage of false positives. We undoubtedly lost a substantial number of true, low-level A-to-I RNA editing events in the process, and users of these RNA-Seq data might for some purposes want to build a more inclusive and less certain list. Another caveat that applies to our pipeline is that our method allows for a maximum of three edits per 75 bp within our reads. Thus, if there was a 75-bp window that had significantly more than three edits, our pipeline would only detect the edits on the periphery of this hyper-edited domain. We could allow more mismatches during the alignment step, but then this would exacerbate the problem of incorrectly mapped reads. To date, the RNA-seq field has emphasized aggressive read mapping to maximize sensitivity and new candidate discovery. This inevitably comes at the expense of specificity, and this effect is greater in complex genomes with extensive paralogous gene and repeat families such as those of mammals. Our current view is that there is no single correct threshold for the sensitivity specificity tradeoff: It has to be selected to match the objective of a given study and the future use of each analysis. When an analysis focuses on the small portion of sequence reads that imply differences from the majority that match well to the appropriate genome/transcriptome models, a general trend toward greater conservatism seems justified. We show here that an early sequencing chemistry issue and a problem with accurate read mapping over splice junctions are two specific contributors to false-positive RNA editing candidates. We have greatly reduced these in our pipeline and present a conservative set for the ENCODE cell lines. In doing so, we demonstrated that there is no persuasive evidence in favor of noncanonical editing. A-to-G(I) edits strongly dominated our data except for three cell lines with modest evidence of C-to-T(U) SNVs, and no unknown edit was above the noise level.

It is a sobering caution that, even when reads were mapped simultaneously against the genome and known splice junctions, we still had obvious splice-mapping artifacts. We were therefore forced to exclude all calls that mapped within a few bases of splice junctions; this, in turn, dramatically reduced non–A-to-G(I) calls. There remain a few noncanonical calls that pass our criteria, but we expect these to be primarily undetected private SNVs, regions with complicated mapping issues due to paralogous genes and pseudogenes in the genome, or uncharacterized splice junctions. This does not preclude the possibility of some very specific APOBEC-like, noncanonical editing of bases but calls into question their previously reported widespread occurrence (Li et al. 2011), which has been recently called into question (Kleinman and Majewski 2012; Lin et al. 2012; Pickerell et al. 2012).

While GM12878 had been "deeply" resequenced by 2010 standards, we find that filtering with ChIP-seq was actually more effective. While this may seem paradoxical at first, ChIP-seq signal as well as background reads are most likely to come from open chromatin within transcribed genes. Although the 1000 Genomes project has greater coverage throughout the genome, we had in fact greater overall coverage of the candidate RNA SNVs in the pooled ChIP-seq data. While the amount of ChIP-seq data in GM12878 and other ENCODE Tier 1 and Tier 2 cell lines is exceptional, they highlight the issue that really high coverage is necessary to detect a subset of SNVs. The fact that private genomic SNVs need to be accounted for and filtered out for assaying RNA editing suggests that we can use RNA-seq to identify SNVs in expressed genes. In effect, the depth of coverage in RNA-seq over medium to highly expressed genes achieves many of the same benefits of whole-exome sequencing in rare variant discovery. In the future, deconvolving rare genomic variants that are detectable by RNA-seq from true RNA editing events will be done more optimally by simultaneous analysis of the raw reads from RNA-seq and genome-resequencing events at even higher coverage than readily available today.

Methods

Reads for each biological replicate data set were mapped to an expanded genome consisting of the human reference (GRCv37 / UCSC hg19) plus GENCODE v7 splice junctions and added spike sequences using Bowtie, version 0.12.7 (Langmead et al. 2009), with at most three mismatches; reporting in SAM format up to one valid alignment per read; suppressing all alignments for a particular read if more than one reportable alignment exist for it; and using only those alignments that fell into the best stratum. We used Bowtie to map reads instead of TopHat because the SNVs that were called using TopHat did not have a significant difference in SNV distribution when using RNA-seq reads or DNase-seq reads. Read ends were mapped separately and pooled afterward, without taking into account pairing information. The resulting SAM files were then stripped of spliced reads; converted to BAM files using samtools, version 0.1.17; sorted; and indexed; variants were called using the pileup command. We called an SNV when at least three nonidentical reads support a nonreference variant, and the variant is present at a minimum frequency of 10% and is supported by at least one read per strand. We discarded sites with more than one type of SNV call at the same location. In addition to the SNV calls in each full data set, a parallel set of analyses was done with potentially duplicated reads removed using the rmdup option of samtools to create the collapsed set.

The intersection of SNV calls in biological replicates from the full BAM file and the intersection of SNV calls in the collapsed BAM files were intersected to create a list of candidate SNVs. Known SNPs from dbSNP132 that were not annotated as based on cDNA and sites lying outside the "comprehensive set" of GENCODE v7 protein coding gene boundaries were set aside, and the remaining novel SNVs within genic regions were corrected for strand sense. These novel genic candidates were then annotated using ANNOVAR (Wang et al. 2010) with a splicing threshold of 5. Gene Ontology analysis was performed using GREAT, version 1.8.2 (McLean et al. 2010). Sites that were annotated by ANNOVAR as splicing were filtered out. To filter out genomic SNVs, genomic alignments were obtained from the 1000 Genomes project for the CEU GM trio, and ChIP-seq alignments were obtained from ENCODE ChIP-seq data from HudsonAlpha and histone modification data from the Broad Institute. Samtools mpileup was used to look at the nucleotide composition over the SNVs, and sites with any evidence of a genomic SNV were filtered out. Hierarchical clustering for editing frequency of individual sites and number of edits for genes was done using Cluster 3.0 (de Hoon et al. 2004), with the centroid linkage hierarchical clustering option of both sites/genes and cell types. The heatmap was viewed using TreeView-1.1.5 (Page 2002).

Figure 2-1.

RNA SNV calling strategy. (A) Flowchart of analysis: 75-bp paired-end RNA-seq reads were mapped onto an extended genome (genome + known splice junctions + spikes) using Bowtie. Reads mapping onto splice sites and spikes were set aside, and reads mapping onto hg19 were used to call single nucleotide variants (SNVs). A parallel set of analyses was done using a collapsed set of reads with unique coordinates, and the intersections of SNVs from the uncollapsed and collapsed treatments were obtained. Known SNPs annotated in dbSNP132, sites outside gene boundaries, and intronic sites within 5 bp of splice junctions were removed. For the GM trio, any candidate with evidence of a private genomic variation was also removed. (B) Example of candidate editing site. Purple arrows pointing to the left represent reads on the (−) strand, while blue arrows pointing to the right represent reads on the (+) strand. The blocks represent variants between the reference DNA and the RNA-seq. A SNV is kept when at least three nonidentical reads support the SNV, with a minimum SNV frequency of 10%, and at least one edit per strand. (C) Intersection strategy for two replicates. For cell types with two replicates, the SNVs remaining after collapsing were intersected between the replicates. (D) The number of SNVs remaining after collapsing for the prefiltered sites. Number of SNVs that are only in the uncollapsed set are in blue; the intersection, purple; and collapsed set, red. (E) Collapsing increases the relative amount of A-to-G SNVs and also increases the relative number of transitions. Number of SNVs that are only in the uncollapsed set are in blue; the intersection, purple; and collapsed set, red. (F) The fraction of dbSNP is highest in the intersection of the full and collapsed sets. The relative amount of calls found in dbSNP132, novel genic SNVs, and other SNVs in the uncollapsed set are at the left; the collapsed set, right; and the intersection of the two, middle.

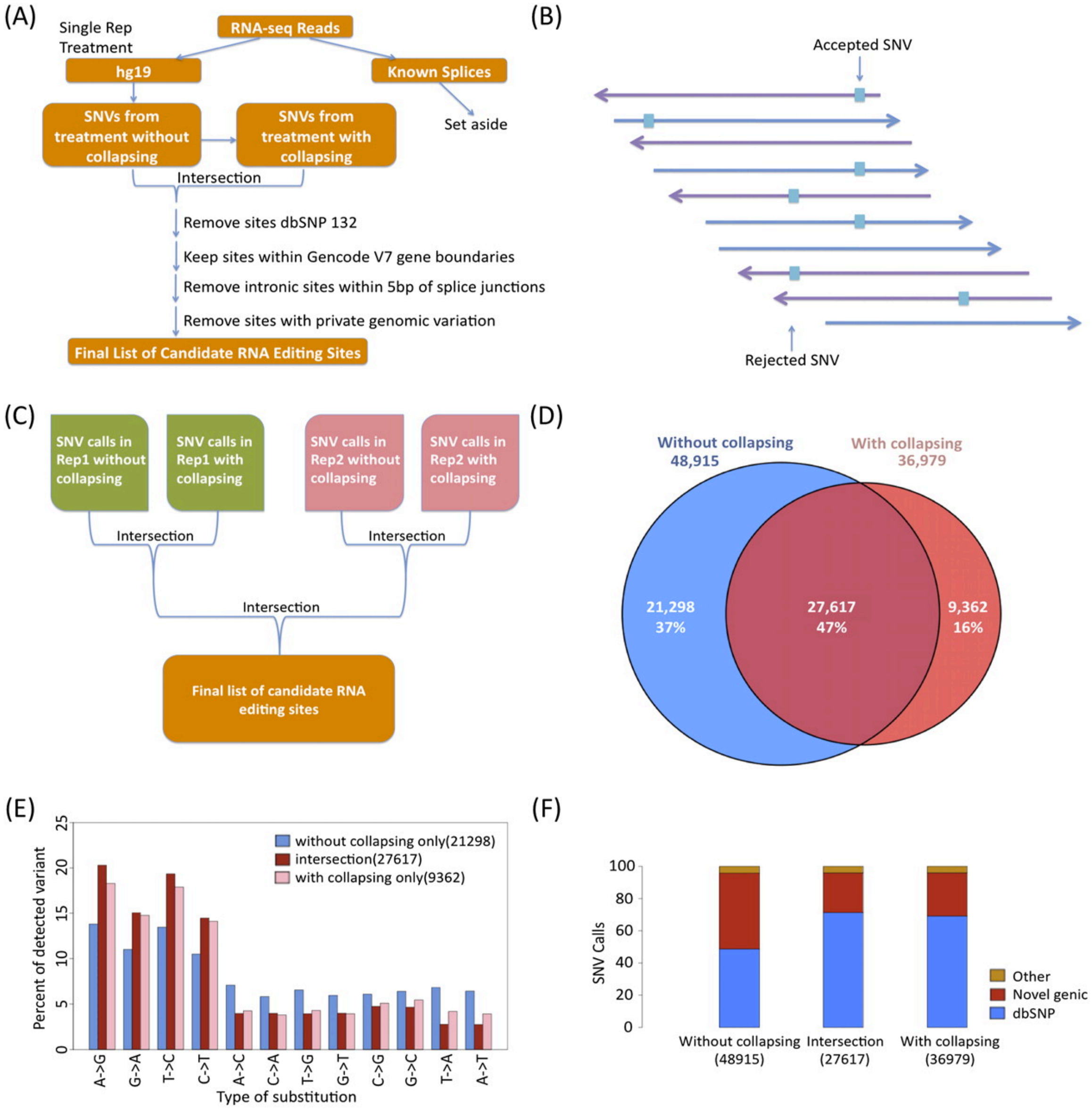

(A)
Single Rep Treatment
RNA-seq Reads
hg19
Known Splices
SNVs from treatment without collapsing
SNVs from treatment with collapsing
Set aside
Intersection
Remove sites dbSNP 132
Keep sites within Gencode V7 gene boundaries
Remove intronic sites within 5bp of splice junctions
Remove sites with private genomic variation
Final List of Candidate RNA Editing Sites
(B)
Accepted SNV
Rejected SNV
(C)
SNV calls in Rep1 without collapsing
SNV calls in Rep1 with collapsing
SNV calls in Rep2 without collapsing
SNV calls in Rep2 with collapsing
Intersection
Intersection
Intersection
Final list of candidate RNA editing sites
(D)
Without collapsing 48,915
With collapsing 36,979
21,298 37%
27,617 47%
9,362 16%
(E)
Percent of detected variant
0
5
10
15
20
25
without collapsing only(21298)
intersection(27617)
with collapsing only(9362)
A->G
G->A
T->C
C->T
A->C
C->A
T->G
G->T
C->G
G->C
T->A
A->T
Type of substitution
(F)
SNV Calls
0
20
40
60
80
100
Other
Novel genic
dbSNP
Without collapsing (48915)
Intersection (27617)
With collapsing (36979)

Figure 2-2.

RNA editing calls in GM12878. (A) Most non–A-to-G SNVs are near splicing boundaries. The distribution relative to gene boundaries of A-to-G SNVs (left) versus non–A-to-G SNVs (right). (B) Example of reads mapped incorrectly across a known splice junction. Overhanging RNA-seq reads are mapped incorrectly into the intron when the correct position is in the adjacent exon, even though the splice junction was provided to the read mapper. (C) Distribution of SNVs at different steps in the pipeline. Prefiltered SNVs defined by having at least three nonidentical reads support the SNV, with a minimum SNV frequency of 10%, at least one edit per strand, and no more than one type of SNV for the same position in blue. SNVs annotated in dbSNP132 are red, SNVs that are not in dbSNP132 and within gene boundaries are green, SNVs that are not in dbSNP132 and within gene boundaries without splicing sites are purple, SNVs that had no matching 1000 Genome sequencing reads are in light blue, and SNVs passing ChIP filtering are in orange. (D) Frequency distribution of SNVs primarily reflects expression of homozygous and heterozygous SNPs. The SNVs that were found in dbSNP132 are in blue; the novel genic SNVs, red. (E) Most nonsplice adjoining SNPs are A-to-G. The nonsplicing novel genic A-to-G calls in filtered calls are in blue; nonsplicing novel genic A-to-G calls, red; nonsplicing novel genic non–A-to-G, brown; nonsplicing novel genic non–A-to-G in filtered calls, purple; and splicing-only novel genic, light blue. (F) Distribution of gene expression versus coverage of exonic sites are in red and intronic sites are in blue for genic SNVs. SNVs in more lowly expressed genes are primarily on exons, due to our minimum depth of coverage requirements.

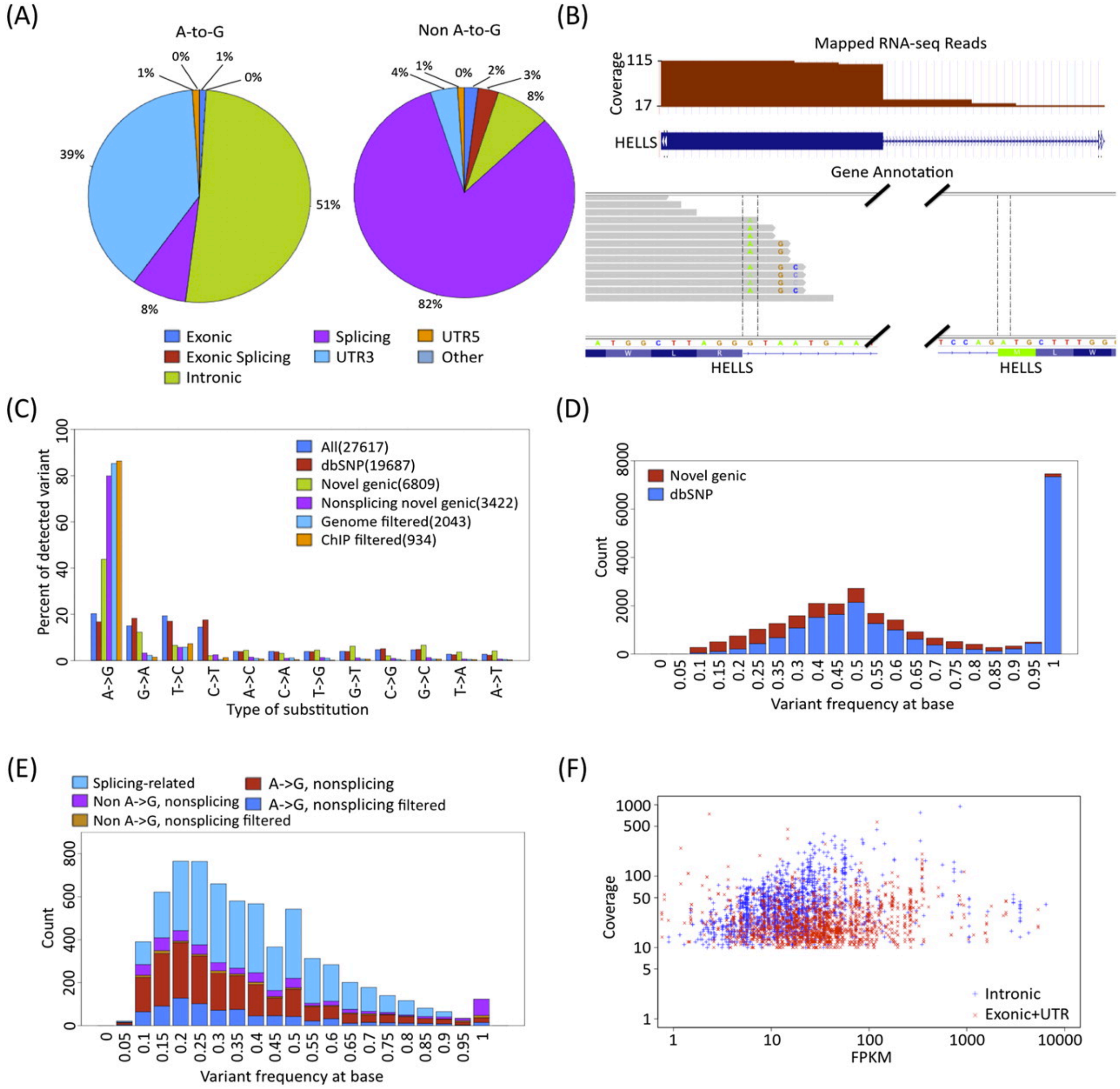

(A)
A-to-G
0%
1%
1%
0%
39%
51%
8%
Non A-to-G
4%
1%
0%
2%
3%
8%
82%
Exonic
Exonic Splicing
Intronic
Splicing
UTR3
UTR5
Other
(B)
Mapped RNA-seq Reads
Coverage
115
17
HELLS
Gene Annotation
HELLS
HELLS
(C)
Percent of detected variant
100
80
60
40
20
0
All(27617)
dbSNP(19687)
Novel genic(6809)
Nonsplicing novel genic(3422)
Genome filtered(2043)
ChIP filtered(934)
A->G
G->A
T->C
C->T
A->C
C->A
T->G
G->T
C->G
G->C
T->A
A->T
Type of substitution
(D)
Count
8000
6000
4000
2000
0
Novel genic
dbSNP
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
Variant frequency at base
(E)
Splicing-related
Non A->G, nonsplicing
Non A->G, nonsplicing filtered
A->G, nonsplicing
A->G, nonsplicing filtered
Count
800
600
400
200
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
Variant frequency at base
(F)
Coverage
1000
500
100
50
10
5
1
Intronic
Exonic+UTR
1
10
100
1000
10000
FPKM

Figure 2-3.

Survey of SNV calls across ENCODE cell lines. (A) Distribution of nonsplicing novel genic SNVs for all data sets. (B) In every cell type, the percentage of A-to-G SNVs increase and the number of candidate sites decrease (red) after filtering for private SNVs using ChIP-seq. GM12878 calls were filtered with 1000 Genomes or ChIP-seq reads are labeled with G or C, respectively. (C) Relatively few non–A-to-G synonymous SNVs (purple), non–A-to-G nonsynonymous SNVs (green), A-to-G synonymous SNVs (red), A-to-G nonsynonymous SNVs (blue) are found in ORFs.

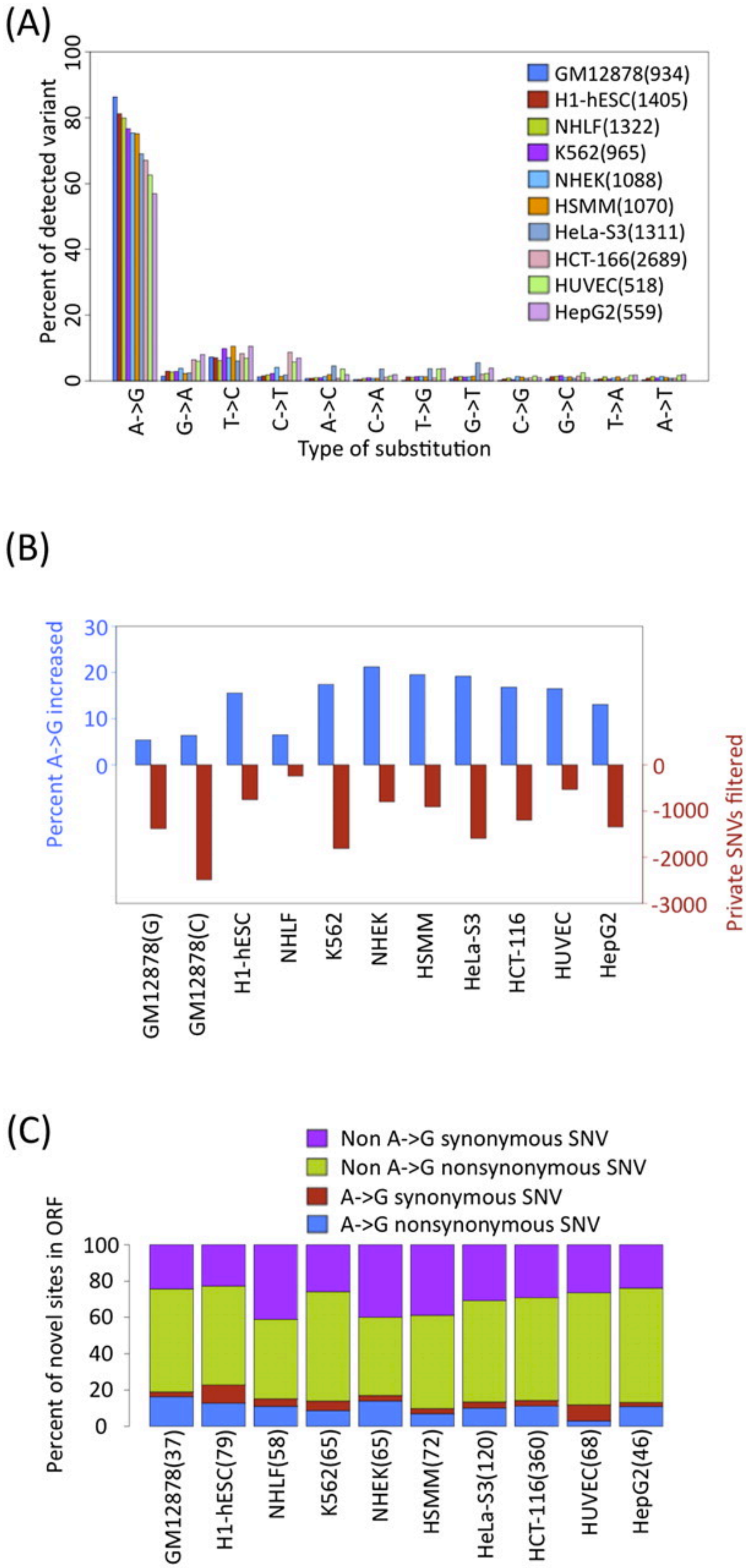

(A)
Percent of detected variant
0
20
40
60
80
100
GM12878(934)
H1-hESC(1405)
NHLF(1322)
K562(965)
NHEK(1088)
HSMM(1070)
HeLa-S3(1311)
HCT-166(2689)
HUVEC(518)
HepG2(559)
A->G
G->A
T->C
C->T
A->C
C->A
T->G
G->T
C->G
G->C
T->A
A->T
Type of substitution
(B)
Percent A->G increased
30
20
10
0
Private SNVs filtered
0
-1000
-2000
-3000
GM12878(G)
GM12878(C)
H1-hESC
NHLF
K562
NHEK
HSMM
HeLa-S3
HCT-116
HUVEC
HepG2
(C)
Non A->G synonymous SNV
Non A->G nonsynonymous SNV
A->G synonymous SNV
A->G nonsynonymous SNV
Percent of novel sites in ORF
100
80
60
40
20
0
GM12878(37)
H1-hESC(79)
NHLF(58)
K562(65)
NHEK(65)
HSMM(72)
HeLa-S3(120)
HCT-116(360)
HUVEC(68)
HepG2(46)

Figure 2-4.

Gene level analysis of RNA editing after private SNV filtering. (A) Hierarchical clustering of the editing frequency of the 33.5% (1905 out of 5695 possible) individual A-to-G candidate editing sites occurring in at least two distinct cell types. (B) Hierarchical clustering of the number of edits in the 47.4% (662 out of 1395 possible) of genes edited in at least two distinct cell types. (C) RNA editing in genes cluster in the UTR or in the introns with few genes having edits in both UTR and introns. Percentage of genes with only UTR edits are in green; intronic edits, blue; and edits in both introns and UTR, red. (D) Reproducibility of calling RNA edits for human H1 ES cells. Scatter plot of RNA edit calls for rep 1,2 versus rep 3,4 is on a log2-log2 scale with a pseudocount of 1. A Gaussian noise was added to points to visualize density. (E) Venn diagrams of A-to-G candidate edits in lymphoblastoid cells from a hapmap trio. The Venn diagram of the individual sites (left) and edited genes (right); 35.8% of the union of edited sites are found in two or more cell types, while 54.2% of the union of edited genes are found in two or more cell types.

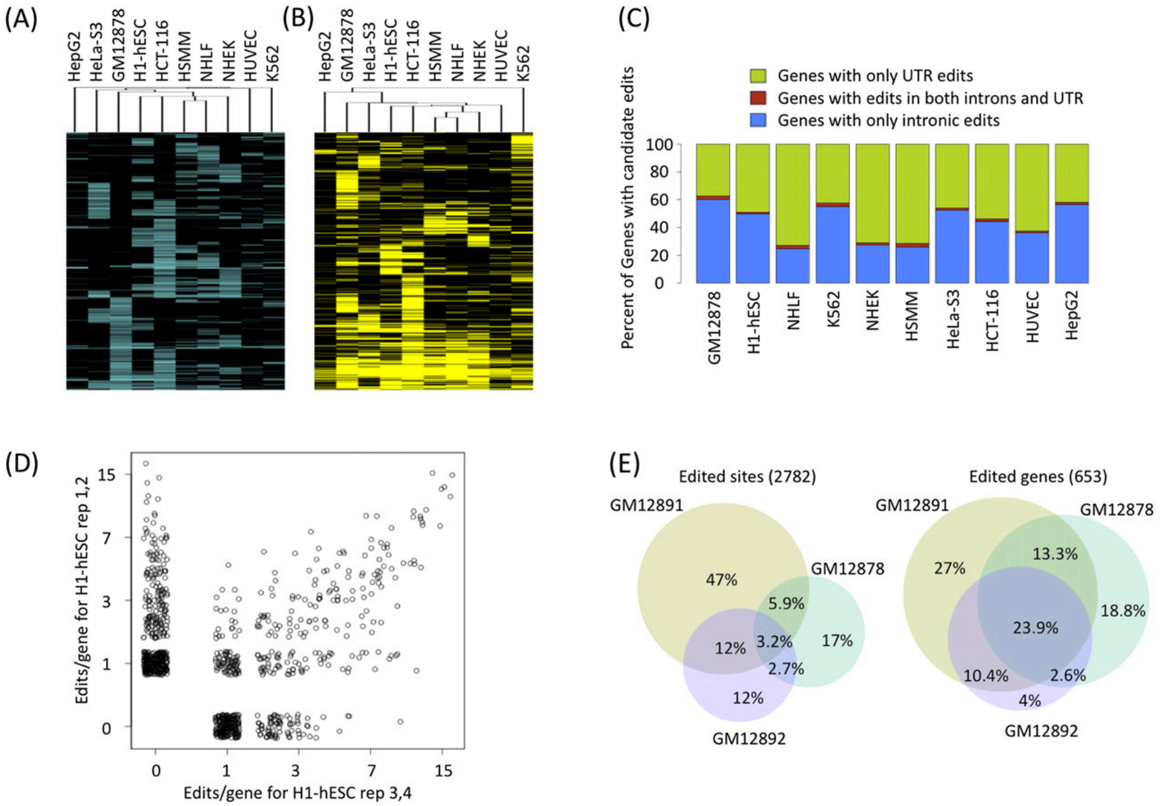

(A)
HepG2
HeLa-S3
GM12878
H1-hESC
HCT-116
HSMM
NHLF
NHEK
HUVEC
K562
(B)
HepG2
GM12878
HeLa-S3
H1-hESC
HCT-116
HSMM
NHLF
NHEK
HUVEC
K562
(C)
Genes with only UTR edits
Genes with edits in both introns and UTR
Genes with only intronic edits
Percent of Genes with candidate edits
100
80
60
40
20
0
GM12878
H1-hESC
NHLF
K562
NHEK
HSMM
HeLa-S3
HCT-116
HUVEC
HepG2
(D)
Edits/gene for H1-hESC rep 1,2
15
7
3
1
0
0
1
3
7
15
Edits/gene for H1-hESC rep 3,4
(E)
Edited sites (2782)
GM12891
GM12878
GM12892
47%
5.9%
12%
3.2%
17%
2.7%
12%
Edited genes (653)
GM12891
GM12878
GM12892
27%
13.3%
18.8%
23.9%
10.4%
2.6%
4%

