A Survey of RNA Editing in the Human Brain

Matthew James Blow

Darwin College

October 2004

This dissertation is submitted for the degree of Doctor of Philosophy

DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. The dissertation does not exceed the word limit set by the Biology Degree Committee.

ACKNOWLEDGEMENTS

I wish to sincerely thank Mike Stratton for his enthusiasm, patience and wisdom throughout this project. I have been fortunate to have him as a supervisor. Thanks also to the entire cancer genome project for their help in many ways in the work that led to this thesis, but mostly for their good company. I am particularly grateful to my supervisors Richard Wooster and Andy Futreal for their guidance and sound advice. In addition, I would like to thank many other members of the Sanger institute, Nav Navaratnam from the MRC Clinical Sciences Centre at Hammersmith Hospital and Nick Coleman from the Hutchison/MRC Research Centre at Addenbrooke's Hospital for their useful comments. Finally, I would like to thank Trish and my family, especially Mum, Dad and Sarah, for their enduring love and encouragement.

ABSTRACT

RNA editing is a post-transcriptional modification of RNA that occurs in prokaryotes, plants and animals. It occurs by a range of mechanisms including nucleotide insertions and deletions and base substitutions. The aim of these studies was to provide an extensive and systematic survey of the classes and distribution of editing in human mRNA. More than 3Mb of sequence from a human brain cDNA library were compared to genomic DNA sequences from the same individual and to the reference human genome sequence. Approximately 1 in 2,000 nucleotides in the RNA sample from which the library was constructed were shown to be edited. All edits were adenosine to inosine (A > I), predominantly in Alu repeats in intronic and non-coding RNAs. No edits were found in coding sequence. Analysis of the genome in the vicinity of edited sequences strongly supports the notion that formation of intramolecular double-stranded RNA (dsRNA) by inverted sequence copies underlies most A>I editing. The likelihood of editing is increased by presence of the two inverted copies within the same intron, proximity of the two copies and a high local density of inverted copies. A > I editing exhibits some sequence specificity, and is less likely at an adenosine 3' to a guanosine and more likely at an adenosine 5' to a guanosine. Simulation of the dsRNA molecules that underlie known edits indicates that there is a greater likelihood of A > I editing at A:C mismatches than at other mismatches or at A:U matches. However, because A:U matches in dsRNA are more common than all mismatches, overall the likely effect of editing is to increase the number of mismatches in dsRNA. The potential functions of A>I RNA editing have been considered in the light of this survey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

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).

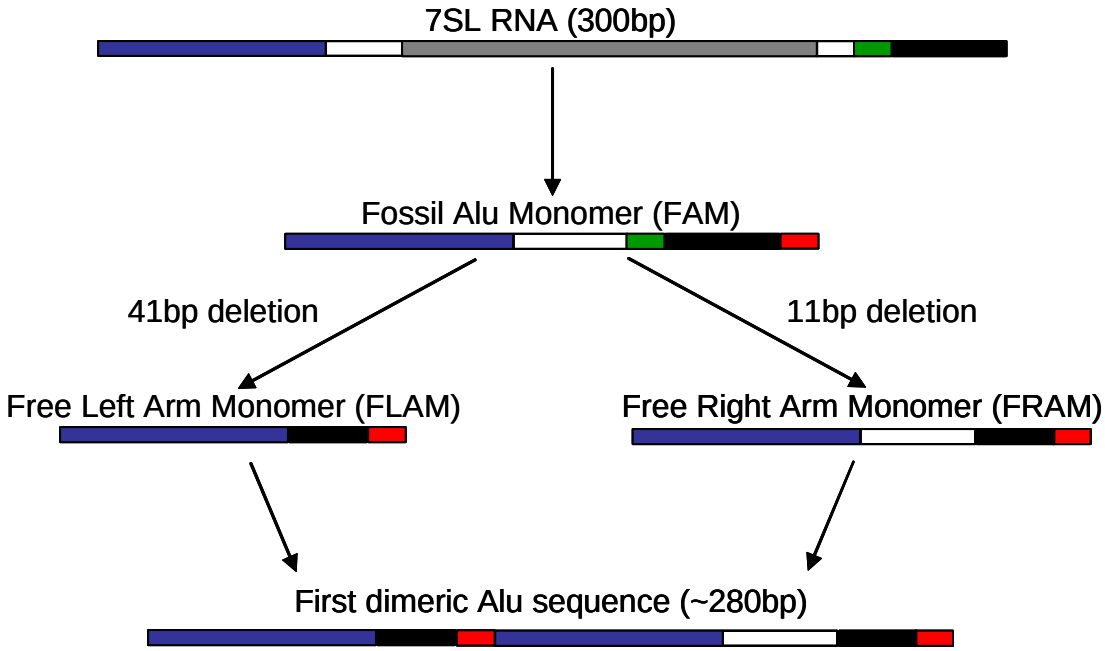


Figure 1-1 Evolution of Alu sequences. Retroposition of 7SL RNA generated the fossil Alu monomer (FAM). This in turn underwent duplication and diversification to generate the free left and free right Alu monomers (FLAM and FRAM respectively). Combination of a FLAM with a FRAM created the ancestral Alu 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 $tRNA_{Arg2}$, 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).

