1 INTRODUCTION

1.1 GENERAL INTRODUCTION

In 1944, it was confirmed that DNA is the material of inheritance. It subsequently became clear that while DNA is located in the nucleus, proteins are synthesised at discrete sites in the cytoplasm. In 1952, James Watson accounted for this discrepancy by proposing the ‘central dogma’ that genetic information is copied from DNA to RNA, and that RNA encodes protein synthesis (Gesteland, 1999). The role of RNA as the messenger molecule was confirmed by Brenner and colleagues, with the discovery of transcription of DNA into messenger RNA (mRNA) (Brenner, 1961). The sites of protein synthesis in the cytoplasm were identified as ribosomes, and shown to contain ribosomal RNA (rRNA) (Crick, 1958). Finally, Francis Crick’s ‘adaptor’ hypothesis of 1958 predicted the existence of additional RNA molecules acting as mediators between the genetic code and the encoded amino acid. Shortly afterwards these were identified and named transfer RNAs (tRNAs) (Hoagland, 2004).

In 1986, Thomas Cech demonstrated that RNA could act as a catalytic molecule (Garriga et al., 1986). This provided evidence to support the ‘RNA world’ hypothesis that life emerged from a world in which RNA was both the genetic and catalytic material (Joyce, 2002). Further support for this hypothesis was provided by the discovery that the RNA rather than the protein components of the ribosome catalyse peptide bond formation during translation (Yusupov et al., 2001). As translation is a highly conserved process, this suggests a central role for RNA in biology from the very earliest stages of evolution on earth.

It is now known that the roles of RNA extend far beyond those of mRNAs, tRNAs and rRNAs in protein synthesis (Eddy, 2001). For example, RNAs have been identified which function in RNA splicing, RNA modification and protein transport, while other RNAs actively shape the human genome by the process of retrotransposition. Of prominence among these recent discoveries is the role of double-stranded RNA in biology, in particular in gene silencing and translational repression by RNA interference (RNAi). In addition to the expanding functions ascribed to RNA, it was first noticed in the 1980s that nucleotides in RNA are subject to modification by RNA editing, and that these changes can profoundly alter the properties of the RNA (Benne et al., 1986). The process of RNA editing has subsequently been shown to be widespread in biology, and involves modification of an RNA sequence by nucleotide substitution, insertion or deletion, such that it no longer resembles that of the DNA from which it was transcribed. In this thesis, I describe a survey of the types and patterns of RNA editing in the human brain.

1.2 THE HUMAN GENOME

The human genome consists of 3.2 billion base pairs of DNA on 22 autosomal chromosomes and the sex chromosomes (X and Y). The chromosomes vary in size from the largest, chromosome 1 (279 Megabases, Mb), to the smallest, chromosome 22 (48Mb). The total number of protein coding genes in the human genome remains elusive. Estimates have fallen from 35,000 in the initial analysis of genome draft sequence to more recent estimates of 24,500 (Pennisi, 2003). The characteristics of protein coding genes in the human genome are summarised in Table 1-1. Based on the estimate of 24,500 genes, approximately 22% of the human genome is transcribed into known protein coding genes. As the average gene consists of only approximately 5% coding sequence (Table 1-1), this suggests that only 1-2% of the human genome sequence is protein coding, and that intronic RNA is by far the major transcriptional product of the genome.

Sequence class	**Genome-wide Median**	**Genome-wide Mean**
Internal exon length	122 bp	145 bp
Number of exons	7	8.8
Intron length	1,023 bp	3,365 bp
3’UTR length	400 bp	770 bp
5’UTR length	249 bp	300 bp
Coding sequence length	1,100 bp	1,340 bp
CDS length	367 aa	447 aa
Genomic extent	14kb	27 kb

Table 1-1 Characteristics of human protein coding genes (Lander et al., 2001).

It is increasingly clear that non-protein coding RNAs constitute a large portion of the transcriptional output of the human genome. The level of transcription from human chromosomes 21 and 22 is an order of magnitude higher than can be accounted for by known or predicted exons (Kapranov et al., 2002), and thousands of putative non-coding RNAs have been identified in cDNA libraries from mouse (Numata et al., 2003) and human (Ota et al., 2004).

The human genome is approximately 20 - 30 times larger than that of the invertebrates *Drosophila melanogaster* (137Mb) and *Caenorhabditis elegans* (97Mb) and over 200 times larger than that of the yeast *Saccharomyces cerevisiae* (12Mb). There is only a small increment in gene number compared with *Drosophila* (~14,000) and *C. elegans* (~19,000), and five times the number of genes in *S. cerevisiae* (~6,300). The human genome is more similar in size and gene number to other mammalian genomes. For example the 2.5Gb mouse genome contains about 30,000 genes, with 99% having direct counterparts in humans (Waterston et al., 2002).

1.2.1 Transposable elements in the human genome

Repetitive DNA accounts for at least 50% of the human genome sequence. The majority of this sequence (approximately 45% of the genome) is derived from transposable elements, with the remaining repetitive sequence from simple repeats, and large scale segmental duplications of DNA (Lander et al., 2001). Transcribed repeat elements are associated with RNA editing (Morse et al., 2002), and therefore are described here in some detail.

Transposable elements are DNA sequences which are capable of replication and insertion at new locations in the genome. There are four classes of mobile elements in the human genome (Table 1-2). Three of these transpose through RNA intermediates and are classed as retrotransposons. These are long interspersed elements (LINEs), short interspersed elements (SINES) and long terminal repeat (LTR) retrotransposons. In contrast, DNA transposons have a DNA intermediate.

Repeat Class	Length (bp)	Copy number	Fraction of genome
LINE	6,000 – 8,000	850,000	21%
SINE	100 – 300	1,500,000	13%
DNA	6,000 – 11,000	450,000	8%
LTR	2,000 – 3,000	300,000	3%

Table 1-2 The repeat composition of the human genome (Lander et al., 2001).

The full length LINE repeat is ~6kb in length. The LINE repeats encode an endonuclease and a reverse transcriptase which are sufficient for insertion of novel LINE elements in the human genome (Deininger and Batzer, 2002). Of 3 LINE subfamilies detectable in the genome (L1-L3) only the most recent (L1) appears to be actively retrotransposing, and accounts for 17% of the genome. It is estimated that there are 80-100 active L1 repeats per diploid genome (Brouha et al., 2003), with one novel insertion occurring every 100-200 births (Deininger and Batzer, 2002). LINE elements are roughly four-fold enriched in AT rich regions, which is consistent with their AT rich insertion sites (Lander et al., 2001). LINES are underrepresented in gene rich regions of the genome (Medstrand et al., 2002).

SINE repeats encode no proteins and rely on the LINE / L1 encoded endonuclease and reverse transcriptase for mobility in the genome (Dewannieux et al., 2003). SINEs account for 13% of the genome with a copy number of 1,500,000. Of the three subclasses of SINEs (Alu, MIR and MIR3), Alus are the most numerous in humans, with over one million copies accounting for 10.6% of the genome (Lander et al., 2001).

The Alu repeat element is derived from the 7SL non-coding RNA component of the signal recognition particle (SRP) involved in transport of proteins to the endoplasmic reticulum. The series of sequence duplication, deletion and recombination events that led to the formation of the modern Alu sequence are illustrated in Figure 1-1. Alus are classified into subfamilies according to age. Alu(J) are the oldest, Alu(S) are intermediate, and Alu(Y) are the youngest. Each subfamily has characteristic mutations observed in all members, whilst individual Alus contain random point mutations, which accumulate over time (Batzer et al., 1996).

Genome-wide, Alu density is higher in genes (12.5% of DNA sequence) than in intergenic regions (9.6%), and within genes, density is higher in introns (12.8%) than in exons (1.6%). The reason for the accumulation of Alus in gene rich DNA is unclear. The human genome contains approximately 190 full length Alus with the potential to retrotranspose, and the rate of retrotransposition is estimated to be similar to that of Line / L1s (Deininger and Batzer, 2002).

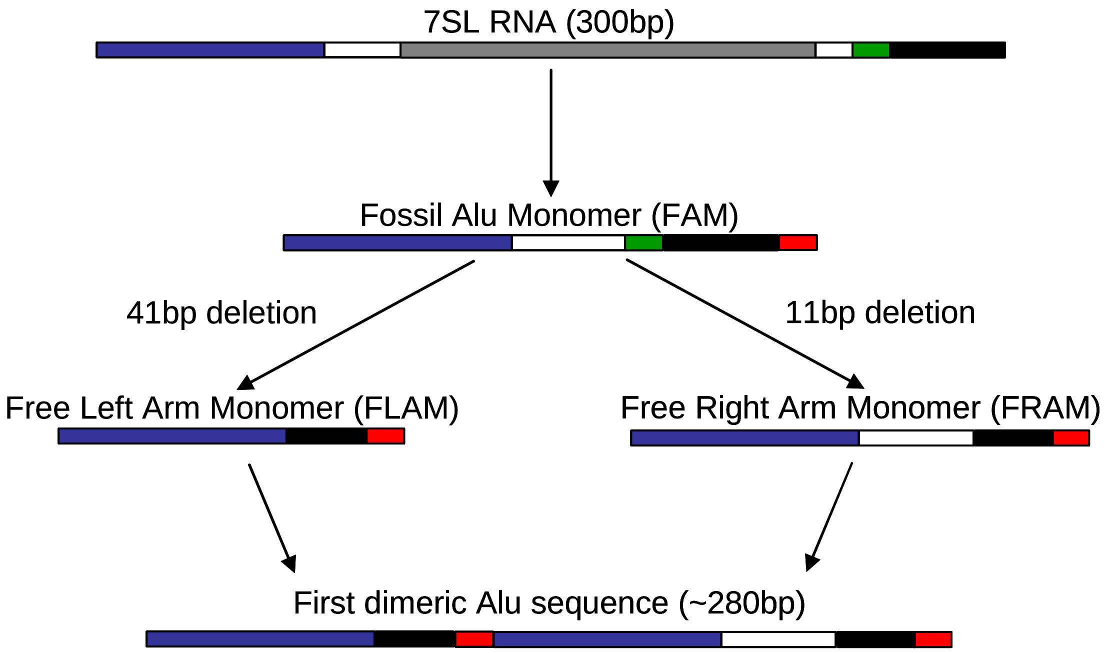
7SL RNA (300bp)
Fossil Alu Monomer (FAM)
41bp deletion
11bp deletion
Free Left Arm Monomer (FLAM)
Free Right Arm Monomer (FRAM)
First dimeric Alu sequence (~280bp)