Figure 2-S1

S1

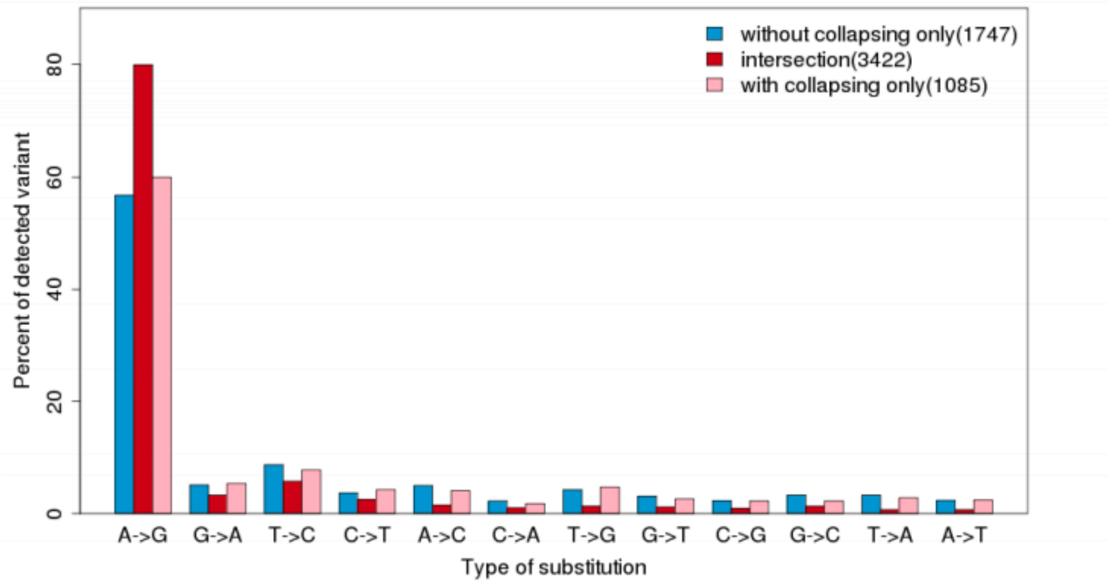


Figure 2-S2

S2

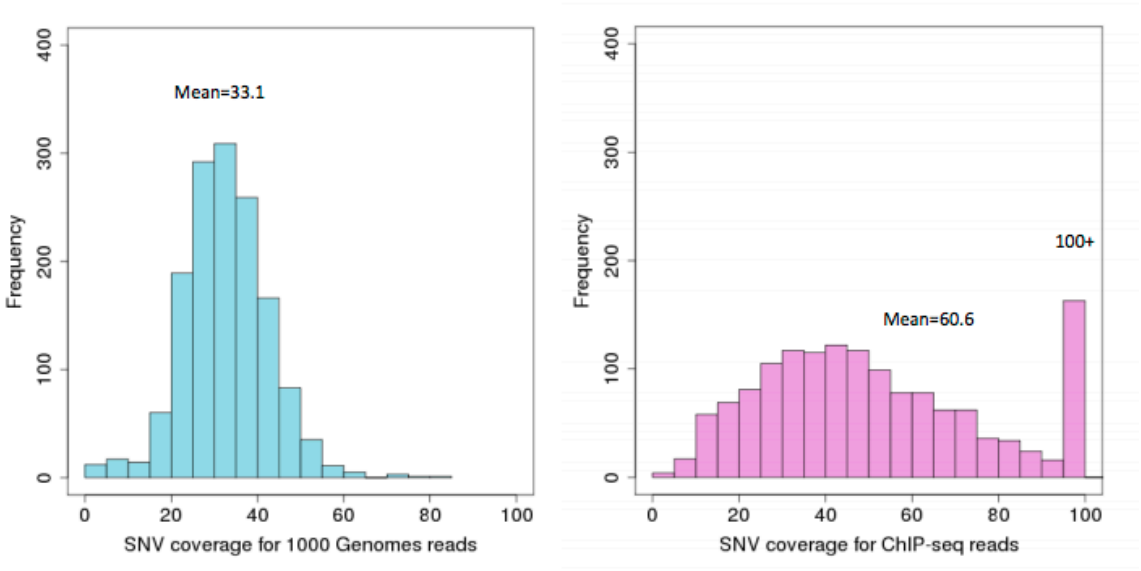


Figure 2-S3

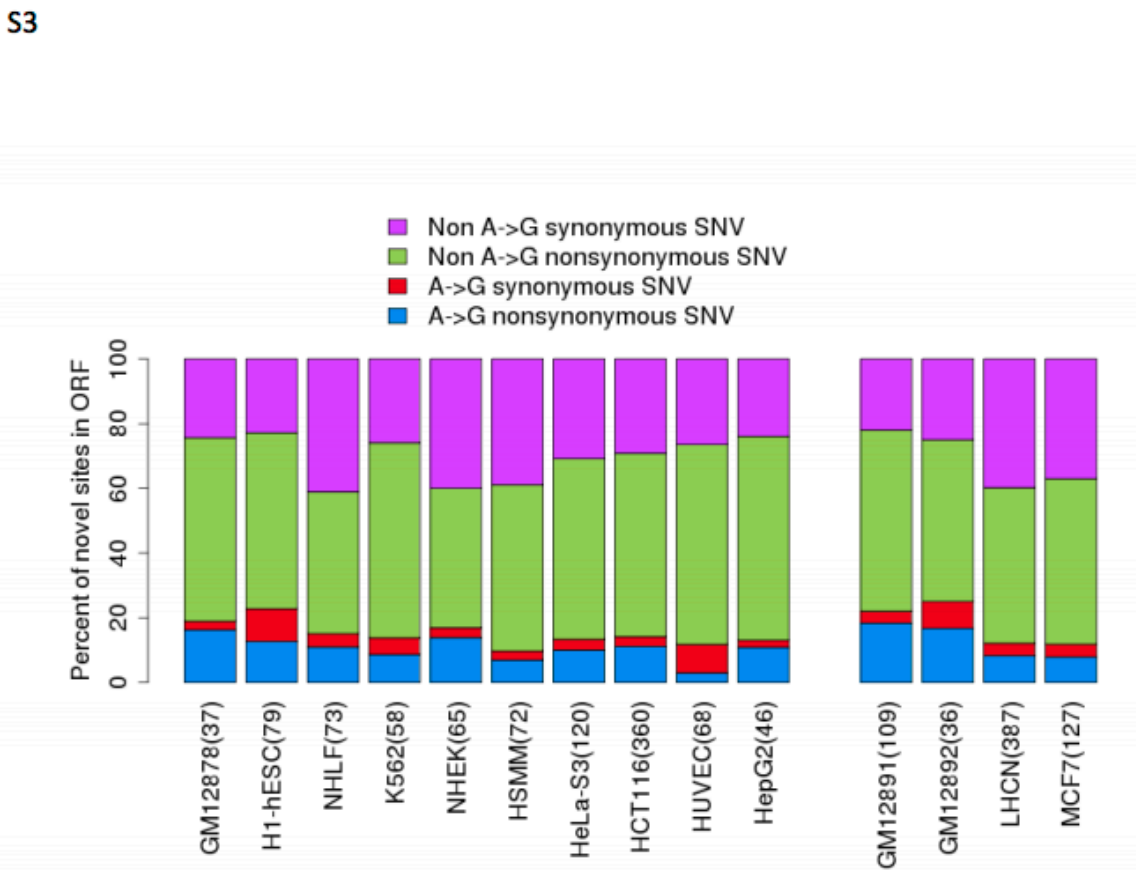

S3
Non A->G synonymous SNV
Non A->G nonsynonymous SNV
A->G synonymous SNV
A->G nonsynonymous SNV
Percent of novel sites in ORF
0
20
40
60
80
100
GM12878(37)
H1-hESC(79)
NHLF(73)
K562(58)
NHEK(65)
HSMM(72)
HeLa-S3(120)
HCT116(360)
HUVEC(68)
HepG2(46)
GM12891(109)
GM12892(36)
LHCN(387)
MCF7(127)

Figure 2-S4

S4

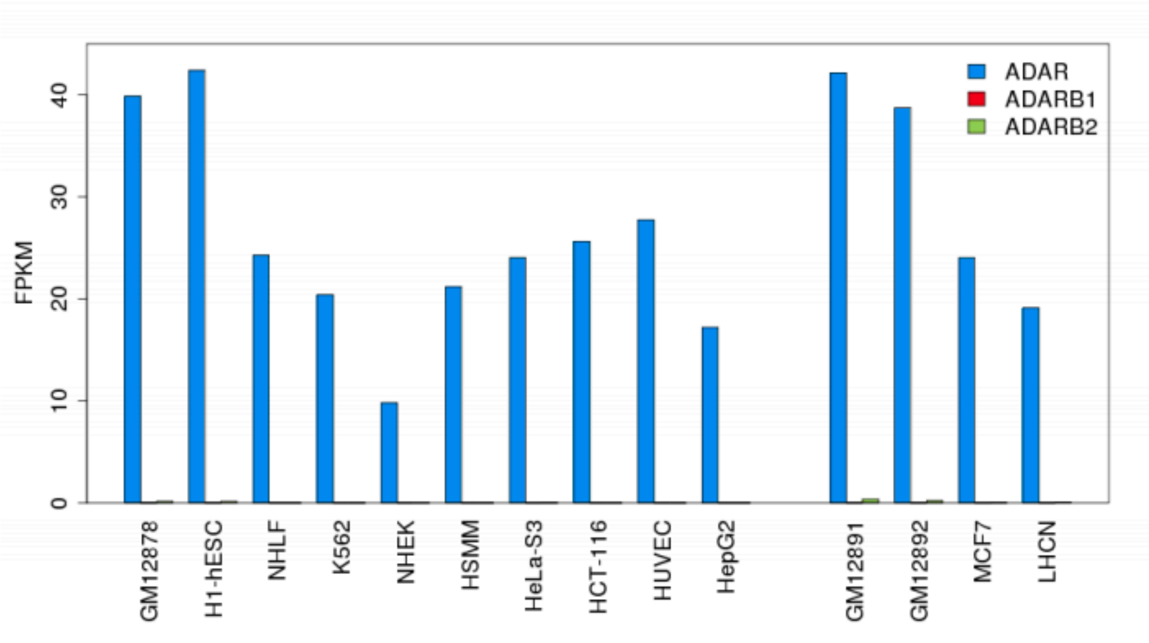


Figure 2-S5

S5

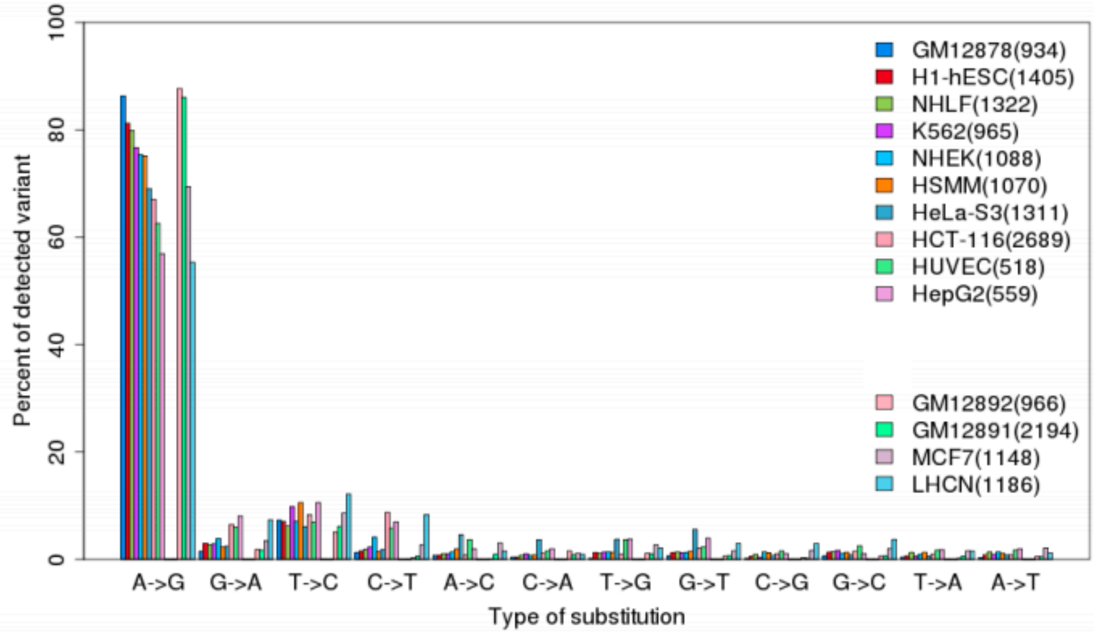


Figure 2-S6

S6

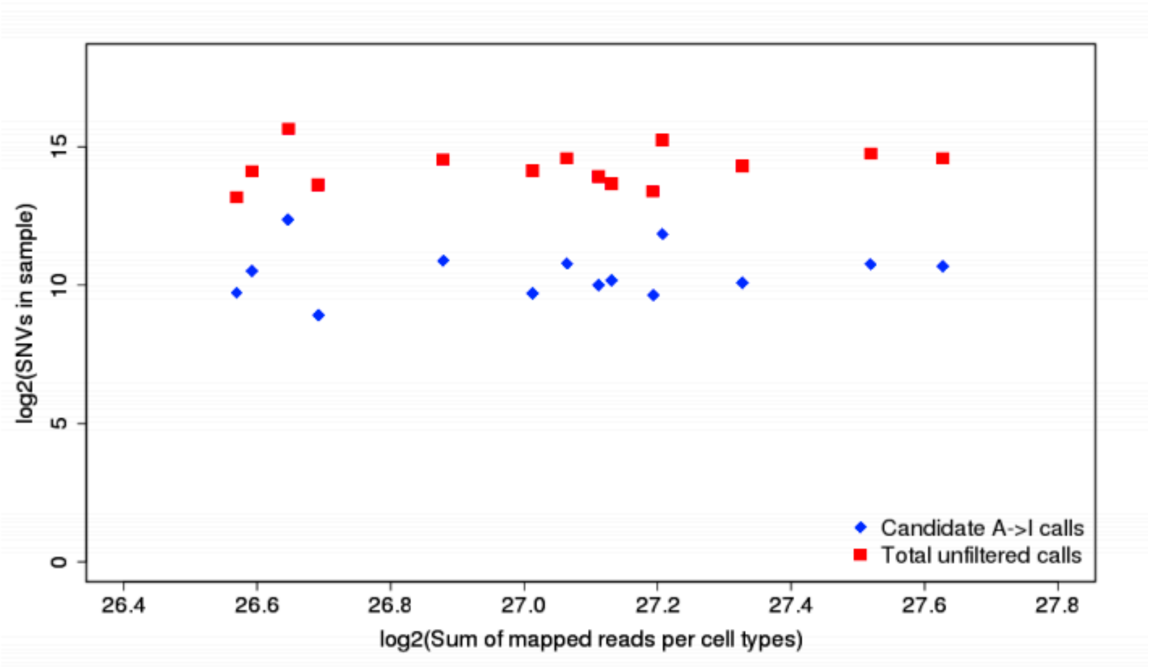


Figure 2-S7

S7

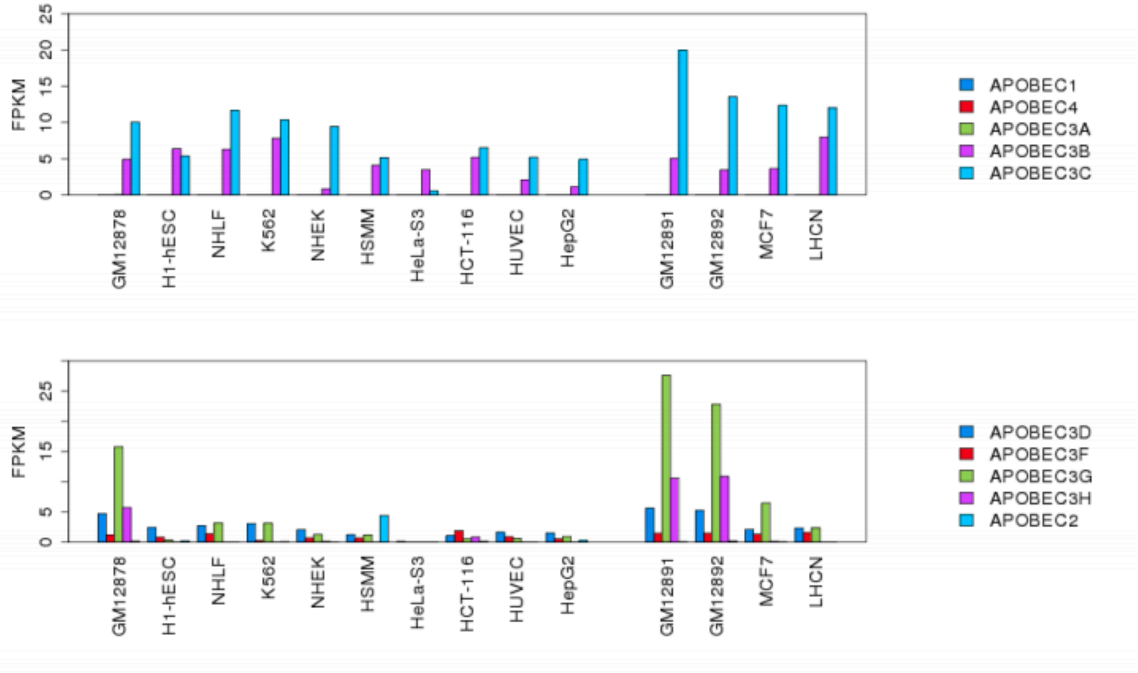


Figure 2-S8

S8

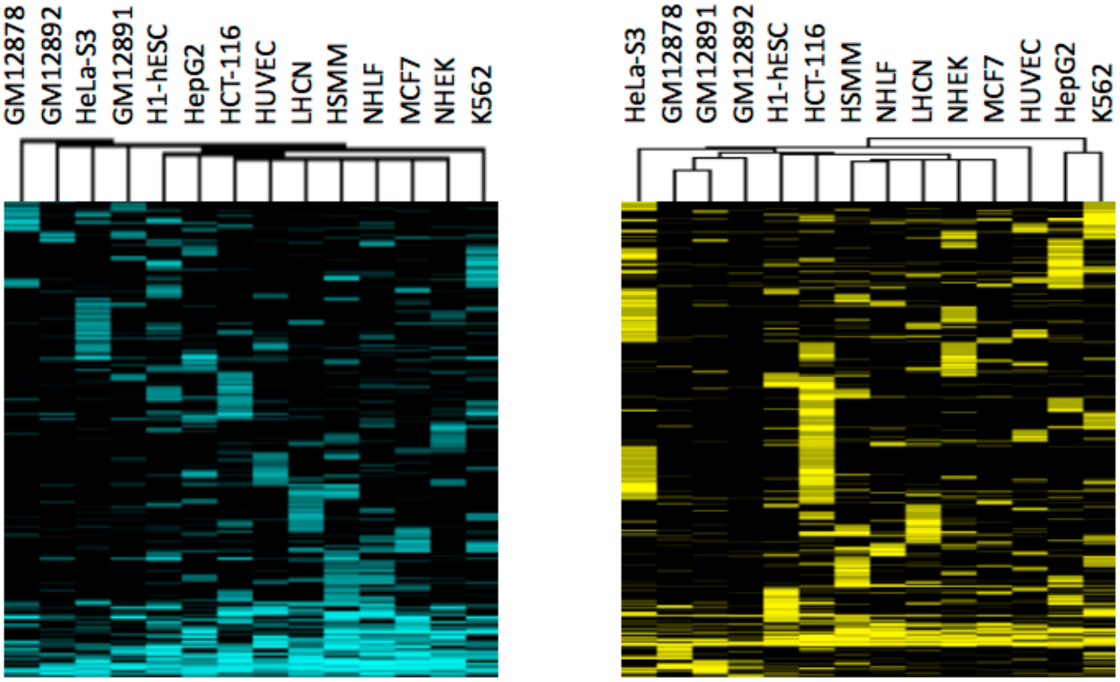


Fig 2-S9

S9

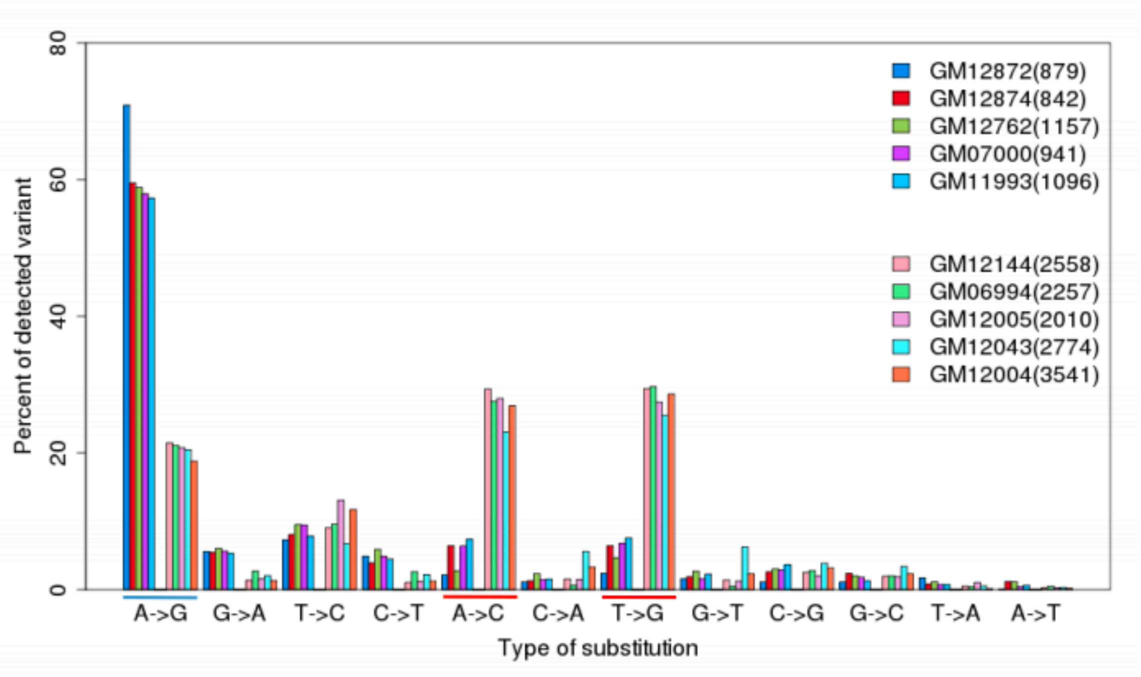

80
60
40
20
0
Percent of detected variant
A->G
G->A
T->C
C->T
A->C
C->A
T->G
G->T
C->G
G->C
T->A
A->T
Type of substitution
GM12872(879)
GM12874(842)
GM12762(1157)
GM07000(941)
GM11993(1096)
GM12144(2558)
GM06994(2257)
GM12005(2010)
GM12043(2774)
GM12004(3541)

Fig 2-S10

S10

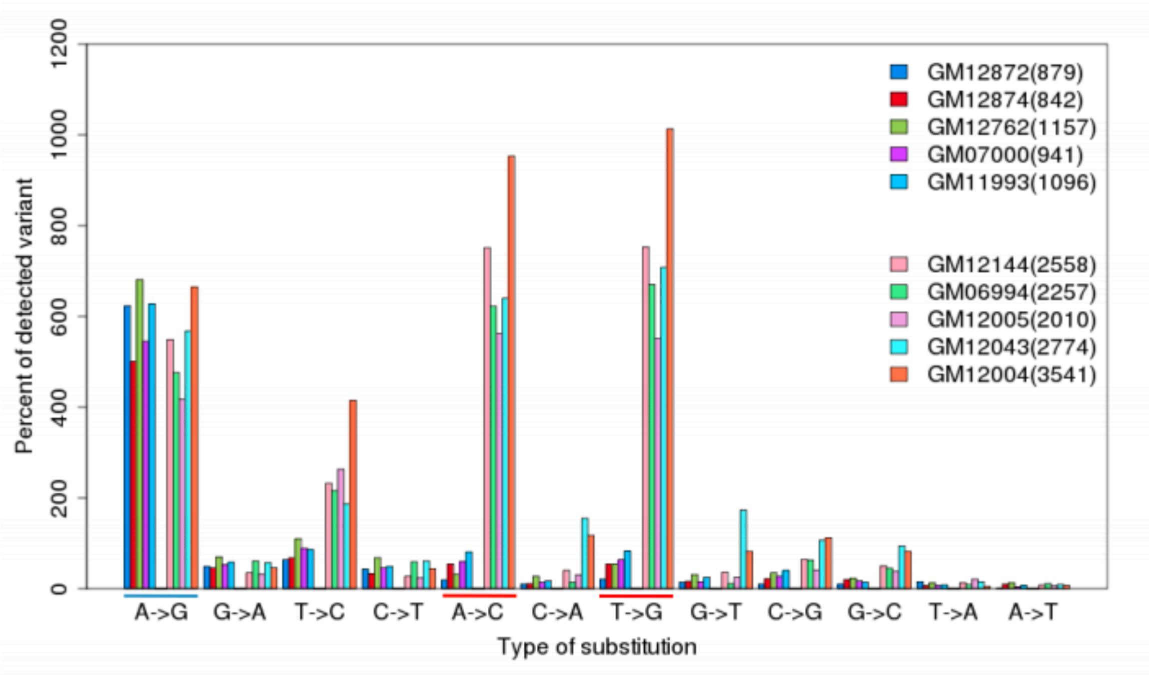

Percent of detected variant
0
200
400
600
800
1000
1200
GM12872(879)
GM12874(842)
GM12762(1157)
GM07000(941)
GM11993(1096)
GM12144(2558)
GM06994(2257)
GM12005(2010)
GM12043(2774)
GM12004(3541)
A->G
G->A
T->C
C->T
A->C
C->A
T->G
G->T
C->G
G->C
T->A
A->T
Type of substitution

Fig 2-S11

S11

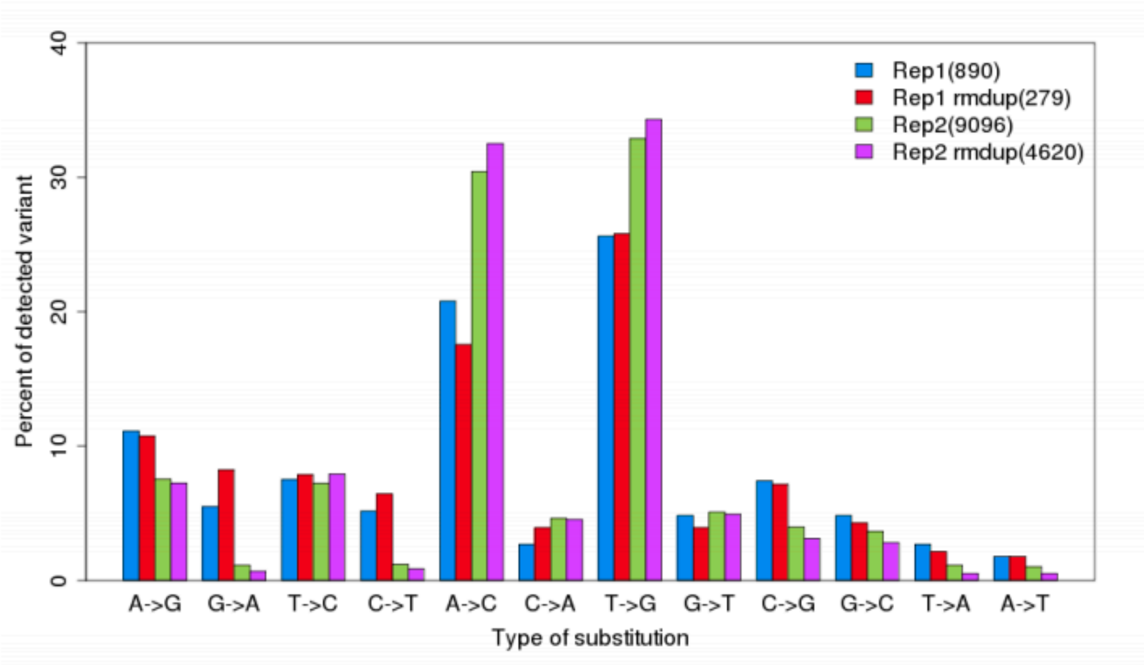
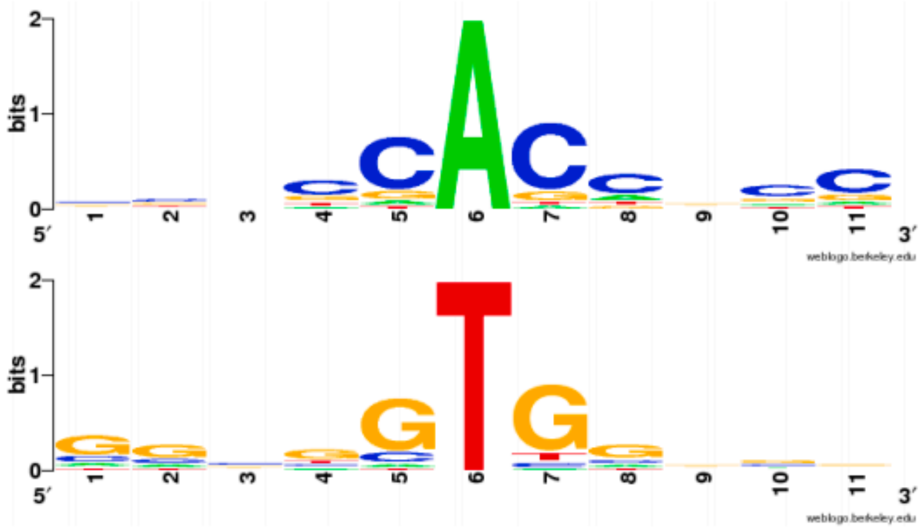

Rep1(890)
Rep1 rmdup(279)
Rep2(9096)
Rep2 rmdup(4620)
Percent of detected variant
0
10
20
30
40
A->G
G->A
T->C
C->T
A->C
C->A
T->G
G->T
C->G
G->C
T->A
A->T
Type of substitution

Fig 2-S12

S12


2
bits
1
0
5′
1
2
3
4
5
6
7
8
9
10
11
3′
weblogo.berkeley.edu
2
bits
1
0
5′
1
2
3
4
5
6
7
8
9
10
11
3′
weblogo.berkeley.edu

Chapter 3

Genome-wide identification and functional analysis of Apobec-1-mediated C-to-U RNA editing in mouse small intestine and liver

Abstract

Background

RNA editing encompasses a post-transcriptional process in which the genomically templated sequence is enzymatically altered and introduces a modified base into the edited transcript. Mammalian C-to-U RNA editing represents a distinct subtype of base modification, whose prototype is intestinal apolipoprotein B mRNA, mediated by the catalytic deaminase Apobec-1. However, the genome-wide identification, tissue-specificity and functional implications of Apobec-1-mediated C-to-U RNA editing remain incompletely explored.

Results

Deep sequencing, data filtering and Sanger-sequence validation of intestinal and hepatic RNA from wild-type and Apobec-1-deficient mice revealed 56 novel editing sites in 54 intestinal mRNAs and 22 novel sites in 17 liver mRNAs, all within 3′ untranslated regions. Eleven of 17 liver RNAs shared editing sites with intestinal RNAs, while 6 sites are unique to liver. Changes in RNA editing lead to corresponding changes in intestinal mRNA and protein levels for 11 genes. Analysis of RNA editing in vivo following tissue-specific Apobec-1 adenoviral or transgenic Apobec-1 overexpression reveals that a subset of targets identified in wild-type mice are restored in Apobec-1-deficient mouse intestine and liver following Apobec-1 rescue. We find distinctive polysome profiles for several RNA editing targets and demonstrate novel exonic editing sites in nuclear preparations from intestine but not hepatic apolipoprotein B RNA. RNA editing is validated using cell-free extracts from wild-type but not Apobec-1-deficient mice, demonstrating that Apobec-1 is required.

Conclusions

These studies define selective, tissue-specific targets of Apobec-1-dependent RNA editing and show the functional consequences of editing are both transcript- and tissue-specific.

Introduction

There is considerable interest in understanding both the repertoire of and mechanisms for RNA-DNA differences reported from deep sequencing (RNA-seq) of mammalian transcriptomes (Gu 2012, Lagarrigue 2013, Bahn 2012, Danecek 2012, Peng 2012, Mortazavi 2008). Among the mechanisms for RNA-DNA differences is RNA editing, in which genomically templated RNA sequences are enzymatically altered. The most prevalent type of editing involves a base change from adenosine to inosine (A-to-I), mediated by adenosine deaminases acting on (double-stranded) RNA (ADARs) (Nishikura 2010). A second, much less prevalent type of RNA editing involves deamination of cytidine to uridine (C-to-U) in single-stranded RNA, mediated by Apobec-1, a member of the APOBEC family of cytidine deaminases (Smith 2012). The prototype for mammalian C-to-U RNA editing is apolipoprotein B (apoB) RNA, where Apobec-1-mediated deamination of a CAA codon introduces a translational termination (UAA) codon in the edited transcript. ApoB mRNA editing is a critical adaptive pathway for lipid transport in both the mouse intestine and liver, and exhibits distinctive developmental and metabolic regulation (Blanc 2011), mediated via the expression and stoichiometric interactions of two dominant trans-acting proteins, Apobec-1 and Apobec-1 complementation factor (ACF), although other proteins are implicated (Blanc 2011, Chen 2007, Lellek 2000, Mehta 2000).