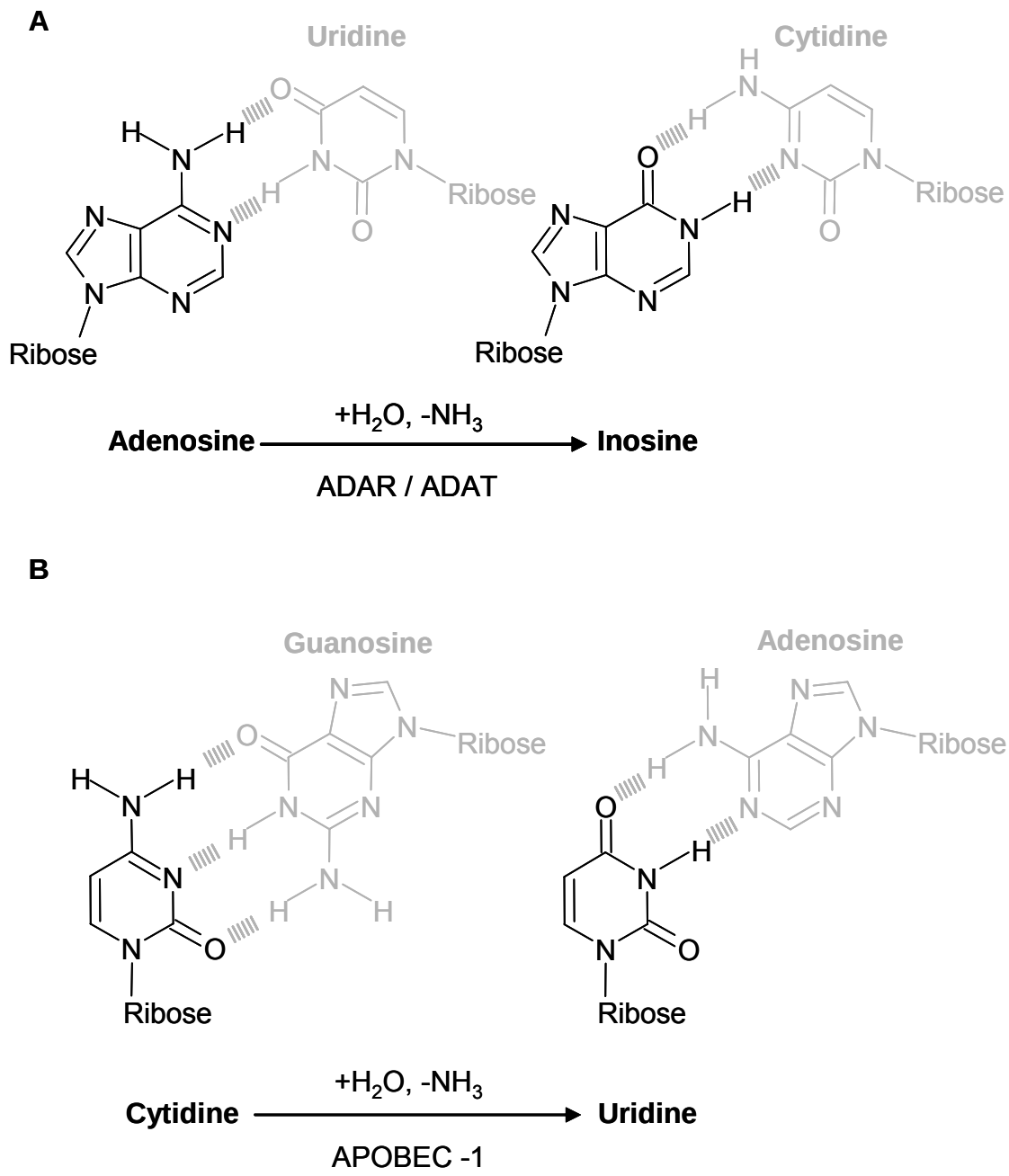


Figure 1-2 The effect of RNA editing on base pairing in RNA. **A.** Adenosine to inosine RNA editing catalysed by adenosine deaminases acting on RNA (ADARs) and tRNA (ADATs). **B.** Cytidine to uridine RNA editing catalysed by APOBEC-1. Base pairing is indicated by grey structures, dashed lines indicate 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	ADAR 2
		R > G	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	ADAR 1 / 2
		I > V	ADAR 1 / 2
		Y > C	ADAR 1 / 2
Serotonin receptor	A > I	I > V	ADAR 1 / 2
		I > M	ADAR 1 / 2
		N > D	ADAR 1 / 2
		N > S	ADAR 1 / 2
		N > G	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 negative-

strand 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 c-

myc 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.

2 METHODS

2.1 LABORATORY METHODS

2.1.1 Construction of a human cerebral cortex cDNA library

All procedures were approved by Cambridge Addenbrooke's Local Research Ethics Committee. Ethical approval was given to isolate nucleic acids from a sample from the cerebral cortex of a 67 year old male who had died following cardiac failure and a chest infection (LREC approval number 01/116).

Genomic DNA was extracted from 400mg outer grey matter of cerebral cortex in 20ml lysis buffer (75mM NaCl, 24mM EDTA pH8.0) plus 2ml SDS (10% w/v in water) and 200μl proteinase K (20mg/ml in water) at 37^{o}C overnight. Protein was precipitated and removed by addition of 8ml NaCl (5M) and centrifugation (3000rpm, 4^{o}C, 30 minutes). DNA was precipitated from the supernatant by addition of 30ml ice cold ethanol (100%) to 15ml supernatant, and retrieved by centrifugation (3000rpm, room temp, 1hour). The precipitated DNA was resuspended in 0.1x TE buffer.