Figure 1-1 Evolution of Alu sequences. Retroposition of 7SL RNA generated the fossil Alu monomer (FAM). This in turn underwent duplication and diversification to generate the free left and free right Alu monomers (FLAM and FRAM respectively). Combination of a FLAM with a FRAM created the ancestral Alu sequence. Internal coloured blocks indicate regions of sequence that are deleted at various stages in Alu evolution. The red blocks indicate Poly-(A) sequences (Mighell et al., 1997).

1.3 INTRODUCTION TO RNA

DNA is copied into RNA by the process of transcription. Like DNA, RNA is a linear polymer of nucleotide subunits. However, RNA differs from DNA in a number of ways. First, the sugar component of nucleotides in RNA is *ribose* rather than *deoxyribose*. Ribose contains an additional hydroxyl group which is modified in some RNAs. Second, RNA contains uridine (U) instead of thymidine (T) in DNA. Although U (like T in DNA) base pairs with adenosine (A), it may occasionally base pair with guanosine (G) in RNA. Third, whereas DNA occurs in cells as a double stranded helix, RNA is single stranded and folds into a variety of shapes. This allows various RNAs to have structural or catalytic functions.

RNA can broadly be categorised as messenger RNA (mRNA), which codes for protein, or non-coding RNA (ncRNA), in which the transcribed RNA is the final product. There are several distinct classes of non-coding RNA (Table 1-3), and many non-coding RNAs with diverse or unknown functions in the cell. RNA is synthesised by one of three RNA polymerases (pol). RNA pol I synthesizes the large ribosomal RNA; RNA pol II synthesizes mRNAs mRNA-like ncRNAs and micro RNAs; and RNA pol III synthesizes small non-coding RNAs including transfer RNAs (Paule and White, 2000).

RNA	**Full name**	**Function**
mRNA	Messenger RNA	Protein coding
tRNA	Transfer RNA	Adaptor molecule in protein synthesis
rRNA	Ribosomal RNA	Catalytic component of protein synthesis
snRNA	Small nuclear RNA	Component of spliceosome
snoRNA	Small nucleolar RNA	Guided modification of rRNA and snRNA
miRNA	Micro RNA	Regulation of RNA stability and translation
siRNA	Short interfering RNA	Targeted degradation of RNA

Table 1-3 The major families of non-coding RNA found in eukaryotic cells

1.3.1 Messenger RNA (mRNA)

Messenger RNAs (mRNAs) are protein coding transcripts. The initial transcript consists of exons, which contain the protein coding sequence, and introns which are non-coding. RNA splicing removes introns from the newly synthesised RNA sequence and joins together adjacent exons. The combination of different exons by alternative splicing allows multiple mRNAs to be made from the same initial transcript. It is estimated that up to 74% of human multi-exon genes are subject to alternative splicing (Johnson et al., 2003).

The spliced mRNA comprises a 5’ untranslated region (5’ UTR), a central protein coding region and a 3’ untranslated region (3’ UTR). During transcription, the 5’ end of mRNAs is covalently modified by a 7-methylguanosine ‘cap’. This protects mRNA from degradation by 5’ exoribonucleases, and facilitates translation by binding to the protein eIF4E which recruits the 40s ribosome subunit to the 5’ end of the RNA (Shuman, 2002). The 3’ ends of mRNAs are modified by addition of a poly-adenosine (poly-(A)) tail of around 200 nucleotides. This is bound by poly-(A) binding proteins which influence mRNA stability, translational efficiency and export of the mRNA to the cytoplasm (Colgan and Manley, 1997). Other than RNA editing (which is discussed in more detail below), the only reported modification of internal nucleotides of mRNAs is N6-methyladenosine (m6A), which is estimated to occur at three to five residues per mRNA (Wei et al., 1976). The functional consequences of adenosine methylation are unknown.

Introns are released from the splicing reaction as a loop of RNA called a lariat, and are subsequently cleaved into linear introns (Kim et al., 2000). The fate of excised introns is unclear but they are widely assumed to be non-functional and rapidly degraded. This may not be the case as some excised introns have been shown to be stable and perhaps subject to trafficking to subcellular compartments (Clement et al., 2001). Other introns undergo processing to produce functional non-coding RNAs including snoRNAs (Smith and Steitz, 1998) and miRNAs (Bartel, 2004).

1.3.2 Ribosomal RNA (rRNA)

The ribosome is a large ribonucleoprotein structure which catalyses the translation of mRNAs into proteins. It is composed of two ribosomal RNA (rRNA) species and many proteins. The large subunit of the ribosome contains 28S and 5.8S rRNAs (collectively the large subunit RNAs, (LSU)) and a 5S rRNA. The small subunit contains an 18S rRNA (the small subunit RNA (SSU)). The LSU and SSU rRNA occur in the human genome as a 44kb tandem repeat unit of which there are estimated 150-200 copies. The 5S rRNA also occurs in tandem arrays, and there are estimated to be 200 – 300 copies in the genome (Lander et al., 2001).

The ribosomal RNA precursor is transcribed and modified in the nucleolus. These modifications include conversion of approximately 100 uridines to pseudouridine and methylation of sugar 2' hydroxyl groups at a further 100 nucleotides (Maden, 1990). These modifications are 'guided' by small nucleolar RNAs (described below). The precise function of the modifications is unknown, but they are concentrated at sites of importance for translation, and therefore may benefit ribosome function (Decatur and Fournier, 2002). The pre-ribosomal RNA also undergoes methylation of bases at 10 locations (Maden, 1990). Following modification, the pre-rRNA undergoes a number of cleavage reactions to generate the mature RNA components which are assembled into the ribosome (Fatica and Tollervey, 2002).

1.3.3 Transfer RNA (tRNA)

Transfer RNAs (tRNAs) are the adapter molecules of protein synthesis. Initial analysis of the human genome identified 497 tRNA genes encoding 38 different tRNA species (Lander et al., 2001). The primary transcripts of tRNAs may contain a 5' leader sequence, introns and a 3' trailer sequence, which are trimmed and spliced by a number of proteins to generate the mature tRNAs of ~80 nucleotides (Hopper and Phizicky, 2003).

All tRNAs are subject to nucleotide modifications. Over 80 modifications have been described from various organisms, including methylation of sugar 2' hydroxyl groups and conversion of uridine to pseudouridine, along with other residues which are the target of more than one kind of modification. It is unclear how the modifications are specified, and whether snoRNAs are involved. The functions of the modifications are similarly unclear, though several are required for efficient translation (Hopper and Phizicky, 2003).

1.3.4 Spliceosomal RNAs (snRNAs)

Small nuclear RNAs (snRNAs) are components of the spliceosome which catalyses the splicing of introns from mRNAs. There are five snRNAs (U1, U2, U4, U5, and U6) involved in splicing of the majority of mRNAs. Each snRNA is approximately 200 nucleotides in length, and complexes with proteins to form a small nuclear ribonucleoprotein complex (snRNP). snRNAs direct the splicing reaction by base pairing with the snRNA components of other ribonucleoprotein complexes, and with highly conserved sequences at the boundaries between introns and exons in mRNA. Also, it is snRNAs rather than proteins that form the catalytic core of the spliceosome.

snRNAs themselves are subject to modification by methylation of sugar 2' hydroxyl groups and conversion of uridine to pseudouridine. As with ribosomal RNAs these modifications are guided by snoRNAs and take place in the nucleolus. The modifications are in the regions of snRNAs involved in base pairing with other RNAs and therefore may regulate splicing (Bachellerie et al., 2002).

1.3.5 Small nucleolar RNAs (snoRNAs)

Small nucleolar RNAs (snoRNAs) are small non-coding RNAs (60 – 140nt) which assemble into ribonucleoprotein complexes (snoRNPs) and guide the modification of nucleotides in rRNA, snRNA and potentially mRNA through complementary base-pairing (Bachellerie et al., 2002). There are two main classes of snoRNAs (Fatica and Tollervey, 2003). The box C / D snoRNPs catalyse methylation of sugar 2' hydroxyl groups, and the box H / ACA snoRNPs guide conversion of uridine to pseudouridine (Ψ). In both cases, modifications are directed by base pairing between short sequences (3 – 20 nucleotides) in the guide snoRNA, and complementary sequences in the target RNA.