Although much is known about the regulation and functional consequences of apoB mRNA editing, remarkably little is known about the range of other targets of C-to-U RNA editing. A recent transcriptome-wide analysis of mouse enterocytes identified 32 novel (non-apoB) Apobec-1-dependent editing targets, all within 3′ untranslated regions (3′ UTRs) (Rosenberg 2011a). These newly identified RNA targets share features with apoB

RNA, including a preference for cytidines embedded in AU-rich regions along with variations of a downstream 11-nucleotide cassette referred to as a 'mooring sequence' with the consensus sequence WRAUYANUAU (Rosenberg 2011a, Backus 1991). Those findings raise the corollary questions of whether any of the novel editing targets identified in mouse intestine are also modified in other tissues expressing Apobec-1, particularly the liver, and, if so, are they modified at the same site and to the same extent, and do these editing events lead to differences in mRNA or protein levels?

Here we used stringent filtering and sequence validation to reveal multiple new sites of Apobec-1-dependent C-to-U RNA editing, with examples of both tissue-specific and common targets (Fig. 3-1). We show that RNA editing led to corresponding changes in mRNA and protein expression in a subset of mRNAs. We also find enrichment in the edited forms of certain mRNAs in cytoplasmic compared to nuclear fractions. We further show that mRNA editing regulates polysome distribution of a subset of targets. We demonstrate editing of some but not all novel targets using cell-free extracts from wild-type (WT) but not Apobec-1-deficient mice, demonstrating that Apobec-1 is necessary for RNA editing. Taken together, our findings demonstrate that C-to-U RNA editing exerts distinct tissue-specific consequences, including a spectrum of outcomes on protein expression.

Results

Overview

We undertook a comprehensive comparison of Apobec-1-dependent C-to-U RNA editing, starting with transcriptome-wide analyses of small intestine mucosa and liver from WT and Apobec-1-/- mice (Fig. 3-1). We then extended those analyses to other lines of mice with low or high transgenic intestinal overexpression of Apobec-1 in either an Apobec-1-/- (that is, Apobec-1Int/O) or WT background (Apobec-1Int/+) (Blanc 2012) (Fig. 3-1). We further studied livers from Apobec-1-/- mice following adenoviral delivery of Apobec-1 (ad-Apobec-1) or a LacZ control virus (Fig. 3-1). This strategy permitted an evaluation of tissue-specific (that is, intestine versus liver) and dose-dependent (that is, WT versus Apobec-1Int/+and Apobec-1Int/OLo versus Apobec-1Int/OHi) changes in C-to-U RNA editing at different levels of Apobec-1 expression (Blanc 2012).

Identification of novel intestinal and hepatic Apobec-1-dependent editing targets

The first task was to examine the 70 to 200 million RNA-seq reads for intestine and liver from WT and Apobec-1-/- mice identifying mRNA sequences with C-to-U differences. C-to-U mismatches found in both WT and Apobec-1-/- mice as well as sites with less than three reads were excluded from further analysis. Results for WT intestine revealed a total of 438 putative editing sites (including apoB), 372 (85%) of which were located in the 3′ UTR, and with the remainder residing in 5′ UTR (7; 1.6%), exonic (7; 1.6%) or intergenic regions (52; 12%). We selected an arbitrary cutoff of 30% C-to-U editing in 3' UTR calls and then validated 56 of 70 calls (80% true positive) in 54 RNAs (App mRNA was edited at two sites) by Sanger sequencing, including cohort validation of a subset of 23 of the 31 RNA targets identified by Rosenberg et al. (Rosenberg 2011b) (74% true positive). Of the seven exonic sites (six RNAs), two were in apoB (one novel), two others were previously unreferenced SNPs, and the remaining three were false positives based on Sanger sequencing. C-to-U RNA editing efficiency among the novel 3′ UTR targets ranged from 31 to 84%. Together the results identify 54 validated Apobec-1-dependent RNA editing targets from mouse intestine, 32 of which have not been reported previously.

We attempted to account for discordances between our results and those of Rosenberg et al. (Rosenberg 2011a). In two instances, miscalled bases reflected the spurious mapping of reads with errors to a small region ('island') of otherwise unexpressed paralogs of an unedited expressed gene. In one instance the location of the editing site was within a homopolymeric stretch of six thymidine residues, known to be vulnerable to nucleotide insertions (Minoche 2011). Four targets (BC003331, Ptpn3, Rb1 and Abcb7) were below our 30% editing threshold, but Sanger sequencing nevertheless validated these mRNAs as Apobec-1. Finally, six additional targets were originally identified in isolated enterocytes (Rosenberg 2011a), rather than from mucosal RNA as in the current study. We then investigated whether the cellular origin of the RNAs might account for these discordances. Sites in Casp6 and Atf2 were sequence-validated using isolated enterocyte RNA. The other four targets were not validated, for reasons that remain to be determined.

Turning to hepatic RNA targets, we identified a total of 39 putative editing sites, of which 27 were located in 3′ UTRs, with the remainder located in 5′ UTR (2; 5%), exonic (6; 15%) and intergenic regions (4; 10%). Because our filtering algorithms indicated fewer putative editing targets in the liver compared to the small intestine, we undertook sequence validation of the entire set and confirmed 22 of 27 3′ UTR sites (81% true positive) distributed across 17 novel RNA targets. Of these 17 liver targets, 11 were also verified by sequence analysis in small intestine, and 6 were unique to liver (Fig. 3-1). Of the 11 RNAs edited in both liver and small intestine, all revealed lower levels of editing in liver (Serinc1: 60 to 66% in intestine by Sanger/RNA-seq versus 9.5 to 38% in liver by Sanger/RNA-seq; Cd36: 85 to 84% in intestine by Sanger/RNA-seq versus 23 to 24% in liver by Sanger/RNA-seq). Most of the shared liver-intestine targets (7/11) were below our threshold for RNA-seq, although Sanger sequencing revealed editing ranging from 4 to 32%. Of the six putative exonic editing sites, two (apoB and a novel site in BC005561), were Sanger sequence validated, while four, not validated by Sanger sequencing, were considered as false positives.

Sequence context features for C-to-U RNA editing

Prompted by findings that a close or exact match to the mooring sequence in apoB RNA was present in almost every other Apobec-1-dependent editing site (Rosenberg 2011b), we examined the flanking sequence of editing sites identified above for features that might explain why some RNAs are edited at much higher efficiency than others. We found the region flanking edited 3′ UTRs to be significantly more AU-rich than a random set of 3′ UTRs in both intestine and liver, which was confirmed by examination of a 101-nucleotide region overlapping the edited sites. Nearest-neighbor nucleotide analysis revealed a strong preference for adenosine and uridine both upstream (-1) and downstream (+1) of the editing site for both intestinal and liver targets. However, mismatches in the mooring sequence, which are required for apoB RNA editing (Backus

1991), did not correlate with intestinal target editing efficiency. For example, Rab1 RNA contained a perfect match to the consensus mooring site and demonstrated 32% editing, while Reps2 RNA contained two mismatches yet exhibited 75% editing. Thus, the immediate sequence context favors Apobec-1-dependent C-to-U RNA editing, but does not distinguish editing targets by tissue type and does not explain the differences in editing efficiency.

Apobec-1 abundance modulates tissue-specific editing efficiency

Previous work demonstrated that transgenic liver overexpression of Apobec-1 produced additional editing sites (so called 'hyperediting') in apoB mRNA and in other targets (Yamanaka 1995, Yamanaka 1997). In order to understand the importance of Apobec-1 expression levels in editing target selection and efficiency, we generated intestinal Apobec-1 transgenic mice on either a WT or Apobec-1-/- background (Blanc 2012) and compared editing efficiencies at different levels of transgene expression among shared RNAs from the indicated genotypes. Specifically, we compared editing efficiencies of shared targets between WT and Apobec-1Int/+ and editing efficiencies of RNA targets shared between Apobec-1Int/OLo and Apobec-1Int/OHi. Among Apobec-1-dependent editing targets in WT mice, a subset demonstrated increased RNA editing in response to increasing levels of Apobec-1 expression. For example, ATP6ap2 demonstrated 28 to 30% editing in WT and 57 to 62% with transgenic overexpression (Apobec-1Int/+), but no detectable editing in Apobec-1-/- mice. Similarly, editing efficiency of ATP6ap2 increased in Apobec-1Int/Hi versus Apobec-1Int/Lo mice. The fold increase observed was variable among RNAs, ranging from 1.2- (Usp25) to 4-fold (Rab1) in WT versus Apobec-1Int/+ mice and from 3 to 80 fold in Apobec-1Int/Lo versus Apobec-1Int/Hi mice. Occasional discordance was found for editing efficiency as inferred from RNA-seq versus Sanger sequencing. For example, Atp6ap2 in Apobec-1Int/Lo mice demonstrated 52% editing by RNA-seq but only 4.5% by Sanger sequencing (1/22 clones edited). Overall, most but not all RNA targets demonstrated increased editing efficiency with increasing Apobec-1 expression.

Examination of eight hepatic RNA editing targets identified in both WT and Apobec-1-/- mice following ad-Apobec-1 transduction revealed increased editing efficiencies for all shared targets from two- (Tmem30a) to nine-fold (Serinc1). Additional C-to-U editing sites (hyper-editing) were also detected; among the eight shared targets, seven RNAs exhibited from one to nine additional editing sites. In addition, as noted above, alignment of nucleotides flanking these edited sites revealed a strong preference for A or U immediately upstream (96%) and downstream (92%) of the edited site, respectively and, as noted above, alignment with the mooring sequence failed to reveal a predictive correlation with hepatic editing efficiency.

In vitro validation of Apobec-1-dependent RNA editing

Because C-to-U RNA editing of a synthetic apoB RNA template can be accomplished using recombinant Apobec-1 and ACF, we asked if editing of these novel targets might also be replicated in an in vitro system. We used a cell-free in vitro editing assay in which RNA from Apobec-1-/- liver was incubated with tissue S100 extract and analyzed by poisoned primer extension analysis (Blanc 2011). This strategy was employed on two candidate RNAs, selected based on their prior identification in small intestine (Roseberg 2011b) and independently in brain (Danecek). We found that Dpyd was approximately 30% edited (Fig 3-2A), while Tmbim6 site 99239051 demonstrated almost complete editing with increasing amounts of WT extracts. For both RNAs, editing was absent in extracts prepared from Apobec-1-/- mice (Fig. 3-2B). C-to-U RNA editing could not be replicated using recombinant Apobec-1 and ACF alone (Fig. 3-2B), conditions previously shown to support in vitro RNA editing of apoB (Blanc 2011). We note that other targets, including Cmtm6, Sh3bgrl, Serinc1 and Cyp4v3, failed to replicate C-to-U editing in this cell-free system (data not shown). Together these findings show that Apobec-1 is required for C-to-U RNA editing and suggest that other factors in addition to ACF may be required for target selectivity and in vitro C-to-U deamination.

Nucleo-cytoplasmic distribution of edited RNAs

Earlier studies demonstrated that apoB RNA undergoes post-transcriptional RNA editing in the nucleus of rat liver (Lau 1991). Those findings demonstrated that C-to-U RNA editing was virtually complete on spliced, polyadenylated intranuclear apoB RNA and that little if any additional editing took place in the cytoplasmic compartment (Lau 1991). We confirmed that >90% intestinal apoB RNA was edited at the canonical site (6666) in both nucleus and cytoplasm, but in addition observed several subpopulations of edited apoB RNAs with distinctive nucleo-cytoplasmic distributions (Fig. 3-3A). Specifically, intestinal nuclear apoB RNA contained a cluster of C-to-U sites distributed between positions 6702 and 6968 in addition to the canonical 6666 site (Fig. 3-3A). None of these sites was edited in liver RNA (Fig. 3-3B). Unexpectedly, intestinal nuclear apoB RNA also demonstrated extensive (>90%) exonic C-to-U editing at positions 6583 and 6659. These sites were again not edited in liver RNA (Fig. 3-3B). RNA editing at position 6583 modifies an ACA to an AUA codon, resulting in a threonine to isoleucine substitution, while editing at position 6659 (UAC to UAU) is a silent modification (Tyr-Tyr) (Fig. 3-3A). In addition, a much lower proportion (19 to 33%) of cytoplasmic apoB RNA contained these two additional edited exonic sites (6583, 6659) compared with what was observed (approximately 90%) in the nucleus.We next turned to the nucleo-cytoplasmic distribution for other editing targets. For Atp6ap2, we identified two edited sites (positions 12193607 and 12193524; Fig. 3-3C). Atp6ap2 RNA edited at both sites was detected only in the cytoplasm and at low frequency (4%, 1/22 clones edited). By contrast, Atp6ap2 RNA containing only the edited site 12193607 was abundantly represented in cytoplasm (45%, 10/22 edited clones) compared to nucleus (23%, 5/22 edited clones sequenced). For Usp25, we identified only a single RNA population edited at site 77116537 and found 68% of the transcripts containing the edited site in cytoplasm (13/19 edited clones) but only 18% editing in nuclear transcripts (4/22 edited clones). Among the testable hypotheses to account for these observations is that RNA editing of Atp6ap2 and Usp25 may favor cytoplasmic export or influence the pathways modulating turnover of the edited RNA. The extent to which other edited RNAs show differences in subcellular distribution remains to be determined.

Apobec-1-mediated changes in mRNA abundance and microRNA seed sites

We next asked whether RNA editing exerts functional effects on the modified transcripts. We undertook transcriptome-wide comparison of intestinal mRNA abundance of the 58 validated editing targets, of which 32 were significantly down-regulated in Apobec-1-/- intestine (lower FPKM (fragments per kilobase of exon per million), as inferred from RNA-seq alignment frequency; see Materials and methods) and a subset of these same samples were validated with quantitative PCR. The remainder showed either no change or (in a single case, Dek) a trend to increased mRNA abundance (more than two-fold) in Apobec-1-/- mice. Similar analysis of liver RNA revealed one target (Cd36) down-regulated (more than two-fold) in Apobec-1-/-, but the majority of targets (11/16) showed no change in expression. Among the 335 differentially expressed mRNAs (Fig. 3-4A), a subset of 17 demonstrated C-to-U RNA editing, although there was no correlation between the extent of editing and mRNA abundance.

Several studies show that A-to-I RNA editing modifies microRNA (miRNA) sites and influences mRNA abundance (Peng 2012, Borchert 2009, Chen 2013). Accordingly, we investigated the possibility that C-to-U editing might create, eliminate or change the affinity of miRNA seed sequences that in turn might influence gene expression. For intestinal targets, the Siglec 5 editing site is contained within four miRNA seed motifs. Interestingly, loss of Siglec5 RNA editing in Apobec-1-deficient mice resulted in a nine-fold decrease in mRNA abundance and not only eliminates four of those miRNA sites (from WT mice), but simultaneously creates five new seed motifs. By contrast, C-to-U editing creates miRNA seed motifs in five other RNA targets (App, Cnih, B2m, Mtmr2 and Sh3bgrl) that show no change in mRNA expression. For liver samples, loss of CD36 editing in Apobec-1-/- mice led to a two-fold mRNA decrease compared to WT samples, yet simultaneously eliminated a miRNA seed motif. Furthermore, RNA editing created miRNA seed motifs in three other hepatic targets whose mRNA abundance either increased in Apobec-1-/- mice or remained unchanged. Taken together, the findings reveal no consensus mechanism by which C-to-U editing within the 3′ UTR alters miRNA binding sites and influences mRNA abundance.

Apobec-1-dependent C-to-U RNA editing influences protein abundance

Since we did not observe a consensus mechanism by which RNA editing regulates mRNA abundance, we asked if RNA editing might influence translational efficiency. We turned to a proteome-wide approach using mass spectrometry-based shotgun proteomics in conjunction with metabolic labeling for quantification to identify 893 proteins that were differentially expressed in small intestine from WT versus Apobec-1-/- mice. Comparison with our transcriptome-wide analyses revealed 26 differentially expressed proteins encoded by an RNA target of Apobec-1 dependent C-to-U editing (Fig. 3-4A). Using a two-fold change in protein expression as a cutoff, we demonstrated a concordant increase in both mRNA and protein expression in WT compared to Apobec-1-deficient mice in 10 targets. One additional target (Ido1) showed a decrease in both RNA and protein abundance in WT compared to Apobec-1-deficient mice. We confirmed this pattern of differential intestinal protein expression for two targets, Cd36 (which showed the greatest magnitude of C-to-U RNA editing, 84%) and Ido1 (Fig. 3-4B, Fig. 3-4C). Cd36 RNA was demonstrated to be approximately two-fold down-regulated in Apobec-1-/- intestine (FPKM and quantitative PCR). Western blot analysis showed an approximately four-fold decrease of Cd36 protein expression in Apobec-1-/- intestine (Fig. 3-4B). Analysis of Ido1 revealed a trend towards increased protein expression in Apobec-1-/- intestine, consistent with the findings from the proteomic survey (Fig. 3-4C).

In seeking an explanation for the changes in protein expression, we considered the possibility that RNA editing influenced mRNA translation by shifting transcript distribution within translating ribosome subfractions. WT intestinal extracts revealed 95% apoB RNA segregated into polysomal fractions while apoB RNA from Apobec-1-deficient mice was distributed into both polysome and monosome fractions (Fig. 3-5A, Fig. 3-5B). We extended this analysis to editing targets that demonstrated alterations in both mRNA and protein abundance. Cyp2c65 RNA from WT mice fractionated predominantly into polysomes but Apobec-1-deficient mice showed distinctive populations of RNA associated with monosome fractions (Fig. 3-5C). By contrast, intestinal Hpgd mRNA revealed virtually overlapping profiles in WT and Apobec-1-/- mice (Fig. 3-5D). Intestinal Cyp2j6 mRNA associated mostly with high molecular weight polysome fractions in WT animals but revealed a shift into lighter fractions in Apobec-1-/- mice (Fig. 3-5E). Intestinal Ido1 mRNA demonstrated a shift into monosome-associated fractions in Apobec-1-/- mice (Fig. 3-5F). These findings together suggest that Apobec-1 and C-to-U RNA editing individually influence polysome loading of a subset of target RNAs (including apoB), and (with the exception of Ido1 whose protein abundance was increased in Apobec-1-deficient mice) suggest a plausible mechanism whereby editing might selectively influence protein expression.

Discussion

Here we report a comprehensive, comparative analysis of Apobec-1-dependent C-to-U RNA editing in mouse intestine and liver and show that the functional effects are both transcript- and tissue-specific. These tissues were selected because they represent the dominant sites of expression of both Apobec-1 as well as its canonical target, apoB. Our approach included Sanger sequence validation to reinforce the confidence of the findings (74 to 81% true positive), an important consideration in view of recent transcriptome-wide analyses reporting approximately 49% false discovery rates for non A-to-I RNA editing (Peng 2012). Given that we restricted our analysis to 3′ UTR targets and for the small intestine to targets showing at least 30% C-to-U RNA editing, the findings represent a conservative view of the scale of Apobec-1-dependent C-to-U RNA editing and its functional implications.

We validated most but not all the findings of transcriptome-wide Apobec-1-mediated RNA editing in enterocytes (Rosenberg 2011b). Some of the discordances were accounted for by differences in the optimized parameters (Park 2012) used in the current report, including filters for sequence quality, strand bias, distance to end of reads, paired-end reads and genomic single nucleotide variants. But it remains possible that other, cell-specific events, including nutritional or circadian factors, might contribute to the differences noted. In addition, the current findings show a restricted subset of shared RNA editing targets between intestine and liver. Other work showed 25% overlap in RNA editing targets in mouse liver and adipose (Lagarrigue 2013), while another study found approximately 53 to 61% concordance in RNA editing (overwhelmingly A-to-I) in seven mouse tissues (including brain and liver but not small intestine) (Danecek 2012).

Nevertheless, among those studies and in the present report, there was conservation in the editing sites identified within each target. The demonstration of fewer C-to-U editing targets in the liver (27) compared to small intestine (372), as well as the reduced range of hepatic (<45%) versus intestinal (approximately 85%) editing efficiency, further emphasize tissue-specific requirements for target selection and cytidine deamination by the hepatic editing machinery. In keeping with this suggestion, only a single edited site was detected for apoB RNA editing in both nuclear and cytosolic hepatic RNA, compared to eight additional sites in intestinal apoB.

Examination of nuclear and cytoplasmic RNA targets revealed unanticipated results. We found that nuclear apoB RNA from WT intestine (but not liver) exhibited extensive C-to-U editing at two exonic sites upstream of the canonical site 6666, one of which (6583) introduces a threonine to isoleucine coding change. There were additional RNA editing sites in nuclear apoB RNA, predominantly 3′ of the canonical site, which were detectable at much lower levels in cytoplasmic RNA. These findings suggest nuclear transcriptomes are relatively enriched with C-to-U edited targets, as suggested recently for A-to-I RNA editing (Chen 2013). Among the possibilities to account for the observed differences in nuclear versus cytoplasmic distribution and efficiency of apoB RNA editing, it is tempting to speculate that nuclear apoB transcripts edited at the canonical site may be preferentially exported to the cytoplasm and/or that C-to-U RNA editing influences nucleo-cytoplasmic transport of apoB RNA in a site-specific manner. These possibilities will require formal evaluation in future studies. In this regard, it is worth noting that both Apobec-1 and its RNA binding cofactor ACF have been shown to shuttle between nuclear and cytoplasmic compartments (Chester 2003, Blanc 2001). In addition, it should be emphasized that the physiological relevance of compartmentalization of editing targets remains unresolved, with some studies showing A-to-I edited RNAs to be retained in the nucleus (Zhang 2001) while others found A-to-I edited mRNAs preferentially distributed in cytoplasmic translating polysome fractions (Hundley 2008).

The finding that editing sites were concentrated in 3′ UTRs suggests a regulatory role in the transport, stability, translation or other function of these targeted RNAs. Elimination of A-to-I RNA editing in ADAR-null flies resulted in upregulation of hundreds of RNAs (St Laurent 2013). By contrast, we found that mRNA abundance of the majority of edited mRNAs was either unchanged or decreased in Apobec-1-deficient mouse intestine. In addition, while other work has demonstrated ADAR-mediated editing of both miRNAs and mRNAs (Borchert 2009), we found no evidence for C-to-U editing of miRNAs from WT small intestine (data not shown). That said, it is possible that either Apobec-1 binding and/or editing affect the stability of the target mRNA-polysome complexes and selectively modulates translational efficiency. For example, RNAs bound to a subset of yeast RNA binding proteins interact with RNA recognition elements located in the 3′ UTR that, in turn, regulate translation (Hogan 2008). It is worth noting that the 26 differentially expressed proteins encoded by Apobec-1 RNA targets represent approximately 3% of the 893 differentially expressed proteins (Fig. 3-4A). By contrast, the 54 Apobec-1 C-to-U RNA editing targets identified by RNA-seq represent approximately 1.7% of the total proteins identified in our proteomic survey (Materials and methods), suggesting a two-fold enrichment of proteins encoded by Apobec-1 RNA targets within the pool of differentially expressed proteins (P-value 0.0163).

The search to understand the functional implications of RNA editing led to another intriguing observation: a subset of 10 targets exhibited downregulation of both mRNA and protein abundance while a single edited target, Ido1, was upregulated in Apobec-1-deficient intestine. We confirmed that another highly edited intestinal Apobec-1-dependent target, Cd36, also showed concordant decreases in RNA and protein abundance in Apobec-1 null mice. The functional implications of these changes will require formal confirmation but targets including Rfk, Tes, Pde5a, Yme1l1 and Ido1 have been implicated in tumorigenesis (Hirano 2011, Segditsas 2008, Bianchini 2006, Tobias 2001, Cherayil 2009, Tinsley 2009). This possibility is intriguing in view of findings that Apobec-1 deletion attenuates the tumor burden in ApcMin/+ mice (Blanc 2007) while deficiency of Deadend1 (Dnd1), a paralog of ACF, increases intestinal polyposis susceptibility in ApcMin/+ mice (Zechel 2013). Among the down-regulated targets in Apobec-1-deficient intestine, Cyp3a11, Cyp2c65, Abcd3, Cyp4v3 and Pde5a are either directly or indirectly modulated by lipid mediators and it is possible that the changes observed are a secondary consequence of alterations in lipid flux rather than a direct effect of eliminating RNA editing (Tinsley 2009, den Bosch 2002, Makishima 2002, Norkina 2004). The consequences for intestinal lipid metabolism of the changes in the fatty acid translocase Cd36 (Drover 2005) are particularly intriguing and will be the focus of future investigation. Alternatively, and by analogy to events described with ADAR-mediated RNA editing, it is conceivable that the changes in protein expression in targets undergoing Apobec-1-dependent C-to-U RNA editing could reflect subtle protein-RNA interactions that influence polysome distribution and in turn modulate gene expression (St Laurent 2009). The current findings demonstrate that Apobec-1-dependent C-to-U RNA editing exerts broad functional effects in a tissue-specific manner, beyond its canonical target apoB and in most cases unrelated to a restricted role in chylomicron assembly.

Methods

Animals

All studies were performed using C57BL/6 from JAX (C57BL/6J) or Apobec-1-/- mice (both genders) backcrossed for >12 generations onto a C57BL/6 background. Apobec-1Int/O mice and intestinal Apobec-1 transgenic mice (Blanc 2012) were on a mixed background (C57BL/6 and 6xCBA). Apobec-1-/- mice were injected with 6 × 108 plaque-forming units of recombinant adenovirus encoding either β-galactosidase (Lac-Z) or rat Apobec-1 (ad-Apobec-1) resulting in hepatic Apobec-1 overexpression. Mice were 8 to 10 weeks old and fed an ad libitum chow diet. All animals were treated following National Institutes of Health guidelines and all protocols (#20130037) were approved by the Washington University Institutional Animal Care and Use Committee.

Accession numbers

RNA sequencing data from deep sequencing are available in the Gene Expression Omnibus under the accession number [GEO:GSE57910]. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (EMBL-EBL) via the PRIDE partner repository (Vizcaino 2014) with the dataset identifier PXD001007.

RNA-seq library

Total RNA was extracted from intestinal mucosa from WT, Apobec-1-/- and Apobec-1Int/O mice and from livers isolated from WT, Apobec-1-/-, Apobec-1-/- + ad-LacZ and Apobec-1-/- + ad-Apobec-1 mice (three mice per genotype), using TRIZol reagent (Invitrogen, Grand Island, NY, USA). DNAse-free RNAs were used for cDNA preparation as previously described (Mortazavi 2008). Pooled RNA (10 μg) was subjected to oligo(dT) selection. After chemical fragmentation RNA was reverse transcribed using random hexamer and sequencing adapters (Illumina) ligated to each end of double-stranded cDNA. The fragments were then PCR-amplified using linker-specific primers (Illumina). All libraries were diluted to 10 nM and an equal volume of each sample was combined to form the final sequencing pool that was run on an Illumina HiSeq2000.

RNA-seq analysis

RNA-seq reads for each genotype were mapped to the mouse reference genome (NCBI37/mm9) and single nucleotide variants were called using a modified version from (Park 2012). Reads from each sample were mapped with Bowtie version 0.12.8, with at most three mismatches, suppressing all alignments for a particular read if more than one reportable alignment exist for it, and using only those alignments that fell into the best stratum The alignment files were sorted and indexed using Samtools version 0.1.18 (Li 2011). Variants were called using the mpileup command. We called a single nucleotide variant when at least three independent reads support a non-reference variant, and the variant is present at a minimum frequency of 10% with minimum coverage of 10 reads and is supported by at least one read per strand. Sites were removed if they had three or more different observed nucleotide variants and a minimum frequency greater than 1.5%. Editing candidate sites were required to have no more than a 5% variant frequency in Apobec-1 knockout genotypes. Known SNPs from dbSNP128 that were not annotated as based on cDNA and sites lying outside of the 5′ and 3′ gene boundaries were set aside, and the remaining sites were corrected for strand sense. These sites were then annotated using ANNOVAR (Wang 2010) with a splicing threshold of 5.

Sanger sequencing validation of Apobec-1-dependent editing sites

Genomic DNA and total RNA were isolated from intestine and liver of WT, Apobec-1-/-, Apobec-1Int/O and Apobec-1-/- + ad-Apobec-1 mice. Genomic DNA was prepared as follows: 100 ng of liver tissue was incubated at 55°C overnight in 600 µl cell lysis solution (QIAGEN, Valencia, CA, USA). After protein removal, DNA was precipitated and resuspended. Total RNA was TRIzol-extracted and subjected to cDNA synthesis using random hexamers and MultiScribe Reverse Transcriptase from High Capacity cDNA Reverse Transcription kit (Applied Biosystems, Foster city, CA, USA). Both isolated genomic DNA and cDNA were used as templates to amplify sequences containing RNA-seq-identified Apobec-1-dependent putative editing sites. PCR amplifications were performed using Pfultra II DNA polymerase (Agilent Technologies, Santa Clara, CA, USA). Quality-controlled PCR products were then cloned into pCR-Blunt II-TOPO vector (Invitrogen) following the manufacturer's recommendations. Twenty individual clones were sequenced using Applied Biosystems BigDye terminator mix version 3.1. C-to-U calls are referred to as true positives when validated by Sanger sequencing. By contrast, C-to-U calls made from RNA-seq but not verified by Sanger sequencing are referred to as false positives.

Nuclear-cytoplasmic RNA isolation

Intestines were harvested from three to four mice per genotype. Preparation of nuclear and cytoplasmic RNAs was undertaken as described (Holden 2009). Briefly, scraped intestinal mucosa was resuspended in ice-cold buffer B (10 mM tris pH 7.4, 140 mM NaCl, 1.5 mM MgCl2, 0.5% NP-40, 1 mM DTT, 20 units/µl RNAse inhibitor (Promega Madison, WI, USA) and 1× protease inhibitor) homogenized and centrifuged at 7,000 g for 10 minutes at 4°C. Supernatant was saved as cytoplasmic fraction. Nuclear pellets were resuspended in 2 volumes of buffer B supplemented with one- tenth volume detergent (3.5% sodium deoxycholate (w/v) and 6.6% Tween 20 (v/v)) incubated for 30 minutes at 4°C and centrifuged at 1,000 g for 5 minutes. Supernatant was combined with the previous cytoplasmic fraction and nuclear pellet was rinsed once in buffer B. Cytoplasmic and nuclear RNAs were extracted using TRiZol (Invitrogen) following the manufacturer's protocol, treated with DNAse (Ambion Life Technology, Grand Island, NY, USA) and subjected to cDNA synthesis as described above. Targets of interest (apoB, Usp25 and ATP6ap2) were then PCR amplified using specific primers. PCR products were cloned and sequenced as described above.

Protein extraction and western blotting

Scraped mucosa was homogenized in tissue lysis buffer containing 20 mM Tris (pH 8), 0.15 M NaCl, 2 mM EDTA, 1 mM sodium vanadate, 0.1 M sodium fluoride, 50 mM β-glycerophosphate, 5% glycerol, 2× protease inhibitor (Roche Applied Science Indianapolis, IN, USA), 1% Triton, and 0.1% SDS. Aliquots of homogenate (60 µg protein) were resolved by SDS-PAGE, transferred to PVDF membrane, and probed with goat anti-CD36 antibody (AF2519, R&D Minneapolis, MN, USA), mouse anti-IDO1 (BioLegend, San Diego, CA, USA). Equal loading was verified using a rabbit anti-α-actin antibody (Sigma-Aldrich St. Louis, MO, USA).