Isolation of total RNA from the same tissue, and all subsequent stages of cDNA library synthesis from this material, were performed by Cytomyx Ltd. Briefly, total RNA was isolated using TRIzol reagent (Life Technologies) according to the manufacturer's instructions. Poly $(A)^{+}$ RNA was twice purified on oligo (dT)-cellulose columns. First strand cDNA was synthesized by

random primed reverse transcription of poly $(A)^+$ RNA using Stratascript reverse transcriptase (Stratagene). cDNA was cloned into EcoRI digested pUC19 plasmid using EcoR1 adapters, and transformed into ultracompetent *E. coli* cells from Stratagene. The percentage of clones containing cDNA inserts was estimated to be 83%, with insert size ranging from 0.4kb to 3kb and an average insert size of approximately 700bp. An amplified library of 8x 10^8 cells / ml was provided as a glycerol stock.

2.1.2 Sequencing of cDNA clones

2.1.2.1 *Reagents*

SOC: *SOB + 200 µl 20% glucose*.

SOB: *20 g tryptone, 5 g yeast extract, 10 ml 1M sodium chloride, 0.5 g potassium chloride, sterile water added up to 1 litre.*

LB agar: *10g bacto-tryptone, 5 g yeast extract, 10 g NaCl (pH7.4), sterile water added up to 1 litre.*

TY: *15g bacto-tryptone, 10g yeast extract, 5g NaCl (pH 7.4), sterile water added up to 1 litre.*

3M KOAc (pH5.5): *60 ml 5 M potassium acetate, 11.5 ml glacial acetic acid, 28.5 ml sterile water pH 4.8*

IPTG: *40 mg/ml in DMSO. Sterilised by filtration and stored at -20°C.*

Xgal: *50 mg/ml in ddH20. Sterilised by filtration and stored at -20°C.*

GTE: *50 mM Glucose, 25 mM Tris (pH7.5), 10 mM EDTA*

NaOH / SDS: *0.2M NaOH, 1% (w/v) SDS*

2.1.2.2 *Preparation of plasmid DNA*

Aliquots of the cDNA library glycerol stock were diluted 1 / 9,000 and 1 / 27,000 in SOC medium (see above), and aliquots of 100μl were then spread onto LB agar plates (with final concentrations of 50μg / ml ampicillin, 2mg / ml X-Gal, 4mg / ml IPTG). Plates were grown at 37^{o}C overnight (17 hours) then placed at 4^{o}C for 2 hours to allow the blue white screen to develop. Recombinant colonies were picked by hand, and used to inoculate 1ml 2xTY media (see above, with 50μg / ml ampicillin) in 20 x 96 deep well plates. Cells were grown in suspension at 37^{o}C overnight (22hrs) then collected by centrifugation (4000rpm, 3minutes) and media discarded. Cells were resuspended in 80μl GTE (with 250μg / ml RNaseA), lysed by addition of 80μl NaOH / SDS, and then neutralised with 80μl KOAc (3M). Bacterial genomic DNA was precipitated by addition of 120μl isopropanol, and removed along with cell debris by filtration under vacuum. Precipitated plasmid DNA was collected from the filtrate by centrifugation (4000rpm, 30 minutes) and washed twice by addition of 100μl Ethanol (70% v / v in sterile water) followed by centrifugation (4000rpm, 4^{o}C, 15 minutes) and removal of the supernatant. Plasmid DNA was dissolved in 60μl sterile water.

2.1.2.3 *Plasmid DNA sequencing*

Sequencing of plasmid DNA was carried out in 10μl reaction volumes in 96 well plates. Each plasmid DNA was sequenced once using the M13 forward primer (5'-CACGACGTTCTAAAACGACGGC-3'). Sequencing reactions were composed of 1μl primer (6 pmoles), 1μl BigDye mix, 3μl BigDye buffer, 2μl

sterile water and 3µl plasmid DNA. Thermocycling was performed on an MJ-Research PTC-225 thermal cycler. Following an initial activation step (96^{o}C for 30 seconds), were 34 cycles of denaturation (92^{o}C for 5 seconds), annealing (50^{o}C for 5 seconds) and extension (60^{o}C for 2minutes). DNA was then precipitated by addition of 10µl water and 50µl precipitation mix (see above), before centrifugation (4000rpm, 4^{o}C, 25minutes). Precipitated DNA was washed twice by addition of 100µl Ethanol (70% v / v in sterile water) followed by centrifugation (4000rpm, 4^{o}C, 4minutes) and removal of the supernatant. The precipitated DNA was allowed to dry and then dissolved in sterile water. Sequencing was performed using ABI Prism 3700 DNA analyzer (Applied Biosystems).

2.1.3 Sequencing of PCR and RT-PCR products

Matched total RNA and genomic DNA from the individual from whom the cDNA library was constructed were provided Cytomyx Ltd (see above). Additional matched genomic DNA and total RNA samples from human brain were obtained from BioChain Ltd.