There is accumulating evidence that the role of snoRNAs may extend beyond the modification of rRNA and snRNAs described above. A recent survey of small RNAs from a mouse cDNA library identified 83 novel snoRNAs, including 25 which lacked anti-sense elements for rRNAs or snRNAs and have been termed ‘orphan’ snoRNAs (Huttenhofer et al., 2001). Another study identified novel snoRNAs in human and mouse brain which were expressed specifically in the brain, from an imprinted locus (Cavaille et al., 2000). One of these, brain specific C / D box snoRNA HB11-52, is transcribed from an intron in the serotonin 2C receptor, and has an 18 nucleotide phylogenetically conserved region of complementarity to the RNA editing site of the serotonin $5HT_{2C}$ receptor mRNA, with the putative target site for methylation corresponding precisely to an edited adenosine.

1.3.6 Miscellaneous non-coding RNAs

In addition to the non-coding RNAs listed above (Table 1-3), there are a large number of RNAs with apparently diverse roles in the genome that do not yet fall into clear families of transcripts with related function. Many RNAs act as components of ribonucleoprotein complexes. For example, 7SL RNA the ancestral sequence of Alu retrotransposons is a component of the signal recognition particle (SRP), and plays a role in protein translocation across the endoplasmic reticulum membrane (Walter and Blobel, 1982). BC1 and BC100 are transcribed specifically in neurons and are both derivatives of retrotransposed RNA (tRNA-ala and Alu respectively). They assemble into ribonucleoprotein complexes and bind to poly-(A) binding protein which functions in translational regulation (Muddashetty et al., 2002). XIST RNA is involved in gene silencing. It is transcribed from the inactive X-chromosome, and binds to that chromosome guiding heterochromatin formation. XIST RNA itself is apparently regulated by a ncRNA anti-sense transcript TSIX (Avner and Heard, 2001). It has recently been demonstrated that the stability of the transcript Makorin-1, is regulated by an expressed homologous pseudogene (Hirotsune et al., 2003). Although the mechanism of regulation is currently unknown, this discovery may indicate a functional role for a proportion of the 20,000 pseudogenes in the human genome.

1.3.7 Double-stranded RNA (dsRNA)

In human cells, double-stranded RNA (dsRNA) can arise endogenously by base pairing of separate sense and anti-sense transcripts or by intramolecular base pairing of inverted repeats. Alternatively, dsRNA can arise exogenously, for example by infection with viruses that have dsRNA genomes (Yelin et al., 2003, Kumar and Carmichael, 1998). DsRNAs are known substrates of the RNA editing enzymes, the adenosine deaminases acting on RNA (ADARs) (Bass, 2002). Other cellular processes which act on dsRNA may therefore influence RNA editing by ADARs, and are described in more detail.

1.3.7.1 *Non-specific responses to dsRNA*

Cytoplasmic dsRNA encountered during viral infections, stimulates the potent interferon response and RNA-dependent protein kinase (PKR) (Kumar and Carmichael, 1998). In the cytoplasm, dsRNA binds to and activates PKR and a number of other proteins which stimulate the expression of interferons. The interferons are secreted from the infected cell and bind to interferon receptors on the surface of neighbouring cells. This in turn initiates a signal transduction cascade in these cells, leading ultimately to apoptosis. Activated PKR can also phosphorylate eukaryotic initiation factor 2α (eIF2α) and inhibit initiation of protein synthesis (Kumar and Carmichael, 1998). Approximately 20% of cellular PKR is located in the nucleus, mainly in the nucleolus. This suggests that it may potentially interact with endogenously transcribed dsRNAs. Cytoplasmic dsRNA can also activate the 2’,5’-Oligoadenylate Synthetase / RNaseL pathway. This results in cleavage of both viral and cellular RNAs (Kumar and Carmichael, 1998).

1.3.7.2 *Gene silencing by RNA interference (RNAi)*

The process of gene silencing by RNA interference was originally discovered in plants and has subsequently been identified in other eukaryotic organisms including humans (Tijsterman et al., 2002). The endonuclease Dicer cleaves exogenous cytoplasmic dsRNAs into double stranded short interfering RNAs (siRNAs) of approximately 21 nucleotides in length. A single strand of these duplexes is then assembled into the RNA induced silencing complex (RISC). This complex degrades mRNAs which contain sequences that are complementary or nearly complementary to the single stranded siRNA. The natural role of RNAi is uncertain. However, several lines of evidence indicate that RNAi may function as a defence mechanism against dsRNA viruses or retrotransposons (Gitlin and Andino, 2003, Sijen and Plasterk, 2003).

The microRNA (miRNA) genes are a source of endogenous dsRNA (Meister and Tuschl, 2004, Bartel, 2004). The primary miRNA transcript is a conserved stem-loop structured RNA which is processed in the nucleus by the ribonuclease Drosha to generate miRNA precursors. The miRNA precursors are then exported to the cytoplasm where they are cleaved by Dicer into mature double stranded miRNAs of approximately 21 nucleotides in length. A single strand of the miRNA duplex is then assembled into a miRNA ribonucleoprotein complex (miRNP). It is not currently known how the RNAi machinery distinguishes between exogenous RNAs (giving rise to siRNAs) and endogenous sources of dsRNA (giving rise to miRNAs), or how the miRNP complex differs from the RISC complex.

Some miRNAs act in a similar manner to siRNAs by directing cleavage of transcripts with completely complementary sequences (Zeng et al., 2003). However, the majority miRNAs in animals appear to bind to partially complementary sequences in the 3’ UTR of target mRNAs, where they regulate gene expression by repression of translation (Bartel, 2004). It is estimated that there are 250 miRNA genes in mammalian genomes. To date, only one mammalian miRNA gene, miR-181, has been characterised biologically. This miRNA is highly expressed in bone marrow and thymus and appears to regulate the development of B-Cells and T-cells (Chen et al., 2004). Currently, no specific gene targets of mammalian miRNAs have been identified.

In addition to the post transcriptional gene silencing effects described above, there is accumulating evidence that siRNAs generated from dsRNAs formed by endogenously transcribed repeat sequences are able to silence transcription by stimulating heterochromatin formation in DNA. In *Arabidopsis* for example, 95% of siRNA is derived from transposons and tandem repeats. (Lippman and Martienssen, 2004). It is not currently clear whether a similar process occurs in mammalian cells, however it has recently been shown that synthetic siRNA directed to CpG islands of gene promoters can induce DNA and histone methylation, resulting in transcriptional silencing (Kawasaki and Taira, 2004).

1.3.7.3 *Other dsRNA binding proteins*

There are many proteins, in addition to those described above, that contain one or more dsRNA binding domains (Saunders and Barber, 2003). In principle, these proteins may compete with the ADAR RNA editing enzymes by binding to dsRNA substrates. Cytoplasmic dsRNA binding proteins include TAR RNA binding protein (TRBP) which regulates translation, and Staufen which may transport mRNAs to sites of translation. Nuclear dsRNA binding proteins include nuclear factor associated with dsRNA (NFAR), which interacts with proteins involved in splicing and RNA helicase A (RHA) which unwinds dsRNA in a 3’ to 5’ direction and is associated with RNA polymerase II. Testis nuclear RNA binding protein (TENR) also has an inactive adenosine deaminase domain, suggesting a role in regulating RNA editing by sequestering substrates.

1.4 GENERAL INTRODUCTION TO RNA EDITING

RNA editing can be broadly defined as any site specific alteration of an RNA sequence yielding a product differing from that encoded by the DNA template. This excludes splicing, polyadenylation and capping of mRNAs and the various other modifications of RNA following transcription that were reviewed in the previous section.

RNA editing has been identified in a variety of organisms including viruses, bacteria, fungi, plants, invertebrates and mammals. The mechanisms of RNA editing are similarly diverse and include nucleotide insertions and deletions, and base substitutions. Across the range of species, there are examples of editing of all three major classes of RNA, transcribed from both nuclear and organellar genomes (Table 1-4).

In this section the types and targets of RNA editing in various organisms is described. Some classes of RNA editing appear to be restricted to a small number of organisms. For example, guided insertion and deletion of nucleotides has only been reported in the trypanosomes. Other classes of RNA editing are more widespread. In particular, the process of adenosine to inosine (A > I) editing of tRNA by Adenosine deaminases that act on tRNA (ADATs) is observed in many organisms including bacteria and humans.

Organism	RNA origin	RNA class	RNA editing
Escherichia coli	Genomic	tRNA	A > I
Paramyxoviruses*	ssRNA genome	mRNA	G insertion
Trypanosomes	Kinetoplastid	mRNA	U insertion U deletion
Slime mould	mitochondrion	mRNA tRNA rRNA	N Insertion NN insertion C > U
Yeast	Nuclear genomic	tRNA	A > I
Plant	Organelles	mRNA tRNA	U > C C > U
Worm	Nuclear genomic	mRNA tRNA	A > I C > U
Fruit fly	Nuclear genomic	mRNA tRNA	A > I
Squid	Nuclear genomic	mRNA tRNA	A > I
Frog	Nuclear genomic	mRNA	A > I
Mouse	Nuclear genomic	mRNA tRNA miRNA	A > I C > U
Human	Nuclear genomic	mRNA tRNA miRNA	A > I C > U

Table 1-4 Overview of the dominant types and targets of RNA editing. *A number of other viral RNAs are subject to A > I editing. However, these processes are catalysed by the RNA editing machinery of the host organism rather than by viral encoded editing machinery.