Polysome isolation

Each polysome isolation used three to four mice with two to five isolations per genotype. Intestinal mucosa was prepared in ice-cold phosphate-buffered saline supplemented with 100 μg/ml cyclohexamine (Sigma, St Louis, MO, USA) was incubated in 1 ml lysis buffer (25 mM Tris-HCl pH 7.5, 250 mM NaCl, 5 mM MgCl2, 0.5 mM PMSF, 200 μg/ml heparin (Sigma), 5 mM dithiothreitol, 20 U/ml RNAsin, 100 μg/ml cycloheximide, 1% Triton X-100, 1× protease inhibitor). Scraped mucosa was homogenized and centrifuged at 10,000 g for 10 minutes at 4°C. The supernatant was loaded onto a 10 to 50% sucrose gradient and centrifuged at 40,000 rpm for 2.25 h at 4°C using an SWT41i rotor (Beckman Brea, CA, USA). Fractions (900 μl) were collected from the bottom of the gradient and 260 nm absorbance monitored by spectrophotometry. RNA was phenol/chloroform extracted from each fraction, precipitated, resuspended in 20 μl H2O and used for cDNA synthesis followed by PCR amplification of specific targets (apoB, Usp25, Atp6ap2). PCR products were cloned and sequenced as described above.

In vitro editing analysis by poisoned primer extension

Total hepatic RNA was isolated from Apobec-1-/- mice and treated with DNA-free reagent (Ambion). Resulting RNA (1 μg) was incubated for 3 h at 30°C with variable amount of hepatic S100 extract prepared from either WT or Apobec-1-/- mice liver (Blanc 2011). Following incubation with S100 extracts, the RNA was phenol/chloroform extracted, precipitated and resuspended in cDNA synthesis reaction mix (High Capacity cDNA Reverse Transcription kit (Applied Biosystems). Single-stranded DNA was then subjected to PCR amplification using primers specific for a Sanger-validated Apobec-1-dependent RNA target followed by poisoned primer extension using γ-ATP 5′ end-labeled primer annealing approximately three to six nucleotides downstream of the identified editing site as previously described (Blanc 2011). Extension products were separated by electrophoresis on a 7 M urea-acrylamide gel and analyzed by autoradiography.

Proteomics analysis

Total proteins were isolated from three WT and three Apobec-1-/- intestine samples using a buffer containing 2% SDS, 30 mM Tris pH 8, supplemented with protease inhibitors (Complete EDTA-free, Roche), phosphatase inhibitors (PhosStop, Roche) and benzonase (25 U/μl, Sigma). Proteins were methanol/chloroform precipitated and resuspended in urea/thiourea buffer (6 M/2 M, 30 mM Tris, pH 8). Protein concentration was estimated using Bradford. Unlabeled samples were mixed with a lys6-labeled SILAC standard (analogously extracted from intestine from lys6-labeled mice; Silantes GmbH, Munich, Germany) at a ratio of 1:1. Samples were in-solution digested (Liberski 2013) using Lys-C (Wako Richmond, VA, USA) only. Peptides (200 μg) were separated by in-solution isoelectric focusing (Offgel fractionator, Agilent) into 12 fractions over a pH range of 3 to 10. Fractionation was performed according to the manufacturer’s protocol with adaptations. Glycerol and ampholytes in the separation buffer were reduced to 0.3% (original, 6%) and 0.1% (1%), respectively. Peptides were focused for 20 kVhr and harvested including a well washed with 50 μl 50:49:1 methanol:MilliQ:TFA for 15 minutes. Fractionated peptides were dried down with a speedvac (Eppendorf Hauppauge, NY, USA) prior to desalting using C18 StageTips according to (Rappsilber 2003). The fractions obtained were individually submitted to liquid chromatography (LC) coupled to mass spectrometry (MS) (Liberski 2013). After trimming to avoid ampholyte interference with data analysis using RecalOffline (ThermoFisher Scientific Waltham, MA, USA), mass spectrometry data were analyzed using the MaxQuant suite of algorithms (version 1.3.0.5; Cox and Mann, 2008). The data were searched against the Mus musculus UniProtKB protein sequence database (as of 8 May 2013) consisting of 79,342 entries, including canonical and isoform sequences. Search parameters were set as follows. Lys-C was selected with a maximum of two missed cleavages. Precursor mass tolerance was set to 20 ppm for the first search and to 6 ppm for the main search. Oxidized methionines and amino-terminal protein acetylation were allowed as variable, carbamidomethylation as a fixed modification. The false discovery rate for peptide and protein identification was set to 1%. Minimum peptide length was set to 7 with no maximum. Peptide identification by chromatography alignment and ID transfer ('match between runs') was enabled and led to identification of 3,210 proteins. Differentially expressed genes were identified by t-test (significance cutoff of 0.1) in R, a language and environment for statistical computing and graphics (The R Project for Statistical Computing).

Apobec-1-dependent editing sites: analysis of flanking sequence features

Analysis of bases flanking the editing sites was performed by aligning 10 nucleotides surrounding the editing sites (5 nucleotides immediately upstream and 5 nucleotides immediately downstream). Frequency plots and logos were generated using the WebLogo application (WebLogo, Crooks 2004). Identification of consensus mooring sequence was performed by aligning 100 nucleotides surrounding the editing sites and looking for the consensus mooring sequence previously identified (Rosenberg 2011b). To determine the AU content of the targeted 3′ UTRs, the average AU content of both the full length 3′ UTR and a 101-bp window surrounding each editing site were compared to the distribution of 100,000 random sets of 101-bp windows in 3′ UTRs and whole 3′ UTRs of equivalent size.

Gene expression analysis

Differential gene expression analysis was performed using the Tuxedo suite of tools (Trapnell 2012). RNA-seq reads were mapped onto the transcriptome (mm9 UCSC knownGene) using Bowtie version 2.0.0b7 (Langmead 2009) and TopHat version 2.0.5 (Langmead 2009). Differentially expressed genes were called using fragments per kilobase of exon per million fragments mapped (FPKM) and reported as a measure of relative transcript abundance using Cufflinks version 2.0.2 (Trapnell 2010).

Statistical analysis

Degree of enrichment of the Apobec-1 targets was represented by the difference in hypergeometric distribution using one-tailed Fisher's exact test. Correlation between editing frequency and fold protein expression is reflected by Spearman's rho (ρ) rank correlation coefficient. Statistical significance was set at a P value <0.05. All analyses were performed using Graphpad Prism 4.0 (GraphPad Software, Inc. La Jolla, CA, USA).

Figure 3-1.

RNA-seq identification of Apobec-1-dependent RNA-editing targets. (A) RNA-seq procedure and analyses of 3' UTR C-to-U calls identified in wild-type (WT) small intestine and liver. Five murine lines with distinctive Apobec-1 expression profiles were used for intestinal transcriptome analysis. Apobec-1-/- mice exhibit no intestinal or hepatic apoB RNA editing. Apobec-1Int/+, intestine-specific Apobec-1 transgenic mice [15], were crossed with Apobec-1-/- mice generating Apobec-1Int/OHi and Apobec-1Int/OLo transgenic mice, with high (Hi) and low (Lo) levels of Apobec-1 expression [15]. WT hepatic transcriptomes were compared to Apobec-1-/- mice. Apobec-1-/- + ad-Apobec-1 or ad-LacZ indicates Apobec-1-/- mice injected with adenovirus expressing Apobec-1 or Lac Z. Overexpression of Apobec-1 in the liver restores apoB RNA editing. Uniquely mapped reads were aligned to the C57BL/6 mouse genome (NCBI37/mm9) containing 23,334 reference genes. To minimize false positive calls, sites identified in both WT and Apobec-1-/- mice, known SNPs from dbSNP128 and sites lying outside the gene boundaries were excluded. The remaining sites were corrected for strand sense and qualified when supported by 3 minimum non-identical reads, a minimum frequency of 10% with a minimum coverage of 10 reads. An arbitrary cutoff of 30% editing frequency was set to sequence-validate calls identified in the intestine. Due to the low number of calls identified in WT liver, all calls (27) were sequenced. (B) Numbers of C-to-U editing events and RNAs Sanger-sequence-validated (SSV). Blue circles represent the 56 3' UTR C-to-U calls identified in 54 WT intestine RNAs. Red circles show the 22 validated C-to-U sites identified in 17 hepatic RNAs. The shaded regions represent the 11 C-to-U sites or RNAs identified in both small intestine and liver. Forty-five sites were specific to the intestine, 11 were liver-specific.

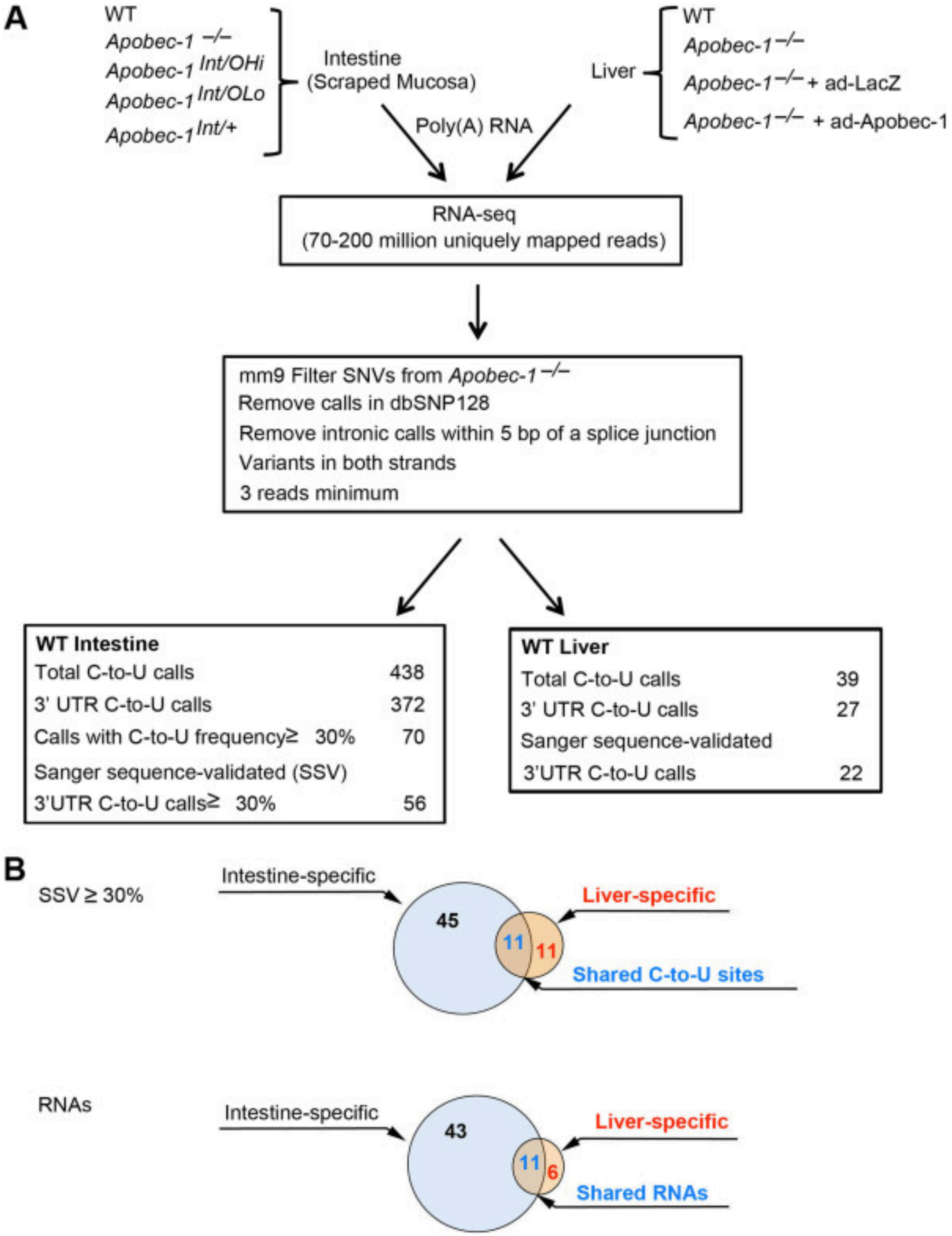

A
WT
Apobec-1 −/−
Apobec-1 Int/OHi
Apobec-1 Int/OLo
Apobec-1 Int/+
Intestine
(Scraped Mucosa)
Liver
WT
Apobec-1 −/−
Apobec-1 −/− + ad-LacZ
Apobec-1 −/− + ad-Apobec-1
Poly(A) RNA
RNA-seq
(70-200 million uniquely mapped reads)
mm9 Filter SNVs from Apobec-1 −/−
Remove calls in dbSNP128
Remove intronic calls within 5 bp of a splice junction
Variants in both strands
3 reads minimum
WT Intestine
Total C-to-U calls 438
3' UTR C-to-U calls 372
Calls with C-to-U frequency≥ 30% 70
Sanger sequence-validated (SSV)
3'UTR C-to-U calls≥ 30% 56
WT Liver
Total C-to-U calls 39
3' UTR C-to-U calls 27
Sanger sequence-validated
3'UTR C-to-U calls 22
B
SSV ≥ 30%
Intestine-specific
45
11
11
Liver-specific
Shared C-to-U sites
RNAs
Intestine-specific
43
11
6
Liver-specific
Shared RNAs

Figure 3-2.

In vitro editing assay of 3' UTR targets. Total hepatic RNA from Apobec-1-/- mice was incubated with increasing amounts of WT hepatic S100 extract. RNA was used for cDNA synthesis followed by PCR amplification of Apobec-1 3′ UTR targets using specific targets. (A) Endogenous Dpyd RNA editing of cytidine 119134696 was determined by poisoned primer extension. The relative mobility of the unedited (C 4696) and edited product (U 4690) is indicated to the right. Vertically is shown the sequence surrounding the editing site. The targeted cytidine is indicated in red. Upon editing, the primer extension reaction proceeds until the next C (represented in green). The 32P-labeled primer is shown in blue. (B) Endogenous Tmbim6 RNA editing of cytidine 99239051. Total hepatic RNA from Apobec-1-/- mice was incubated with recombinant Apobec-1 and ACF or with increasing amounts of hepatic WT S100 extract. C-to-U editing of cytidine 9051 was determined by poisoned primer extension. To the right is shown the sequence surrounding the editing site. The edited cytidine (9051) is shown in red. Cytidine 9043 also appears to be targeted, resulting in an extension product terminating at cytidine 9035.

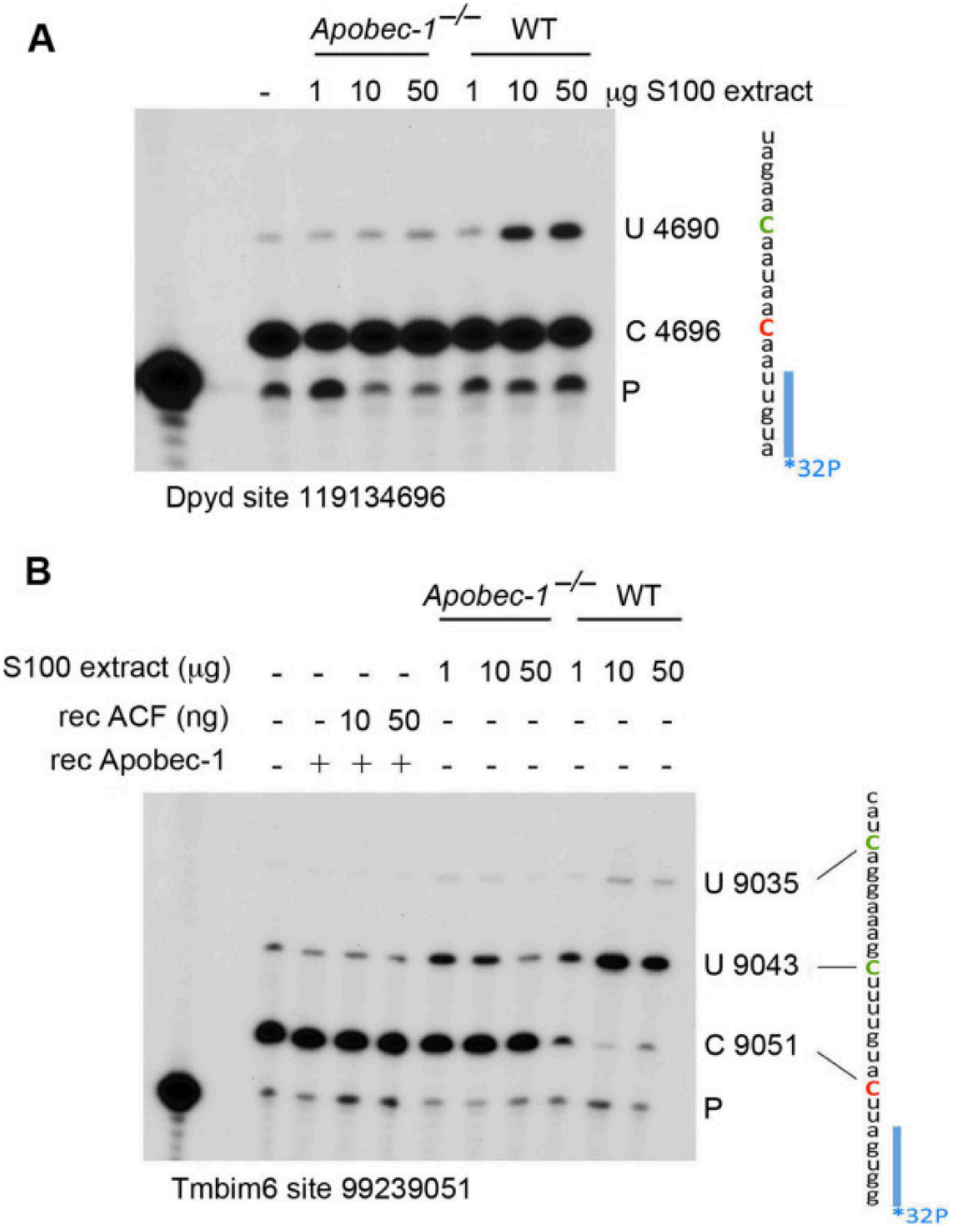

A
Apobec-1−/−
WT
- 1 10 50 1 10 50 µg S100 extract
U 4690
C 4696
P
u a g a a C a a u a a C a a u u g u a
*32P
Dpyd site 119134696
B
Apobec-1−/−
WT
S100 extract (µg) - - - - 1 10 50 1 10 50
rec ACF (ng) - - 10 50 - - - - - -
rec Apobec-1 - + + + - - - - - -
U 9035
U 9043
C 9051
P
c a u C a g g a a a g C u u u u g u a C u u a g u g g
*32P
Tmbim6 site 99239051

Figure 3-3.

Nucleo-cytoplasmic distribution of Apobec-1-dependent mRNA editing targets. (A,B) Distribution of WT small intestine (A) and hepatic (B) edited apoB RNA. A 738 bp amplicon (nucleotides 6,508 to 7,246) from nuclear and cytoplasmic apoB mRNA was cloned and sequenced. Twenty-two clones from each subcellular fraction (from three independent nuclear-cytoplasmic isolations) were analyzed. Left panel: graphic representation of percentage of edited clones in nuclear and cytoplasmic apoB RNA. Right panel: targeted cytidines identified in nuclear apoB RNA are indicated with green circles; cytidines identified in cytoplasmic apoB RNA are represented by blue circles. All cytidines are aligned with the nucleotide position to the left. (C) Nuclear-cytoplasmic distribution of intestinal Apobec-1 3′ UTR targets identified by RNA-seq and validated by Sanger sequencing. A 550 bp (ATP6ap2) and a 667 bp (Usp25) amplicon were generated from nuclear and cytoplasmic RNA and analyzed by sequencing 19 to 22 clones. For both ATP6Ap2 and Usp25 RNAs, the edited RNA is predominantly exported to the cytoplasm.

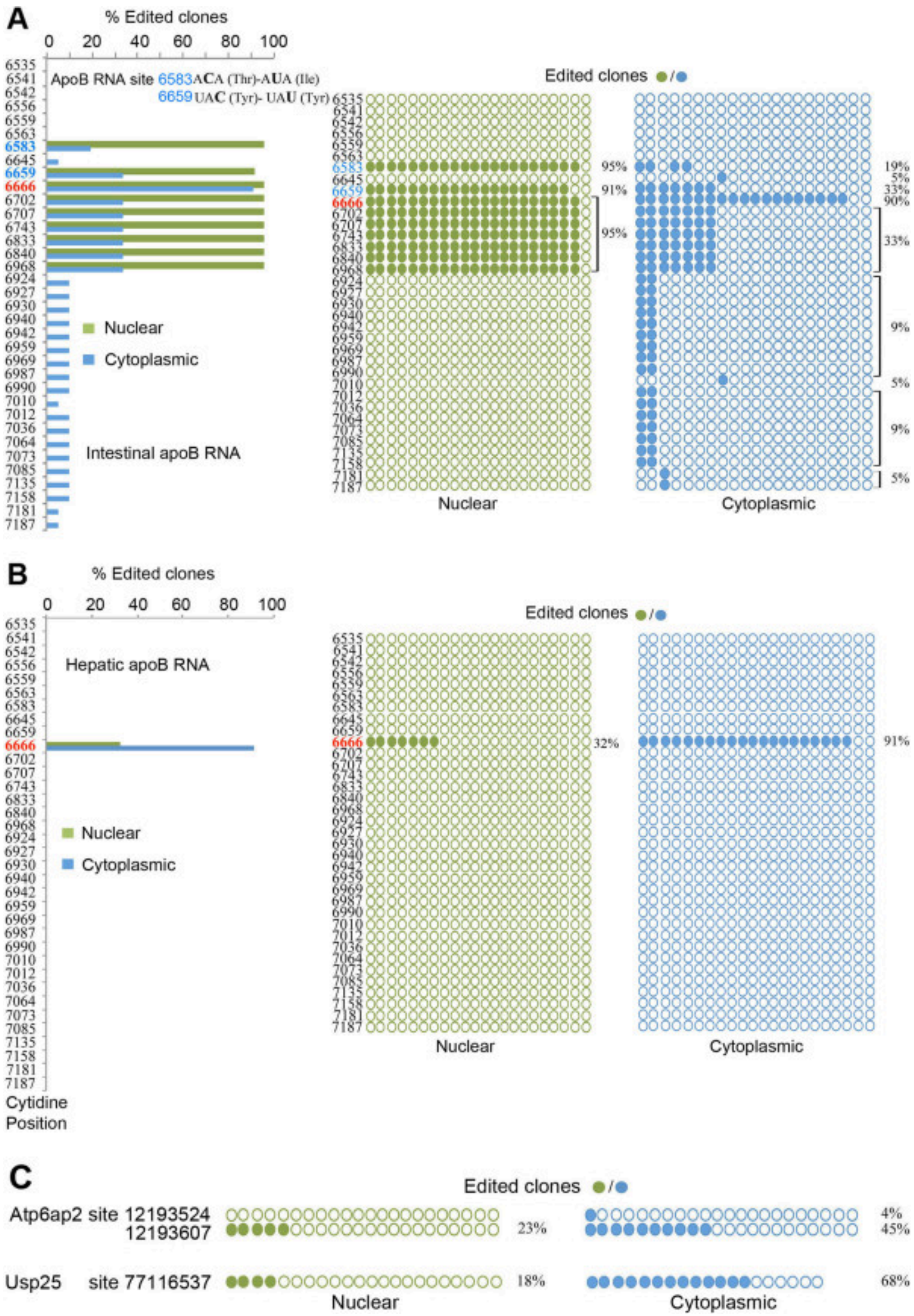

A
% Edited clones
0 20 40 60 80 100
ApoB RNA site 6583 ACA (Thr)-AUA (Ile)
6659 UAC (Tyr)- UAU (Tyr)
6535 6541 6542 6556 6559 6563 6583 6645 6659 6666 6702 6707 6743 6833 6840 6968 6924 6927 6930 6940 6942 6959 6969 6987 6990 7010 7012 7036 7064 7073 7085 7135 7158 7181 7187
Nuclear
Cytoplasmic
Intestinal apoB RNA
Edited clones ●/●
95%
91%
95%
19%
5%
33%
90%
33%
9%
5%
9%
5%
Nuclear
Cytoplasmic
B
% Edited clones
0 20 40 60 80 100
Hepatic apoB RNA
Nuclear
Cytoplasmic
Edited clones ●/●
32%
91%
Nuclear
Cytoplasmic
Cytidine Position
C
Edited clones ●/●
Atp6ap2 site 12193524
12193607
23%
4%
45%
Usp25 site 77116537
18%
68%
Nuclear
Cytoplasmic

Figure 3-4.

Apobec-1 editing targets in relation to RNA and protein expression. (A) Schematic representation of Apobec-1-dependent editing targets in relation to RNA and protein expression. Total proteins were extracted from WT Apobec-1-/- intestine and submitted for proteomic analysis (Materials and methods). The relative expression and editing status of the RNAs encoding the differentially expressed proteins were analyzed in parallel. Data comparison between WT and Apobec-1-/- data sets revealed 238 Apobec-1 RNA editing targets (blue circle), 335 differentially expressed RNAs (green circle) and 893 differentially expressed proteins (orange circle). Overlapping these three groups led to the identification of only 11 edited RNAs showing altered expression concomitant with altered protein level: 10 RNAs and proteins were up-regulated in WT (blue upward arrow) and one RNA and its protein product were down-regulated in WT compared to Apobec-1-/- (red downward arrow). (B) Reduced expression of Cd36 in intestinal extracts from Apobec-1-/- mice. Total cell lysates from three individual WT mice and four individual Apobec-1-/- animals were separated by SDS-PAGE probed with an anti-Cd36 and anti-α-actin antibody. * Indicates $p < 0.05$ for difference in protein abundance (C) Trend to increased expression of Ido1 protein expression in western blots of intestinal extracts from two individual Apobec-1-/- mice and two individual WT mice, normalized to α-actin as a loading control. Error bars represent mean ± se of relative protein abundance by genotype.

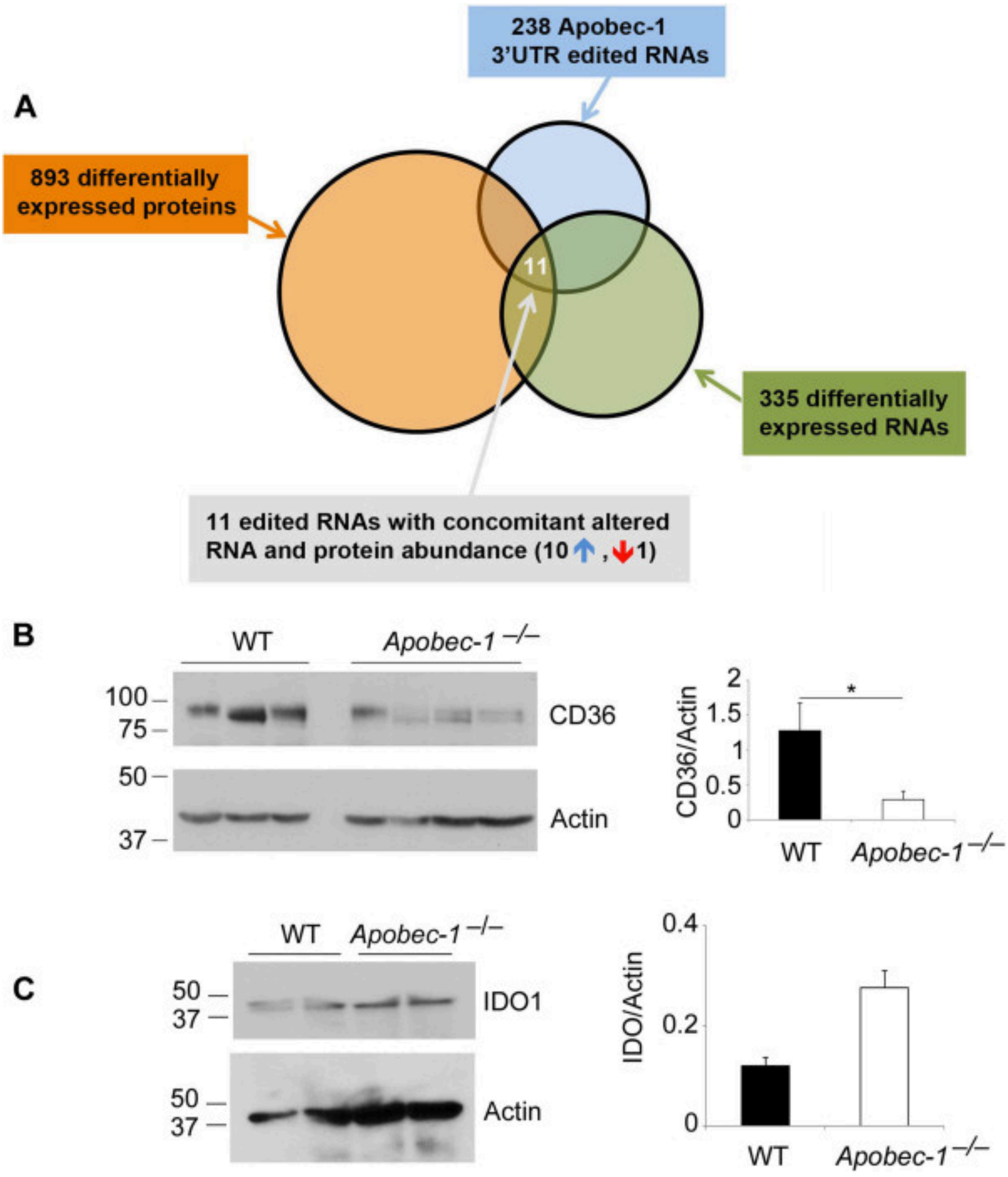
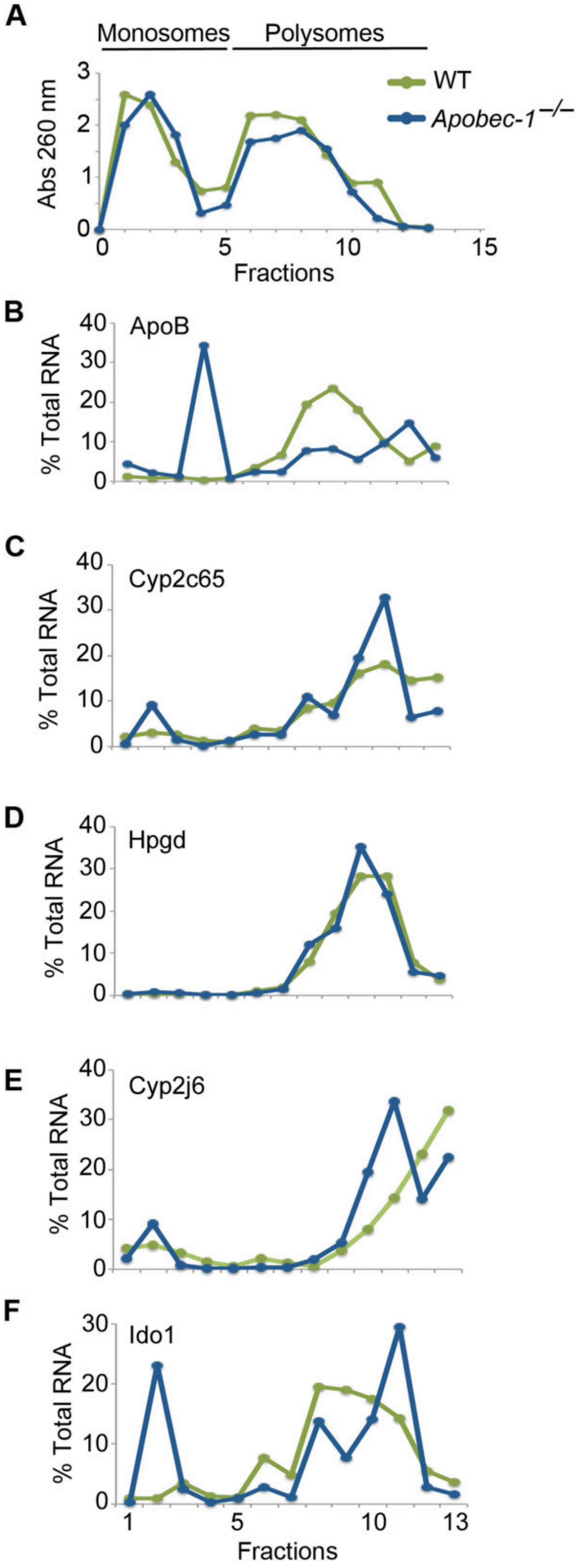

A
238 Apobec-1 3'UTR edited RNAs
893 differentially expressed proteins
11
335 differentially expressed RNAs
11 edited RNAs with concomitant altered RNA and protein abundance (10 ↑, ↓1)
B
WT
Apobec-1 −/−
100
75
50
37
CD36
Actin
CD36/Actin
2
1.5
1
0.5
0
*
WT
Apobec-1 −/−
C
WT
Apobec-1 −/−
50
37
IDO1
50
37
Actin
IDO/Actin
0.4
0.2
0
WT
Apobec-1 −/−

Figure 3-5.