2.1.3.1 Reagents

Exo / AP (per reaction): *1µl reaction buffer, 1µl dilution buffer, 0.05µl Exonuclease I (20U / µl, New England biolabs), 0.2µl Antartic Phosphatase (5U / µl, New England biolabs), 7.75µl sterile water.*

Exo / AP reaction buffer (stock): *100ml Tris (1M, pH 8.0), 50ml $MgCl_2$ (1M), 350ml sterile water.*

Exo / AP dilution buffer (stock): *25ml Tris (1M, pH 8.0), 475ml sterile water.*

BigDye terminator cocktail (stock): *2.9ml BigDye terminator V3.1 (Applied Biosystems), 17.1ml 5x BigDye reaction buffer (Applied Biosystems), 20ml sterile water.*

Precipitation mix: *500ml Ethanol, 10ml Sodium acetate (3M, pH 5.0), 20ml EDTA (0.1mM).*

2.1.3.2 *Reverse Transcription*

Total RNA was treated with DNAseI (Sigma) according to the manufacturer's instructions. Reverse transcription of total RNA was performed using Superscript III RNaseH$^-$ reverse transcriptase (invitrogen) and primed using random nonamers (Sigma). To 5µl DNaseI treated total RNA (100ng / µl) was added 2µl random nonamers (250ng / µl), 1µl dNTPs (10mM each) and 5µl sterile water. This mixture was heated to 65^{o}C for 5 minutes and then placed on ice for 1 minute. 4µl first strand reaction buffer, 1µl DTT (100mM), 1µl RNaseOUT ribonuclease inhibitor (Invitrogen, 40U / µl) and 1µl Superscript III reverse transcriptase (200U / µl) were added to the mixture. This was incubated at room temperature for 5 minutes, followed by 60 minutes at 50^{o}C and 15 minutes at 70^{o}C. 1µl cDNA was used in subsequent PCR reactions.

2.1.3.3 *Primer design*

The custom Perl programs create_design_tempate.pl and create_masked_design_template.pl were used to create primer design templates from the repeat masked or unmasked genome sequence

respectively in the vicinity of candidate RNA edits. Primer design was performed using a local copy of the Primer3 software (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi). The program was configured to design primers for PCR products as close to 200bp as possible, centred on the candidate RNA edit. In order to avoid non-specific amplification, first attempts at primer design were made using repeat masked sequence templates. If this failed, primers were designed using unmasked sequences. Primers were synthesised in house or by Sigma-Genosys.

2.1.3.4 *PCR*

PCR of genomic DNA and cDNA was carried out in 15µl reaction volumes in 96 well plates. To 1µl genomic DNA (20ng / µl), or 1µl cDNA was added 7.5µl primers (4ng / µl), 1.5µl dNTPs (2mM each), 1.5µl GeneAmp 10x reaction buffer (Applied Biosystems), 0.09µl Thermostart Taq (5U / µl, Abgene) and 3.4µl sterile water. Cycling was performed on an MJ-Research PTC-225 thermal cycler. Following an initial denaturation step of heating to 95^{o}C for 15 minutes, were 40 cycles of denaturation at 95^{o}C for 30 seconds, annealing at 60^{o}C for 30 seconds and extension at 72^{o}C for 30 seconds and a final extension step at 72^{o}C for 10 minutes. PCR products were evaluated by electrophoresis of 4µl aliquots on a 2% agarose gel (containing 0.2µg / ml ethidium bromide). To the remaining 11µl PCR products was added 10µl Exo / AP mix (see above), followed by incubation at 37^{o}C for 30 minutes and 80^{o}C for 15 minutes to remove residual primers and unreacted dNTPs.

2.1.3.5 *PCR product sequencing*

Sequencing of PCR products was carried out in 8µl reaction volumes in 384 well plates. For each PCR product, forward and reverse sequencing reactions were performed in duplicate. To 2µl sense or anti-sense primer (15ng / µl) and 4µl BigDye terminator cocktail (see above) was added 2µl Exo / AP treated PCR product. Thermocycling was performed on an MJ-Research PTC-225 thermal cycler. Following an initial activation step of heating to 96^{o}C for 30 seconds, were 44 cycles of denaturation at 92^{o}C for 5 seconds, annealing at 50^{o}C for 5 seconds and extension at 60^{o}C for 2 minutes. DNA was then precipitated by addition of 25µl precipitation mix (see above), and centrifugation (4000rpm, 4^{o}C, 25minutes). Precipitated DNA was washed twice by addition of 30µl Ethanol (70% v / v in sterile water) followed by centrifugation (4000rpm, 4^{o}C, 4minutes) and removal of the supernatant. The precipitated DNA was allowed to dry and then dissolved in 10µl EDTA (0.1mM). Sequencing was performed using ABI 3730 DNA analyzer (Applied Biosystems).

2.2 COMPUTATIONAL METHODS

2.2.1 Programs and databases

Several freely available programs were used extensively in this thesis. Sequence traces were visualised using Trev and Gap4 which are part of the Staden package (http://staden.sourceforge.net/). cDNA clone sequences were aligned to the genome using web-based and locally installed copies of BLAST

(http://www.ncbi.nlm.nih.gov/BLAST/), and BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat), and visualised in the EnsEMBL genome browser (http://www.ensembl.org/) and UCSC genome browsers (http://genome.ucsc.edu) respectively. Pairwise comparisons were made using BLAST 2 sequences (http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html).

The human genome reference sequence used for these analyses was the NCBI_v34 'golden path', consisting of a single FASTA sequence for each of the chromosomes. The sequences were not repeat-masked, as RNA edits are known to occur in repeat sequences (Morse et al., 2002). The 44kb ribosomal RNA repeat unit reference sequence (U13369) which encodes the 28s, 5.8s and 18s rRNAs of the ribosome, and the human mitochondrial genome sequence reference (NC_001807) were appended to the database sequence. Annotation of the human genome reference sequence was obtained from EnsEMBL version 19 (http://www.ensembl.org/), using custom Perl programs (see below).