1.4.1 RNA editing of tRNA in *Escherichia coli*

The *E. coli* tadA protein catalyses the conversion of A > I at adenosine 34 in $\text{tRNA}_{\text{Arg2}}$, and is the only known prokaryotic RNA editing enzyme (Wolf et al., 2002). The edited nucleotide is at the first position in the tRNA anticodon (the “wobble” position). Edited tRNAs are able to recognise multiple codons in the mRNA by base pairing of I34 with C, A or U at the third position of the codon in mRNA. This allows the same tRNA to insert its amino acid at different codons in the mRNA.

E. coli TadA is currently the most ancient example of the family of Adenosine Deaminases that act on tRNA (ADATs). However, inosine is found at the wobble position of tRNAs in many organisms ranging from archaea to humans indicating that even more ancient ADAT enzymes may exist (Grosjean et al., 1996).

1.4.2 RNA editing of Paramyxovirus RNA by polymerase stuttering

The Paramyxoviruses are a large family of viruses which infect vertebrates, and include Measles and Mumps viruses. The genomes of Paramyxoviruses are single stranded RNA molecules encoding 6 mRNAs. The P gene encodes the P protein (phosphoprotein) which is involved in binding and packaging of the viral RNA genome. The P genes of many paramyxoviruses overlap with one or more genes in a different reading frame. To access these alternate reading frames, the viral RNA polymerase “stutters” at a G-rich sequence found at the transition from the P Gene to the out of frame overlapping genes. This results in the insertion of one or more non-coded Gs and consequently a shift in the reading frame such that the overlapping genes are translated as a fusion protein with the N-terminal of the P protein (Haussmann et al., 2001).

1.4.3 Guided uridylate insertion and deletion RNA editing in Trypanosome kinetoplasts

The Trypanosomatids are parasitic protozoans including *Trypanosoma brucei* which is transmitted by tsetse flies and causes African sleeping sickness. The Trypanosomatids have a single mitochondrion, containing a giant network of concatenated DNA 'minicircles' and 'maxicircles' called the kinetoplast. There are approximately 10,000 minicircles and 50 maxicircles of DNA per kinetoplast. Kinetoplast RNAs undergo extensive RNA editing by multiple insertion and deletion of uridylate residues (Simpson et al., 2003). For example, the ATP synthase 6 subunit (A6) is edited by the insertion of 447 and deletion of 28 uridylate residues. The scale of editing means that some RNAs contain more nucleotides from editing than from transcription.

RNA editing of kinetoplast RNA is directed by small RNA molecules (~1kb) called guide RNAs, the majority of which are encoded on the minicircle (Blum et al., 1990). Guide RNAs interact with the RNA to be edited, by base pairing at two sequences spanning the editing site. The target RNA is cleaved between these sites and uridylates are inserted or deleted according to the sequence of the guide RNA. A single round of editing is complete when the guide RNA base pairs completely with the edited transcript. However, several rounds of editing directed by guide RNAs are required for the complete editing of a transcript. Insertion / deletion editing directed by one guide RNA often creates the binding site of the next resulting in an overall 3' to 5' direction of RNA editing (Simpson et al., 2003).

An RNA editing complex of ~1600 kDa with 20 major protein components has been isolated, and shown to have many of the enzymatic activities required for the editing process (Panigrahi et al., 2001). However, the mitochondrial proteins and complexes involved in catalysis of guided RNA editing are currently the subject of research (Simpson et al., 2004).

1.4.4 Nucleotide insertion and nucleotide substitution RNA editing in *Physarum polycephalum* mitochondria

Physarum polycephalum (slime mould) is unique in using RNA editing by both nucleotide insertion and substitution. Edits include specific insertion of nucleotides (C or U) and dinucleotides (CU, UA, GU, AA and GC) and C to U base substitution. RNA editing by nucleotide insertion occurs at approximately 1,000 sites in mitochondrial mRNAs, tRNAs and rRNAs. RNA editing of these sequences restores complete reading frames, and effects coding changes, and is essential for the expression of functional protein products and structural RNAs (Gott, 2000). RNA editing by nucleotide insertion appears to be a co-transcriptional process as nucleotides are added to the 3' end of nascent RNA (Cheng et al., 2001). It is not currently known how the site of insertion and the type of nucleotide or dinucleotide to be added is specified. In addition to RNA editing by nucleotide insertions, C > U substitutions have been observed in the mitochondrial cytochrome *c* oxidase subunit 1 mRNA (Gott et al., 1993). C

> U editing does not occur co-transcriptionally, but by some other pathway proposed to be a base deamination reaction similar to that in mammals.

1.4.5 Nucleotide substitution RNA editing in yeast

A > I editing of tRNAs by ADATs was first identified at the wobble position (I34) in yeast. In contrast to *E. coli* which has a single ADAT, eukaryotes have two ADATs (called Tad2 and Tad3 in yeast), which form heterodimers and catalyse A > I editing at the wobble position in a number of tRNAs (Gerber and Keller, 1999), and a third ADAT (called Tad1p in yeast), which catalyses A > I editing at position 37 in tRNA (Gerber et al., 1998). The function of the modification at position 37 is unclear. A yeast cytidine deaminase (CDD1) has recently been identified and shown to have C > U RNA editing activity (Dance et al., 2001). The *in vivo* substrates of this enzyme are unknown.

1.4.6 Nucleotide substitution RNA editing in Plant organelles

RNA editing in plants is by C > U and, to a lesser extent, U > C substitution in mitochondrial and chloroplast RNAs. There are no reports of A > I editing and there is no evidence of editing of nuclear transcripts. The relative abundance of C > U and U > C edits and the relative extent of editing in the two organelles is variable between species of plants (Bock, 2000). The catalytic component of RNA editing in plants has not been identified. Deletion studies have shown that trans acting factors and sequences in the target mRNA are essential (Bock and Koop, 1997, Bock et al., 1996).

RNA editing of a number of plant organelles has been examined by systematic sequencing of cDNA, and comparison with genomic DNA. Analysis of RNA editing in the mitochondria of the model higher plant *Arabidopsis thaliana* showed 456 C > U but no U > C conversions (Giege and Brennicke, 1999). In a similar analysis of RNA editing in the chloroplast of the model lower plant *Anthoceros formosae*, 509 C > U and 433 U > C conversions were identified (Kugita et al., 2003). In both cases, there is a predominance of editing in the first two positions of a codon, indicating selection for biologically relevant RNA edits. Consequently, the vast majority of RNA edits result in conversion of codons to a conserved form required for the translation of functional protein products. The amino acid changes resulting from RNA editing are predicted to increase the hydrophobicity of mitochondrial proteins.

1.4.7 Nucleotide substitution RNA editing of *Caenorhabditis elegans* RNAs

In addition to A > I editing of tRNAs, the nematode worm *C. elegans* exhibits A > I editing of mRNAs by adenosine deaminases acting on RNA (ADARs). The worm has two *ADAR* genes (*adr1* and *adr2*) which are distantly related to the vertebrate *ADAR*s (Keegan et al., 2004).

Using a technique to identify inosine containing transcripts (Morse and Bass, 1997, Morse and Bass, 1999, Morse et al., 2002), ten novel RNA editing substrates were identified in poly(A)+ RNA from *C. elegans*. These comprised 7 from 3’UTR, 1 from 5’ UTR, 1 from a non-coding RNA, and 1 from intron. Only four targets were of known function, three of which are important for proper function of the nervous system. The substrates identified were all predicted to form dsRNA by base pairing of transposon derived inverted repeat sequences. Currently, there are no reported A > I edits in coding sequences in *C. elegans*.

Recently, C > U editing of *GLD2* mRNA was reported. GLD2 encodes an atypical poly-(A) polymerase that controls the mitosis / meiosis decision in the germ line. C > U editing is predicted to result in a proline to leucine change. The enzyme responsible for this change is currently unknown. *C. elegans* contains nine putative cytidine deaminases. However, none of these has confirmed C > U RNA editing activity and none are homologous to the human RNA cytidine deaminase APOBEC-1 (Wang et al., 2004a).

1.4.8 Nucleotide substitution RNA editing in *Drosophila melanogaster*

Drosophila has a single *ADAR* gene (*dADAR1*) which is expressed in the adult central nervous system and shares homology with human *ADAR2* (Palladino et al., 2000a). A > I editing in Drosophila appears to be important for the regulation of a number neuronal transcripts and is predicted to alter the protein coding sequence of the voltage gated sodium channel (para) (Hanrahan et al., 2000), the calcium channel subunit (cacophony) (Smith et al., 1998) and a glutamate gated chloride channel (Semenov and Pak, 1999). Furthermore, *Drosophila* mutants lacking *dADAR1* showed altered nervous system function (Palladino et al., 2000b), and increased sensitivity to oxygen deprivation in conjunction with a lack of editing at the known editing sites (Ma et al., 2001). dADAR1 also edits its own transcript, resulting in a serine to glycine substitution in the catalytic domain which may alter enzyme specificity (Palladino et al., 2000a).

A > I editing substrates in *Drosophila* are predicted to form dsRNA between the edited exon and complementary sequences in adjacent introns. Selective pressure to retain these dsRNAs means that exonic sequences near editing sites are more highly conserved than at non-editing sites (Hoopengardner et al., 2003). This property was used to carry out a comparative analysis of candidate editing substrates from two *Drosophila* species, and revealed novel RNA editing sites in 16 transcripts involved in rapid electrical and chemical neurotransmission, many of which encoded functionally important amino acid changes. The human orthologue of one of these targets, the potassium channel KCNA1, shows conservation of editing of an isoleucine codon to a valine codon in the pore lining domain (Hoopengardner et al., 2003).

