14

Editing Informational Content of Expressed DNA Sequences and Their Transcripts

Harold C. Smith

Overview

A preliminary annotation of eukaryotic genomes has suggested that there are far fewer genes encoding mRNA than predicted from the number of proteins expressed in cells (the proteome). In fact, the coding capacity of genomes is expanded through conditionally activated mechanisms. These mechanisms are regulated in species- and tissue-specific manners and include, for example, mutation and recombination of DNA, use of alternative promoters, alternative pre-mRNA splicing, RNA editing, alternative polyadenylation, and mRNA turnover. It is likely that a substantial fraction of the genome encodes processes that diversify expressed sequences. The increasing awareness and acceptance that a simple linear analysis of DNA sequences is not sufficient to annotate the genome's full coding capacity represents a significant change in the scope of hypotheses that will drive research in the twenty-first century.

This chapter discusses select aspects of RNA (and DNA editing) with a goal of providing the reader with a sense of the exciting new research frontiers that have opened due to developments in this area. RNA editing is defined as a co- or post-transcriptional process that changes the nucleotide sequence in RNA from that encoded in DNA, through mechanisms that involve either base modification, substitution, deletion, or insertion.

Discovery of RNA Editing

Once the table of codons was described in the 1960s, researchers assumed that they could simply translate a DNA sequence into the sequence of amino acids in proteins. This view of the informational content of the genome was shaken up by the discovery of intervening sequences in the late 1970s. However, once introns were incorporated into our thinking, exon splicing and the removal of intervening noncoding

intronic sequences was considered by and large the major means of diversifying the proteome. In this mind-set, once splice sites were identified in a gene, all of the protein-coding information could be translated from the linear DNA sequence. Yet the mechanism of coding for several proteins or protein variants remained enigmatic until mRNA editing was discovered. Unlike numerous covalent modifications of the sugar or base moiety of nucleotides in mRNA, ribosomal RNA, and transfer RNA (known generally as RNA modification) that already were known at this time, RNA editing had the potential to directly change the sequence and/or half-life of the protein encoded by the mRNAs.[1,2]

The potentially broad significance of the discovery by Rob Benne and colleagues[3] of RNA editing in flagellated protozoa known as kinetoplastids (referred to as Trypanosomes)was not immediately appreciated, although this discovery demonstrated unprecedented posttranscriptional uridine nucleotide insertions in mitochondrial mRNA. These edited nucleotides were not explicitly encoded by the DNA sequence that was transcribed into mRNA, yet they were absolutely required to induce frame shifts that established the correct reading frame of several mRNAs. Within the next few years, Stuart and colleagues[4] made the startling and widely noted discoveries that in some cases 50% of the protein coding sequence in mRNAs from Trypanosome mitochondria were added through editing, and in fact some of the DNA-encoded uridines were deleted. Mitochondrial genomic DNA revealed few or no full-length sequences corresponding to the mature mitochondrial mRNA sequences that encoded several essential proteins in the respiratory chain of enzymes. Neither were these proteins encoded in nuclear genomic DNA. Instead, mature mRNAs encoding full-length and functional protein sequences were constructed from partial or rudiments of mRNA encoded in the mitochondrial genome (genomic partial genes known as crypto genes) that were expressed and subsequently processed by multiple U insertions and deletions. Furthermore, each mRNA contained numerous editing sites, and each site was specified by a unique *trans*-acting small RNA (referred to as guide RNA, gRNA) containing complementary sequence to the mRNA just 3′ of the editing site, a region in the mRNA known as the anchor sequence (the mechanism is summarized in figure 14.1). The mitochondrial genome of Trypanosomes consists of cantenated maxi and mini circular DNAs[5]. Crypto genes (and most of the mitochondrial transcribed sequences) are encoded within the maxi circular genome whereas guide RNAs are encoded largely on the mini-circles.

Fewer than ten years after Benne's seminal work, the broad scope of RNA editing was evident as it had been discovered to affect numerous RNAs in phylogenetically diverse organisms including: the mRNA encoding mammalian transmembrane glutamate-gated ion channels and apolipoprotein B lipid carrying protein, plant mRNAs from chloroplasts and mitochondria, RNA viral genomes and numerous classes of RNAs in slime molds, amoeba, and yeast (select mechanisms are summarized in table 14.1 and figure 14.1). These editing events were in many instances more subtle than the extensive insertions seen in kinetoplastid mitochondrial mRNAs, and in most cases involved modification or substitution of individual nucleotides, resulting in various nucleotide transitions and transversions.[1,2]