2.2.2 Custom Perl programs

Several custom computer programs written in the Perl programming language were used for cDNA clone sequence analysis. In particular, programs based around the EnsEMBL API (application programming interface) were written to query the EnsEMBL genome annotation database (version 19). A tutorial explaining EnsEMBL API was obtained from http://cvsweb.sanger.ac.uk/cgi-bin/cvsweb.cgi/ensembl/docs/tutorial/ensembl_tutorial.pdf. This document provides an overview of the EnsEMBL annotation

database structure, instructions for the installation of the EnsEMBL Perl modules and BioPerl modules, and examples of how to use these modules in simple Perl programs to connect to the EnsEMBL annotation databases and retrieve information. Custom Perl programs also made extensive use of EnsEMBL Perl modules (http://www.ensembl.org/Docs/Pdoc/ensembl/) and BioPerl modules (http://www.ensembl.org/Docs/Pdoc/bioperl-live/). Several books were also referred to extensively when developing custom Perl programs (Tisdall, 2001, Christiansen and Torkington, 2003). The main Perl programs used in this thesis are included on the CD attached to this thesis. The following are brief descriptions of these programs.

2.2.2.1 *cDNA sequence variant detection and annotation*

The following scripts were run sequentially, with the output of one program used as the input for the next.

parse_pslx.pl: This script processes the 'pslx' format BLAT alignments of cDNA clones to the genome reference sequence (Figure 2-1B). For each alignment, a BLAT score and percentage identity score is determined using the following calculations from the web-based BLAT program (http://genome.ucsc.edu/FAQ/FAQblat):

1. BLAT score = Number of matches – (Number of mismatches +Number of gaps in the query sequence + number of gaps in the database sequence)
2. Percentage score = 100 – (Millibad x 0.1)
3. Millibad = (1000 (M + QI + 3log(1+QA – HA)) / (M +MM + RM)

Where, M = Matches, QI = inserts in the query sequence, QA = Query alignment length, HA = hit alignment length, MM = mismatches, RM = repeatmatches. The millibad value is a measure of mismatches in parts per thousand. This value uses logarithms to allow for large insertions in the alignment (i.e. introns). For each cDNA clone, the program returns the highest scoring BLAT alignment, along with the genomic coordinates of any sequence variants (Figure 2-1C).

verify_variants.pl: Takes the list of sequence variants generated by the previous script, and determines the trace quality in the vicinity of sequence variants by reference to the sequence trace quality file. Trace quality is given by 'q-scores' which are generated by the phred base-calling algorithm. Only variants that are in high quality sequence are returned (see section 4.2.1).

annotate_SNPs.pl: Takes the list of high quality sequence variants generated by the previous script and uses their genomic coordinates to query the dbSNP database (http://www.ncbi.nlm.nih.gov/SNP/). Known SNPs are indicated (Figure 2-1C).

annotate_exons.pl: Uses the genomic coordinates of the cDNA clone alignments to retrieve the coordinates of all overlapping exons from known genes and predicted genes in EnsEMBL. It then compares the coordinates of each 'exon' of the cDNA clone with each exon retrieved from the database. If all 'exons' of the cDNA clone align to exons of the same gene from EnsEMBL, then that gene is taken to be the one from which the cDNA clone is derived. The gene name and the genomic coordinates of any intron / exon boundaries that overlap with the cDNA clone are returned (see section 3.2.6.1).

annotate_coding.pl: Compares the cDNA clone with the gene from which it was derived and returns the start and end of any 3'UTR, coding sequence and 5'UTR in genomic coordinates.

annotate_repeats.pl: Compares the genomic coordinates of the cDNA clone with all repeat sequences on the overlapping segment of the genome, and returns the repeat family name, repeat class name and the genomic start and end coordinates of any repeat elements which overlap with the cDNA clone.

annotate_variants.pl: Annotates variants using a two letter code. Known SNPs (KS) were previously identified (see annotate_SNPs.pl). Assumed hyperedits (AH) were identified by comparing the number of each class of variants in a cDNA sequence. If a sequence had more than three variants of a single type (eg A>G, T>C or C>T, G>A) that accounted for more than 75% of all variants, it was classed as hyperedited and all variants of that type were assumed edits (see 4.2.2.1). Other variants were annotated following experimental evaluation (Confirmed edit = CE, novel SNP = NS, artefact = CA, unknown = UK).

annotation_summary.pl: Calculates the total amount of cDNA library sequence from various sequence categories (e.g. the total amount of intronic sequence), from the annotation of individual clones.

2.2.2.2 *Analysis of edited Alu sequences*

alu_anlysis.pl: Identifies Alus present in cDNA clones that are from the introns of known genes. All other Alu sequences from that intron are retrieved from the EnsEMBL, and their position in relation to the reference Alu is recorded in genomic coordinates. For each 1kb window of sequence either

side of the reference Alu the number of overlapping bases between i) the reference Alu and flanking same-sense Alus, and ii) the reference Alu and flanking anti-sense Alus, is calculated.

nearest_Alus.pl: Creates a file containing the sequences of the edited Alu, the nearest same-sense Alu, the nearest anti-sense Alu, and the position of RNA edits in the edited Alu.

opposing_base.pl: Aligns the edited Alu to the nearest same-sense Alu and the nearest anti-sense Alu using a locally installed copy of blast2sequences (see above), then identifies edited bases in the alignment and returns the total number of edited and unedited adenosines at matched bases and at each class of mismatched base.