A > I editing of another *Drosophila* transcript appears to involve intermolecular dsRNA formation between complementary sense and anti-sense transcripts rather than the intramolecular base pairing described above (Peters et al., 2003). *4f-rnp* and *sas10* are closely adjacent genes on opposite strands of DNA. The developmentally regulated *sas10* transcript base pairs with *4f-rnp* resulting in A > I editing and a reduction in *4f-rnp* RNA (Peters et al., 2003).

Recently, U > C RNA editing was reported in a cockroach sodium channel and subsequently in the *Drosophila* orthologue. The U > C edit results in a

Phe > Ser amino acid substitution and altered ion channel properties (Song et al., 2004, Liu et al., 2004).

1.4.9 Nucleotide substitution RNA editing in squid

Two potassium channel subunits from squid have been shown to undergo extensive A > I editing. (Patton et al., 1997, Rosenthal and Bezanilla, 2002). In both cases, the density of A > I editing in coding sequences is extremely high. For example, SqKv1.1 mRNA is edited at 14 adenosines, of which 13 result in amino acid changes (Rosenthal and Bezanilla, 2002). The function of the edits are unknown but are predicted to result in changes to the conductance of the ion channel. The enzymes responsible for A > I editing in squid have yet to be identified.

1.4.10 Nucleotide substitution RNA editing in *Xenopus laevis*

Adenosine to inosine editing of dsRNAs by ADARs was first discovered in *Xenopus* as a dsRNA unwinding activity which introduces A > I changes in the RNA substrate (Bass and Weintraub, 1988). *Xenopus* has three *ADAR* genes (*ADAR1a*, *ADAR1b* and *ADAR2*) which are equivalent to mammalian *ADAR1* and *ADAR2* (Keegan et al., 2004).

1.4.11 Nucleotide substitution RNA editing of mammalian RNAs

The dominant forms of RNA editing in mammals are C > U substitutions in mRNA catalysed by cytidine deaminases and A > I substitutions in mRNA and tRNA catalysed by ADARs and ADATs respectively. The types and targets of

RNA editing appear to be broadly conserved between humans and other mammals and therefore are discussed in the following sections on RNA editing in humans.

1.5 RNA EDITING IN HUMANS

In humans, there are two predominant forms of RNA editing. Adenosine to inosine (A > I) editing is known to occur in mRNA, tRNA and miRNA, and cytidine to uridine (C > U) editing is known to occur in mRNA. The editing reactions involve deamination of the nucleotide base, and in both cases the product of RNA editing has altered base pairing properties compared to the unedited base. Adenosine base pairs with uridine in RNA whereas inosine has similar properties to guanine and base pairs with cytidine. Cytidine base pairs with guanine, whereas uridine base pairs with adenosine (Figure 1-2).

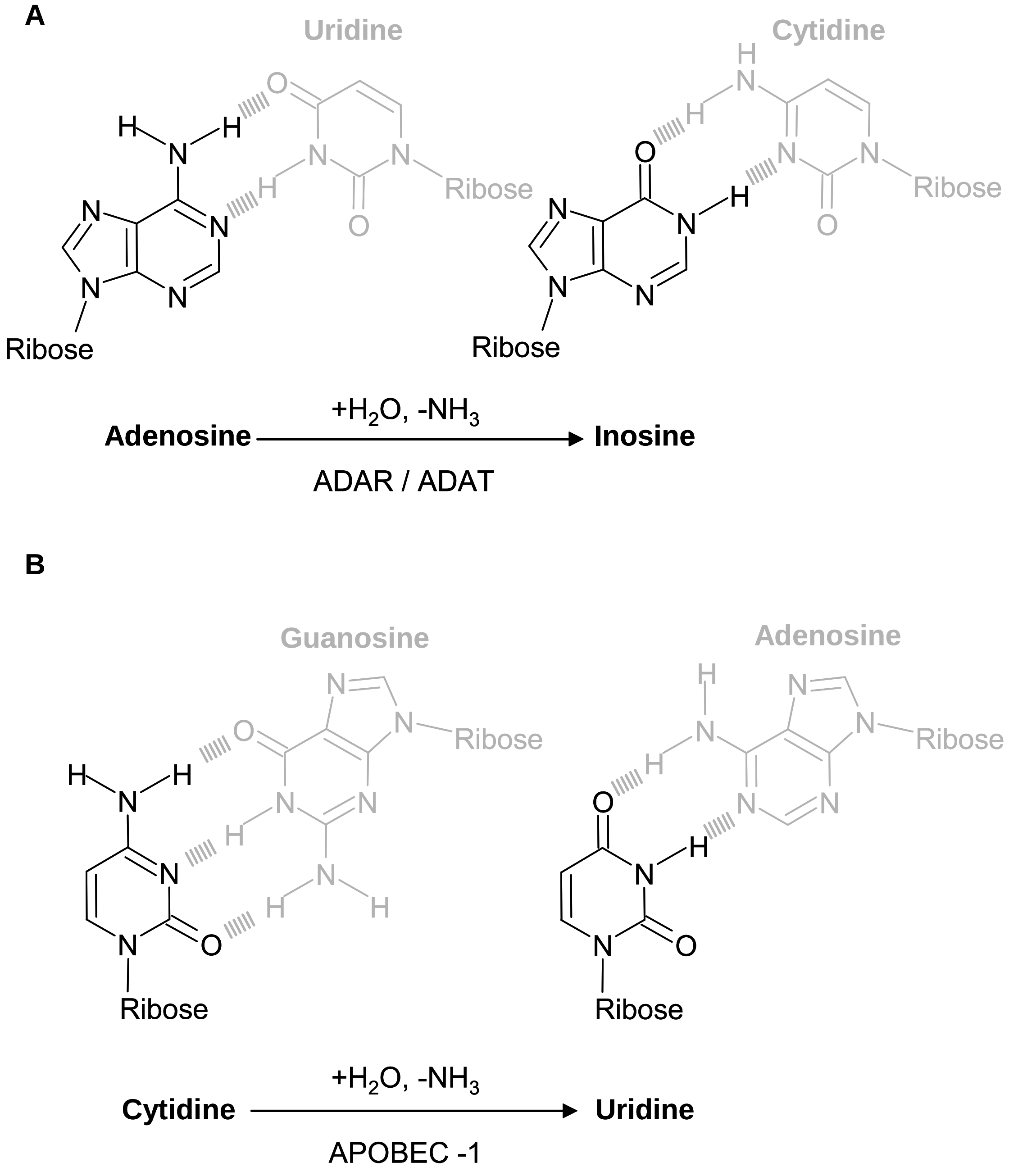

Figure 1-2 The effect of RNA editing on base pairing in RNA. **A.** Adenosine to inosine RNA editing catalysed by adenosine deaminases acting on RNA (ADARs) and tRNA (ADATs). **B.** Cytidine to uridine RNA editing catalysed by APOBEC-1. Base pairing is indicated by grey structures, dashed lines indicate hydrogen-bonds.

The first reported example of mRNA editing in humans was C > U editing of the apolipoprotein mRNA (Powell et al., 1987, Chen et al., 1987). This was followed by the discovery of A > I editing in the transcripts of glutamate receptors (Sommer et al., 1991). In both cases, RNA editing was discovered serendipitously by comparison of cDNA sequences with genomic DNA. Although further examples of both classes of edit have since been identified (Table 1-5), it is only recently that systematic approaches have begun to reveal the extent to which RNA editing modifies the transcriptome. In addition to C > U and A > I edits, a small number of other classes of RNA edit have been reported (Table1-5). The enzymes responsible for these other classes of RNA edit have not been identified, and in most cases these edits are known only by a single example.

Transcript	**Edit**	**Codon change**	**Enzyme**
GluR-B	A > I	Q > R R > G	ADAR 2 ADAR 1 / 2
GluR-C	A > I	R > G	ADAR 1 / 2
GluR-D	A > I	R > G	ADAR 1 / 2
GluR-5	A > I	Q > R	ADAR 1 / 2
GluR-6	A > I	Q > R I > V Y > C	ADAR 1 / 2 ADAR 1 / 2 ADAR 1 / 2
Serotonin receptor	A > I	I > V I > M N > D N > S N > G	ADAR 1 / 2 ADAR 1 / 2 ADAR 1 / 2 ADAR 1 / 2 ADAR 1 / 2
K^+ channel	A > I	I > V	ADAR 2
HDV antigenome	A > I	W / Amber	ADAR 1
Non-coding RNA*	A > I	Hyperediting	ADAR 1 / 2
Viral RNA$^{\#}$	A > I	Hyperediting	ADAR 1 / 2
ApoB mRNA	C > U	Q > Stop	APOBEC-1
NF1	C > U	Q > Stop	Unknown
IL12 R2beta	C > U	A > V	Unknown
GluR7	G > A	R > Q	Unknown
GluR7	U > G	S > A	Unknown
Alpha-galactosidase	U > A	F > Y	Unknown
WT1	U > C	L > P	Unknown
APP	2nt deletion	Frameshift	Unknown
ubiquitin B	2nt deletion	Frameshift	Unknown

Table 1-5 Known RNA edits in human transcripts. *The data presented in this thesis and in recent publications (Levanon et al., 2004, Kim et al., 2004) indicate the presence of several thousand A > I editing sites in the introns and

UTRs of mRNAs and in intergenic transcripts. #RNAs of several viruses undergo extensive A > I editing in human cells. See text for details.