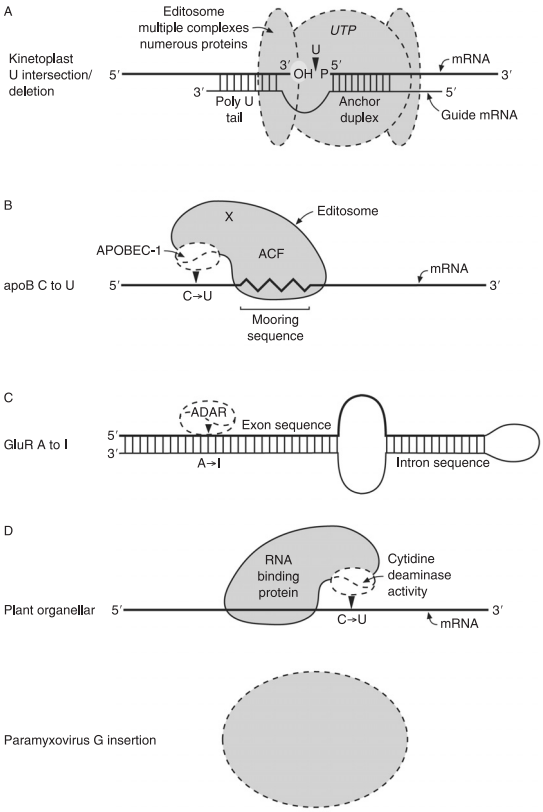

Figure 14.1. Models of select mRNA editing mechanisms. Macromolecular complexes involved in RNA editing are shown for a few mechanisms: only the most general aspects of each editing mechanism are indicated. For most editing mechanisms, the protein composition of the editosomes has not been fully characterized. For each model, an example of an edited RNA substrate or the organelle in which a group of RNAs are edited is stated to the left. (a) The editing complex or editosome for U insertion or deletion consists of multiple subcomplexes, each containing several proteins (suggested as gray ovals) and involving distinct enzymatic activities for insertion and deletion editing. The anchor duplex determines the site of editing, and mismatches between the guide RNA and the substrate (looped out region) are thought to determine the actual nucleotide position of editing. For other editing mechanisms (b–d), the part of the editosome involved in editing site recognition and binding to the catalytic subunit are shown in gray and the catalytic subunit is indicated separately. For apoB mRNA editing (b), the RNA binding protein ACF (see figure 14.2) binds to the 11 nucleotide mooring sequence and binds and positions APOBEC-1 for editing the appropriate C. The deaminases for C to U and A to I editing function as dimers. ADARs bind to double-stranded RNA (c) and deaminate A to I within duplex regions (shown here as exon sequence in duplex with an adjacent and intron sequence). ADARs have autonomous double-stranded RNA binding activity. Therefore, unlike apoB or plant organellar C to U editing (c), A to I editing is believed not to require an auxiliary protein. Bold arrowheads indicate examples of the consequence of editing, which are shown as CU (C to U), A I (A to I), U (uridine insertion), and G (guanidine insertion).

Table 14.1. Examples where RNA is edited.

Type	*Organism*	*Edited Transcript (or Genome)*	*Mechanism*
U insertion/deletion	Kinetoplastids, *Trypanosoma, Leishmani, Crithidia, Bodonids*	mRNAs (m)	gRNA targeting site, U insertion or deletion and ligation
C insertion (also U, AA, CU, GU, GC, UA)	*Physarum*	mRNAs (m), rRNAs, tRNAs (m)	Co-transcriptional C insertion
G insertion	Paramyxoviruses (SV5, Sendai), mumps, measles	P mRNA	Co-transcriptional G insertion
A insertion	Ebola viruses	Glycoprotein mRNA	Unknown
GA deletion	Rats	vasopressin mRNA (n)	Unknown
C to U	Plants	mRNAs (c), (m), numerous mRNAs at multiple sites	C-deamination
	Physarum	cox1 mRNA (m)	C-deamination
	Mammals	Gly→Asp tRNA anticodon	C-deamination
	Mammals	ApoB mRNA (n), Gln→STOP NF-1 mRNA (n), Arg→STOP	C-deamination
	Mammals	$tRNA^{Asp}$ (n) (adjacent to the anticodon loop)	C-deamination
U to C	Land plants	mRNAs (c), (m)	U-deamination
	Mammals	WT-1 mRNA (n), Leu→Pro	Unknown
		$tRNA^{Asp}$ (n) (adjacent to the anticodon loop)	Unknown
		C18 ORF 1 mRNA 5′ UTR	Unknown
A to I	Vertebrates, fly	GluR-B,5,6 (n), Gln→Arg	A-deamination
		GluR-B,C,D (n), Arg→Gly	

(Continued)

Table 14.1. (Continued)

Type	*Organism*	*Edited Transcript (or Genome)*	*Mechanism*
		GluR-6 (n), Tyr→Cys (n), Ile→Val	
		5-HT_{2C}R (n), Ile→Val, Asn→Ser	
		PTPN6 phosphatase, ablates splicing branch site	A-deamination
		Endothelin B receptor Glu→Arg	A-deamination
		5′ and 3′ UTRs alu sequences	A-deamination
	Hepatitis delta virus	Antigenome, STOP→Trp	A-deamination
	Mammals, squid, fly	Kv2 K^+ channel mRNA	A-deamination
	Rats	α-2,6-Sialyltransferase, Tyr→Cys	Unknown
	Bee, fly, moth, worm	Numerous exon and intron sequences	A-deamination
C to A, A to G, U to G, U to A	*Acanthomoeba*	tRNAs (m)	Unknown

Examples of RNA editing in organisms and viruses were taken from the cited literature in the text (refs 1–11, 15, 17, 21, 25–27, 33, 43). (c), chloroplast; (m), mitochondria; (n), nucleus.

The discovery of plant mRNA editing is of particular note as it brought to light tens to hundreds of editing sites within mRNAs from chloroplasts and mitochondria, respectively. So extensive were these editing events that prior to the discovery of mRNA editing, the disparity between mRNA (cDNA) and organelle genomic sequences had led researchers to speculate that plant organelles used a different genetic code. With the discovery of RNA editing, comparisons of expressed homologous mRNA sequences from either chloroplasts or mitochondria in different species suggested that editing frequently served to generate amino acid substitutions necessary for functional proteins.[2,6–8]

The discovery of mRNA editing in mammalian tissues had the additional effect of establishing that protein expression could be regulated not only through the control of transcription, translation, and mRNA half-life but also through mRNA editing. Perhaps most remarkable was the example of the glutamate-gated calcium channels of the central nervous system (controlling virtually all levels of human cognitive and motor activity).[9,10] Each receptor protein serves as one of the five

subunits that interact to establish a transmembrane channel for calcium within the postsynaptic membrane. These channels are regulated (gated) by the neurotransmitter glutamate. A direct translation of the genomic DNA sequence for the subunits positions a glutamine at a key position within the channel. Channels with glutamine in this position are leaky to calcium even in the absence of glutamate signaling. A to I (inosine) editing (changing CAG to CIG, which is read as CGG) changes this glutamine to arginine, thereby placing a positive charge in the channel. A positive charge in this position helps to exclude calcium, thus closing the channel. Signaling by glutamate during synaptic activity opens the channel by inducing appropriate conformational changes.[1,9–11] Other sites of receptor subunit mRNA editing have been identified that affect the rate with which membranes returned to their resting potential following an action potential.