seq_context.pl: For every edited and unedited adenosine from all cDNA clones returns the 10bp of sequence from either side of that adenosine

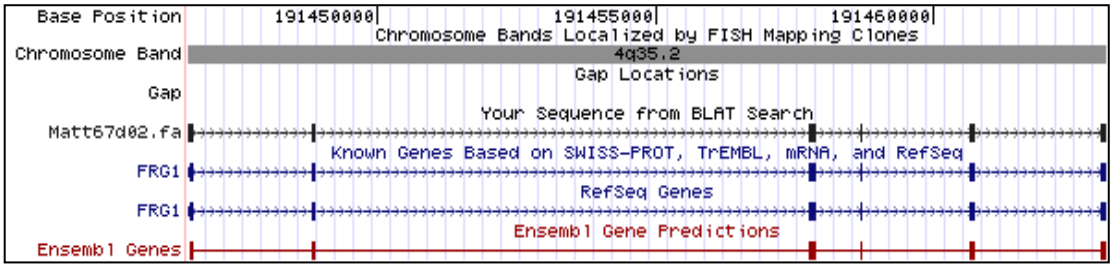
2.2.3 Detection of high quality sequence variants

The cDNA clone sequences were processed using the analysis software ASP (http://www.sanger.ac.uk/Software/sequencing/docs/asp/). ASP uses the Phred base-calling algorithm, and converts sequence traces into SCF (standard chromatogram format). The program also produces a 'quality' file for each clone, consisting of the Phred-called nucleotide sequence of the clone, along with the numerical phred quality score (q-score) for each nucleotide. Bases which had phred quality scores of less than 15 were masked using the custom Perl program caf_to_fa.pl and clone sequence derived from the cloning vector or adapter sequences was masked using the alignment

program cross_match (Green, unpublished). The cDNA clone sequences were then combined into a single file in FASTA format.

For each alignment, the parse_pslx.pl was used to compare the two aligned sequences base by base and generate a list of sequence variants, and their coordinates in the cDNA clone and the genome. High quality sequence variants were evaluated by reference to quality score files using the Perl program *verify_variants.pl*, and known SNPs identified using the program *annotate_SNPs.pl*. The cDNA clone sequences were then annotated using the custom Perl programs *annotate_exon.pl*, *annotate_coding.pl* and *annotate_repeats.pl* (see above). Candidate hyperedited sequences were identified as sequences that had more than three variants of a single type (eg A>G, T>C or C>T, G>A) that accounted for more than 75% of all variants. An example of the output of these programs, compared with the original BLAT alignment '*pslx*' output, is shown in Figure 2-1. The custom Perl program *annotation_summary.pl* was used to calculate sequence composition of the whole cDNA library from the fully annotated files (e.g. Figure 2-1D).

A



B

```
440[1]  20[2]  0[3]  1[4]  2[5]  10[6]  5[7]  18916[8]  +[9]  Matt67d02.fa[10]  631[11]  56[12]  527[13]  9.1-
134455819[14]  134455819[15]  63297740[16]  63317117[17]  7[18]  89,68,5,126,58,63,52,[19]
56,154,222,227,353,411,475,[20]
63297740,63299983,63302090,63314253,63315160,63317002,63317065,[21]
ccggcctcagcctctccgcgcagaagttgcccggagccatggccgagtactcctatgtgaagtctaccaagctcgtgctcaagggaacc,agtaa
gaagaaaaagagcaaagataagaaaagaaaaagagaagaagatgaagaaacccagcttgatat,tgttg,gaatctggtggacagtaacaaacttt
tggtgaaatttcaggaaccatagccattgaaatggataagggaacctatatacatgcactcgacaatggtctttttaccctgggagctccacaca
aagaag,ttgatgagggccctagtcctccagagcagtttacggctgtcaaattatctgattccag,aattgccctgaagtctggctatggaaaat
atcttggtataaattcagatggacttgttgttgg,cgttcagatgcaattggaccangagaacaatgggaaccagtctttcaaaatg,[22]
ccggcttcagcctctccgcgcagaagtctcccggagccatggcctagtattcttatgtgaagtctaccaagcttgtgctcaagggaacc,agtaa
gaagaaaaagagcaaagataagaagagaaaaagagaagaagatgaagaaacccagcttgatat,tgttg,gaatctggtgaacagtaacaaacttt
tggtgaaatttcaggaaccatagccattgaagtggatgagggaacctatatacatgcactcaacaatggtctttttaccctgggagctccacaca
aagaag,ttgatgagggccctagtcctccagagcagtttatggctgtcaaattatctaattccag,aatcgccctgaaacctggctatggaaaat
accttagtataaattcagatgaacttgttgttgg,cgttcagatgcaattggaccaagagaacaatgggaaccagtctttcaaaatg,[23]
```

C

```
Matt67d02[24] 4,+,6,47,631,191446570,191463111[25]
*1,47,156,191446570,191446680,V,*2,157,227,191448811,191448881,*3,228,353,191457771,191457896,*
4,354,411,191458677,191458734,*5,412,526,191460646,191460760,*6,527,631,191463007,191463111,[26]
1,0,2,0,0,0,0,0,1,0,0,0,0,0,1[27]  g,85,t,191446609[28]  g,85,t,191446609 t,112,c,191446636  -
,50,1,191446573[29]
```