1.5.1 Human A > I RNA editing enzymes

The human genome encodes three adenosine deaminases that act on tRNA (ADAT1 – 3), and two dsRNA specific adenosine deaminases that act on RNA (ADAR1 and ADAR2). Both families of enzymes are characterised by zinc-containing adenosine deaminase domains. It is believed that the ADARs evolved from ADATs, by acquisition of dsRNA binding domains (dsRBDs). ADATs in turn are believed to descend from cytidine deaminases. (Keegan et al., 2004).

Both ADAR1 and ADAR2 form homodimers, and interactions between the two monomers may confer editing site selectivity (Cho et al., 2003, Gallo et al., 2003). Once the ADAR is bound to dsRNA through its dsRBD, it flips the nucleotide into the active site. Based on similarities with the cytidine deaminase (CDA) it is believed that the active site harbours a zinc binding domain, and that a metal-bound hydroxide ion attacks the purine ring to form a tetrahedral intermediate which decomposes to the inosine containing RNA and ammonia.

1.5.1.1 *ADATs*

Adenosine deaminase that acts on tRNA 1 (*ADAT1*) was identified in humans as an orthologue of the yeast tRNA editing gene *Tad1p* (Maas et al., 1999), and encodes a protein with an adenosine deaminase but no dsRBDs that acts at position I34 (Maas et al., 2001a). Human homologues *ADAT2* and *ADAT3* are present in the human genome as homologues of tad, though evidence of their expression is yet to be presented.

1.5.1.2 *ADAR1*

ADAR1 was the first ADAR gene to be identified (Kim et al., 1994, O'Connell et al., 1995), and is transcribed in two forms. The full length transcript encodes a 150kDa protein, and is produced from an interferon inducible promoter. The carboxy-terminal region of this protein contains a catalytic deaminase domain, three dsRNA binding domains (dsRBDs) and a nuclear localization signal (Eckmann et al., 2001). The amino-terminal region contains two Z-DNA binding domains (Herbert et al., 1997), and an overlapping nuclear export signal (Poulsen et al., 2001). The 110kDa shorter form of *ADAR1* is constitutively expressed. This form lacks the amino terminal 295 amino acids, which includes the Z-DNA binding domain and nuclear export signal (George and Samuel, 1999).

ADAR1 is widely expressed, but is most abundant in the brain and least abundant in skeletal muscle (Kim et al., 1994, O'Connell et al., 1995). The interferon inducible form of ADAR1 is predominantly cytoplasmic and appears to be responsible for hyperediting of viral dsRNAs *in vivo* (Patterson and Samuel, 1995). In contrast, the constitutively expressed short form of ADAR1 is predominantly nuclear and appears to be the enzyme responsible for editing the HDV RNA (Wong and Lazinski, 2002). Within the nucleus, ADAR1 has been shown to accumulate in the nucleolus, dependent on binding to rRNA (Desterro et al., 2003, Sansam et al., 2003). ADAR1 is capable of selectively editing the serotonin receptor and a number of other substrates *in vitro*, but is incapable of editing the glutamate receptor Q / R site.

Two recent studies have demonstrated that ADAR1 null mutations lead to embryonic lethality in mice (Wang et al., 2004b, Hartner et al., 2004). The ADAR deficient embryos were characterised by widespread apoptosis in cells derived from various tissues, associated with a decrease in the expression of anti-apoptotic genes (Wang et al., 2004b). Embryos also suffered liver degeneration along with severe defects in haematopoiesis, and ADAR deficient stem cells failed to contribute to the development of a number of non-neuronal tissues. Analysis of RNA editing substrates from cloned neuronal cells of ADAR1 deficient mice indicate that ADAR1 is responsible for *in vivo* A > I editing of three adenosines leading to coding changes in the serotonin receptor transcript (Hartner et al., 2004). The ADAR substrates responsible for the severe phenotypes observed in these experiments are unknown.

1.5.1.3 *ADAR2*

ADAR2 was isolated as the enzyme responsible for editing of the glutamate receptor Q / R site (Melcher et al., 1996b, O'Connell et al., 1997). The protein has a carboxy-terminal with 50% homology to ADAR1, a central region with two dsRBD and a short amino-terminal, lacking Z-DNA binding domains. Alternative splicing yields 4 isoforms, resulting from variable inclusion of an

Alu cassette insert, and long or short carboxy-terminal sequences (Gerber et al., 1997, Lai et al., 1997). An additional splice site is generated in rat brain by the action of ADAR2 on its own mRNA. An AA dinucleotide is edited to an AI dinucleotide which functions as an AG splice acceptor (Rueter et al., 1999).

ADAR2 is widely expressed, but is most abundant in the brain and least abundant in skeletal muscle (Kim et al., 1994, Melcher et al., 1996b). Within the brain, ADAR2 expression varies developmentally (Paupard et al., 2000), and regionally. For example, RNA editing by ADAR2 is lower in white matter than in grey matter (Kawahara et al., 2003). ADAR2 is located in the nucleus and, like ADAR1, accumulates in the nucleolus. The active ADAR2 enzyme is a homodimer, and is capable of site-specific editing of the Q / R site of the glutamate receptor, the serotonin receptor and the potassium channel RNAs. ADAR2 is also able to bind and edit other substrates *in vitro*.

ADAR2 deficient mice were prone to seizures and died young (Higuchi et al., 2000). This was associated with substantially reduced editing at most of the known RNA editing sites. However, the impaired phenotype reverted to normal when the edited alleles for just one site, the Q / R site in the Glutamate receptor B subunit transcript, were encoded genomically. This suggests that physiologically, this is the most important substrate of ADAR2 (Higuchi et al., 2000).

1.5.1.4 *Other ADARs*

There are two additional ADAR related proteins with unknown function. ADAR3 encodes a protein with a similar structural arrangement to ADAR2. It is expressed exclusively in the mammalian brain, but as yet no adenosine deaminase activity has been described, leading to speculation that it regulates A > I editing by the other ADARs (Melcher et al., 1996a, Chen et al., 2000). TENR is a testis specific dsRNA binding protein with a deaminase motif identified in mouse and with a homologue in human. No RNA editing activity has been demonstrated (Hough and Bass, 1997).

1.5.2 Human A > I editing substrates

All known ADAR substrates are dsRNAs, which are recognised by the dsRNA binding domains of ADAR editing enzymes. The edited nucleotides may be in protein coding or non-coding sequences (Table 1-5). The majority of RNA edits in coding sequences are dsRNAs formed between the exon sequence and complementary sequence in a flanking intron. For example, in the GluR-B transcript Q / R site editing is in a region of dsRNA formed between the edited exon and an inverted repeat in the downstream intron (Higuchi et al., 1993). However, the recently identified editing site in the intronless potassium channel RNA is a dsRNA formed exclusively from coding exon sequence (Bhalla et al., 2004). In non-coding RNA, editing substrates are predicted to form dsRNA between pairs of inverted high copy repeat sequences in the same transcript (Morse et al., 2002).

Imperfections within the dsRNA substrate may be important for selecting adenosines for deamination. Whereas long, perfectly base paired dsRNA is extensively edited (~60% of all adenosines), the introduction of mismatches and bulges effectively breaks the RNA into a series of substrates (Lehmann and Bass, 1999). Consistent with this, long hairpins formed by inverted Alus of human substrates were edited at multiple sites in both strands, whereas sequences for which no secondary structure could be easily predicted were infrequently edited (Morse et al., 2002).

In vitro studies using artificial substrates indicate that ADAR1 has a 5' neighbour preference of U = A > C > G. ADAR2 has the similar preference U ≈ A > C = G (Lehmann and Bass, 2000). ADAR2 also has a 3' neighbour preference of U = G > C = A. Both ADAR1 and ADAR2 edit more efficiently at A:C mismatches than at an A:A or A:G mismatch or an A:U base pair *in vitro* (Wong et al., 2001). Analysis of the limited number of previously known *in vivo* editing substrates indicates that editing occurs preferentially at adenosines in A:C mismatches, whereas adenosines in A:A and A:G mismatches are unedited (Kallman et al., 2003). The analyses of larger datasets of A > I edits presented in this thesis, and in recent publications are consistent with these sequence preferences (Kim et al., 2004, Levanon et al., 2004).

1.5.2.1 *A > I editing of translated exons*

A > I editing is known to edit the coding sequences of a number of transcripts expressed in the central nervous system (Table 1-5). The first to be discovered was the Q / R site of the glutamate receptor B subunit mRNA in which a glutamine codon (CAG) is edited to an arginine codon (CIG) resulting in an amino acid substitution change at a conserved residue within the pore of the glutamate receptor ion channel (Sommer et al., 1991). The edited nucleotide is present in more than 99% of transcripts in adult rat brain, and results in reduced permeability of the ion channel to Ca^{2+} ions, regulation of the rate of formation of glutamate receptor tetramers, and trafficking of GluR-B from the endoplasmic reticulum (Greger et al., 2003). A Q / R editing site is also present in the related glutamate receptor subunits GluR-5 and GluR-6, However, these editing sites are not functionally equivalent to the Q / R editing site of the GluR-B mRNA.

The transcripts of the glutamate receptor subunits GluR-B, C and D also undergo an arginine (AGA) to glycine (IGA) edit (Lomeli et al., 1994), while the glutamate receptor subunit GluR-6 also contains an isoleucine (ATT) to valine (ITT) edit, and a tyrosine (TAC) to cysteine (TIC) edit. The effects of these latter edits are not well characterized but appear to regulate calcium permeability (Kohler et al., 1993).