Flies and worms also require A to I mRNA editing of their homologous channel receptor subunits for the neuronal activity necessary to coordinate motor functions and food foraging.[12,13] Discoveries such as these underscored the underappreciated dependence of organisms on mRNA editing for appropriate protein function.

At about the same time as the discovery of A to I mRNA editing, the mRNA encoding apolipoprotein B (apoB) in mammals was discovered to be C to U edited.[14,15] Virtually 100% of all apoB mRNA is edited within the epithelial cells (enterocytes) that line the small intestines of all mammals and a variable and regulated proportion of apoB mRNA is edited in the liver (hepatocytes) of some species.[16] Editing converts a cytidine at nucleotide 6666 of a CAA glutamine codon to a UAA stop codon, thereby enabling both full-length (apoB100) and truncated (apoB48) variants of apoB protein to be expressed from a single gene.

ApoB48 is stored in the enterocyte and assembled with dietary lipids as the structural protein core of chylomicrons. These are secreted into the lymphatic ducts draining the small intestine and enter the bloodstream, from which they are rapidly taken up by the liver. Chylomicron derived lipids are reassembled in the liver as very low density lipoproteins (VLDLs) on apoB100 protein, which are secreted into the circulation for peripheral tissue utilization. In several mammals, apoB mRNA editing also occurs in liver[16] where, unlike intestine, apoB mRNA editing is regulated to determine the proportion of edited apoB mRNA as well as the amount of secreted B48 VLDLs.[17,18] B48- and B100-containing particles differ greatly in the amount of lipid that they can transport (B48-containing particles have a significantly higher capacity and hence it is the protein of choice for transporting dietary lipid from the intestine). Hepatic VLDLs are assembled and secreted only with B100 protein cores in humans (or with B100 and B48 in other species).[16] B48 lacks a low density lipoprotein (LDL) receptor binding domain, and therefore the body "manages" VLDLs that contain B48 differently than those containing B100. VLDLs assembled with B100 have a longer half-life in the blood stream and as a consequence are digested by liver and bloodstream lipases, rendering them to protein and cholesterol rich LDL. Elevated abundance of LDL in the blood is an atherosclerotic risk factor. ApoB48 VLDL is cleared from the blood more rapidly than apoB100 VLDL and is not metabolized to LDL.[19] For this reason, hepatic apoB mRNA editing has been considered as a means of reducing the risk of atherogenic disease.

ApoB mRNA editing catalytic subunit 1 (APOBEC-1) is the sole cytidine deaminase responsible for editing apoB mRNA.[20,21] Although APOBEC-1 can bind and deaminate free cytidine nucleoside or nucleotide substrates, as well as bind weakly to AU-rich RNA sequences, it cannot bind specifically to, nor under physiological temperature and salt concentrations edit, *apoB* RNA.[22,23] In cells, site-specific apoB mRNA editing requires an editing complex (or C to U editosome) consisting minimally of an APOBEC-1 homodimer[24] interacting with a single-stranded-RNA binding protein known as "APOBEC-1 complementation factor" (ACF), which binds to an RNA editing site recognition motif 3′ of C6666 (figure 14.1).[25–26]

Throughout the 1990s, many more examples of base modification, substitution, insertion, and deletion RNA editing were brought to light. In addition, the protein coding capacity of some viruses with RNA genomes were found to be altered by RNA editing.[1,2,11,28-30] Table 14.1 lists a few examples of RNA editing in organisms and tissues (described in greater detail below and in reviews cited in this chapter). The term "RNA editing" was extended conceptually to include those modifications of nucleotides in tRNA that resulted in a change in the amino acid coded during translation. In some instances editing modified nucleotides in the acceptor stem of tRNAs and thereby changed the specificity of amino acylation (amino acid charging of the tRNA), while in other instances amino acylation remained the same after editing but the anticodon sequence was edited resulting in altered codon recognition.

Distinguished from editing events that result in a change in protein translation are mechanistically related processes that result in nucleotide modifications in mRNA, rRNA, and tRNA. Modification typically affects RNA stability, processing, secondary structure, interaction with other RNAs or proteins and/or affects RNA subcellular localization. As these are generally considered as RNA modification, not editing (i.e., they do not change the protein-coding specificity of the mRNA), they will be discussed only to a limited extent in this chapter. The interested reader is referred to recent texts that broadly addresses RNA and nucleoside/nucleotide modifications.[1,2,11]

mRNA Editing Throughout Time

A simple statement concerning the selective forces acting on mRNA editing is unlikely to be accurate because of the diverse mechanisms involved and the breadth of species with one or more forms of mRNA editing. However, it seems likely that mRNA editing mechanisms must have their roots at the very origin of life due to their apparent relationship to nucleotide modification, which some believe dates back to the "RNA World." (The "RNA world" hypothesis,[31] originally proposed by Gilbert, is that RNA preceded DNA as the genetic material.) A case will be made, from the findings described below, that the machinery for mRNA editing emerged through gene duplication and divergence of preexisting purine and pyrimidine base and nucleoside/nucleotide modifying enzymes.[1,2,32]

Genomic mutations that result in impaired protein function will either be deleterious to an organism and become limited within populations, or will be tolerated because the function, while decreased, remains in an acceptable range or is compensated by redundant pathways. It might be speculated that mRNA editing would

render these mutations neutral by "correcting" them at the level of the transcriptome. The possibility also has to be considered that editing is not specifically a 'repair' process, but instead enables protein variants to be expressed with a range of activities and, as discussed below, this might have conferred a more robust phenotype to organisms during their evolution.