D

```
Matt67d02[30]  4,+,6,51,631,191558011,191574547[31]
*1,51,156,191558011,191558116,V,*2,157,227,191560247,191560317,*3,228,353,191569207,191569332,*
4,354,411,191570113,191570170,*5,412,526,191572082,191572196,*6,527,631,191574443,191574547,[31]
ENSG00000109536[32]  KNOWN_GENE[33]
FRG1_PROTEIN_(FSHD_REGION_GENE_1_PROTEIN)._[Source:SWISSPROT;Acc:Q14331][34]  1[35]  191558011-
191558055, 191558056-191574547, 0-0[36]  191558011-191558116,191560247-191560317,191569207-
191569332,191570113-191570170,191572082-191572196,191574443-191574547,[37] [38]
Low_complexity,191560249,191560299,0 dust,191560249,191560300,0 dust,191560249,191560300,0[39]
0,0,2,0,0,0,0,0,1,0,0,0,0,0,1[40]  g,85,t,191558045:KS t,112,c,191558072:UK[41]
```

Figure 2-1 Automated detection and annotation of sequence variants. Annotation of a single cDNA clone sequence is shown **A.** Alignment of a cDNA clone sequence to the human genome reference sequence using the web based BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat). **B.** Alignment of the same sequence to the human genome reference sequence using a locally installed copy of the BLAT set to output results in pslx format. Output values are: [1]number of matched query sequence bases in the alignment, [2]number of

mismatched query sequence bases in the alignment, [3] number of repetitive DNA elements in sequence (returns zero as repeat masking is not used), [4]number of Ns, [5]number of gaps in query sequence, [6]number bases in gaps in query sequence, [7]number of gaps in hit sequence, [8] number bases in gaps in hit sequence, [9]strand, [10]name of query sequence, [11]length of query sequence, [12]start of alignment in query sequence, [13]end of alignment in query sequence, [14]name of database sequence, [15]length of database sequence, [16]start of alignment in hit sequence, [17]end of alignment in hit sequence [18]number of blocks of alignment (a "block" of alignment generally refers to an exon, however an ins/del polymorphism between two sequences will result in an exon being broken into two blocks of alignment), [19]lengths of blocks of alignment, [20]start of each block of alignment in query sequence, [21]start of each block of alignment in hit sequence, [22]query sequence of each block of alignment (comma separated), [23]hit sequence of each block of alignment (comma separated). **C.** Output following analysis with custom Perl programs parse_pslx.pl, verify_variants.pl and annotate_snps.pl. [24]cDNA clone, [25]coordinates of the alignment (chromosome number, chromosome strand, number of 'exons', start in cDNA sequence, end in cDNA sequence, start on chromosome, end on chromosome), [26]The coordinates of each exon (as for coordinates of alignment), [27]Number of high quality variants of each categories (total number of insertions, total number of deletions, total number of substitutions, number of A > C, number of A >G variants, number of A > T variants, number of C > A variants, number of C >G variants, number of C >T variants, number of G > A variants, number of G > C variants, number of G > T variants, number of T > A variants, number of T > C variants, number of T >

G variants), [28]Known SNPs (nucleotide in cDNA clone, position in cDNA clone, nucleotide on chromosome, position on chromosome), [29] All high quality variants (as for known SNPs). **D**. Output following analysis with custom Perl programs *annotate_exons.pl*, *annotate_coding.pl*, *annotate_repeats.pl* and *annotate_variants.pl*. [29]cDNA clone, [30]coordinates of the alignment (as for **C**), [31]The coordinates of each exon (as for coordinates of alignment), [32] EnsEMBL gene ID, [33] EnsEMBL classification, [34]EnsEMBL gene description, [35]Strand of Gene, [36]Genomic coordinates of coding sequence (5'UTR start in clone - 5'UTR end in clone, coding start in clone – coding end in clone, 3'UTR start in clone – 3'UTR end in clone), [37]Coordinates of exonic sequence (exon start in clone – exon end in clone), [38]Coordinates of intronic sequence (intron start in clone – intron end in clone), [39]Coordinates of repeat sequence (repeat class, start in clone, end in clone), [40] (as for [27]), [41]annotated variants (as for [28] except annotated by 2 letter code).

2.2.4 Analysis of edited Alu sequences

Full length Alu sequences corresponding to repeats sequenced as part of cDNA clones were obtained from EnsEMBL using the Perl script *alu_analysis.pl*. For all studies of edited and unedited Alus (Figures 5-4 to 5-8), only Alus for which at least 80% of their genomic extent was sequenced as part of a cDNA clone were used as reference Alus in the analyses. For studies of the patterns of Alu elements in the same intron as edited and unedited Alus (Figures 5-4 to 5-6), only Alu elements from cDNA clones which aligned to the introns of EnsEMBL known genes were used as reference Alus in the analyses. Intron sizes and the orientation and genomic coordinates of

flanking Alus were obtained from the EnsEMBL genome annotation database using the genomic coordinates of reference Alus as queries.

Reference Alus were aligned to neighbouring Alus using BLAST (http://www.ncbi.nlm.nih.gov/blast/bl2seq/) (Table 6-1). The positions of mismatches in the alignments were recorded and compared with the positions of edited bases in the reference sequence. BLAST is not generally considered an algorithm for simulating RNA duplexes. However, we compared the base pairing produced by BLAST to that generated by MFOLD, a program designed to simulate RNA secondary structure and found that for the 32 edited bases evaluated, the predicted base pairing was identical using the two methods. We therefore used BLAST for this purpose.