Polysomal distribution of Apobec-1 mRNA editing targets. (A) Absorbance profile (A260) of fractions harvested from WT (green) and Apobec-1-/- (blue) mouse small intestine cytoplasmic extracts separated on sucrose gradients. Cytoplasmic extracts (two to five preparations) were prepared, each with three to four animals per genotype. (B) Sucrose gradient fractionation of apoB RNA from WT (green) and Apobec-1-/- small intestine cytoplasmic extracts (blue). ApoB RNA content in each fraction was analyzed in triplicate by quantitative PCR. Data were normalized to the expression of 18S mRNA and expressed as percentage of total apoB RNA. Data represent the mean of four to five separate isolations. (C-F) Polysomal distributions of Cyp2c65, Hpgd, Cyp2j6 and Ido1 RNAs, respectively, evaluated by quantitative PCR as described above. WT distribution (green), Apobec-1-/- distribution (blue).


A
Monosomes
Polysomes
WT
Apobec-1−/−
Abs 260 nm
3
2
1
0
0
5
10
15
Fractions
B
ApoB
C
Cyp2c65
D
Hpgd
E
Cyp2j6
F
Ido1
% Total RNA
40
30
20
10
0
1
5
10
13
Fractions

Chapter 4

Analysis of canonical A-to-I RNA editing in Drosophila using ICE-seq

Abstract

The predominant form of RNA editing in animals is the enzymatic conversion of adenosine to inosine that is sequenced in cDNA as guanine. While RNA editing should be identifiable from RNA-seq data alone, genomic SNPs as well as sequencing and mapping errors result in a high false-positive rate. We used Inosine Chemical Erasing (ICE) with deep sequencing (ICE-seq) to identify a high-quality set of RNA editing events in the head transcriptome of Drosophila melanogaster and six Drosophila yakuba wild-caught isofemale lines using the Illumina HiSeq and NextSeq platforms. We measured the level of conservation between D. melanogaster and D. yakuba and found that some of the best-studied editing sites in D. melanogaster are also edited in D. yakuba. We also detect a significant amount of species-specific editing in genes with different Gene Ontology enrichments that are expressed in both sets of head transcriptomes, which suggests that the biological function of RNA editing in these two Drosophila species is diverging.

Introduction

RNA editing is a post-transcriptional process where RNA bases are modified to other bases in transcripts. The most frequent type of RNA editing in metazoans is the deamination of adenosine to inosine (A-to-I) that is mediated by the ADAR family of enzymes. RNA editing can result in nonsynonymous protein coding substitutions, alternative splicing, nuclear retention of mRNA, or alterations of microRNA seed regions (Nishikura 2010). Mice lacking ADAR1 die embryonically at E11.5 (Wang et al. 2000; Wang et al. 2004) while mice lacking ADAR2 die shortly after birth due to seizures (Higuchi et al. 2000). In Drosophila, dADAR knockout causes paralysis, discoordination, and neurodegeneration (Palladino et al. 2000); however, this phenotype can be rescued with the expression of human ADAR2 (Keegan et al. 2011). Previous RNA editing studies in Drosophila focused mainly on reporting the locations and levels within Drosophila melanogaster, with two brief comparative analyses of editing conservation between Drosophila species (Chen et al. 2014; Ramaswami et al. 2013); in both cases, the focus of the manuscripts was not on in-depth comparative analysis. To the best of our knowledge no RNA editing analysis of variation between strains within same drosophila species has been published. A second type of RNA editing is the deamination of cytosine to uracil (C-to-U) that is mediated by APOBEC1 (Blanc et al. 2010); this form of editing occurs less frequently, is believed to be mammalian-specific and is found predominantly in a limited number of cell-types (Blanc et al. 2014). Inosines are sequenced as guanine sequence variants because of their ability to base-pair with cytosines during cDNA synthesis. Uracils are sequenced as thymine through reverse transcription of library generation. A-to-I (A->G) and C-to-U (C->T) RNA editing are referred to as canonical RNA editing because these types of modifications to RNA molecules have been well documented and the enzymes/co-enzymes involved have been characterized.

In recent years, there has been a surge in the number of RNA editing sites reported (Ramaswami et al. 2013; Ju et al. 2011; Bahn et al. 2012; Park et al. 2012), primarily due to high-throughput sequencing assays such as RNA-seq (Mortazavi et al. 2008). However, extensive problems with initial reports of non-canonical RNA editing (editing other than A-to-I or C-to-U) show that there are multiple challenges associated with calling RNA editing events from RNA-seq data, which include genomic SNPs, sequencing artifacts, and mapping artifacts (Li M et al. 2011; Lin et al. 2012; Pickrell et al. 2012; Kleinmn et al. 2012; Schrider et al. 2011). The consensus is that most non-canonical RNA editing calls can be attributed to technical artifacts in the data or in the analysis and can therefore be used as a measure of false positives to test the robustness of computational RNA editing identification methods.

Inosine Chemical Erasing (ICE) is a technique that has been shown to be an efficient way to validate RNA editing using Sanger sequencing (Sakurai et al. 2010). The ICE method exploits the ability of acrylonitrile to selectively react with inosine to form cyanoethylinosine, which sterically blocks reverse-transcription and effectively "erases" inosine-containing fragments from the sequencing results. This method has been coupled with deep sequencing (ICE-seq) in human (Sakurai et al 2014). We have applied ICE-seq to the adult female head transcriptome of Drosophila melanogaster as well as six wild-caught isofemale lines of Drosophila yakuba that have been resequenced (Rogers et al. 2014). We compared our results with other studies done in Drosophila melanogaster and assessed the extent of conservation of RNA editing between species and across strains within the same species using the Illumina HiSeq and NextSeq platforms.

Results

Integrating ICE with RNA-seq

We have used ICE-seq to detect and validate A-to-I RNA editing in two species of Drosophila (Fig. 4-1A). Total RNA is extracted from the tissue of interest and mRNA is isolated by poly-A selection, fragmented and split into two samples for parallel analysis: ICE and mock treatment. One half is treated with acrylonitrile (ICE treatment) while the other half is used as a control (mock treatment). Since RNA secondary structure is known to affect the efficiency of cyanoethylation (Yoshida et al. 1968), the acrylonitrile treatment is done after fragmentation to minimize this effect. The ICE and mock samples are reverse-transcribed; libraries are built from the cDNA and sequenced. RNA editing sites are determined by comparing the differences between the ICE treatment and the mock treatment. Cyanoethylinosine blocks reverse transcription and therefore true RNA editing sites have fewer G substitutions in the ICE treatment than in the mock treatment and coverage decreases in the ICE treatment over the editing sites.

Analysis strategy and software

There are three main types of single nucleotide variants in RNA sequences: genomic SNPs, sequencing/mapping artifacts, and RNA editing events. In theory, genomic SNPs and sequencing/mapping artifacts should have the same variant frequency and coverage in both the ICE and mock treatments. Conversely, only RNA editing events should have a decrease in variant frequency and coverage in the ICE treatment. Thus, ICE-seq is robust to false-positives introduced by genomic SNPs and sequencing/mapping errors and this can be leveraged to identify RNA editing events.

We devised a computational strategy to analyze ICE-seq data with RNA-seq data that maximizes our sensitivity and specificity. Briefly, sequencing reads are mapped to the reference genome and a set of preliminary variants is called and filtered to pass a minimal set of criteria (Fig. 4-1B). From there, the Irreproducible Discovery Rate (IDR) is calculated and a threshold at which the replicates stop agreeing with each other is determined (Fig. 4-1C). The IDR framework (Li Q et al. 2011) was originally developed to analyze replicated ChIP-seq data to determine reproducibility and find a threshold below which the ChIP-seq signal failed to be reproducible. We have used IDR in our analysis to set a similar threshold. The fraction of A-to-G plus T-to-C calls (the sum of the A-to-G and T-to-C calls divided by the total number of transition substitutions) is used as a rough estimate of specificity while the number of A-to-G plus T-to-C calls (the sum of the A-to-G and T-to-C) is used as an indicator of sensitivity. We include T-to-C calls because the libraries were built with an unstranded library protocol. Non-(A-to-G plus T-to-C) calls are typically used as an internal negative control for analyzing for RNA editing computational pipelines; however, we used the G-to-A plus C-to-T calls as a negative control because of a sequencing error bias in reads sequenced on the NextSeq. Note that ICE treatment should not affect C-to-U changes, so G-to-A plus C-to-T calls can be used as a negative control without consequence in Drosophila, since there is no documented evidence of C-to-U editing in flies. A full description of the computational method is described in the methods.

Comparison of four channel SBS with two channel SBS

The sequencing for this study was done using both the Illumina HiSeq and NextSeq. The HiSeq uses a four-channel sequencing by synthesis (SBS) where each base is sequenced with one independent color. The NextSeq v1 chemistry uses a two-channel SBS where two colors are used to determine four bases; two of the colors are assigned to one base each (red for C, green for T), a third base (A) is determined by the mixture of both colors, and the fourth base (G) is determined by the absence of both colors.

We compared reads from two RNA-seq libraries sequenced on both platforms. We found that the two types of sequencing methods have distinct error profiles for single nucleotide variants (Fig. 4-1D, 4-S1 Fig., 4-S2 Fig.). The HiSeq displayed a relatively uniform distribution of errors across the twelve possible types of single nucleotide variants (SNV). The NextSeq had a slightly higher overall error rate and had a noticeable error bias. The current two-channel SBS chemistry has a higher error rate for a (C or T)-to-A or A-to-(C or T) because the sequencer has difficulty distinguishing one color from two colors or vice-versa, presumably when neighboring clusters mix. Fortunately for RNA editing analysis, A-to-G errors have a relatively low error rate; a cluster with no color is very distinguishable from a cluster with two colors. Similarly, T-to-C errors are also made infrequently because a red cluster is distinguishable from a green cluster.

Identifying RNA editing events

RNA from the heads of female adult flies from the reference strain of Drosophila melanogaster was sequenced using RNA-seq and ICE-seq. We first examined RNA editing in the quiver locus (Fig. 4-2A). quiver is a gene that is known to be edited and encodes a membrane protein involved in the regulation of synaptic transmission. We find the four highly edited sites reported by the Drosophila modENCODE project (Graveley 2011) as well as additional secondary sites that are edited less frequently. As expected, there is a statistically significant reduction in guanine substitutions for these sites. Additionally, there is a sharp decrease in coverage in the ICE treatment when compared to the mock treatment (Fig. 4-2B). Similar changes in SNV frequency and coverage are seen for isolated RNA editing sites (4-S3 Fig.) but not for sites with sequencing errors (4-S4 Fig.) or sites with genomic SNPs or mapping artifacts (4-S5 Fig.). Coverage differences between ICE and mock treatments are proportional to editing level. This drop of coverage is compounded when editing sites cluster together. Coverage alone is a poor indicator of editing when a site is isolated and infrequently edited, because there is a relatively large variance in the signal-to-noise ratio under these conditions. However, our pipeline takes advantage of this in conjunction with the other indicators.

Global properties of edited sites

We identified 1238 RNA editing events in 384 genes for Drosophila melanogaster. We noticed that there is a power law distribution (4-S6 Fig.), which is often seen in many biological and non-biological systems, for the number of edits per gene (i.e. many genes with few editing sites and few genes with many editing sites). Most of the editing events occur in the introns (33%) and 3’ UTR (41%) (4-S7 Fig.), but a significant fraction are in coding exons (21%). The sites within coding regions have a larger editing rate than noncoding regions (4-S8 Fig.), which suggest a functional role for these RNA editing events. We find that the edited genes had a higher expression level compared to all genes (Fig. 4-2C). Although it would be tempting to concluded that RNA editing is associated with higher expression levels, this observation can also be explained by the possibility that RNA editing events are more detectible in genes that are more highly expressed.

There are 48 possible codon changes that could result from A-to-I RNA editing, 33 (69%) of which are non-synonymous (4-S9 Fig.) and would result in 16 possible unique amino acid changes. We identified 197 (78%) sites in 91 genes that resulted in a nonsynonymous substitution and two sites that created a stop-loss in Drosophila melanogaster (Fig. 4-2D). Of the two RNA editing sites that created a stop-loss change, one occurred in DopEcR (Chr3L:4369162), which is edited at a rate of more than 70%; however, this editing event is neutralized by genomic stop codons to prevent deleterious effects (STOP-S-STOP-I-STOP -> W-S-STOP-I-STOP). The second stop-loss change occurred in lawc (chrX:8303787) which is edited at a rate of less than 10%. This stop-loss results in a doubling of the protein sequence, which could be better tolerated because of a lower editing rate. We compared our editing calls with three other studies: Drosophila modENCODE (Graveley et al. 2011), deep sequencing using the now-defunct Helicos technology (St Laurent et al. 2013), and Nascent-seq (sequencing the nascent RNA) of adult D. melanogaster heads (Rodriguez et al. 2012). We find a 52% agreement (Fig. 4-2E) with other studies at the level of individual sites and a 76% agreement at the level of genes (Fig. 4-2F). We limited ourselves to sites in the nonrepetitive regions for stringency purposes. The gene ontology of edited genes (Fig. 4-2G) had an enrichment of neuronal related terms such as neurological systems process and ion channel complex as expected from the literature. We also see a non-uniform distribution of amino acid changes that result from RNA editing (4-S9 Fig.).

ADAR edited bases have been reported to have 5’ and 3’ neighbor preferences. ADAR1 targets have a 5′ neighbor preference of (A = U > C > G) but no apparent 3′ neighbor preference (Polson et al. 1994), whereas ADAR2 targets have a 5' neighbor preference of (A

≈ U > C = G) and also show a 3′ neighbor preference of (U = G > C = A) (Lehmann et al. 2000). Drosophila Adar is considered to be most similar to human ADAR2 (Keegan et al. 2011); however, we find that our edited sites display neighbor preferences that are slightly more similar to ADAR1 sites (4-S10 Fig.).

RNA editing comparison between Drosophila melanogaster and Drosophila yakuba

We sequenced the female adult head transcriptome of six isofemale lines of Drosophila yakuba (CY21B3, NY66-2, NY81, NY56, NY48, and CY08) that were derived from natural populations from Nairobi, Kenya and Nguti, Cameroon through nine or more generations of sibling mating (Rogers et al. 2014). We found 1226 RNA editing events in Drosophila yakuba and also found that a large number of sites are called as being edited reproducibly (4-S11 Fig., 4-S12 Fig.).

124 sites that showed strain-specific editing with a p-value of 0.01 or less. One intronic site in the potassium channel Shaker (chrX:16454932) showed moderate levels of editing in five strains and no detectable levels of editing in the fourth strain (4-S13 Fig.). RNA editing can also be abrogated with a non-A, non-G genomic SNP abolishing ADAR editing, as in the case for the insect AuTophaGy homolog Atg1 (chr3L: 12849242) where a cytosine has replaced an intronic adenosine (4-S14 Fig.).

We used UCSC's LiftOver tool to translate the Drosophila yakuba coordinates to the Drosophila melanogaster genome. We find that there is a 17% overlap between edited sites, a 25% overlap between edited genes, and a 53% overlap of GO terms for these edited genes (Fig. 4-3A). We see a similar degree of overlap for edited protein domains (4-S15 Fig.). 79 RNA editing GO terms are shared between the two species, including expected ones such as synaptic transmission and gated channel activity (Fig. 4-3A, Fig. 4-3B). In particular, genes that are shared between both species, but only detected as edited in one seem to fall into categories that lead to specific GO term enrichments, that include specific GO term relating to chemosensory behavior in D. melanogaster and multi-organism behavior in D. yakuba. Of the RNA editing sites that are shared between the two species, we see that the editing rates are most similar between the samples from the same species (Fig. 4-3C, 4-S16 Fig.). In addition, we see that some sites are diverging between the species and between strains (4-S17 Fig.). For example, an edited site in shi (chrX:15791887) is highly edited in Drosophila melanogaster and lowly edited in Drosophila yakuba. Similarly a site in rg (chrX:5140152) is highly edited in Drosophila yakuba, but lowly edited in Drosophila melanogaster. Lastly, the editing rate for sites that are species-specific tend to be slightly lower than the editing rate for sites that are shared (Fig. 4-3D).

Discussion

We have used ICE-seq in two species of Drosophila to distinguish true RNA editing events from other potential artifacts of RNA-seq sequencing and mapping. We compare the reproducibly detectable editing events in the female head transcriptome of Drosophila melanogaster and compare it to the corresponding female head transcriptomes in Drosophila yakuba and find both shared and species-specific genes with distinct gene ontology enrichments.

Integrating genome sequencing with RNA-seq is the most common method for finding RNA editing sites (Ju et al. 2011); however, since only a fraction of the genome is transcribed within a given sample, it is not a cost-efficient method. In addition, due to splicing events, RNA reads mappers work differently than DNA read mappers, which could lead to a large number of reproducible, false-positive calls near splice junctions as well as in paralogs and retrotransposed pseudogenes even when the underlying genome has been resequenced. Dozens of RNA-seq experiments from distinct individuals can be used to distinguish RNA editing events that occur frequently in a population from genomic SNPs that only occur within a subset of a population (Ramaswami et al. 2013); however, this would require sequencing at least one to two orders of magnitude more than other methods if done without an enrichment scheme and miss the fraction of RNA editing events that only occur within a subset of the population. A method using inosine specific cleavage has also been developed using glyoxal to protect guanosine from RNAse T1 treatment (Cattenoz et al. 2013); however, this method would have difficulty in distinguishing editing events that are highly clustered, which is often the case for many RNA editing sites. ICE-seq therefore offers advantages over other methods by using a direct RNA-to-RNA comparison where the only difference between the ICE treatment and the mock treatment is the nucleotide frequency of RNA edited sites. Furthermore, ICE-seq can be used to look at within-species variation and strain-specific editing without the need to resequence all individuals. ICE-seq is robust to incomplete sequence assembly and can be combined with de novo transcriptome assembly of its companion RNA-seq data to analyze samples from unsequenced genomes to get a wider census of the scope and extent of ADAR-mediated editing.

Our analysis of genes that are edited in the comparative transcriptomes of Drosophila melanogaster and Drosophila yakuba revealed that, in addition to the shared core set of synaptic genes that are edited in both genomes and are presumably under negative selection for editing, over 75% of edited sites and genes appear to be species specific. While false positives and false negatives may reduce concordance across species, it is far less likely that our method could have systematically missed numerous genes related to single fundamental processes such as "anatomical structure morphogenesis" that is specifically enriched in D. melanogaster or "courtship behavior" that is specifically enriched in D. yakuba. These results suggest that there is likely divergence in the extent to which molecular processes and pathways are affected by RNA-editing in the two species. Further, among the six strains of D. yakuba assayed, we identify 124 sites that have strain-specific editing, as well as hundreds of sites with population-level variation. Thus, populations house high levels of variation in RNA editing that in certain cases are observed to alter amino acid sequences. Thus, differences in RNA editing could lead to phenotypic variation among individuals in a single population. Whether this process of divergence in editing pattern at the gene level has an adaptive function on something as complex as behavior or is the consequence of drift remains to be determined, but it certainly represents an additional dimension of variation available for the evolution of each organism.

Methods

RNA preparation

Total RNA was obtained from two biological replicates of Drosophila melanogaster (y; cn, bw, sp strain) and each Drosophila yakuba strain. Approximately 50 fly heads were used to obtain total RNA (5 μg) using TRIzol reagent (Life) as per manufacturer's instruction. RNA integrity was evaluated on the BioAnalyzer (Aligent). Oligo(dT) selection was performed by using Dynal magnetic beads (Invitrogen) according to the manufacturer's protocol. The mRNA was fragmented by addition of 5x fragmentation buffer (200 mM Tris acetate, pH 8.2, 500 mM potassium acetate and 150 mM magnesium acetate), heated at 94 °C for 90 seconds, and then transferred to ice and run over a Sephadex-G50 column (USA Scientific) to remove the fragmentation ions.

Cyanoethylation of fragmented mRNA and cDNA preparation

The fragmented mRNA was incubated in 38 μl CE solution (50% (v/v) ethanol, 1.1 M triethylammoniumacetate (pH 8.6)) with 1.6 M acrylonitrile at 70 °C for 1 hr (ICE treatment). A parallel mock treatment was preformed without 1.6 M acrylonitrile. The RNA was purified with ethanol precipitation and reverse transcribed with random hexamers, added to prime first-strand reverse transcription according to the manufacturer's protocol (Invitrogen cDNA synthesis kit). After the first strand was synthesized, a custom second-strand synthesis buffer (Illumina) was added, and dNTPs, RNase H and Escherichia coli polymerase I were added to nick translate the second-strand synthesis for 2.5 h at 16 °C. The reaction was then cleaned up on a QiaQuick PCR column (Qiagen) and eluted in 30 μl EB buffer (Qiagen).

Sequencing and mapping of reads

HiSeq reads were sequenced as single-ended 50-mers and NextSeq reads were sequenced as 86-mers. The reads were mapped onto the Drosophila melanogaster genome (BDGP R5/dm3) or Drosophila yakuba genome (WUGSC 7.1/droYak2) using TopHat version 2.0.6 (Kim et al. 2013) / Bowtie version 2.0.2 (Langmead et al. 2012) using the following options: --read-realign-edit-dist 0 --read-edit-dist 4 --read-mismatches 4 -G <gene annotation file> -x 1 -g 1. Unmapped reads were allowed to remap using bsmap version 2.74 (Xi et al. 2009) using the following options: -v 3 -M GA. RNA-seq and ICE-seq reads are deposited in GEO with accession GSE60851.

RNA editing Analysis Pipeline

With the alignments, we obtained an initial set of candidate sites using the mpileup function of samtools (Li et al. 2009) version 0.1.19 using the options -d 100000 -f <genome file>. Using a python script, we filtered the candidate sites by requiring variants to be seen in at least two RNA-seq replicates and a minimum editing rate of 5%. We also required the following for at least one replicate: a p-value less than 0.01 for the variant frequency between the ICE and RNA samples, less coverage in the ICE-seq than the RNA-seq, and a lower variant frequency in the ICE-seq than the RNA-seq. The Storer-Kim method (Storer et al. 1990; Wilcox 2003) was used to determine the p-value because it is considered one of the better statistical tests, in terms of power and Type I error, particularly for small and/or unequal sample sizes. The p-value was computed using the R script written by (Wilcox 2003). We limited ourselves to A-to-G, T-to-C, G-to-A, and C-to-T variants because of the error bias in the NextSeq. The A-to-G and T-to-C variants represented our potential set of RNA editing sites (T-to-C variants were included because we used an unstranded library protocol). The G-to-A and C-to-T variants were used as a negative control because these types of variants had a comparable error profile, the ICE treatment should not affect C-to-U RNA editing, and C-to-U RNA editing has not been described in Drosophila. We computed the IDR using the R script that was written by (Li Q et al. 2011) for each site using the change in the number of G's from RNA-seq to ICE-seq as the signal value. The IDR method works by comparing two replicates. For Drosophila melanogaster, the two paired ICE/mock replicates were used. For Drosophila yakuba, in lieu of biological replicates, a pairwise comparison between all strains was used to perform the IDR analysis. The sites were sorted by IDR and a 0.85 fraction of A-to-G/T-to-C sites was used to determine a cutoff threshold. Of the remaining sites, we removed sites outside gene boundaries, sites that were not A-to-G in the direction of transcription, and sites in repetitive regions.

Comparison of RNA-seq editing calls in two sequencing platforms

We sequenced two biological replicates of RNA-seq libraries derived from Drosophila melanogaster heads. The libraries were sequenced on a HiSeq 2500 and NextSeq 500. The resulting reads were normalized for read number and read length and mapped onto the dm3 assembly using tophat2 (Kim et al. 2013) version 2.0.6. Variants were called using samtools mpileup. The samtools mpileup output provides a list of bases over each position in the genome. For the analysis of sequencing errors, any site with at least 1 read supporting a variant was considered for downstream analyses in order to enrich for sequencing errors. We reasoned that true biological variants and mapping errors were shared between the two sequencers while sequencing errors should be in the outersects. Variants that were sequencer specific were determined to be sequencing errors. The variants were normalized to errors/1000 bases.

Gene ontology analysis

Gene ontology analysis was done using WebGestalt (Wang et al. 2013). A significance level of 0.01 adjusted with the Benjamini-Hochberg procedure was used. A minimum of 5 genes per category was required. Genes expressed more than 1 FPKM was used as a background list for each species. Drosophila melanogaster orthologs were used for Drosophila yakuba genes.

Figure 4-1.

ICE-seq workflow and error profile of two types of SBS chemistry.

(A) The ICE-seq workflow. Isolated mRNA is fragmented and is split into two parallel samples: ICE and mock treatment. The ICE sample is treated with acrylonitrile while the mock sample is used as a control. Both samples are reverse-transcribed, sequenced, and mapped. SNVs are called from the alignments and a stringent set of filters and statistical analyses are preformed to obtain a final list of RNA editing calls. (B) The computational method. Two replicates of paired ICE-seq/RNA-seq datasets were used. Each variant had to be seen in two replicates with an editing rate greater than 5%. For at least one replicate, the p-value is required to be less than 0.01, the coverage in the ICE sample is required to be less than the RNA sample, and the editing rate in the ICE sample is required to be less than the RNA sample. (C) Irreproducible discovery rate of ICE-seq calls. Candidate RNA editing sites were sorted by IDR. The IDR (blue) and the cumulative true positive rate (purple) are plotted. The true positive rate was measured as the fraction of potential editing sites (the sum of the A-to-G and T-to-C divided by the total number of transition substitutions). The cutoff threshold is represented with a dashed black line. (D) Error profile of HiSeq and NextSeq. Sequence variants that were specific for either HiSeq (red) or NextSeq (blue) platforms were obtained and normalized per 1000 bases.

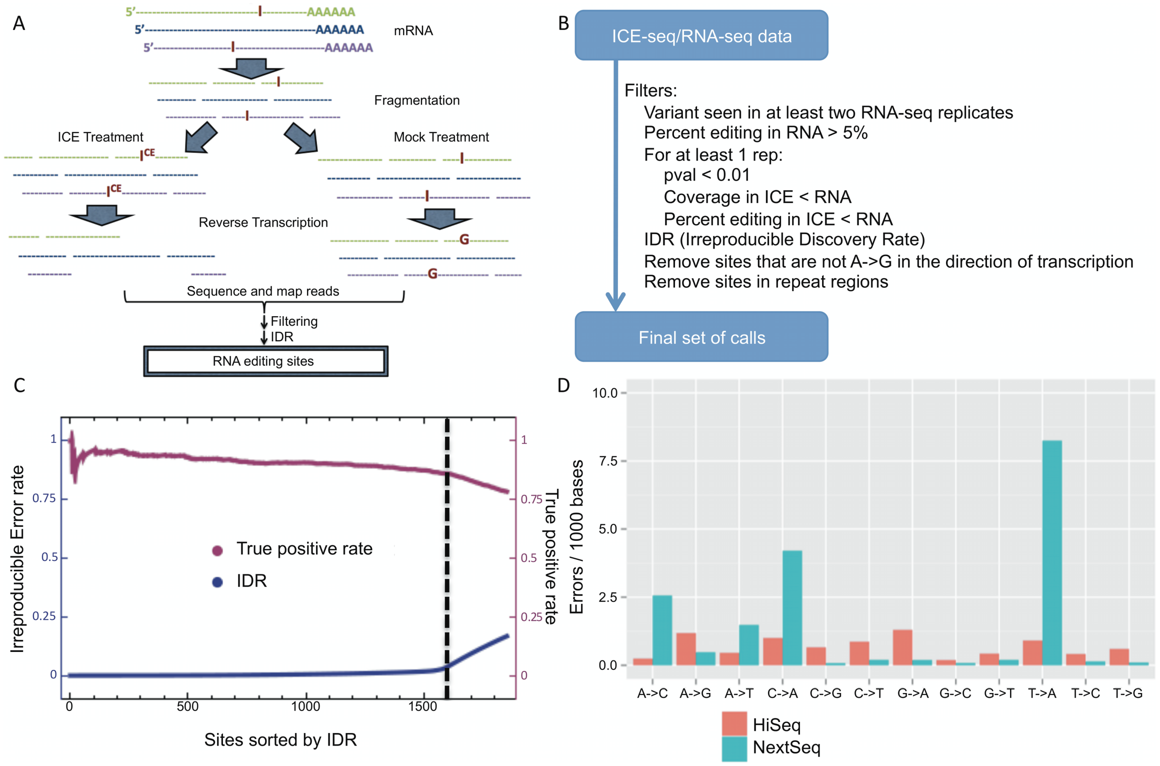

A
5'
AAAAAA
mRNA
Fragmentation
ICE Treatment
Mock Treatment
ICE
I
Reverse Transcription
G
Sequence and map reads
Filtering
IDR
RNA editing sites
B
ICE-seq/RNA-seq data
Filters:
Variant seen in at least two RNA-seq replicates
Percent editing in RNA > 5%
For at least 1 rep:
pval < 0.01
Coverage in ICE < RNA
Percent editing in ICE < RNA
IDR (Irreproducible Discovery Rate)
Remove sites that are not A->G in the direction of transcription
Remove sites in repeat regions
Final set of calls
C
Irreproducible Error rate
True positive rate
● True positive rate
● IDR
1
0.75
0.5
0.25
0
0
500
1000
1500
Sites sorted by IDR
D
Errors / 1000 bases
10.0
7.5
5.0
2.5
0.0
A->C
A->G
A->T
C->A
C->G
C->T
G->A
G->C
G->T
T->A
T->C
T->G
🟥 HiSeq
🟦 NextSeq

Figure 4-2.

ICE-seq in Drosophila Melanogaster.