The bias inherent in each mRNA-editing enzyme for substrate recognition and site-selective editing that we see today (e.g., nearest neighbor nucleotide preferences of editing enzymes)[33–35] may have been acquired by mutation of modification enzymes and selection of nascent activities with the capacity to edit specific RNA sequences. Presumed in this discussion is that mRNA editing provided a selective advantage in the face of ongoing mutagenesis. In this model, orphan editing activities may have emerged spontaneously and were in some instances, maintained through positive selection. The emergence of genomic mutations that could be corrected by mRNA editing and their propagation throughout populations would have fixed some forms of mRNA editing (or tRNA editing) in modern-day organisms.

Consistent with this "environmental selection pressure" hypothesis, C to U mRNA editing has not been observed in aquatic plants but is evident in the organelles of all dry land based plants. An enriched oxygen environment, desiccation, enhanced radiation exposure, or other changes associated with dry land may have promoted or permitted mutations that could have selected for mRNA editing as a "corrective" capacity. Homologous mRNAs from modern-day monocots and dicots, however, are not edited in all species.[1,2,7,36] A high frequency of C to U editing sites found in the mRNA of one species were genomically encoded as T in other species, while some discrepancies in editing site utilization involved cytidines at the third nucleotide position within codons (wobble base pairing).[7]

The findings suggested a high incidence of genomic nucleotide transitions at positions corresponding to plant organellar editing sites. It is uncertain whether these discrepancies are the result of forward (T to C) or back (C to T) mutations. It is likely, however, that the mutability of these sites maintains selection pressure on the maintenance of an mRNA editing capacity and its evolution as new targets for editing emerge. In this regard, editing site recognition in plants requires unique sequences immediately 5′ of the edited C (mammalian C to U mRNA editing requires the 11 nucleotide mooring sequence immediately 3′ of the edited C) (figure 14.1). This editing site recognition sequence is not the same for chloroplast mRNAs and mitochondrial mRNAs nor can the mRNAs from one organelle be edited when expressed in the other. Additional studies evaluating editing enzyme–substrate relationships will be necessary if we are to understand the selection pressures driving the evolution of C to U editing within plant chloroplasts and mitochondria.

Nucleotide Modifications in RNA

A considerable number and diversity of enzymes are dedicated to modifying nucleotides in tRNA and rRNA in bacteria, archae, and eukaryotes, as is apparent in the observation that there nearly 100 different posttranscriptional RNA modifications of pyrimidines, purines, and of the 2′ hydroxyl moieties of ribose.[1,2] Over half of the

different types of RNA modifications found in bacteria also are observed in various organisms found in the Archaea and Eukarea kingdoms. While the catalytic domains of the enzymes carrying out similar modifications of RNA within these kingdoms show considerable homology and structural conservation, the RNA substrates and the positions of modification within homologous substrates are only occasionally similar. Some RNA modifications of tRNAs affect codon sense (now considered tRNA editing) (table 14.1), while other modifications may stabilize RNA secondary structure and are required for tRNA processing or improve translation efficiency and fidelity.[1,8,32,37–39] There are examples in both modification and insertion editing where one type of modification is a prerequisite for additional modifications (frequently of a different type) at either the same nucleotide or at a another site within the RNA.[1,2,8,40]

Given that each RNA modifying enzyme interacts with unique sites within a limited range of substrates, there must have been multiple occasions during evolution where new catalytic activities emerged (or diverged from existing enzymes) with the capacity for different or broader substrate specificities. Grosjean has observed that over half of the types of RNA modifications in various organisms in Eukarea are unique to this kingdom,[1] suggesting that new modification activities have emerged or that unique selection pressure led to the retention of activities, now lost in organisms from other kingdoms. Genomic mutations that affected the structure and/or function of an essential RNA(s) would have provided selection pressure for emerging modification activities or the maintenance of activities that were until that point under neutral selection. The capacity of all life forms to carry out so many diverse RNA modifications with such a large number of enzymes remains a topic of controversy and discovery.

C/U and A/I Base Modification mRNA Editing is Unique to Eukaryotes

C to U RNA Editing

Have organisms become more or less dependent upon RNA editing over the course of evolution? As editing of mRNA has not yet been reported in Archaea or in Bacteria, it is possible that it is a unique characteristic to eukaryotes and so might have emerged rather recently (either from RNA modification processes or as a new function). Although mRNA editing mechanisms in different organisms are, in some instances, very different reactions, involving unique enzymes and auxiliary factors and occurring in different organelles (figure 14.1), we can ask, for a given mechanism of mRNA editing, how broadly across classes and orders can one find this activity, and in which species?

C to U modification mRNA editing activity has been demonstrated (or implicated by comparisons of DNA and protein sequences) in both lower and higher eukaryotes, including yeast, *Physarum*, all dry-land plants, *Caenorhabditis elegans*, mammals, and marsupials. These editing events have been shown (or are postulated) to be catalyzed by cytidine deaminases active on RNA (CDARs). Phylogenetic studies[41] and structural modeling studies[24] have suggested that CDARs are related to cytosine and cytidine deaminase, which

are found in all forms of life that use free pyrimidines or nucleoside/nucleotide as substrates.

These enzymes have homologous domains for the coordination of zinc, which is used as a Lewis acid for hydrolytic deamination of cytidine to form uridine, and a glutamic acid residue for proton shuttling during the reaction.[24,42] Crystal structure analysis suggested that nucleotide deaminases must function as dimers or tetramers because each catalytic center is composed of the residues for deamination in one subunit and a substrate coordination "flap" contributed by the other subunit.[24] Whereas deaminases that are active only on free nucleosides or nucleotides have long and inflexible flap domains, those that have the capacity to deaminate nucleotides within RNA (or DNA) have short flexible flaps. Structural studies have suggested that the evolution of CDARs must have involved changes in the flap domain that enable these enzymes to accommodate nucleic acids as substrates.