The transcript of the serotonin receptor, 5-HT_{2C}R (a G-protein coupled receptor) is edited at five sites (Burns et al., 1997). RNA editing alters three amino acids in the second intracellular loop of the receptor, leading to a conformational change and disruption of the G-protein interaction. This results in a 10 to 15-fold reduction of signalling by phosphoinositide hydrolysis in response to serotonin binding, and silencing of constitutive activity (Visiers et al., 2001).

The human potassium channel is edited by ADAR2 at a single adenosine leading to a isoleucine (ATT) to valine (ITT) substitution (Bhalla et al., 2004). The potassium channel transcript is intronless and is the first example of RNA editing of a small hairpin formed entirely of exonic RNA. The altered amino acid is in a highly conserved ion-conducting pore of the potassium channel and affects ion channel inactivation.

1.5.2.2 *A > I editing of viral RNA*

Hepatitis delta virus (HDV) is a sub-viral human pathogen, which requires co-infection with the Hepatitis B virus, for production of the HDV coat protein. The HDV genome is a circular RNA of ~1700bp which forms a rod structure through extensive base pairing. A single open reading frame produces the delta antigen (HDAg) in two forms, dependent on RNA editing. Editing of the antigenome results in an extended protein product by specifically converting an amber codon (UAG) to tryptophan (UIG). Whereas the smaller version is essential for genome replication, the edited version inhibits genome replication and is required for viral packaging (Polson et al., 1996).

RNA editing of other viruses is non-selective. For example, transcripts of the polyoma virus may undergo RNA editing at up to half of the adenosines specified by the viral genome (Kumar and Carmichael, 1997). By an unknown mechanism, inosine containing transcripts are preferentially retained in the nucleus where they are isolated from the translation machinery, and are eventually degraded (Kumar and Carmichael, 1997). Similarly, the negativestrand genomic RNA of the measles virus is edited at multiple sites, affecting transcription, translation, stability or function of the viral proteins.

1.5.2.3 *A > I editing of microRNAs*

It has recently been demonstrated that the mammalian microRNA precursor miRNA22 is modified by RNA editing *in vivo* (Luciano et al., 2004). Editing occurs at a low level (approximately 5 – 10% cDNA clones sequenced from human brain), and appears to be catalysed by ADAR1. The function of miRNA editing is unknown; however editing occurs at several adenosines that are present in the mature miRNA and therefore may influence binding of the miRNA to target sequences in mRNAs.

1.5.2.4 *A > I editing of sequences involved in RNA splicing*

In addition to the creation of a splice site in the transcript of ADAR2 by RNA editing (Rueter et al., 1999), there are several examples where RNA editing appears to regulate RNA splicing. Editing within an intron of the PTPN6 transcript destroys a branch site adenosine. An adenosine at this position is required for normal splicing, and RNA editing leads to intron retention, and a premature stop codon (Beghini et al., 2000). A study of the intron-exon dsRNA at the GluR-B R/G editing site revealed that splicing and ADAR2 binding compete with one another *in vitro* but not *in vivo* (Bratt and Ohman, 2003). As RNA editing at this site requires the intron, this conflict could be resolved by coordination of the two processes, with RNA editing preceding splicing.

Consistent with this is the isolation of ribonucleoprotein complexes containing splicing factors and editing activity (Raitskin et al., 2001).

1.5.2.5 *A > I editing of non-coding RNA from human brain*

A systematic method for the identification of A > I edits in RNA has been developed which uses inosine specific cleavage of RNA to enrich for potential editing substrates (Morse and Bass, 1999). Applying this technique to human brain poly (A)+ RNA, 19 novel A > I editing substrates were identified. These included five from introns, three from 3'UTRs and one from a non-coding RNA. No example of coding RNA editing was observed (Morse et al., 2002). Each of the novel edited substrates was found to be edited at multiple adenosines when analysed from total brain RNA. Most sequences contained high copy repeats, which were predicted to form dsRNAs by base pairing with inverted copies of the repeat in the flanking transcript. In nine out of nineteen novel substrates, editing was associated with an Alu repeat, with an inverted copy within 1kb in the flanking transcript (Morse et al., 2002).

The data presented in this thesis, and recent computational analyses of EST and cDNA sequences, have confirmed that A > I editing of Alu sequences is widespread (Kikuno et al., 2002, Kim et al., 2004, Levanon et al., 2004). Together, these results suggest that a major target of A to I editing is non-coding, rather than coding regions of mRNAs.

1.5.3 The function of A > I editing

Clearly, one function of A > I editing is to alter protein coding sequences, and to a lesser extent RNA splicing, in transcripts of the central nervous system. However, the function of the large numbers of A > I edits in non-coding sequence is unclear. One consequence of extensive RNA editing may be to reduce the amount of base-pairing in dsRNA. Editing of perfectly dsRNA will continue until 50-60% of the adenosines are edited and then the reaction stops, apparently because the edited molecule becomes less double stranded and is consequently less tightly bound by the dsRBDs of ADARs (Lehmann and Bass, 1999). A peculiarity of A > I editing is that despite the tendency to reduce mismatches in dsRNAs, ADARs are apparently conformed to edit most efficiently at A:C mismatches which would result in an increase in double-stranded character (Bass, 2002).

Several recent investigations have attempted to establish the interplay between RNA editing and RNA interference, given that both pathways act on long dsRNA (Bass, 2000). Mutation of the *adr* genes of *C. elegans* vastly reduces RNA editing, but is not fatal, and results in chemosensory defects (Tonkin et al., 2002). The *adr* deficient worms, like wild-type worms, do not elicit an RNAi response to dsRNA injected into the cytoplasm of cells. However, the *adr* deficient worms, but not the wild-type worms, exhibit gene silencing in response to nuclear encoded dsRNA. This suggests that in normal cells, RNA editing of endogenous dsRNA prevents it from entering the RNA interference pathway. If the ability to edit dsRNA is lost, for example by mutation of RNA editing enzymes, then dsRNA is able to enter the RNAi pathway and gene silencing occurs (Knight and Bass, 2002). In subsequent experiments it was demonstrated that the chemosensory defects observed in the *adr* deficient worms were rescued by additional inactivating mutations in two genes required for RNAi. This is consistent with the hypothesis that one of the functions of RNA editing is to prevent endogenously transcribed dsRNA from entering the RNAi pathway (Tonkin and Bass, 2003).

It has also been shown *in vitro*, that hyper-editing of dsRNAs by ADAR2 antagonises RNAi, and is accompanied by a decrease in the production of siRNAs (Scadden and Smith, 2001). Taken together, these results suggest a role for RNA editing in the regulation of whether an endogenously synthesised dsRNA enters the RNAi pathway. This regulation requires RNA editing to precede RNAi, achievable either through isolation from cytoplasmic Dicer, or through higher affinity binding of dsRNA to ADARs than to the components of RNAi.

1.5.4 A > I editing and human disease

Aberrant RNA editing has been observed in a variety of neurological disorders. Significantly reduced RNA editing at the GluR-B Q / R site was found in the spinal motor neurons of amyotrophic lateral sclerosis (ALS) patients (Kawahara et al., 2004). Under-editing of the same site was also observed in human brain tumours, and a link was proposed between lowered ADAR2 activity and the occurrence of epileptic seizures associated with malignant gliomas (Maas et al., 2001b). Increased Q/R site-editing of GluR-5 and GluR-6 was observed in brain tissue from patients with epilepsy

(Kortenbruck et al., 2001). Serotonin receptor RNA editing appears to change in mental disorders such as schizophrenia and depression (Niswender et al., 2001, Sodhi et al., 2001) and depressed suicide victims (Gurevich et al., 2002).

Heterozygous ADAR1 mutations have recently been identified as the cause of Dyschromatosis Symmetrica Hereditaria (DSH) (Miyamura et al., 2003). Patients with DSH have a good prognosis, and suffer only from patches of hyperpigmented and hypopigmented skin on the backs of hands and tops of feet. These findings are broadly consistent with the mild phenotypes of mice which are heterozygous for ADAR deficiency.

RNA editing by ADAR1 increases during acute inflammation and results in an increase in the inosine content of total mRNA to approximately 5% of all adenosine (Yang et al., 2003a). This response is associated with alterations in the abundance and intracellular localisation of ADAR1 splice variants (Yang et al., 2003b). The targets and functional consequences of this editing reaction are unknown.

1.5.5 Human C > U RNA editing enzymes

C to U RNA editing is catalysed by cytidine deaminases. The first of these to be identified, and the only which clearly catalyses C > U editing of RNA *in vivo* was APOBEC-1 (Teng et al., 1993). Subsequently the homologues AID, APOBEC-2, and APOBEC-3A to 3G were identified (Muramatsu et al., 1999, Jarmuz et al., 2002, Liao et al., 1999).

1.5.5.1 *APOBEC-1*

The Apolipoprotein B mRNA editing enzyme, catalytic polypeptide 1 (APOBEC1) is currently the only cytidine deaminase with a clear role in cytidine deamination of RNA *in vivo*. APOBEC-1 has catalytic, RNA binding and protein binding domains (Lau et al., 1994). The minimal components of a C > U editing complex are an APOBEC-1 homodimer bound to APOBEC-1 complementation factor (ACF) (Mehta et al., 2000). Another potential component of the C > U editing complex is the glycine-arginine-tyrosine-rich binding protein (GRY-RBP), which binds to and sequesters ACF, reducing RNA editing (Blanc et al., 2001). There is also evidence that APOBEC-1 is regulated by phosphorylation (Chen et al., 2001).