(A) ICE and mock treatment of the qvr locus. The change in variant frequency (right) and the read coverage (left) are shown. One replicate of RNA-seq (mock) and two replicates of the ICE-seq (acrylonitrile treatment) are indicated. The numbers over the bar chart represents the numbers of A's and G's that were observed. (B) 401-bp window centered at editing event in the coding region of qvr. The region in (A) is marked within the black box. The location of the four editing sites is indicated with four orange vertical lines. (C) Gene expression of edited genes. The gene expression for edited genes (left) and all genes (right) in Drosophila melanogaster is represented in Log2(FPKM) using violin plots. The horizontal thickness of the plot represents the frequency for the expression level. A lower bound cutoff of 1 FPKM was used. (D) Pie chart of editing consequences for sites in coding exons. The type of SNV was determined for the 254 RNA editing sites within coding exons and their frequency is shown. The absolute numbers are indicated in parentheses. (E) Venn Diagram of RNA editing sites in Drosophila melanogaster. A comparison of known RNA editing sites from four independent studies is shown in the Venn diagram. The number below each label represents the total number of sites and the percentage in the parentheses represents the percent that overlapped with other studies. We limited ourselves to nonrepetitive regions for higher stringency. (F) Venn Diagram of genes containing RNA editing sites in Drosophila melanogaster. A comparison of known RNA edited genes from four independent studies is shown in the Venn diagram. The number below each label represents the total number of genes and the percentage in the parentheses represents the percent that overlapped with other studies. We limited ourselves to nonrepetitive regions for higher stringency. (G) Gene Ontology analysis of genes edited in D. melanogaster. A list of ten significantly enriched Gene Ontology terms are shown in the plot. The values are the negative log adjusted p-value for each Gene Ontology term.

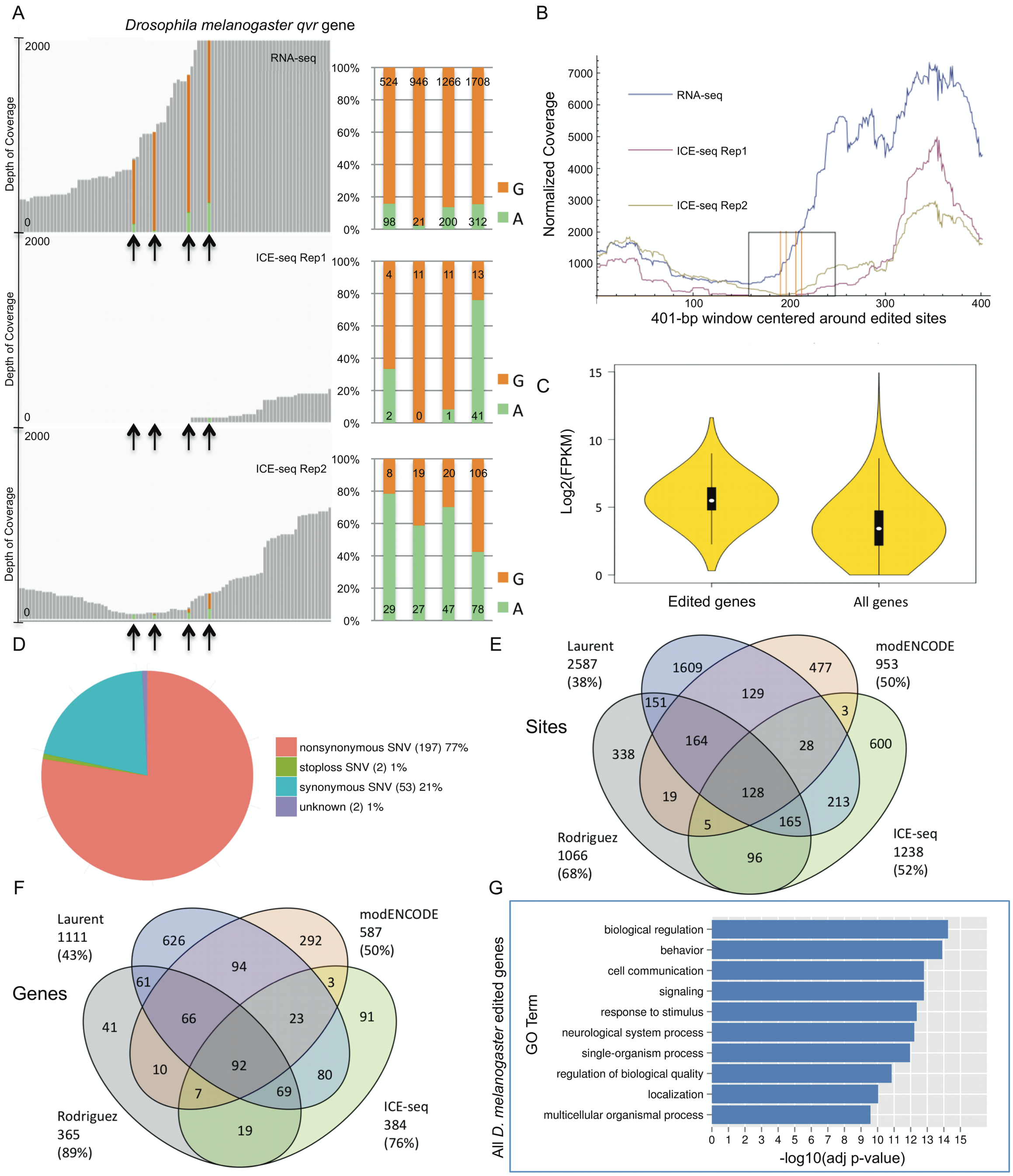
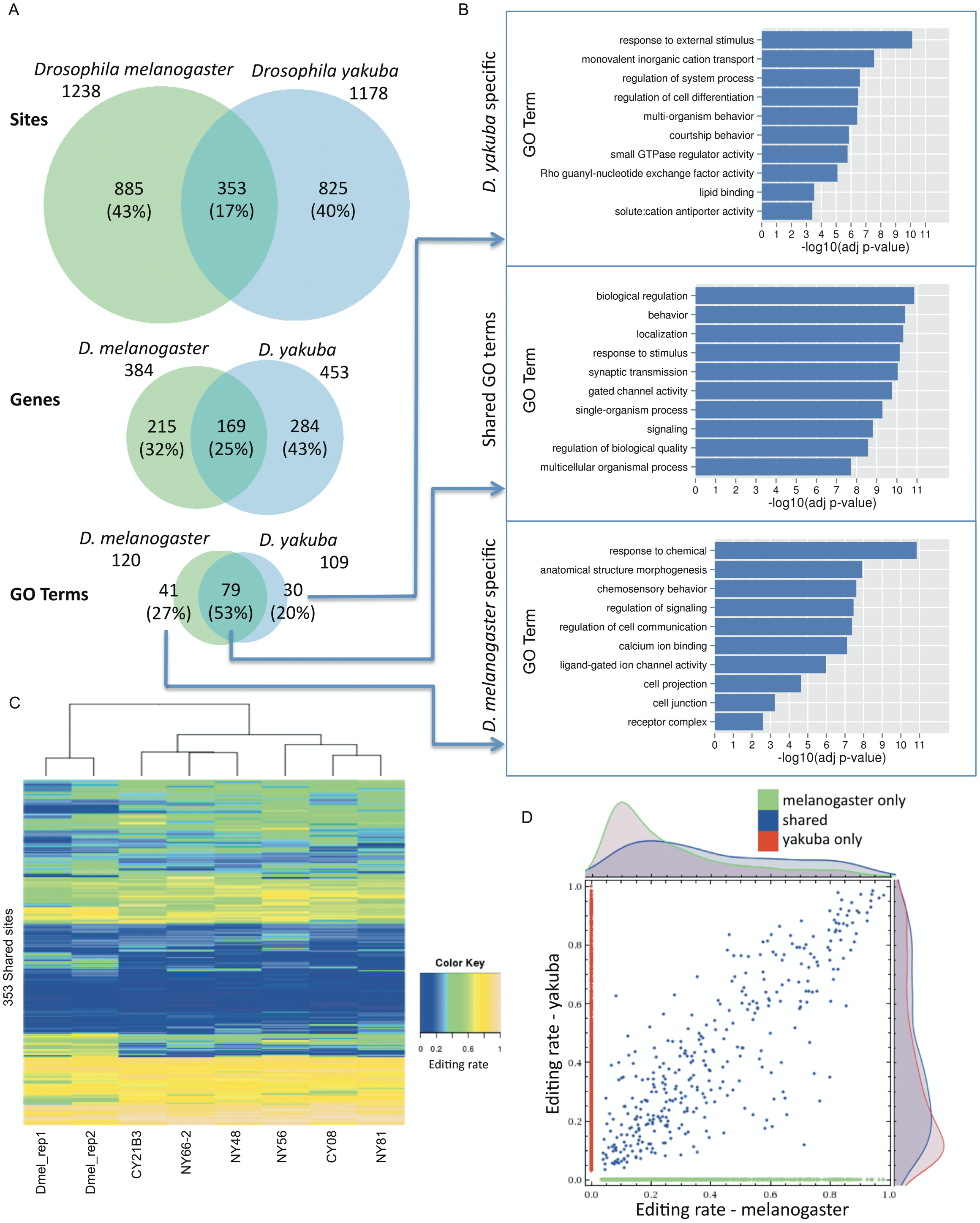

A
Drosophila melanogaster qvr gene
Depth of Coverage
2000
0
RNA-seq
ICE-seq Rep1
ICE-seq Rep2
100%
80%
60%
40%
20%
0%
524 946 1266 1708
98 21 200 312
4 11 11 13
2 0 1 41
8 19 20 106
29 27 47 78
G
A
B
Normalized Coverage
7000
6000
5000
4000
3000
2000
1000
RNA-seq
ICE-seq Rep1
ICE-seq Rep2
100
200
300
400
401-bp window centered around edited sites
C
Log2(FPKM)
15
10
5
0
Edited genes
All genes
D
nonsynonymous SNV (197) 77%
stoploss SNV (2) 1%
synonymous SNV (53) 21%
unknown (2) 1%
E
Sites
Laurent
2587
(38%)
modENCODE
953
(50%)
Rodriguez
1066
(68%)
ICE-seq
1238
(52%)
1609
477
151
129
3
338
164
28
600
19
128
213
5
165
96
F
Genes
Laurent
1111
(43%)
modENCODE
587
(50%)
Rodriguez
365
(89%)
ICE-seq
384
(76%)
626
292
94
61
3
41
66
23
91
10
92
80
7
69
19
G
All D. melanogaster edited genes
GO Term
biological regulation
behavior
cell communication
signaling
response to stimulus
neurological system process
single-organism process
regulation of biological quality
localization
multicellular organismal process
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
-log10(adj p-value)

Figure 4-3.

Comparison of RNA editing in Drosophila melanogaster and Drosophila yakuba.

(A) Comparison of RNA editing at the level of sites, genes, and significant Gene Ontology terms. The Venn diagram of RNA editing of sites (top), genes (genes), and Significant GO terms (bottom) between Drosophila melanogaster (left) and Drosophila yakuba (right) are shown. The numbers shown outside the Venn diagrams represent the total number for each species. The numbers in parentheses represent the percent of each section with respect to the union. (B) Shared and species-specific Gene Ontology analysis. Significant GO terms for Drosophila yakuba-specific(top), shared (middle), and Drosophila melanogaster-specific (bottom) are shown. The values for plots are the negative log adjusted p-value for each Gene Ontology term. (C) Hierarchical clustering analysis of RNA editing rate. For the 353 sites that are shared between the two species, a hierarchical clustering analysis was performed. The rows represent the sites and the columns represent the Drosophila melanogaster replicates and Drosophila yakuba strains. (D) Scatter plot of RNA editing rate. The editing rate for sites that are only within Drosophila melanogaster (green), sites that are only in Drosophila yakuba (red), and shared sites (blue) are plotted. Smooth histograms are shown for the editing rate in Drosophila yakuba (right) and the editing rate in Drosophila melanogaster (top).

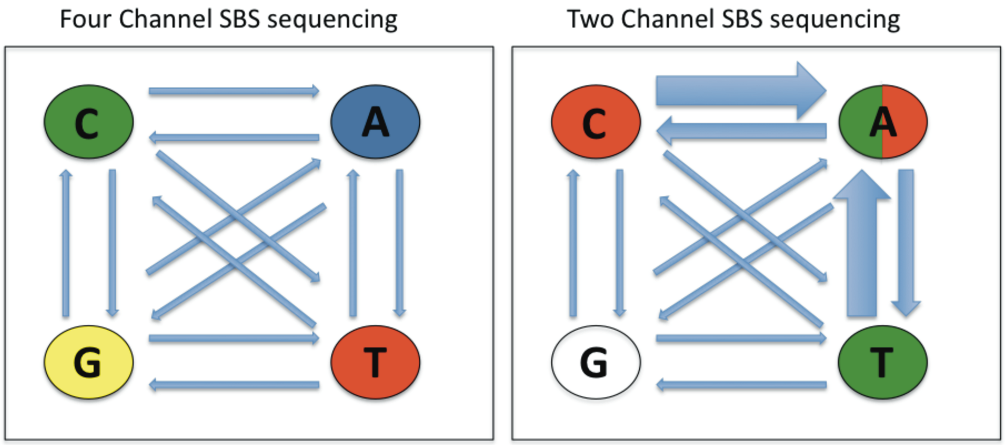

A
Drosophila melanogaster
1238
Drosophila yakuba
1178
Sites
885
(43%)
353
(17%)
825
(40%)
D. melanogaster
384
D. yakuba
453
Genes
215
(32%)
169
(25%)
284
(43%)
D. melanogaster
120
D. yakuba
109
GO Terms
41
(27%)
79
(53%)
30
(20%)
B
D. yakuba specific
GO Term
response to external stimulus
monovalent inorganic cation transport
regulation of system process
regulation of cell differentiation
multi-organism behavior
courtship behavior
small GTPase regulator activity
Rho guanyl-nucleotide exchange factor activity
lipid binding
solute:cation antiporter activity
0 1 2 3 4 5 6 7 8 9 10 11
-log10(adj p-value)
Shared GO terms
GO Term
biological regulation
behavior
localization
response to stimulus
synaptic transmission
gated channel activity
single-organism process
signaling
regulation of biological quality
multicellular organismal process
0 1 2 3 4 5 6 7 8 9 10 11
-log10(adj p-value)
D. melanogaster specific
GO Term
response to chemical
anatomical structure morphogenesis
chemosensory behavior
regulation of signaling
regulation of cell communication
calcium ion binding
ligand-gated ion channel activity
cell projection
cell junction
receptor complex
0 1 2 3 4 5 6 7 8 9 10 11
-log10(adj p-value)
C
353 Shared sites
Color Key
0 0.2 0.6 1
Editing rate
Dmel_rep1
Dmel_rep2
CY21B3
NY66-2
NY48
NY56
CY08
NY81
D
melanogaster only
shared
yakuba only
Editing rate - yakuba
0.0
0.2
0.4
0.6
0.8
1.0
Editing rate - melanogaster
0.0
0.2
0.4
0.6
0.8
1.0

4-S1 Figure.

Diagram of relative error rate of HiSeq and NextSeq sequencing platforms.

The 12 possible substitutions between the 4 bases are shown. The colors for each oval represent the fluorescent colors that are used to sequence the respective base. The relative error rate for the HiSeq (four channel) (left) and the NextSeq (two channel) (right) are represented by the thickness of the arrows.

Supplemental Fig 1


Four Channel SBS sequencing

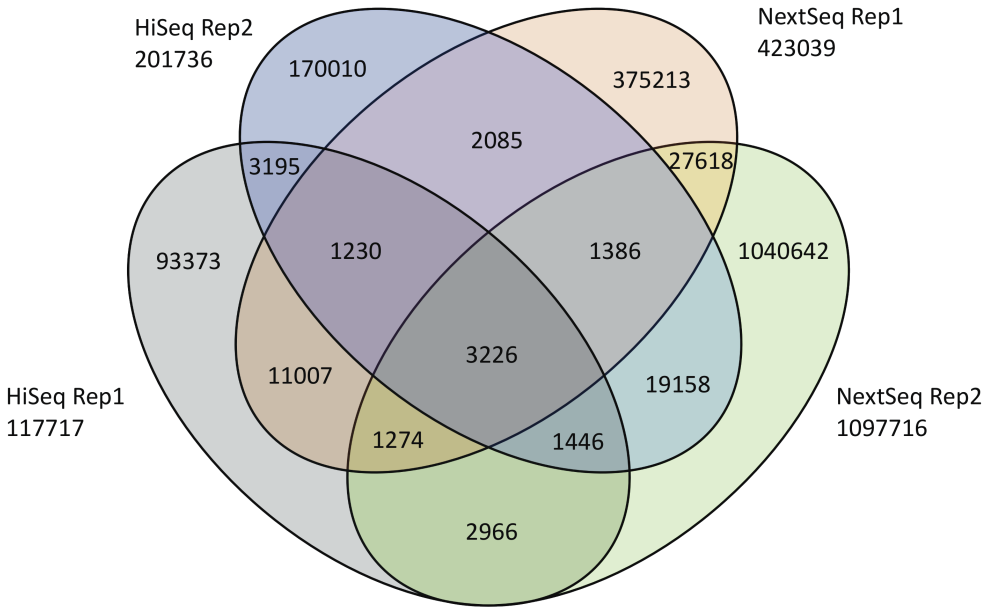
Two Channel SBS sequencing

4-S2 Figure.

SNVs from HiSeq and NextSeq sequencing platforms.

Two RNA-seq libraries were generated and sequenced using both the HiSeq and NextSeq platforms. SNVs were called from two replicates and a Venn diagram of SNVs is shown. The numbers next to the replicate label represents the total number of SNVs called for the sample.

Supplemental Fig 2



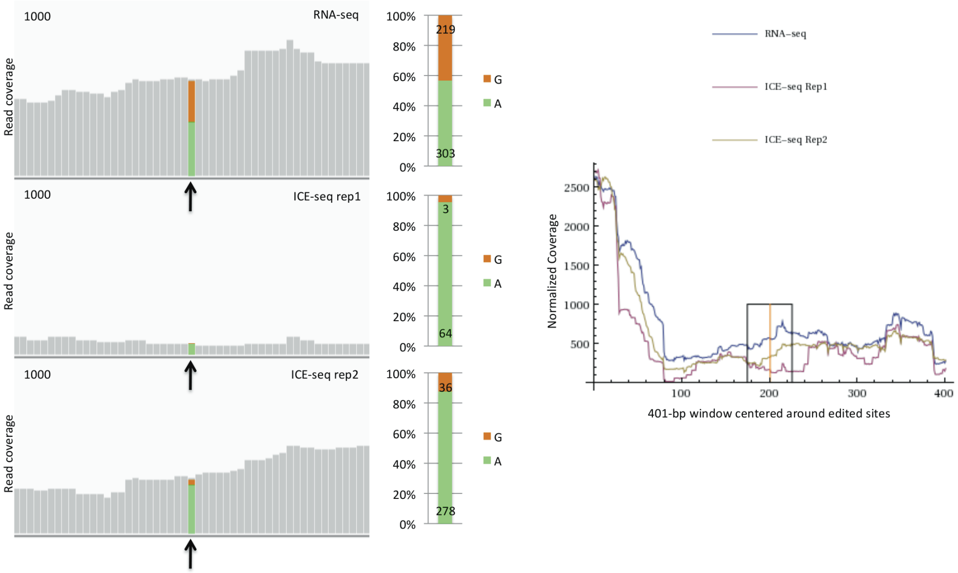
4-S3 Figure.

Additional RNA-seq vs ICE-seq frequency and coverage comparisons for two isolated RNA editing sites.

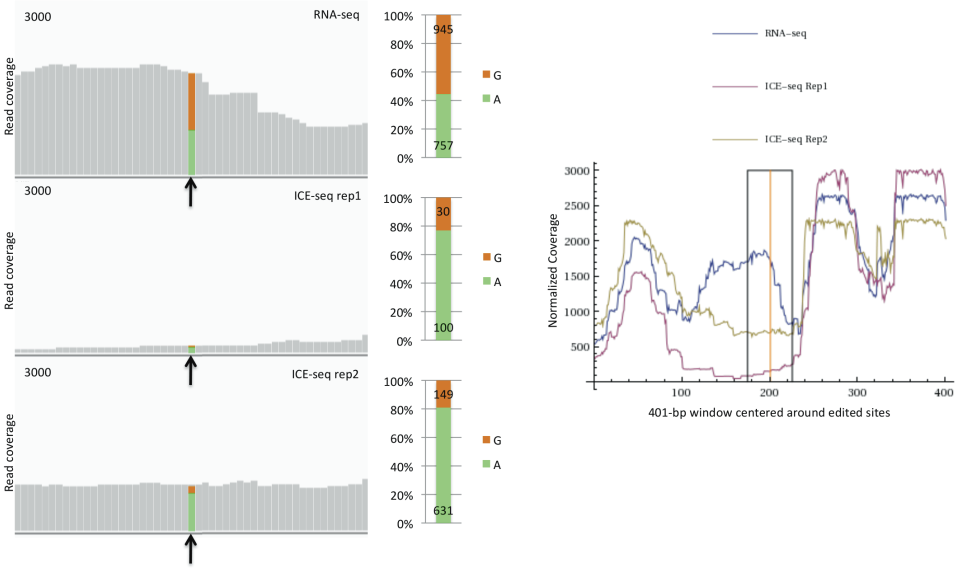
RNA-seq/ICE-seq frequency and coverage comparisons for two isolated RNA editing sites are provided. For each site, the coverage within a 51-bp window (left) and a 401-bp window (right) is shown. The black box within the 401-bp window represents the area of the 51-bp window. The site of interest is represented by the vertical orange bar in the 401-bp window. The variant frequency is indicated by the bar charts (middle) and the numbers over the bars represent the number of reads supporting each variant.

Supplemental Fig 3

RNA editing site Adar chrX:1675873



RNA editing site shi chrX:15791327

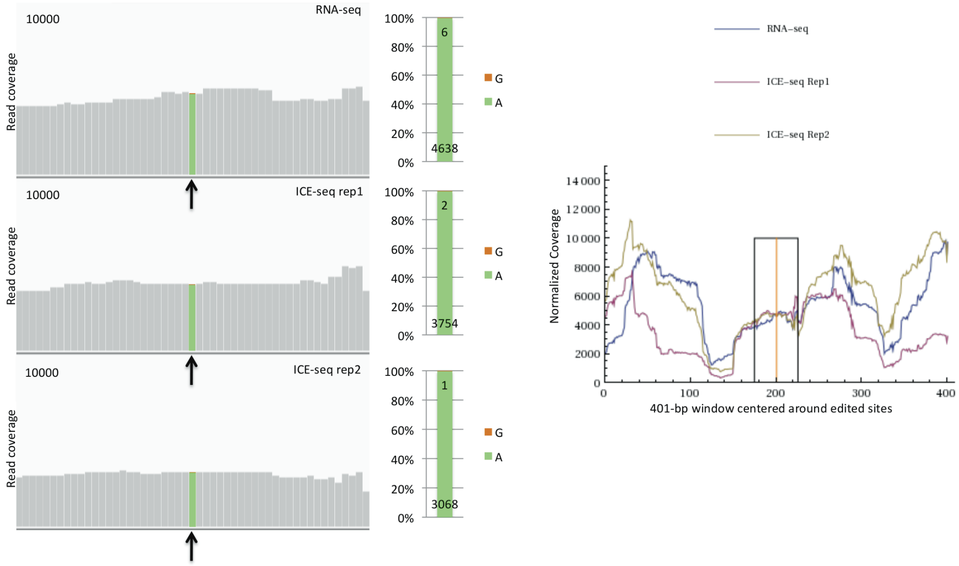


4-S4 Figure.

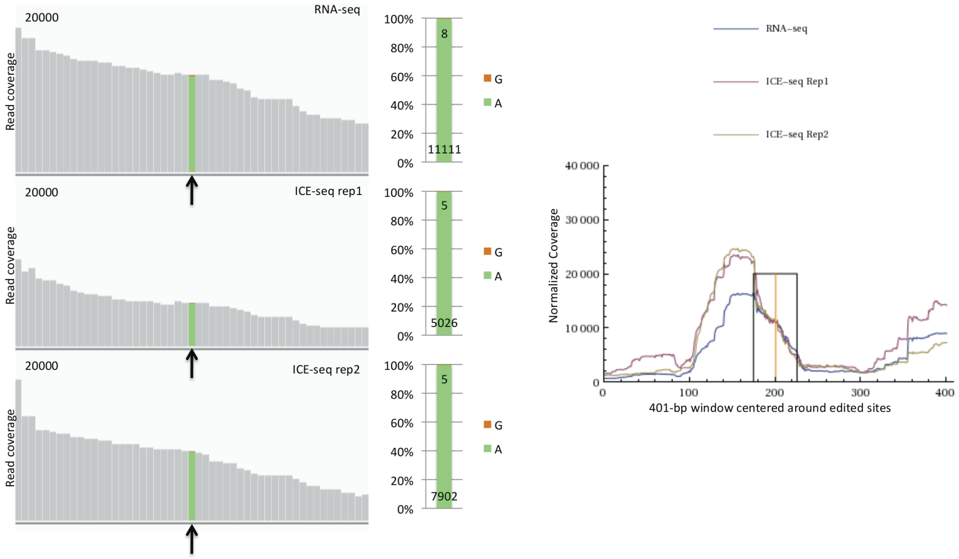
Additional RNA-seq vs ICE-seq frequency and coverage comparisons for sequencing errors. RNA-seq/ICE-seq frequency and coverage comparisons for two sites with sequencing errors are provided. For each site, the coverage within a 51-bp window (left) and a 401-bp window (right) is shown. The black box within the 401-bp window represents the area of the 51-bp window. The site of interest is represented by the vertical orange bar in the 401-bp window. The variant frequency is indicated by the bar charts (middle) and the numbers over the bars represent the number of reads supporting each variant.

Supplemental Fig 4

Sequencing error gapdh2 chrX:15763305



Sequencing error Appl chrX:470231



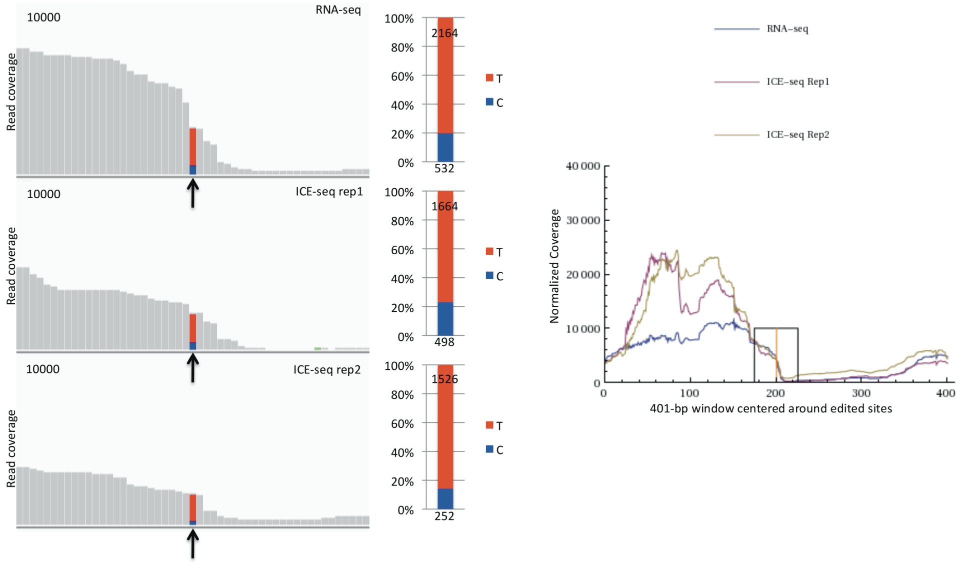
4-S5 Figure.

RNA-seq vs ICE-seq frequency and coverage comparisons for SNPs/mapping artifacts.

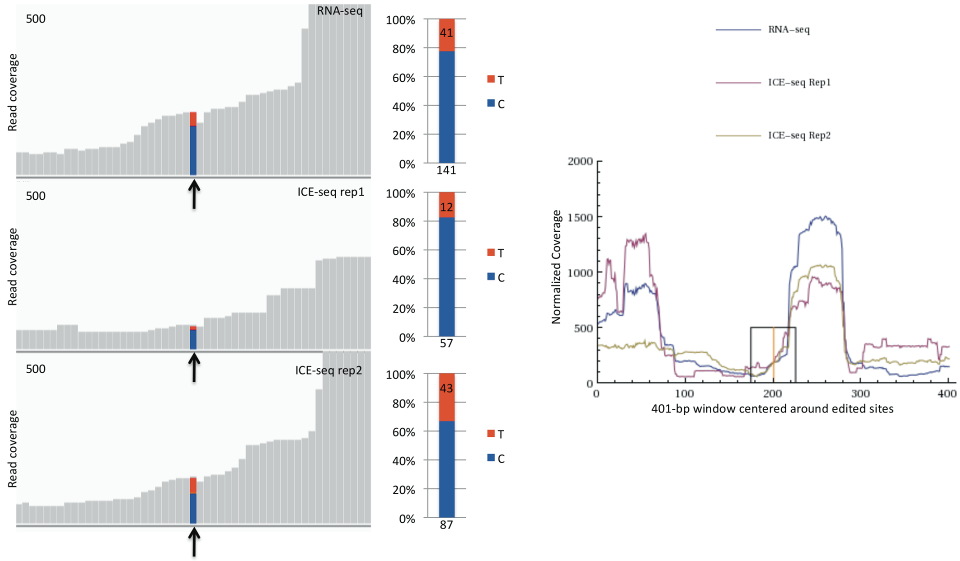
RNA-seq/ICE-seq frequency and coverage comparisons for two sites with SNPs or mapping artifacts are provided. For each site, the coverage within a 51-bp window (left) and a 401-bp window (right) is shown. The black box within the 401-bp window represents the area of the 51-bp window. The site of interest is represented by the vertical orange bar in the 401-bp window. The variant frequency is indicated by the bar charts (middle) and the numbers over the bars represent the number of reads supporting each variant.

Supplemental Fig 5

SNP or mapping artifact Act5c chrX:5798415



SNP or mapping artifact CG42257 chr2R:17956972

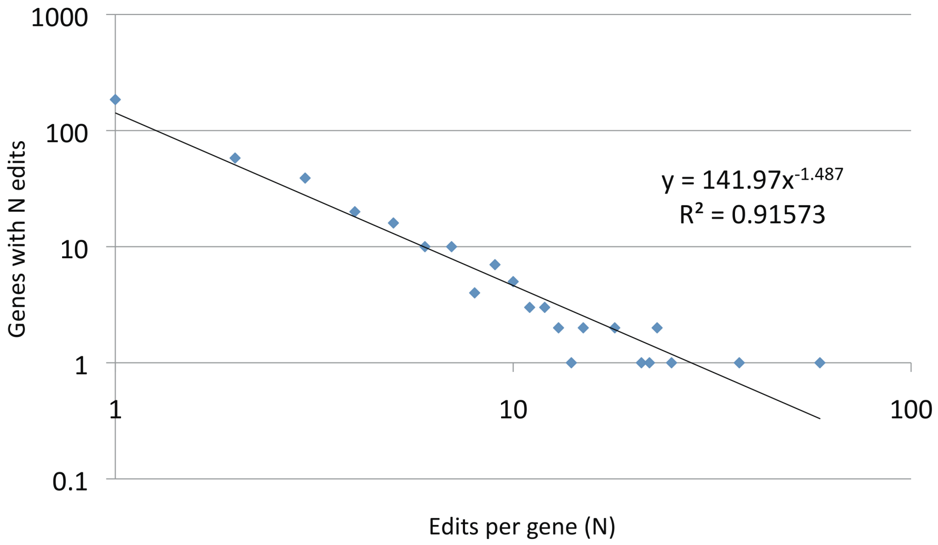


4-S6 Figure.

Power law distribution for the number RNA editing sites within genes.

The number of edits per gene (N) vs genes with N edits is shown. The scales for both axes are shown in log scale. The data is for Drosophila melanogaster.