CDD1, an orphan C to U mRNA editing enzyme from yeast,[43] bears striking structural and functional homology to the mammalian C to U mRNA editing APOBEC-1, which carries out RNA editing of apoB (the major structural protein of low density lipoproteins) mRNA (table 14.1, figure 14.1). Several APOBEC-1 related proteins (ARPs)[42,44] functioning in mammalian cells as deoxycytidine deaminases on genomic[45] and viral[34,46] ssDNAs have catalytic domain folds homologous to APOBEC-1 (figure 14.2). Based upon sequence and structural homology it has been predicted that ARPs also have a flexible flap domains and a distribution of charged and hydrophobic residues within the catalytic cleft that accommodate either single-stranded RNA or single-stranded DNA substrates. A role for ARPs in mRNA editing, while highly likely, remains hypothetical.[42,45,47]

A to I RNA Editing

A to I mRNA editing is catalyzed by a family of zinc-dependent enzymes known as "adenosine deaminases active on RNA" (ADARs), which edit a large variety of mRNAs expressed in *Xenopus*, *Drosophila*, *C. elegans*, squid, and all mammals[9] (table 14.1, figure 14.1), and may function as interferon-inducible antiviral deaminases.[30] ADARs may have evolved from a primordial cytosine/cytidine deaminase.[41] The catalytic domains of both CDARs and ADARs bear greater homology to that of *E. coli* cytidine deaminase and C to U mRNA editing enzymes than they do to adenosine deaminases (figure 14.2). Similarly, adenosine deaminases active on tRNA (ADATs) have been identified in yeast and mammals.[9,48] These enzymes are homologous in their catalytic domains to ADARs and therefore also may have evolved divergently from a primordial cytidine deaminase.

As previously mentioned, ADAR editing can change the sense of codons because I base-pairs to C, and it can generate new open reading frames through the generation of translation start codons (by editing ATA to ATI) or the alteration of mRNA splice site and branch point signals.[9] ADARs have domains that require double-stranded RNAto bind, which restricts these enzymes to the targeting of adenosines within RNA secondary structure (figure 14.2). When ADAR editing sites occur within protein-coding regions of primary transcripts, the editing site typically is situated within an exon sequence that forms a duplex with its 3′ intron sequence (figure 14.1).

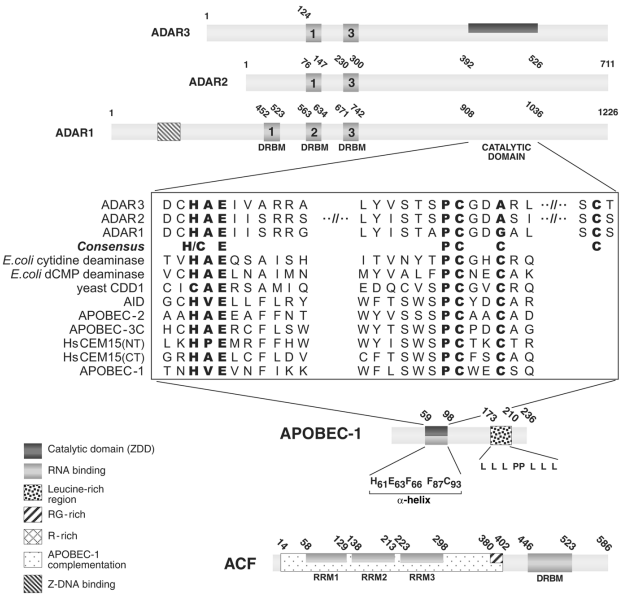

Figure 14.2. Structure-based alignments and the distribution of functional domains of mammalian editing factors. Conserved residues within the zinc-dependent deaminase domain (catalytic domain) are shown for the ADARs (adenosine deaminases active on RNA), APOBEC-1, ARPs (APOBEC-1 related proteins), and *Escherichia coli* deaminases. The catalytic domain of APOBEC-1 is characterized by three zinc coordinating amino acids (each of which can be either histidine or cysteine), a glutamic acid, a proline residue, and a conserved primary sequence spacing (key amino acids, shown in bold type as the "consensus"). The spacing of the terminal cysteine in the primary sequence of ADARs is greater than that seen in cytidine (shown within the consensus as a fourth C in bold type, but note that C to U and A to I deaminases each coordinate zinc with only three amino acids). Shown in comparison to APOBEC-1 and the consensus sequence are: the catalytic domains of deaminases that use free nucleosides/nucleotides as substrates (*E. coli* cytidine deaminase and dCMP deaminase), nucleosides/nucleotides and/or RNA as substrates (CDD1); and those of the ARPs that currently are only know to act as DNA editing enzymes (AID, CEM15) or have no known substrates (APOBEC-2 and APOBEC-3C). ADARs bind to their editing sites through double-stranded RNA binding domains (DRBMs). The indicated residues in the catalytic site of APOBEC-1 bind AU-rich RNA with weak affinity. The leucine-rich region of APOBEC-1 has been implicated in APOBEC-1 dimerization and shown to be required for editing and may be involved in interactions with APOBEC-1 complementation factor (ACF). ACF is an ssRNA binding protein that is required biologically for APOBEC-1 to find and edit apoB mRNA. The three RNA recognition motifs (RRMs) are required for mooring sequence-specific RNA binding, and these domains plus the sequences flanking them are required for APOBEC-1 interaction and complementation. APOBEC-1 complementation activity minimally depends on ACF binding to both APOBEC-1 and ACF binding to the mooring sequence. A broad APOBEC-1 complementation region is indicated on ACF that is inclusive of all regions implicated in this activity. The complete protein sequence is modeled with numbering to indicate key amino acid positions at the borders of domains and at the end of each structure, indicating the total number of amino acids in each protein.