APOBEC-1 expression is restricted exclusively to the small intestine of humans (Teng et al., 1993), and the editing complex is located in the nucleus by virtue of a nuclear localization signal in ACF (Blanc et al., 2003). RNA editing takes place post-transcriptionally in the nucleus (Lau et al., 1991). Unlike APOBEC-1, ACF is widely expressed in human tissues suggesting that it may be involved in other RNA editing events.

1.5.5.2 *APOBEC-2*

APOBEC-2 on chromosome 6 was identified through sequence homology to APOBEC-1, and is evolutionarily conserved (Liao et al., 1999). *In vitro*, it shows weak intrinsic cytidine deamination activity but no RNA editing of the

APOBEC-1 substrate (ApoB RNA). It is expressed abundantly in the heart and skeletal muscles suggesting a role in RNA modification in these tissues. However, no natural substrate has been identified. APOBEC2 binds to and inhibits APOBEC1, suggesting that its *in vivo* role may be to regulate RNA editing by APOBEC1 (Anant et al., 2001).

1.5.5.3 *APOBEC-3 A - G*

A series of seven sequences with homology to APOBEC-1 were identified on human chromosome 22, and designated APOBEC3A to 3G as potential C > U RNA editing enzymes. (Jarmuz et al., 2002). However, recent research suggests that these enzymes are likely to catalyse C > U changes in DNA rather than RNA. For example, APOBEC3G appears to be responsible for G > A hypermutation of the HIV-1 RNA genome by C > U deamination of the minus strand DNA (Zhang et al., 2003). Other members of the APOBEC3 family may also play an antiviral role and may also contribute to the accumulation of mutations during the evolution of organisms or in cancer (Neuberger et al., 2003).

1.5.5.4 *Activation induced deaminase (AID)*

Activation-induced deaminase (AID) is another homologue of APOBEC-1. It has intrinsic cytidine deaminase activity, but no ApoB mRNA editing and is responsible for two processes which generate antibody diversity (Muramatsu et al., 1999, Muto et al., 2000). First, the process of class switch recombination involves the rearrangement of DNA at the Immunoglobulin (Ig) gene locus, resulting in a switch between antibody classes. Second, the process of somatic hypermutation involves the accumulation of massive numbers of point mutations in immunoglobulin variable genes, giving rise to high affinity antibodies (Muramatsu et al., 2000, Revy et al., 2000, Honjo et al., 2002). It is currently unclear whether AID is a DNA or RNA deaminase. The DNA deamination model for antibody diversification proposes that AID carries out localized deamination of dC to dU in DNA at the Immunoglobulin gene locus. Modified bases in the variable region of the Ig gene may be either copied or subject to error-prone repair giving rise to somatic hypermutation. Modified bases in the class switch region of the Ig gene may initiate strand cleavage and repair by non-homologous end joining, resulting in class switch recombination (Neuberger et al., 2003). In contrast, the RNA editing model proposes that AID acts at cytidine in an unknown mRNA to generate an active protein capable of catalysing class switch recombination (Begum et al., 2004).

1.5.6 Human C > U editing substrates

1.5.6.1 *Apolipoprotein B (apoB) mRNA*

Apolipoprotein B (apoB) mRNA is the only known substrate of APOBEC-1 in normal human tissues (Powell et al., 1987, Chen et al., 1987). In the intestine RNA editing by APOBEC-1 converts C > U at position 6666 of the apoB mRNA. This changes a glutamine codon (CAA) to a stop codon (UAA), and results in expression of a truncated protein product. The full length (apoB100) and truncated (apoB48) proteins assemble into lipoproteins with different properties and both forms are required for the transport of triglycerides and cholesterol around the body.

Several sequence elements within the apoB mRNA have been identified which are essential for RNA editing and are conserved from marsupials to man. An AU rich ‘mooring’ sequence (Shah et al., 1991) is located 4-5 nucleotides downstream of the editing site and is bound by APOBEC-1. The artificial insertion of this region into other sequences permits C to U editing (Anant et al., 1995). The 4-5 nucleotides separating the mooring sequence from the editing site is also essential and is termed the ‘spacer’ (Backus et al., 1994). Distant sequences flanking the editing site (termed 5’ and 3’ efficiency elements respectively) also play a role (Hersberger and Innerarity, 1998). Secondary structure analysis of the mRNA suggests formation of a stem-loop structure with the edited C6666 within the loop (Hersberger et al., 1999).

1.5.6.2 *C > U editing of NF1 mRNA*

C to U editing has been observed in the tumour suppressor protein neurofibromatosis type 1 (NF1) mRNA, which contains an apoB-like mooring sequence (Skuse et al., 1996). C > U editing is predicted to result in a truncation of NF1 just N-terminal to its GTPase activating domain. Editing at this site is greater in subjects with tumours than in healthy individuals suggesting that a functional loss of tumour suppressor activity could therefore be one consequence of NF1 RNA editing (Liao et al., 1999). NF1 editing shows no response to levels of APOBEC-1 concentration suggesting different editing machinery.

1.5.6.3 *Interleukin 12 Receptor beta subunit 2 (IL-12R beta2) mRNA*

A C > U editing site has been reported in the Interleukin 12 Receptor beta subunit 2 (IL-12R beta2) mRNA, resulting in an amino acid change from alanine to valine (Kondo et al., 2004). C to U RNA editing at this site was not detectable in all individuals, but was more frequent in sufferers of atopy than in healthy individuals. Editing appears to impair the IL12 signalling cascade, and reduces the amount of the signalling molecule interferon-γ released from cells.

1.5.7 C > U editing and disease

Overexpression of APOBEC-1 in mice and rabbits resulted in transgenic animals with liver dysplasia and hepatocellular carcinomas. It was subsequently shown that in these tumours, specificity of RNA editing of the apoB mRNA is lost, and a novel target (NAT1) is subject to aberrant editing by APOBEC-1 (Yamanaka et al., 1997). NAT1 has been renamed eukaryotic translation initiation factor 4g2 (EIF4g2) and is an inhibitor of translation *in vitro*. Since editing of this mRNA alters amino acids and creates stop codons, it was suggested that this would interfere with its repressor function, and could contribute to the tumour formation caused by APOBEC-1 overexpression.

Elevated levels of APOBEC-1 mRNA have been found in a number of human cancers, and overexpressed APOBEC-1 was shown to bind to and stabilize cmyc mRNA, suggesting that altered APOBEC-1 expression may in turn alter the stability of transcripts involved in cancers (Anant and Davidson, 2003)

1.5.8 Rare RNA edits of other classes

There are several reports of RNA editing by mechanisms other than A > I or C > U (Table 1-5). The majority of these edits are known only by a single example, and in no cases has the enzyme responsible for the edit been identified. cDNA clones from GluR-7 (which is not known to be subject to A > I RNA editing) were isolated from a human foetal brain cDNA library and found to, contain G > A and U > G variants resulting in Ser > Ala and Arg > Gln changes to the amino acid sequence respectively (Nutt et al., 1994).

U > A changes in the Alpha galactosidase mRNA were identified in cDNA clones and RT-PCR products derived from human skeletal muscle, cerebellum and a fibroblast cell line (Novo et al., 1995). The edit is predicted to result in a Phe > Tyr substitution in the protein, but the consequence of this change is unknown.

The Wilm's tumour suppressor gene (WT1) transcript was reported to undergo U > C RNA editing in RNA isolated from rat kidney, resulting in a leucine to proline amino acid substitution (Sharma et al., 1994). However, RNA editing at this position was not detected in a study of 15 primary Wilm's tumors from human patients (Gunning et al., 1996).

Transcripts of β-amyloid precursor protein (APP) and Ubiquitin-B mRNAs were found to harbour GA or GT dinucleotide deletions in the vicinity of GAGAG sequence motifs (van Leeuwen et al., 1998). The deletions result in frameshift mutations and altered proteins which are detectable in brain tissue from patients with Alzheimer's disease (AD). Mutant transcripts were present at a very low frequency (on average 6 / 20,000 cDNA clones). It is thought that aging neurons may become susceptible to transcriptional errors, resulting in accumulation of altered proteins which initiate degeneration.

Two of these unusual RNA edits involve unusual pyrimidine to purine conversions (U > G in GluR-7 and U > A in Alpha galactosidase), and those in APP and ubiquitin involve nucleotide deletions. These changes cannot be achieved by deamination reactions in the way that A > I and C > U edits occur, and therefore require novel mechanisms of RNA editing.

1.6 PROJECT INTRODUCTION

Inosine containing RNA has been found to be most abundant in the brain, with one inosine for every 17,000 nucleotides (Paul and Bass, 1998). According to this estimate, RNA editing of the GluR-B mRNA accounts for just 0.06% of A > I edited sites in rat brain, and the other known sites of RNA editing described above clearly do not account for the deficit. Furthermore, the existence of putative A > I editing enzymes with no known substrates, the unknown extent of C > U editing and the unexplained lethal phenotypes associated with a lack of RNA editing suggests that there are many more RNA editing targets to be discovered. In this thesis, a survey of RNA editing in the human brain provides an evaluation of the number, types and distribution of RNA edits associated with various classes of RNA.