Supplemental Fig 6

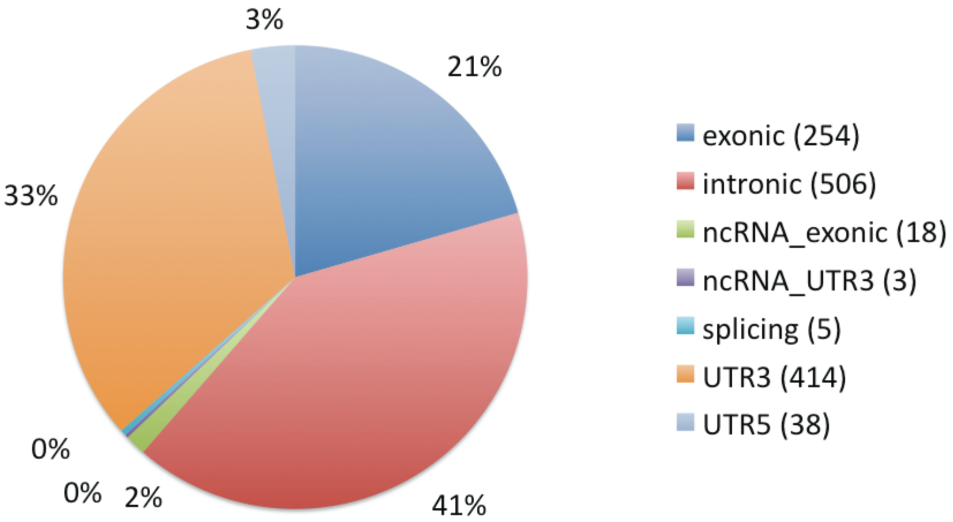


4-S7 Figure.

Genomic regions for RNA editing sites in Drosophila melanogaster.

The genomic regions for the 1238 Drosophila melanogaster sites are shown. The absolute number in each category is indicated with parentheses.

Supplemental Fig 7

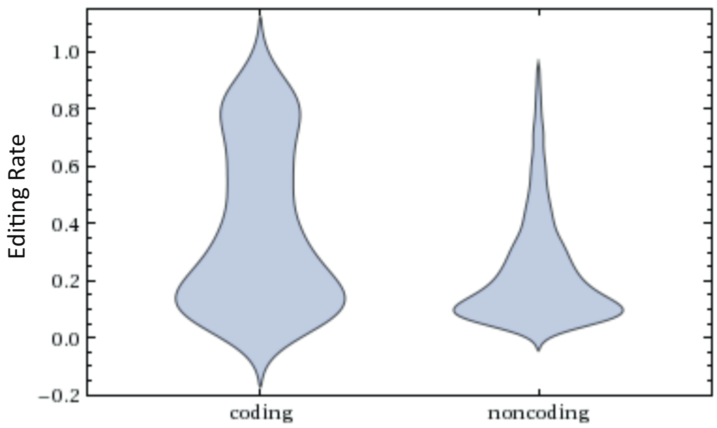


4-S8 Figure.

RNA editing frequency is greater for coding RNA editing sites.

The RNA editing frequencies for coding (left) and noncoding (right) RNA editing events in Drosophila melanogaster are indicated by the two violin plots. The y-axis indicates the editing rate and the thickness of the plot indicates the frequency.

Supplemental Fig 8



4-S9 Figure.

Analysis of amino acid changes induced by RNA editing.

For the 64 codons, 48 possible codon changes are possible through single nucleotide substitutions (left). Of the 48 possible codon changes, 33 are nonsynonymous and 15 are synonymous. There are 16 unique coding changes within the 33 nonsynonymous changes (middle). The counts for Drosophila melanogaster coding changes are shown (top right). The normalized counts for coding changes (number of coding changes / unique change counts) are shown (bottom right).

Supplemental Fig 9

48 possible codon changes

Unedited Codon				Edited Codon			Coding Change
A	A	A	->	G	A	A	K->E
A	A	C	->	G	A	C	N->D
A	A	G	->	G	A	G	K->E
A	A	U	->	G	A	U	N->D
A	C	A	->	G	C	A	T->A
A	C	C	->	G	C	C	T->A
A	C	G	->	G	C	G	T->A
A	C	U	->	G	C	U	T->A
A	G	A	->	G	G	A	R->G
A	G	C	->	G	G	C	S->G
A	G	G	->	G	G	G	R->G
A	G	U	->	G	G	U	S->G
A	U	A	->	G	U	A	I->V
A	U	C	->	G	U	C	I->V
A	U	G	->	G	U	G	M->V
A	U	U	->	G	U	U	I->V

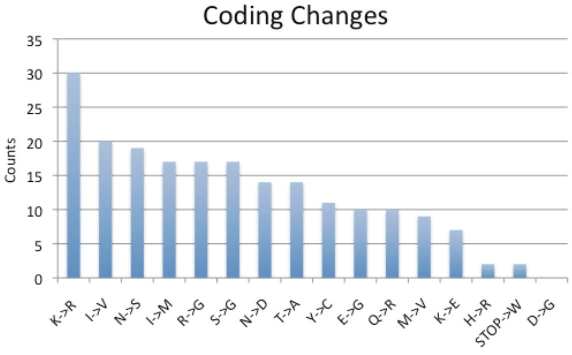
A	A	A	->	A	G	A	K->R
A	A	C	->	A	G	C	N->S
A	A	G	->	A	G	G	K->R
A	A	U	->	A	G	U	N->S
C	A	A	->	C	G	A	Q->R
C	A	C	->	C	G	C	H->R
C	A	G	->	C	G	G	Q->R
C	A	U	->	C	G	U	H->R
G	A	A	->	G	G	A	E->G
G	A	C	->	G	G	C	D->G
G	A	G	->	G	G	G	E->G
G	A	U	->	G	G	U	D->G
U	A	A	->	U	G	A	STOP->STOP
U	A	C	->	U	G	C	Y->C
U	A	G	->	U	G	G	STOP->W
U	A	U	->	U	G	U	Y->C

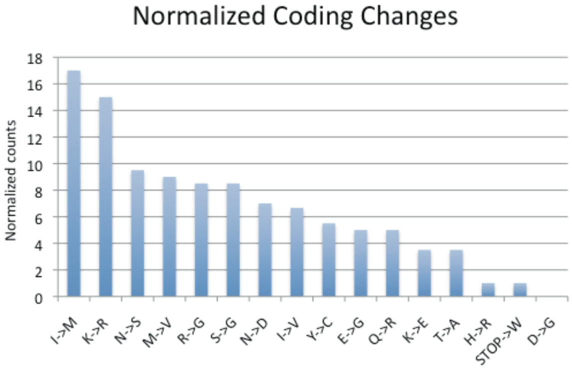
A	A	A	->	A	A	G	K->K
A	C	A	->	A	C	G	T->T
A	G	A	->	A	G	G	R->R
A	U	A	->	A	U	G	I->M
C	A	A	->	C	A	G	Q->Q
C	C	A	->	C	C	G	P->P
C	G	A	->	C	G	G	R->R
C	U	A	->	C	U	G	L->L
G	A	A	->	G	A	G	E->E
G	C	A	->	G	C	G	A->A
G	G	A	->	G	G	G	G->G
G	U	A	->	G	U	G	V->V
U	A	A	->	U	A	G	STOP->STOP
U	C	A	->	U	C	G	S->S
U	G	A	->	U	G	G	STOP->W
U	U	A	->	U	U	G	L->L

16 unique coding changes

Count	Unique change
4	T->A
3	I->V
2	D->G
2	E->G
2	H->R
2	K->E
2	K->R
2	N->D
2	N->S
2	Q->R
2	R->G
2	S->G
2	STOP->W
2	Y->C
1	I->M
1	M->V

synonymous (15)
nonsynonymous (33)



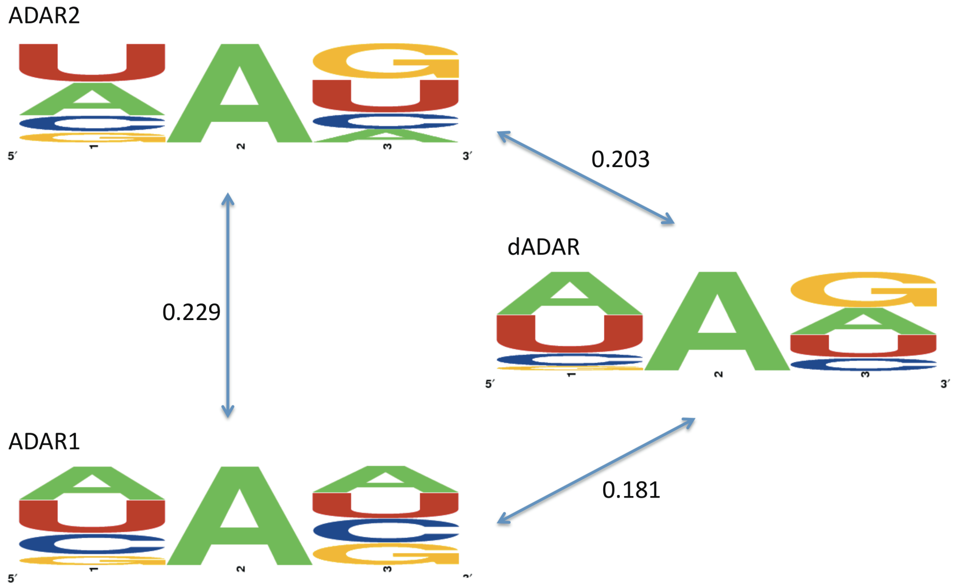


4-S10 Figure.

Frequency plot of 5' and 3' neighbors of editing sites.

The frequency plots for adjacent neighbors of ADAR1, ADAR2, and Drosophila melanogaster Adar are shown. The heights of the bases indicate their respective frequencies. The 2-norm of difference PPM (Position Probability Matrix) is indicated between each frequency plot. The PPM was used as a distance measure between the three neighbor preferences.

Supplemental Fig 10

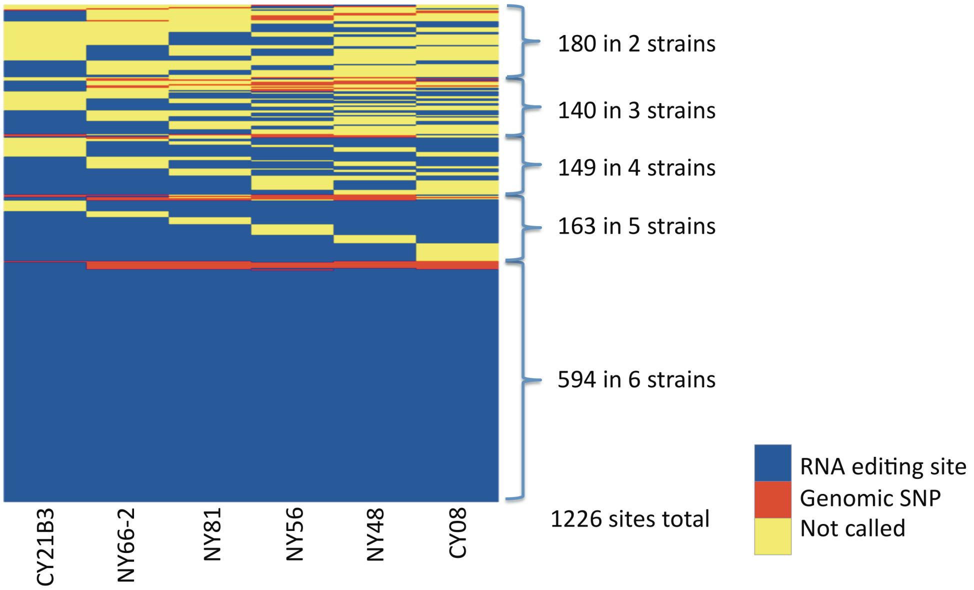


4-S11 Figure.

Heatmap of RNA editing calls for six strains of Drosophila yakuba.

The Drosophila yakuba sites that were called as an RNA editing even are indicated blue for each Drosophila yakuba strain. Genomic SNPs are indicated in red and sites that were not called within a particular strain are indicated in yellow. The rows correspond to each site and the columns correspond to the indicated Drosophila yakuba strain. The sites are grouped and sorted by the number of strains in which the site was observed.

Supplemental Fig 11

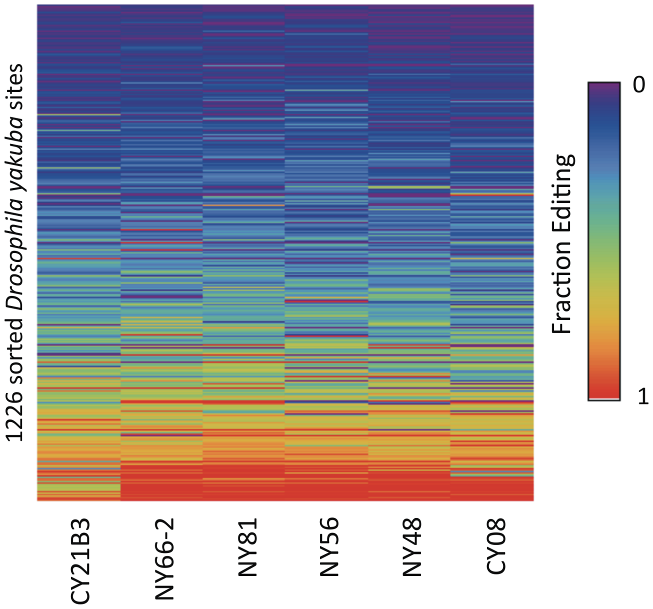


4-S12 Figure.

Heatmap of RNA editing rate for six strains of Drosophila yakuba.

The RNA editing rate for all Drosophila yakuba sites are shown. The rows correspond to each site and the columns correspond to the indicated Drosophila yakuba strain. The sites are sorted by increasing average editing rate.

Supplemental Fig 12

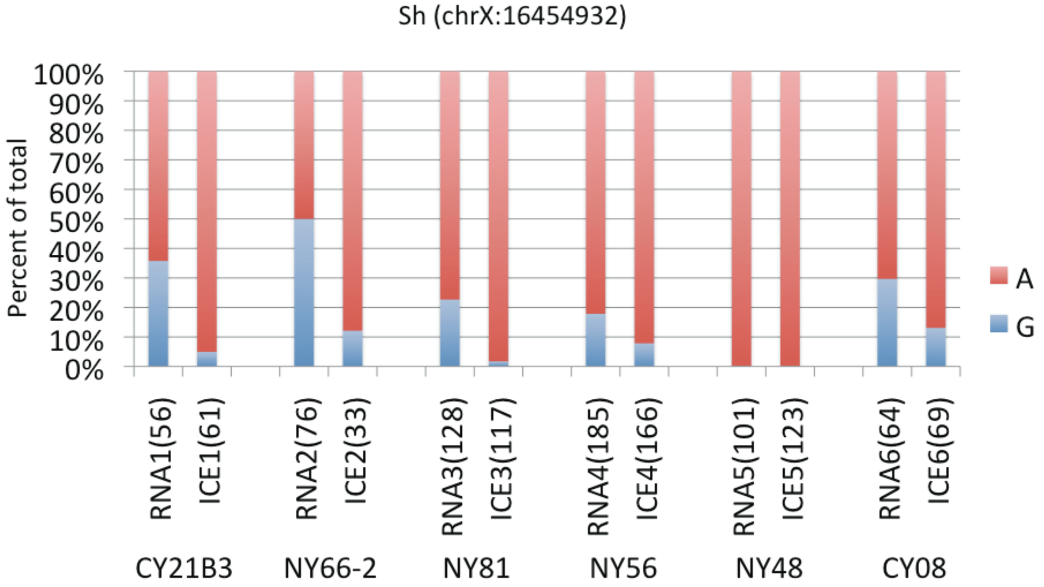


4-S13 Figure.

Example of strain-specific RNA editing in which one strain is not edited.

The variant frequencies in the RNA-seq and ICE-seq reads are shown for each of the Drosophila yakuba strains. The site is within the Sh gene at position 16454932 of chromosome X. Strain NY48 had no detectable levels of RNA editing while the other five strains had variable levels (20%-50%) of RNA editing. The variant frequencies are indicated by the heights of the bars and the coverage for each sample is indicated in parentheses.

Supplemental Fig 13

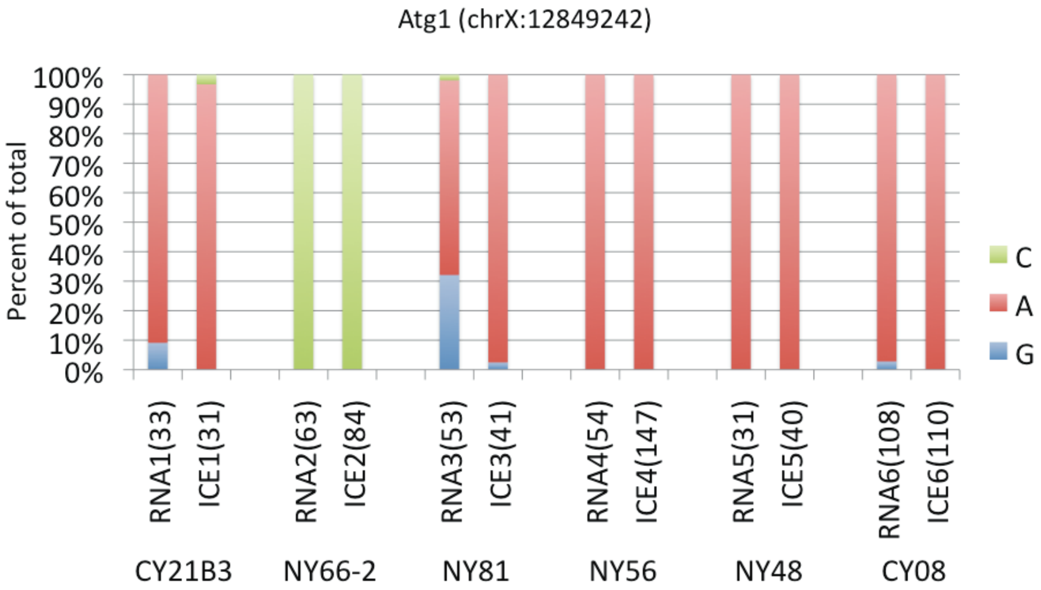


4-S14 Figure.

Example of strain-specific RNA editing in which one strain has a genomic Cytosine.

The variant frequencies in the RNA-seq and ICE-seq reads are shown for each of the Drosophila yakuba strains. The site is within the Atg1 gene at position 12849242 of chromosome X. The NY66-2 strain had a C genomic SNP at that location. Strains CY21B3, NY81, and CY08 had moderate levels of RNA editing. Strains NY56 and NY48 had no detectable levels of RNA editing. The variant frequencies are indicated by the heights of the bars and the coverage for each sample is indicated in parentheses.

Supplemental Fig 14

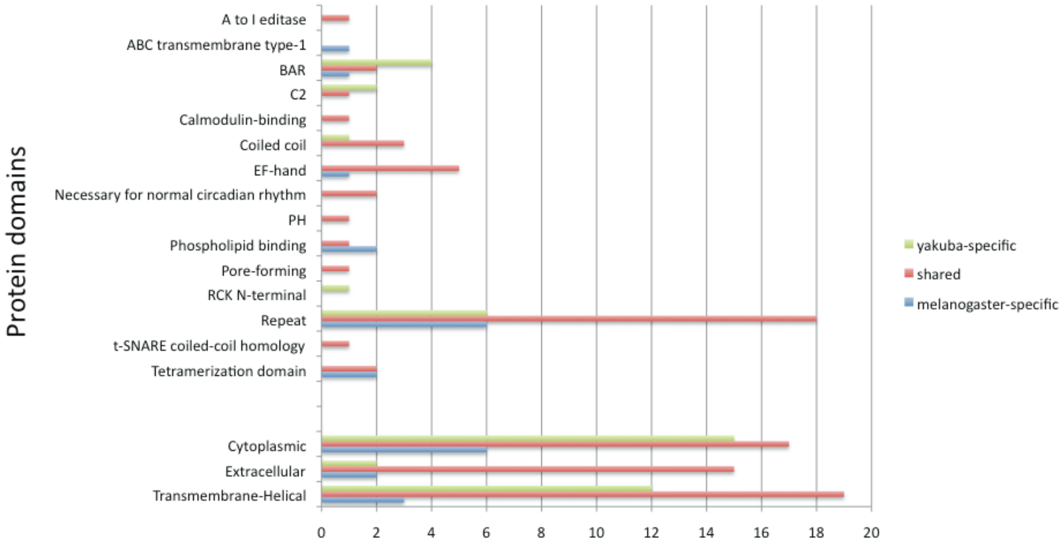


4-S15 Figure.

Comparison of annotated protein domains affected by RNA editing.

The RNA editing sites for Drosophila melanogaster and Drosophila yakuba were mapped onto their respective protein coordinates and checked for known protein domains that were annotated within the uniprot database. For comparability, the Drosophila yakuba sites were converted to Drosophila melanogaster genome coordinates and compared against because the Drosophila melanogaster protein domains were better annotated. Shared domains are labeled in red, Drosophila yakuba specific domains are labeled in green, and Drosophila melanogaster domains are labeled in blue.

Supplemental Fig 15

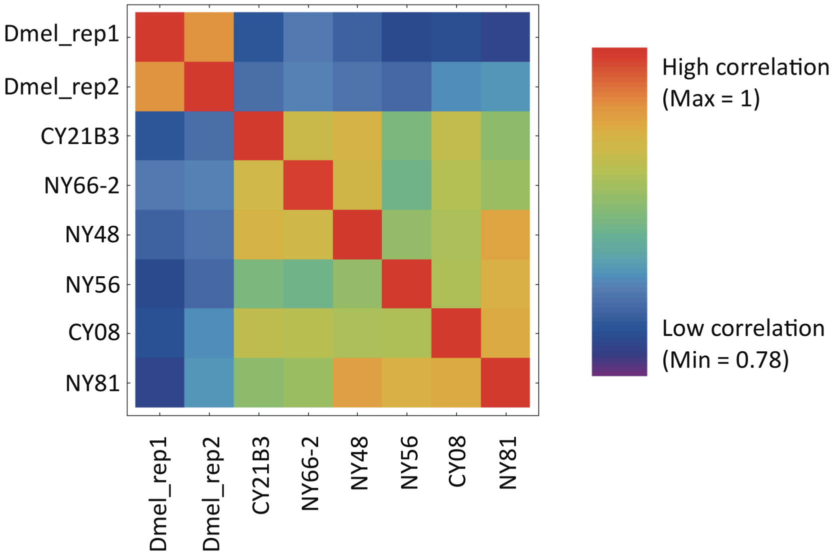


4-S16 Figure.

RNA editing correlation between samples.

For the 353 sites that are shared between Drosophila melanogaster and Drosophila yakuba, the heatmap for the correlation between samples are shown. The color scale represents the correlation range from 1 to 0.78.

Supplemental Fig 16

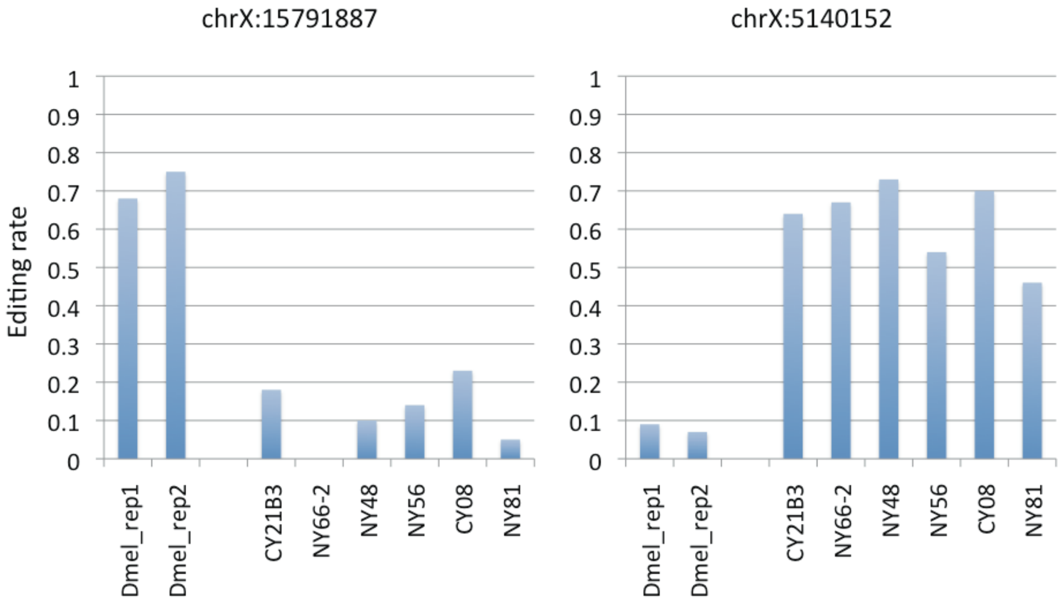


4-S17 Figure.

Examples of diverging RNA editing sites.

Two examples of diverging RNA editing sites are shown. An example of a site with high editing in Drosophila melanogaster chrX:15791887 (left) and an example of a site with high editing in Drosophila yakuba are provided.

Supplemental Fig 17


chrX:15791887

chrX:5140152

Chapter 5

Summary and conclusions

Concluding remarks

Here I have investigated the extent and function of RNA editing within human cell lines, mouse, and fly by using a variety of high-throughput sequencing based approaches. As part of the ENCODE project, I have mined RNA-seq datasets from fourteen human cell lines for RNA editing events. In order to remove genomic SNPs from the data, I repurposed ENCODE ChIP-seq reads as a source of genomic sequence. To determine the extent of C-to-U RNA editing in mice, I analyzed liver and intestine RNA-seq datasets from wild-type and Apobec-1 knockout mice. By comparing RNA sequence variation from the two types of mice, I was able to identify dozens of C-to-U RNA editing events. A subset of the identified C-to-U RNA editing targets had altered RNA and protein levels. In order to evaluate the nature of RNA editing within Drosophila, I compared sequencing reads from ICE-seq and RNA-seq datasets to identify and measure RNA editing events in two species of fly and within six strains within the same species. I found that some of the best characterized RNA editing targets are shared between the two species, but there were a distinguishable subset of targets that were diverging in functional enrichments.

Future directions

In general, the first step in the study of RNA modifications is to identify the sites and the second step is to associate functions to the sites. With respect to the ENCODE RNA editing study in Chapter 2, it would be interesting to interrogate the functions of the RNA editing events within the cell lines using a genome-editing approach such as the CRISPR/Cas9 system. Specific RNA editing events of functional interest can be removed from the transcriptome by altering the edited adenosine or by deleting regions within the underlying genome. Since double stranded RNA is a required substrate for ADARs, RNA editing would be altered by removing the complementary region and thus making the RNA a less desirable substrate for ADAR. Similarly, an exogenous system can be used where transcripts with and without regions susceptible to RNA-editing can be introduced to the cells in order to monitor how the cells treat the RNA in a controlled setting. Within this system, the cellular localization could be tracked and translatability monitored. Using these two approaches, the effect of nuclear localization of edited transcripts can be analyzed. Additionally, the cells can be infected with viruses or stimulated with lipopolysaccaride (LPS) or poly(I:C) to trigger an immune response. This would initiate transcription of the p150 form of ADAR and the changes in RNA editing and gene expression upon infection can be investigated. In addition, it has been suggested that many transcribed ALUs are retained as exons (Zarnack 2013, Lin 2008, Lev-Maor 2007). Since many of the RNA editing sites were found within intronic regions, it is possible that some of these regions might be incorporated within exons and that RNA editing might modulate these exonization events. Lastly, it would be worthwhile to revalidate the previously identified RNA editing events using ICE-seq.

Several RNA-editing targets of Apobec1 have been identified. Additionally, editing of these targets have been shown to alter protein levels by altering polysome profiles of the targets. Overexpression of Apobec1 has been sufficient to induce tumorigenesis in rabbit livers and certain human cancers have been shown to have an overabundance of C-to-T substitutions, indicative of a cytosine deaminase. Thus, it would be interesting to investigate the possible role of APOBEC1 in initiating tumorigenesis. Using an inducible system to overexpress Apobec1 in adult animals, one could perform a time-course experiment to identify the mechanistic changes that are associated with Apobec1 induced tumors. In a similar manner, the same approach can be used to investigate the potential for ADARs to intiate tumorigenesis within model organisms.

ICE-seq is a powerful method to identify RNA editing events because it allows for a direct RNA-RNA comparison, which would control for many potential sources of error. Additionally it does not require a reference genome because a de novo transcriptome assembly based approach can be used. Thus, it would be interesting to apply this method to other species in order to explore the extent and evolution of A-to-I RNA editing. It is clear that the nature of RNA editing has evolved distinct functions and patterns within different organisms. It would be interesting to investigate the extent of divergence across multiple metazoan species.

References

Agranat L, Raitskin O, Sperling J, Sperling R. 2008. The editing enzyme ADAR and the mRNA surveillance protein hUpf1 interact in the cell nucleus. Proc Natl Acad Sci 105: 5028–5033.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjörd JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Jäger N, Jones DT, Jones D, Knappskog S, Kool M, Lakhani SR, López-Otín C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt AN, Valdés-Mas R, van Buuren MM, van 't Veer L, Vincent-Salomon A, Waddell N, Yates LR; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain, Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, Stratton MR. Signatures of mutational processes in human cancer. Nature. 2013 Aug 22;500(7463):415-21.

Alon S, Garrett SC, Levanon EY, Olson S, Graveley BR, Rosenthal JJ, Eisenberg E. The majority of transcripts in the squid nervous system are extensively recoded by A-to-I RNA editing. Elife. 2015 Jan 8;4.

Arnez JG, Steitz TA. Crystal structure of unmodified tRNA(Gln) complexed with glutaminyl-tRNA synthetase and ATP suggests a possible role for pseudo-uridines in stabilization of RNA structure. Biochemistry. 1994 Jun 21;33(24):7560-7.

Athanasiadis A, Rich A, Maas S. 2004. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biol 2: e391. doi: 10.1371/journal.pbio.0020391.

Backus JW, Smith HC. Apolipoprotein B mRNA sequences 3′ of the editing site are necessary and sufficient for editing and editosome assembly. Nucleic Acids Res 1991, 19:6781-6786.

Bahn JH, Lee JH, Li G, Greer C, Peng G, Xiao X. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. Genome Res 2012, 22:142-150.

Bakin A, Ofengand J. Four newly located pseudouridylate residues in Escherichia coli 23S ribosomal RNA are all at the peptidyltransferase center: analysis by the application of a new sequencing technique. Biochemistry. 1993 Sep 21;32(37):9754-62.

Bass BL, Weintraub H. A developmentally regulated activity that unwinds RNA duplexes. Cell. 1987 Feb 27;48(4):607-13.

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

Batista PJ, Molinie B, Wang J, Qu K, Zhang J, Li L, Bouley DM, Lujan E, Haddad B, Daneshvar K, Carter AC, Flynn RA, Zhou C, Lim KS, Dedon P, Wernig M, Mullen AC, Xing Y, Giallourakis CC, Chang HY. m(6)A RNA modification controls cell fate transition in mammalian embryonic stem cells. Cell Stem Cell. 2014 Dec 4;15(6):707-19.

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.

Bianchini M, Levy E, Zucchini C, Pinski V, Macagno C, De Sanctis P, Valvassori L, Carinci P, Mordoh J. Comparative study of gene expression by cDNA microarray in human colorectal cancer tissues and normal mucosa. Int J Oncol 2006, 29:83-94.