In the case of ion channels and transmembrane receptors, A to I editing enables these complexes to form with variable ratios of subunits translated from edited and unedited mRNAs. This is because each channel is composed of five receptor subunits, which may be from either edited or unedited mRNAs and which function together to modulated calcium flux through the pore. This regulation is more like a rheostat than a switch. Editing of these receptors at other sites affects their rate of repolarization following a wave of depolarization. This results in a broader range in an organism's ability to modulate the level of response to signaling and the rate of recovery following a change in membrane potential and thus would be more robust; hence editing would be selected for and spread throughout populations (as long as the genomic sequence corresponded to the unedited version). In this regard, A to I editing of the mRNA encoding glutamate gated calcium channel receptors is ubiquitous in land animals and insects. ADAR gene knockout studies in mice and flies demonstrated that these organisms have become dependent on A to I editing activity not only for central nervous system function but also for the development of several other organ systems.[10,49–52] Interestingly, the editing site within glutamate gated calcium channel receptors are genomically encoded as G in fish but the sodium ion channels in squid require A to I editing. Editing in invertebrates, fish, and amphibians has not been well studied, but it is tempting to speculate (as appears to be the case for plant C to U mRNA editing) that in some instances genomic mutations may have selected for A to I mRNA editing activity as organisms occupied dry land.

The RNA sequences within the RNA secondary structure forming ADAR editing sites are not generic to all edited mRNAs. Consequently, once mRNA is spliced, little remains of the ADAR recognition sequence element. This has made the prediction of novel ADAR-edited mRNAs difficult. However, based upon comparisons of cDNA and genomic DNA sequences, it has been estimated that there may be over 12,000 A to I editing sites in as many as 1600 different genes in the human genome.[53] This is a conservative estimate as mRNAs were not scored as "hits" in this study unless they contained minimally three A to G discrepancies. Many presumptive editing sites were within coding regions, but the majority were within *Alu* repeats within 5′ and 3′ untranslated regions where they may have a function in the control of mRNA secondary structure and stability (see chapter 8 for more discussion of *Alu*-containing sequences). Numerous A to I mRNA editing sites also have been predicted in coding and noncoding regions of *Drosophila* mRNAs.[54] One study that compared editing sites from homologous mRNAs showed that some editing sites, including the flanking sequences that contributed to the secondary structures used by ADARs to bind to the editing sites, were highly conserved in two species of *Drosophila* separated by 61–65 million years, whereas other editing sites differed as to whether they were genomically encoded as G or A.[49]

ADAR binding to RNA secondary structure and A to I editing enables these enzymes to act as double-stranded RNA helicases (unwinding enzymes) due to their ability to alter base pairing within RNA secondary structure.[9,48] As ADARs edit adenosines within duplex regions they can disrupt local RNA secondary structure.[55] Aside from the aforementioned mRNA editing activity, ADAR ability to disrupt RNA duplexes has important implications for the regulation of gene expression

through RNA interference-mediated (RNAi) mRNA depletion.[56,57] Double-stranded RNA is required as the substrate for the enzyme "dicer," which generates through cleavage small interfering RNAs (siRNAs) (see also discussion of siRNAs in chapter 8 and RNAi in chapter 13).

siRNAs must form perfect duplexes with select mRNAs in order to target them for nucleolytic degradation. ADAR editing and unwinding activity can reduce or eliminate RNAi regulation of gene expression and thereby affect dicer activity and perhaps alter the targeting-specificity of siRNA for mRNAs.[9,58] RNAi has been described in eukaryotes ranging from *Tetrahymena*, *Drosophila*, *C. elegans*, and humans, and is involved in tissue differentiation and organism development. The functions of siRNAs now include chromatin remodeling and genomic DNA sequence deletion and reorganization.[59] RNAi is likely an ancient process and as such is another example of a function that could have contributed selection pressure on emergence and maintenance of ADAR activity in the evolution of organisms.

Nucleotide Insertion mRNA Editing: Here to Stay or Gone Tomorrow?

Uridine insertion and deletion editing of mRNAs in Trypanosome mitochondria can be traced to divergence within the phylum Euglenozoa (one of the earliest groups of organisms with mitochondria).[5] Subsequently, the requirement for editing at some sites has been lost within laboratory strains (which have been maintained as stocks for around 60 years). In these instances, the uridines have become encoded genomically and the genomic regions encoding the gRNAs responsible for targeting these editing sites have, in many instances, been selectively lost. For mRNAs that are generated through numerous editing events (pan edited mRNAs), editing of one site frequently (but not always) generates the anchor sequence for the next gRNA and a subsequent editing event. Consequently, uridine insertion and deletion editing often proceeds with a 3′ to 5′ polarity. Interestingly, the regions of edited mRNAs that have become genomically encoded in laboratory strains are mostly derived from the 3′ end of mRNAs, making possible the remaining editing events that will generate the 5′ end of these mRNAs.

It is apparent in these organisms that selection pressure can be exerted at the level of an individual editing site. Ultimately, it is only the gRNAs that are unique to each editing site, while the editosomal enzymes and structural proteins responsible for either uridine insertion or deletion are assembled (depending on the particular process) at multiple editing sites (figure 14.1).[40] On the other hand, the selection pressure that maintains the proteins involved in nucleotide insertion or deletion editing must manifest at the level of the collective requirement to establish functional mRNAs (i.e., as long as editing is required for the activity of an essential protein there will be positive selection for the genes encoding the editosomal components) (see discussion of "second-order selection" in chapter 4). While in laboratory strains that have incorporated the Us into the genome, proteins are encoded explicitly, these proteins are only implied in the genome of natural strains, through the combination of the DNA sequence, the guide RNAs, and the recognition specificities of the editing enzymes.

Insertion RNA editing of tRNA in the mitochondria of amoeba and *Physarum,* and cytidine insertion editing of all forms of mitochondrial RNA from slime molds, were discovered shortly thereafter.[1] Guanidine insertion mRNA editing of the mumps and measles virus (Paramyxo viruses) also is well documented.[60] In contrast to Trypanosome's use of guide RNAs for U insertion/deletion editing, guanidine and cytidine insertion editing are cotranscriptional processes. Polymerase slippage relative to the template strand during transcription appears to account for the insertion of additional G and/or C nucleotides in nascent transcripts. Edited mRNAs contain frame shifts that enable the expression of essential proteins integral to the virus's and organism's ability to encode essential proteins.

In the case of Paramyxo viruses, proteins expressed from edited and unedited mRNAs are believed to contribute to different stages of the viral life cycle. Perhaps more importantly, there is considerable selection pressure to maintain the capacity for editing as viral replication potentially generates irregular genome lengths that cannot be encapsulated into virions were it not for nucleotide insertion editing that restores genome lengths. This genome length restriction stems from the fact that the nucleocapsid protein has a six ribonucleotide binding capacity and viral genomes that are not exact multiples of six do not assemble functional virions (known as "the rule of six").

DNA Mutational Editing

Homologs of APOBEC-1 and ARPs are expressed in fish, *Xenopus*, birds, and all mammals.[42,44,61] At the turn of the Millennium, some of the enzymes in the ARP family were shown to induce genomic mutations by deamination of deoxycytidine to deoxyuridine.[62] This activity will be referred to as DNA editing because, unlike spontaneous dC to dU deamination, DNA editing is regulated and targeted to regions within genes. The mechanism for DNA editing is nucleotide deamination (although there may be other mechanisms yet-to-be described) largely occurring on single-stranded DNA at the sites of transcription. DNA editing is in this sense a mutation-initiating mechanism with sequence selectivity, and thus distinctly different from environmental factors that give rise to mutations due to DNA damage (see chapter 2).

Up to 1999, C deaminating activity was known only for APOBEC-1, which functions physiologically as an mRNA editing enzyme. Subsequently, under experimental conditions, APOBEC-1 itself was shown to have DNA deaminase activity.[63] This stimulated speculation that APOBEC-1 overexpression might, in addition to promiscuous RNA editing,[63] lead to DNA mutation and thereby induce neoplasias. Indeed, liver-specific transgenic overexpression of APOBEC-1 had previously been shown to induce liver carcinoma and dysplastic disease.[36,42,64] At the time, this effect was proposed to be due to hyper mRNA editing and consequent activation of oncogenic activities, but in light of the discovery of ARP DNA editing, APOBEC-1-induced genomic mutations cannot be ruled out.[42] There has been a recent flurry of interest in the possibility that ARP DNA editing may be morewidespread. For example, it recently has been observed that unregulated expression of the DNA editing ARP known as "activation induced deaminase"

(AID) (figure 14.2),which normally targets immunoglobulin genes (see below and chapters 10 and 11), can lead to oncogene activation[65] and the dysfunctional regulation of antibody expression seen in leukemias.[66] Dysregulation of AID activity also has been suggested as the mechanism by which hepatitis C virus induces genomic mutations and neoplasia.[67]

APOBEC-1 and ARPs have distinct substrates. AID does not edit apoB mRNA and overexpression of APOBEC-1 does not induce DNA mutations in immunoglobulin (Ig) genes of mammalian B lymphocytes. These findings suggest that if APOBEC-1 mutates DNA in mammalian cells, its activity on the genome is nonrandom. Given that genomic mutation frequencies are rare, it is believed that the DNA editing activity of all of the ARPs is highly regulated. In fact ARPs are expressed tissue-specifically and their activity is restricted to nucleic acids within select subcellular compartments. In the case of AID, expression is restricted to activated B lymphocytes within germinal centers (in the spleen and lymph nodes) and AID activity is focused on Ig genes in the cell nucleus (see chapters 10 and 11),[45,68] whereas APOBEC-3G/CEM15 and APOBEC-3F are expressed in T lymphocytes and have activity on HIV-1 and HIV-2 during minus strand DNA reverse transcription in the cytoplasm.[34,46]

That AID is not absolutely selective for Ig genes was seen whenexogenous non-Ig reporter genes, recombined randomly throughout the genome, were found to be mutated by AID.[69] The specter of what might happen if AID activity were not regulated properly also is raised by APOBEC-1's and other ARPs' abilities to deaminate deoxycytidine in a wide variety of DNAs under experimental conditions. When APOBEC-1 or ARPs were expressed in *E. coli* (under selection for a DNA mutator phenotype), or reacted with single-stranded DNA in vitro, numerous deoxycytidines were deaminated at a variety of DNA sequences. Although the sites of DNA modification were abundant, their distribution was unique to each enzyme as assessed by nearest neighbor sequence preferences for nucleotides immediately 5′ of the target cytidine (dT for APOBEC-1, dA/dG for AID, and dC for CEM15).[33–35] Not all dCs with appropriate flanking sequences were deaminated, suggesting a broader flanking region recognition requirement and/or that although site selectivity of DNA editing may be determined by the intrinsic bias of each enzyme, other factors determine which deoxycytidines are deaminated in vivo (as discussed in chapter 10).

Current hypotheses propose that targeting specificity results from ARP association with other macromolecular assemblies, such as those involved in DNA recombination, repair, transcription, reverse transcription, and chromatin remodeling.[47,70–73] This model is consistent with the known mechanism for apoB mRNA editing and the role the single-stranded RNA binding protein ACF[74] in determining site-specific C to U editing. It is ACF's ability to bind APOBEC-1 and specific sequences 3′ of the editing site (the mooring sequence) that restrict editing to a specific mRNA substrate (figure 14.1 and figure 14.2).[36]

The requirement for a robust immune defense system may have selected for AID and APOBEC-3G/CEM15 DNA editing. As described in chapters 10 and 11, AID is essential for increasing the repertoire and affinity of the adaptive immune response through somatic hypermutation and class switch recombination. In mammals with no or low AID, both of these processes are impaired, leading to life-threatening immunodeficiency (an autosomal recessive condition known as "hyper IgM2" which

affects one in 106 people).[45,75] Class switch recombination involves not only AID deaminase activity but also the C-terminal domain of the AID protein. The noncatalytic C-terminal domain is thought to be essential for interaction of AID with proteins such as those involved in nonhomologous DNA recombination[76] and transcription[73,77] that facilitate targeting of AID's DNA editing activity and in turn DNA repair and recombination activity to select regions of the genome for CSR and SHM.