Blanc V, Davidson NO. C-to-U RNA editing: mechanisms leading to genetic diversity. J Biol Chem. 2003 Jan 17;278(3):1395-8.

Blanc V, Davidson NO. Mouse and other rodent models of C to U RNA editing. Methods Mol Biol 2011, 718:121-135.

Blanc V, Henderson JO, Kennedy S, Davidson NO. Mutagenesis of apobec-1 complementation factor reveals distinct domains that modulate RNA binding, protein-protein interaction with apobec-1, and complementation of C to U RNA-editing activity. J Biol Chem 2001, 276:46386-46393.

Blanc V, Henderson JO, Newberry RD, Xie Y, Cho SJ, Newberry EP, Kennedy S, Rubin DC, Wang HL, Luo J, Davidson NO. Deletion of the AU-rich RNA binding protein Apobec-1 reduces intestinal tumor burden in Apc(min) mice. Cancer Res 2007, 67:8565-8573.

Blanc V, Park E, Schaefer S, Miller M, Lin Y, Kennedy S, Billing AM, Ben Hamidane H, Graumann J, Mortazavi A, Nadeau JH, Davidson NO. Genome-wide identification and functional analysis of Apobec-1-mediated C-to-U RNA editing in mouse small intestine and liver. Genome Biol. 2014 Jun 19;15(6):R79.

Blanc V, Xie Y, Luo J, Kennedy S, Davidson NO. Intestine-specific expression of Apobec-1 rescues apolipoprotein B RNA editing and alters chylomicron production in Apobec1(-)/(-) mice. J Lipid Res 2012, 53:2643-2655.

Borchert GM, Gilmore BL, Spengler RM, Xing Y, Lanier W, Bhattacharya D, Davidson BL. Adenosine deamination in human transcripts generates novel microRNA binding sites. Hum Mol Genet 2009, 18:4801-4807.

Burns CM, Chu H, Rueter SM, Hutchinson LK, Canton H, Sanders-Bush E, Emeson RB. 1997. Regulation of serotonin-2C receptor G-protein coupling by RNA editing. Nature 387: 303–308.

Bykhovskaya Y, Casas K, Mengesha E, Inbal A, Fischel-Ghodsian N. Missense mutation in pseudouridine synthase 1 (PUS1) causes mitochondrial myopathy and sideroblastic anemia (MLASA). Am J Hum Genet. 2004 Jun;74(6):1303-8.

Cantara WA, Crain PF, Rozenski J, McCloskey JA, Harris KA, Zhang X, Vendeix FA, Fabris D, Agris PF. The RNA Modification Database, RNAMDB: 2011 update. Nucleic Acids Res. 2011 Jan;39(Database issue):D195-201. doi: 10.1093/nar/gkq1028.

Carlile TM, Rojas-Duran MF, Zinshteyn B, Shin H, Bartoli KM, Gilbert WV. Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature. 2014 Nov 6;515(7525):143-6.

Cattenoz PB, Taft RJ, Westhof E, Mattick JS. Transcriptome-wide identification of A > I RNA editing sites by inosine specific cleavage. RNA. 2013 Feb;19(2):257-70.

Cenci C, Barzotti R, Galeano F, Corbelli S, Rota R, Massimi L, Di Rocco C, O'Connell MA, Gallo A. 2008. Down-regulation of RNA editing in pediatric astrocytomas: ADARB1 editing activity inhibits cell migration and proliferation. J Biol Chem 283: 7251–7260.

Charette M, Gray MW. Pseudouridine in RNA: what, where, how, and why. IUBMB Life. 2000 May;49(5):341-51.

Chen K, Lu Z, Wang X, Fu Y, Luo GZ, Liu N, Han D, Dominissini D, Dai Q, Pan T, He C. High-resolution N(6) -methyladenosine (m(6) A) map using photo-crosslinking-assisted m(6) A sequencing. Angew Chem Int Ed Engl. 2015 Jan 26;54(5):1587-90.

Chen L. Characterization and comparison of human nuclear and cytosolic editomes. Proc Natl Acad Sci U S A 2013, 110:E2741-E2747.

Chen LL, Carmichael GG. Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA. Mol Cell. 2009 Aug 28;35(4):467-78.

Chen T, Hao YJ, Zhang Y, Li MM, Wang M, Han W, Wu Y, Lv Y, Hao J, Wang L, Li A, Yang Y, Jin KX, Zhao X, Li Y, Ping XL, Lai WY, Wu LG, Jiang G, Wang HL, Sang L, Wang XJ, Yang YG, Zhou Q. m(6)A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency. Cell Stem Cell. 2015 Mar 5;16(3):289-301.

Chen Z, Eggerman TL, Patterson AP. ApoB mRNA editing is mediated by a coordinated modulation of multiple apoB mRNA editing enzyme components. Am J Physiol Gastrointest Liver Physiol 2007, 292:G53-G65.

Chen ZX, Sturgill D, Qu J, Jiang H, Park S, Boley N, Suzuki AM, Fletcher AR, Plachetzki DC, FitzGerald PC, Artieri CG, Atallah J, Barmina O, Brown JB, Blankenburg KP, Clough E, Dasgupta A, Gubbala S, Han Y, Jayaseelan JC, Kalra D, Kim YA, Kovar CL, Lee SL, Li M, Malley JD, Malone JH, Mathew T, Mattiuzzo NR, Munidasa M, Muzny DM, Ongeri F, Perales L, Przytycka TM, Pu LL, Robinson G, Thornton RL, Saada N, Scherer SE, Smith HE, Vinson C, Warner CB, Worley KC, Wu YQ, Zou X, Cherbas P, Kellis M, Eisen MB, Piano F, Kionte K, Fitch DH, Sternberg PW, Cutter AD, Duff MO, Hoskins RA, Graveley BR, Gibbs RA, Bickel PJ, Kopp A, Carninci P, Celniker SE, Oliver B, Richards S.

Comparative validation of the D. melanogaster modENCODE transcriptome annotation. Genome Res. 2014 Jul;24(7):1209-23.

Cherayil BJ. Indoleamine 2,3-dioxygenase in intestinal immunity and inflammation. Inflamm Bowel Dis 2009, 15:1391-1396.

Chester A, Somasekaram A, Tzimina M, Jarmuz A, Gisbourne J, O'Keefe R, Scott J, Navaratnam N. The apolipoprotein B mRNA editing complex performs a multifunctional cycle and suppresses nonsense-mediated decay. EMBO J 2003, 22:3971-3982.

Conticello SG, Langlois MA, Yang Z, Neuberger MS. DNA deamination in immunity: AID in the context of its APOBEC relatives. Adv Immunol. 2007;94:37-73.

Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo. a sequence logo generator. Genome Res 2004, 14:1188-1190.

Danecek P, Nellåker C, McIntyre RE, Buendia-Buendia JE, Bumpstead S, Ponting CP, Flint J, Durbin R, Keane TM, Adams DJ. High levels of RNA-editing site conservation amongst 15 laboratory mouse strains. Genome Biol. 2012 Apr 23;13(4):26.

Davis FF, Allen FW. Ribonucleic acids from yeast which contain a fifth nucleotide. J Biol Chem. 1957 Aug;227(2):907-15.

de Hoon MJL, Imoto S, Nolan J, Miyano S. 2004. Open Source Clustering Software. Bioinformatics 20: 1453–1454.

den Bosch HM DV-v, Bunger M, de Groot PJ, Bosch-Vermeulen H, Hooiveld GJ, Muller M. PPARalpha-mediated effects of dietary lipids on intestinal barrier gene expression. BMC Genomics 2008, 9:231.

Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J, Amariglio N, Kupiec M, Sorek R, Rechavi G. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature. 2012 Apr 29;485(7397):201-6.

Doria M, Neri F, Gallo A, Farace MG, Michienzi A. 2009. Editing of HIV-1 RNA by the double-stranded RNA deaminase ADAR stimulates viral infection. Nucleic Acids Res 37: 5848–5858.

Drover VA, Ajmal M, Nassir F, Davidson NO, Nauli AM, Sahoo D, Tso P, Abumrad NA. CD36 deficiency impairs intestinal lipid secretion and clearance of chylomicrons from the blood. J Clin Invest 2005, 115:1290-1297.

Edelheit S, Schwartz S, Mumbach MR, Wurtzel O, Sorek R. Transcriptome-wide mapping of 5-methylcytidine RNA modifications in bacteria, archaea, and yeast reveals m5C within archaeal mRNAs. PLoS Genet. 2013 Jun;9(6):e1003602.

EMBL-EBI. ProteomeExchange. [http://www.proteomexchange.org/ webcite]

Fernández IS, Ng CL, Kelley AC, Wu G, Yu YT, Ramakrishnan V. Unusual base pairing during the decoding of a stop codon by the ribosome. Nature. 2013 Aug 1;500(7460):107-10.

Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A. 1992 Mar 1;89(5):1827-31.

Fustin JM, Doi M, Yamaguchi Y, Hida H, Nishimura S, Yoshida M, Isagawa T, Morioka MS, Kakeya H, Manabe I, Okamura H. RNA-methylation-dependent RNA processing controls the speed of the circadian clock. Cell. 2013 Nov 7;155(4):793-806.

Gelinas JF, Clerzius G, Shaw E, Gatignol A. Enhancement of replication of RNA viruses by ADAR1 via RNA editing and inhibition of RNA-activated protein kinase. J Virol. 2011 Sep;85(17):8460-6.

Gerber AP, Keller W. 2001. RNA editing by base deamination: more enzymes, more targets, new mysteries. Trends Biochem Sci 26: 376–384.

Geula S, Moshitch-Moshkovitz S, Dominissini D, Mansour AA, Kol N, Salmon-Divon M, Hershkovitz V, Peer E, Mor N, Manor YS, Ben-Haim MS, Eyal E, Yunger S, Pinto Y, Jaitin DA, Viukov S, Rais Y, Krupalnik V, Chomsky E, Zerbib M, Maza I, Rechavi Y, Massarwa R, Hanna S, Amit I, Levanon EY, Amariglio N, Stern-Ginossar N, Novershtern N, Rechavi G, Hanna JH. Stem cells. m6A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation. Science. 2015 Feb 27;347(6225):1002-6.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, Brown JB, Cherbas L, Davis CA, Dobin A, Li R, Lin W, Malone JH, Mattiuzzo NR, Miller D, Sturgill D, Tuch BB, Zaleski C, Zhang D, Blanchette M, Dudoit S, Eads B, Green RE, Hammonds A, Jiang L, Kapranov P, Langton L, Perrimon N, Sandler JE, Wan KH, Willingham A, Zhang Y, Zou Y, Andrews J, Bickel PJ, Brenner SE, Brent MR, Cherbas P, Gingeras TR, Hoskins RA, Kaufman TC, Oliver B, Celniker SE. The developmental transcriptome of Drosophila melanogaster. Nature. 2011 Mar 24;471(7339):473-9.

Greer EL, Blanco MA, Gu L, Sendinc E, Liu J, Aristizábal-Corrales D, Hsu CH, Aravind L, He C, Shi Y. DNA Methylation on N6-Adenine in C. elegans. Cell. 2015 Apr 29. pii: S0092-8674(15)00422-5.

Gu T, Buaas FW, Simons AK, Ackert-Bicknell CL, Braun RE, Hibbs MA. Canonical A-to-I and C-to-U RNA editing is enriched at 3′UTRs and microRNA target sites in multiple mouse tissues. PLoS One 2012, 7:e33720.

Harper JE, Miceli SM, Roberts RJ, Manley JL. Sequence specificity of the human mRNA N6-adenosine methylase in vitro. Nucleic Acids Res. 1990 Oct 11;18(19):5735-41.

Heiss NS, Knight SW, Vulliamy TJ, Klauck SM, Wiemann S, Mason PJ, Poustka A, Dokal I. X-linked dyskeratosis congenita is caused by mutations in a highly conserved gene with putative nucleolar functions. Nat Genet. 1998 May;19(1):32-8.

Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell. 2013 Apr 25;153(3):654-65.

Higuchi M, Maas S, Single FN, Hartner J, Rozov A, Burnashev N, Feldmeyer D, Sprengel R, Seeburg PH. Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. Nature. 2000 Jul 6;406(6791):78-81.

Hirano G, Izumi H, Yasuniwa Y, Shimajiri S, Ke-Yong W, Sasagiri Y, Kusaba H, Matsumoto K, Hasegawa T, Akimoto M, Akashi K, Kohno K. Involvement of riboflavin kinase expression in cellular sensitivity against cisplatin. Int J Oncol 2011, 38:893-902.

Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. PLoS Biol 2008, 6:e255.

Holden P, Horton WA. Crude subcellular fractionation of cultured mammalian cell lines. BMC Res Notes 2009, 2:243.

Hundley HA, Krauchuk AA, Bass BL. C. elegans and H. sapiens mRNAs with edited 3′ UTRs are present on polysomes. RNA 2008, 14:2050-2060.

Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, de Castro E, Coggill P, Corbett M, Das U, Daugherty L, Duquenne L, Finn RD, Fraser M, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, McMenamin C, Mi H, Mutowo-Muellenet P, Mulder N, Natale D, Orengo C, Pesseat S, Punta M, Quinn AF, Rivoire C, Sangrador-Vegas A, Selengut JD, Sigrist CJ, Scheremetjew M, Tate J, Thimmajanarthanan M, Thomas PD, Wu CH, Yeats C, Yong SY. InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res. 2012 Jan;40(Database issue):D306-12.

Hussain S, Sajini AA, Blanco S, Dietmann S, Lombard P, Sugimoto Y, Paramor M, Gleeson JG, Odom DT, Ule J, Frye M. NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs. Cell Rep. 2013 Jul 25;4(2):255-61.

Iizasa H, Nishikura K. 2009. A new function for the RNA-editing enzyme ADAR. Nat Immunol 10: 16–18.

Jia G, Fu Y, Zhao X, Dai Q, Zheng G, Yang Y, Yi C, Lindahl T, Pan T, Yang YG, He C. N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. Nat Chem Biol. 2011 Oct 16;7(12):885-7.

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.

Karijolich J, Yu YT. Converting nonsense codons into sense codons by targeted pseudouridylation. Nature. 2011 Jun 15;474(7351):395-8. doi: 10.1038/nature10165.

Kariko K, Buckstein M, Ni H, Weissman D. Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA. Immunity. 2005 Aug;23(2):165-75.

Kawahara Y, Zinshteyn B, Sethupathy P, Iizasa H, Hatzigeorgiou AG, Nishikura K. Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. Science. 2007 Feb 23;315(5815):1137-40.

Keegan LP, McGurk L, Palavicini JP, Brindle J, Paro S, Li X, Rosenthal JJ, O'Connell MA. Functional conservation in human and Drosophila of Metazoan ADAR2 involved in RNA editing: loss of ADAR1 in insects. Nucleic Acids Res. 2011 Sep 1;39(16):7249-62.

Khoddami V, Cairns BR. Identification of direct targets and modified bases of RNA cytosine methyltransferases. Nat Biotechnol. 2013 May;31(5):458-64.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013 Apr 25;14(4):R36.

Kim DD, Kim TT, Walsh T, Kobayashi Y, Matise TC, Buyske S, Gabriel A. 2004. Widespread RNA editing of embedded Alu elements in the human transcriptome. Genome Res 14: 1719–1725.

Kim U, Wang Y, Sanford T, Zeng Y, Nishikura K. 1994. Molecular cloning of cDNA for double-stranded RNA adenosine deaminase, a candidate enzyme for nuclear RNA editing. Proc Natl Acad Sci 91: 11457–11461.

Kiran A, Baranov PV. 2010. DARNED: A DAtabase of RNa EDiting in humans. Bioinformatics 26: 772–776.

Kleinman CL, Majewski J. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science. 2012 Mar 16;335(6074):1302; author reply 1302. doi: 10.1126/science.1209658.

Kumar M, Carmichael GC. 1997. Nuclear antisense RNA induces extensive adenosine modifications and nuclear retention of target transcripts. Proc Natl Acad Sci 94: 3542–3547.

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, 193:1107-1115.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357-9.

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009, 10:R25.

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, 266:20550-20554.

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

Lehmann KA, Bass BL. The importance of internal loops within RNA substrates of ADAR1. J Mol Biol. 1999 Aug 6;291(1):1-13.

Lellek H, Kirsten R, Diehl I, Apostel F, Buck F, Greeve J. Purification and molecular cloning of a novel essential component of the apolipoprotein B mRNA editing enzyme-complex. J Biol Chem 2000, 275:19848-19856.

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9.

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.

Li Q, Brown JB, Huang H, Bickel PJ. Measuring reproducibility of high- throughput experiments. The Annals of Applied Statistics. 5, no. 3, 1752-1779.

Liberski AR, Al-Noubi MN, Rahman ZH, Halabi NM, Dib SS, Al-Mismar R, Billing AM, Krishnankutty R, Ahmad FS, Raynaud CM, Rafii A, Engholm-Keller K, Graumann J. Adaptation of a commonly used, chemically defined medium for human embryonic stem cells to stable isotope labeling with amino acids in cell culture. J Proteome Res 2013, 12:3233-3245.

Lin L, Shen S, Tye A, Cai JJ, Jiang P, Davidson BL, Xing Y. Diverse splicing patterns of exonized Alu elements in human tissues. PLoS Genet. 2008 Oct 17;4(10):e1000225.

Lin W, Piskol R, Tan MH, Li JB. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science. 2012 Mar 16;335(6074):1302; author reply 1302.

Liu J, Yue Y, Han D, Wang X, Fu Y, Zhang L, Jia G, Yu M, Lu Z, Deng X, Dai Q, Chen W, He C. A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. Nat Chem Biol. 2014 Feb;10(2):93-5.

Liu N, Dai Q, Zheng G, He C, Parisien M, Pan T. N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. Nature. 2015 Feb 26;518(7540):560-4.

Liu N, Parisien M, Dai Q, Zheng G, He C, Pan T. Probing N6-methyladenosine RNA modification status at single nucleotide resolution in mRNA and long noncoding RNA. RNA. 2013 Dec;19(12):1848-56.

Luciano DJ, Mirsky H, Vendetti NJ, Maas S. 2004. RNA editing of a miRNA precursor. RNA 10: 1174–1177.

Maas S, Kawahara Y, Tamburro KM, Nishikura K. 2006. A-to-I RNA editing and human disease. RNA Biol 3: 1–9.

Maas S. Gene regulation through RNA editing. Discov Med. 2010 Nov;10(54):379-86.

Machnicka MA, Milanowska K, Osman Oglou O, Purta E, Kurkowska M, Olchowik A, Januszewski W, Kalinowski S, Dunin-Horkawicz S, Rother KM, Helm M, Bujnicki JM, Grosjean H. MODOMICS: a database of RNA modification pathways--2013 update. Nucleic Acids Res. 2013 Jan;41(Database issue):D262-7. doi: 10.1093/nar/gks1007.

Makishima M, Lu TT, Xie W, Whitfield GK, Domoto H, Evans RM, Haussler MR, Mangelsdorf DJ. Vitamin D receptor as an intestinal bile acid sensor. Science 2002, 296:1313-1316.

Mannion NM, Greenwood SM, Young R, Cox S, Brindle J, Read D, Nellåker C, Vesely C, Ponting CP, McLaughlin PJ, Jantsch MF, Dorin J, Adams IR, Scadden AD, Ohman M, Keegan LP, O'Connell MA. The RNA-editing enzyme ADAR1 controls innate immune responses to RNA. Cell Rep. 2014 Nov 20;9(4):1482-94.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 28: 495–501.

Mehta A, Kinter MT, Sherman NE, Driscoll DM. Molecular cloning of apobec-1 complementation factor, a novel RNA-binding protein involved in the editing of apolipoprotein B mRNA. Mol Cell Biol 2000, 20:1846-1854.

Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SR. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. Cell. 2012 Jun 22;149(7):1635-46.

Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol 2011, 12:R112.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Larch RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. 2010. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464: 773–777.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008, 5:621-628.

Nishikura K, Yoo C, Kim U, Murray JM, Estes PA, Cash FE, Liebhaber SA. 1991. Substrate specificity of the dsRNA unwinding/modifying activity. EMBO J 10: 3523–3532. Medline

Nishikura K. Editor meets silencer: Crosstalk between RNA editing and RNA interference. Nat Rev Mol Cell Biol. 2006. 7: 919–931.

Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem. 2010;79:321-49.

Norkina O, Kaur S, Ziemer D, De Lisle RC. Inflammation of the cystic fibrosis mouse small intestine. Am J Physiol Gastrointest Liver Physiol 2004, 286:G1032-G1041.

Page RD. 2002. Visualizing phylogenetic trees using TreeView. Curr Protoc Bioinformatics 6.2.1–6.2.15.

Palladino MJ, Keegan LP, O'Connell MA, Reenan RA. dADAR, a Drosophila double-stranded RNA-specific adenosine deaminase is highly developmentally regulated and is itself a target for RNA editing. RNA. 2000 Jul;6(7):1004-18.

Park E, Williams B, Wold BJ, Mortazavi A. RNA editing in the human ENCODE RNA-seq data. Genome Res 2012, 22:1626-1633.

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.

Paz N, Levanon EY, Amariglio N, Heimberger AB, Ram Z, Constantini S, Barbash ZS, Adamsky K, Safran M, Hirschberg A, et al. 2007. Altered adenosine-to-inosine RNA editing in human cancer. Genome Res 17: 1586–1595.

Peng PL, Zhong X, Tu W, Soundarapandian MM, Molner P, Zhu D, Lau L, Liu S, Liu F, Lu Y. 2006. ADARB1-dependent RNA editing of AMPA receptor subunit GluR2 determines vulnerability of neurons in forebrain ischemia. Neuron 49: 719–733.

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, 30:253-260.

Pickrell JK, Gilad Y, Pritchard JK. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science. 2012 Mar 16;335(6074):1302; author reply 1302.

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.

Prasanth KV, Prasanth SG, Xuan Z, Hearn S, Freier SM, Bennett CF, Zhang MQ, Spector DL. 2005. Regulating gene expression through RNA nuclear retention. Cell 123: 249–263.

Ramaswami G, Zhang R, Piskol R, Keegan LP, Deng P, O'Connell MA, Li JB. Identifying RNA editing sites using RNA sequencing data alone. Nat Methods. 2013 Feb;10(2):128-32.

Rappsilber J, Ishihama Y, Mann M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. Anal Chem 2003, 75:663-670.

Rodriguez J, Menet JS, Rosbash M. Nascent-seq indicates widespread cotranscriptional RNA editing in Drosophila. Mol Cell. 2012 Jul 13;47(1):27-37.

Rogers RL, Cridland JM, Shao L, Hu TT, Andolfatto P, Thornton KR. Landscape of standing variation for tandem duplications in Drosophila yakuba and Drosophila simulans. Mol Biol Evol. 2014 Jul;31(7):1750-66.

Rosenberg BR, Dewell S, Papavasiliou FN. Identifying mRNA editing deaminase targets by RNA-Seq. Methods Mol Biol 2011, 718:103-119.

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.

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

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.

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.

Samuel CE. Adenosine deaminases acting on RNA (ADARs) are both antiviral and proviral. Virology. 2011 Mar 15;411(2):180-93.

Schrider DR, Gout JF, Hahn MW. Very few RNA and DNA sequence differences in the human transcriptome. PLoS One. 2011;6(10):e25842.

Schwartz S, Bernstein DA, Mumbach MR, Jovanovic M, Herbst RH, León-Ricardo BX, Engreitz JM, Guttman M, Satija R, Lander ES, Fink G, Regev A. Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. Cell. 2014 Sep 25;159(1):148-62.

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.

Segditsas S, Sieber O, Deheragoda M, East P, Rowan A, Jeffery R, Nye E, Clark S, Spencer-Dene B, Stamp G, Poulsom R, Suraweera N, Silver A, Ilyas M, Tomlinson I. Putative direct and indirect Wnt targets identified through consistent gene expression changes in APC-mutant intestinal adenomas from humans and mice. Hum Mol Genet 2008, 17:3864-3875.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: The NCBI database of genetic variation. Nucleic Acids Res 29: 308–311.

Smith HC, Bennett RP, Kizilyer A, McDougall WM, Prohaska KM. Functions and regulation of the APOBEC family of proteins. Semin Cell Dev Biol 2012, 23:258-268.

Spitale RC, Flynn RA, Zhang QC, Crisalli P, Lee B, Jung JW, Kuchelmeister HY, Batista PJ, Torre EA, Kool ET, Chang HY. Structural imprints in vivo decode RNA regulatory mechanisms. Nature. 2015 Mar 26;519(7544):486-90.

Squires JE, Patel HR, Nousch M, Sibbritt T, Humphreys DT, Parker BJ, Suter CM, Preiss T. Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. Nucleic Acids Res. 2012 Jun;40(11):5023-33.

St Laurent G 3rd, Savva YA, Reenan R. Enhancing non-coding RNA information content with ADAR editing. Neurosci Lett 2009, 466:89-98.

St Laurent G, Tackett MR, Nechkin S, Shtokalo D, Antonets D, Savva YA, Maloney R, Kapranov P, Lawrence CE, Reenan RA. Genome-wide analysis of A-to-I RNA editing by single-molecule sequencing in Drosophila. Nat Struct Mol Biol. 2013 Nov;20(11):1333-9.

Storer BE, Kim C. Exact Properties of Some Exact Test Statistics for Comparing 2 Binomial Proportions. Journal of the American Statistical Association. 85:146-155.

Taylor DR, Puig M, Darnell ME, Mihalik K, Feinstone SM. 2005. New antiviral pathway that mediates hepatitis C virus replicon interferon sensitivity through ADAR. J Virol 79: 6291–6298.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.

The R Project for Statistical Computing [http://www.r-project.org/index.html webcite]

Tinsley HN, Gary BD, Keeton AB, Zhang W, Abadi AH, Reynolds RC, Piazza GA. Sulindac sulfide selectively inhibits growth and induces apoptosis of human breast tumor cells by phosphodiesterase 5 inhibition, elevation of cyclic GMP, and activation of protein kinase G. Mol Cancer Ther 2009, 8:3331-3340.

Tobias ES, Hurlstone AF, MacKenzie E, McFarlane R, Black DM. The TES gene at 7q31.1 is methylated in tumours and encodes a novel growth-suppressing LIM domain protein. Oncogene 2001, 20:2844-2853.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 2012, 7:562-578.

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, 28:511-515.

Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D, Dianes JA, Sun Z, Farrah T, Bandeira N, Binz PA, Xenarios I, Eisenacher M, Mayer G, Gatto L, Campos A, Chalkley RJ, Kraus HU, Albar JP, Martinez-Bartolomé S, Apweiler R, Omenn GS, Martens L, Jones AR, Hermjakob H. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol 2014, 32:223-226.

Wagner RW, Smith JE, Cooperman BS, Nishikura K. 1989. 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 86: 2647–2651.

Wang C, Hanly EK, Wheeler LW, Kaur M, McDonald KG, Newberry RD. Effect of alpha4beta7 blockade on intestinal lymphocyte subsets and lymphoid tissue development. Inflamm Bowel Dis 2010, 16:1751-1762.

Wang J, Duncan D, Shi Z, Zhang B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W77-83.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38: e164. doi: 10.1093/nar/gkq603.

Wang Q, Khillan J, Gadue P, Nishikura K. Requirement of the RNA editing deaminase ADAR1 gene for embryonic erythropoiesis. Science. 2000 Dec 1;290(5497):1765-8.

Wang Q, Miyakoda M, Yang W, Khillan J, Stachura DL, Weiss MJ, Nishikura K. Stress-induced apoptosis associated with null mutation of ADAR1 RNA editing deaminase gene. J Biol Chem. 2004 Feb 6;279(6):4952-61.

Wang Q, Zhang Z, Blackwell K, Carmichael GG. 2005. Vigilins bind to promiscuously A-to-I-edited RNAs and are involved in the formation of heterochromatin. Curr Biol 15: 384–391.

Wang X, Lu Z, Gomez A, Hon GC, Yue Y, Han D, Fu Y, Parisien M, Dai Q, Jia G, Ren B, Pan T, He C. N6-methyladenosine-dependent regulation of messenger RNA stability. Nature. 2014 Jan 2;505(7481):117-20.

WebLogo [http://weblogo.berkeley.edu/ webcite]

Wei CM, Gershowitz A, Moss B. Methylated nucleotides block 5' terminus of HeLa cell messenger RNA. Cell. 1975 Apr;4(4):379-86.

Wilcox RR. Applying contemporary statistical techniques. 1st ed. Amsterdam, Boston: Academic Press.

Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009 Jul 27;10:232. doi: 10.1186/1471-2105-10-232.

Xu C, Wang X, Liu K, Roundtree IA, Tempel W, Li Y, Lu Z, He C, Min J. Structural basis for selective binding of m6A RNA by the YTHDC1 YTH domain. Nat Chem Biol. 2014 Nov;10(11):927-9.

Yamanaka S, Balestra ME, Ferrell LD, Fan J, Arnold KS, Taylor S, Taylor JM, Innerarity TL. Apolipoprotein B mRNA-editing protein induces hepatocellular carcinoma and dysplasia in transgenic animals. Proc Natl Acad Sci U S A. 1995 Aug 29;92(18):8483-7.

Yamanaka S, Poksay KS, Arnold KS, Innerarity TL. A novel translational repressor mRNA is edited extensively in livers containing tumors caused by the transgene expression of the apoB mRNA-editing enzyme. Genes Dev 1997, 11:321-333.

Yao Z, McLeod RS. Synthesis and secretion of hepatic apolipoprotein B-containing lipoproteins. Biochim Biophys Acta. 1994 May 13;1212(2):152-66.

Yoshida M, Ukita T. Modification of nucleosides and nucleotides. 8. The reaction rates of pseudouridine residues with acrylonitrile and its relation to the secondary structure of transfer ribonucleic acid. Biochim Biophys Acta. 157:466-475.

Zarnack K, König J, Tajnik M, Martincorena I, Eustermann S, Stévant I, Reyes A, Anders S, Luscombe NM, Ule J. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. Cell. 2013 Jan 31;152(3):453-66.

Zechel JL, Doerner SK, Lager A, Tesar PJ, Heaney JD, Nadeau JH. Contrasting effects of Deadend1 (Dnd1) gain and loss of function mutations on allelic inheritance, testicular cancer, and intestinal polyposis. BMC Genet 2013, 14:54.

Zhang G, Huang H, Liu D, Cheng Y, Liu X, Zhang W, Yin R, Zhang D, Zhang P, Liu J, Li C, Liu B, Luo Y, Zhu Y, Zhang N, He S, He C, Wang H, Chen D. N6-Methyladenine DNA Modification in Drosophila. Cell. 2015 Apr 29. pii: S0092-8674(15)00435-3.

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, 106:465-475.

Zhao HQ, Zhang P, Gao H, He X, Dou Y, Huang AY, Liu XM, Ye AY, Dong MQ, Wei L. Profiling the RNA editomes of wild-type C. elegans and ADAR mutants. Genome Res. 2015 Jan;25(1):66-75.

Zheng G, Dahl JA, Niu Y, Fedorcsak P, Huang CM, Li CJ, Vågbø CB, Shi Y, Wang WL, Song SH, Lu Z, Bosmans RP, Dai Q, Hao YJ, Yang X, Zhao WM, Tong WM, Wang XJ, Bogdan F, Furu K, Fu Y, Jia G, Zhao X, Liu J, Krokan HE, Klungland A, Yang YG, He C. ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. Mol Cell. 2013 Jan 10;49(1):18-29.

Zheng H, Fu TB, Lazinski D, Taylor J. 1992. Editing on the genomic RNA of human hepatitis delta virus. J Virol 66: 4693–4697.