APOBEC-3G/CEM15, APOBEC-3F, and possibly APOBEC-3B, previously referred to as "phorbolins,"[42,44] are coexpressed in human lymphoid and myeloid cells and, as is the case for APOBEC-1, can form homodimers but also heterodimers.[34] Our current understanding is that these proteins serve in host defense as antiviral deaminases, although their potential for other activities within the cell has not been explored. For example, APOBEC-3G/CEM15 and 3F deaminate deoxycytidine on HIV-1 and HIV-2 minus strand cDNA that satisfies nearest neighbor nucleotide requirements (as discussed previously). These dC to dU modifications template dG to dA mutations on the positive strand during replication, which inactivate multiple proteins essential for viral infectivity.[46,78] Unlike APOBEC-1 and other ARPs, APOBEC-3G/CEM15 and 3F establish a close proximity with viral genomes by becoming integrated within virions during their assembly.[34,79–81] With regard to the deaminase activity, homodimers of either APOBEC-1 and AID are predicted to contain two catalytic centers.[24] APOBEC-3G/CEM15 3F and 3B each have two catalytic centers (both of which have activity).[70,82,83] Homo- and heterodimers of ARPs like APOBEC-3G/CEM15 and 3F are likely to have four catalytic domains and therefore considerable combinatorial substrate targeting potential. This could provide the host cell with an adaptive advantage against a broad spectrum of viruses.

Six phorbolin genes (as well as phorbolin pseudogenes) are clustered on human chromosome 22 (mice have only one phorbolin gene on chromosome 15). Presumably these are the result of gene duplication from a primordial phorbolin gene followed by divergent evolution. The phorbolins known as APOBEC-3A, 3C, 3D, and 3E may be partial gene duplications as they each have single catalytic domain and partial C-terminal sequences like that seen in homologous regions of APOBEC-3G, 3F, 3B, APOBEC-1, APOBEC-2, and AID. The function(s) of phorbolins with one catalytic domain and APOBEC-2 remains to be determined.

HIV-1 and HIV-2 use the accessory protein known as "viral infectivity factor" (Vif) to defeat the ARP host defense. Vif binds to both APOBEC-3G/CEM15 and 3F and targets them to ubiquitination and proteolytic degradation via the proteasome.[84] Vif's interaction with APOBEC-3G/CEM15 occurs in a noncatalytic region that lies C-terminal to first catalytic domain. Interestingly, a single amino acid within this region (an aspartic acid in humans and a lysine in monkeys) provides the essential charge for the interaction of APOBEC-3G/CEM15 with Vif.[81,85] Site-directed mutagenesis has shown that this single amino acid change in an ARP changes host range of a retroviruses.[81,85,86] Due to this single amino acid difference, Vif derived from simian virus (SIV) cannot bind to human APOBEC-3G/CEM15 and vise versa, and consequently there is species-specific exclusion of APOBEC-3G/CEM15 from the virion. Consequently, this region of APOBEC-3G/CEM15, perhaps more than any other, may have constrained the extent to which Vif can mutate and still protect the virus from the ARP-based host defense.

Conclusions

The discovery of RNA editing appeared at first to be esoteric; uncovered in a small number of organisms and thought not to be mechanistically related. The development of in vitro systems helped identify the proteins involved in mRNA and DNA editing. Sequence alignments and structural comparisons of enzymes, together with molecular approaches and transgenic or gene knockout model organisms, have facilitated the identification additional enzymes and substrates, and established the biological requirements for mRNA editing. Persistent efforts by labs across the globe validated not only that mRNAs were edited but that editing was required to add diversity, and in many cases specific functionalities, to the proteome of many organisms and viruses.

Striking examples of the importance of editing are gRNA-dependent uridine insertion and deletion, the requirement of AID for immunoglobulin gene diversification and class switch recombination, the multiplicity of C to U modifications in plant organelle mRNAs, and C insertion editing of all RNAs within *Physarum* mitochondria. In some cases, such as gRNA editing, it is very clear how the genome implies the informational content through selection of U insertion sites and their lengths, while in other systems, such as C insertion editing, we do not understand the form of the genomic information that is required to direct the assembly of the final sequence. It appears that RNA modification systems have been a part of biology from its origin and mRNA editing has played an important role.

The past five years have shown dramatic progress in the areas of editosomal component identification, discerning mechanisms, and identification of novel substrates. Deaminases with genomic DNA editing capacity have taken center stage and are being carefully evaluated for the possibility that they, like APOBEC-1, may also have mRNA editing activity. Discovery of the requirement for editing enzymes in viral life cycles on the one hand, and on the other as host antiviral defense factors, and as the agent required for somatic hypermutation and class switch recombination of immunoglobulin genes during the development of the immune system, provided long sought answers in immunology. We need to learn how to identify genomic sequences that may be substrates for RNA and DNA editing systems, and to understand the selection pressures under which they have evolved. Even given our limited knowledge at this time, it is apparent that at least some C to U editing systems, such as those involved in immunity and viral defense, increase the hardiness of organisms.

The discovery of the APOBEC-1 related protein family and their DNA editing activities also brought new insight, with broad implications, as to how genomic instability can be selectively activated in certain cells to regulate diversity in the proteome or prevent viral proteomes from being expressed. It is likely that additional editing mechanisms and novel substrates will be revealed and that these too will prove to increase our appreciation for the extent to which information is implicitly, in addition to explicitly encoded, in the genome.

Acknowledgments The author acknowledges the many contributions of investigators in the field whose specific work may not have been reference due to space limitations. The author is grateful to Lynn Caporale and Joseph E. Wedekind for their many helpful suggestions in the preparation of this chapter and to Jenny M.L. Smith for the preparation of the figures. The author's efforts on this chapter and contributions to the field of RNA editing have been supported in part by grants from the National Institutes of Health, The Air Force Office of Scientific Research, The Alcoholic Beverage Medical Research foundation, The Council for Tobacco Research, and The Office of Naval Research.