Multiple alignments were constructed from all edited Alu sequences using CLUSTALW. Information from all sequences was used to calculate the percent nucleotide composition at each position in the alignment. Only bases sequenced in this study were used to calculate the proportion of adenosines edited at each position in the alignment.

3 SEQUENCING AND EVALUATION OF A HUMAN BRAIN cDNA LIBRARY

3.1 INTRODUCTION

The aim of this thesis was to utilise high throughput nucleotide sequencing and mutation detection, coupled to the human genome reference sequence to perform a systematic survey of RNA editing. Although many different tissue types have been shown to contain edited RNAs, previous observations suggest that mammalian A > I editing is most abundant in the brain. The inosine content of total RNA from the brain is higher than in total RNA from any other tissue (Paul and Bass, 1998), the known A > I RNA editing enzymes ADAR1 and ADAR2 are most highly expressed in the brain (Kim et al., 1994, Melcher et al., 1996b), and the putative A > I editing enzyme ADAR3 is expressed exclusively in the brain (Chen et al., 2000).

Based on estimates of 1 in 17,000 nucleotides of human brain RNA being edited from A > I (Paul and Bass, 1998), sequencing of 3Mb from a cDNA library would be expected to yield over 150 A > I edits alone. This would provide insight into the genome-wide targets and patterns of RNA editing. Therefore, the cDNA library used for this survey was constructed from human cerebral cortex RNA.

In this chapter, over 3Mb cDNA sequenced at random from a human brain cDNA library was aligned to the human genome reference sequence. As

these alignments were subsequently used to identify novel RNA edits (see Chapter 4), it was important to ascertain whether they were representative of the transcriptome of human brain cells. Therefore, the alignments of cDNA clones to the genome were used to evaluate the quality of the cDNA library with respect to contamination by genomic DNA. The composition of the cDNA library was evaluated by annotation of known genes.

3.2 RESULTS

3.2.1 Construction of a human brain cDNA library

A central requirement of this project was matching RNA and genomic DNA from the same individual, allowing us to easily clarify which of the sequence variants identified through alignment of the cDNA clone sequences to the genome reference sequence were due to SNPs. Matching nucleic acids were isolated *de novo* from a human cerebral cortex tissue sample. RNA was submitted to Cytomyx Ltd (Cambridge, UK) who prepared a cDNA library.

Tissue sections were removed from the cerebral cortex of a male donor, whose cause of death was congestive cardiac failure. The brain tissue had been frozen with a post-mortem delay of 9 hours, and was classified as normal from its appearance under the microscope. Total RNA was analysed by denaturing agarose gel electrophoresis. The 28S and 18S ribosomal RNAs were clearly visible indicating that the RNA was reasonably intact. Poly-$(A)^+$ RNA was isolated by two rounds of purification on an oligo-(dT)-cellulose

column. Analysis by agarose gel electrophoresis indicated that the majority of the ribosomal RNA was removed (Figure 3-1A). For the purposes of identifying SNPs between the tissue donor and the human genome reference sequence, genomic DNA was isolated from tissue adjacent to that used in the preparation of RNA.

To avoid any bias towards the 3' end of mRNAs, cDNA synthesis was primed using random hexamers rather than oligo-dT primers. The primary library contained 3.3 x 10^5 colony forming units (cfu). The library was subject to one round of amplification in semi-solid media, to reduce representational biases. The final titre of the amplified cDNA library was >8 x 10^8 cfu / ml. To estimate cloning efficiency, 30 individual colonies were picked at random. Plasmid DNA was isolated and subject to *EcoRI* digestion prior to electrophoresis on a 1% agarose gel (Figure 3-1B). cDNA inserts were found in 83% of the clones, with the insert sizes ranging from 0.4kb to 3kb (data provided by Cytomyx Ltd.).

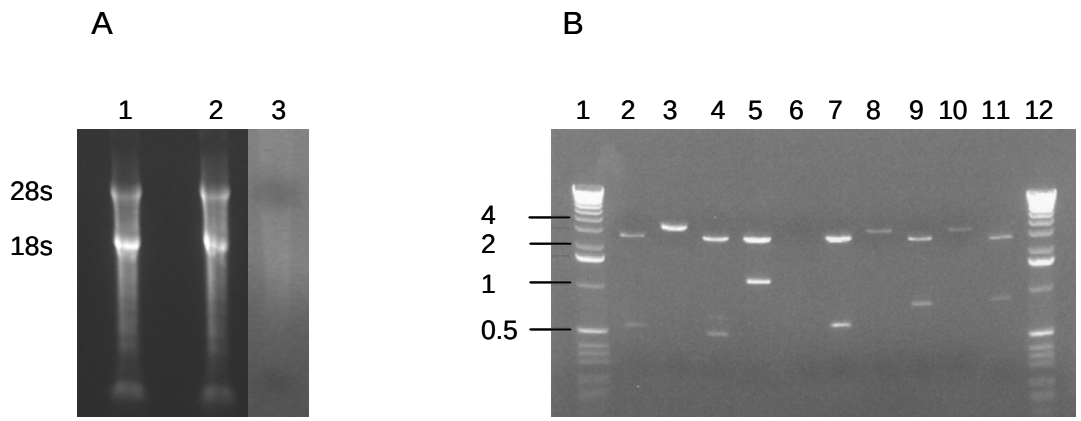


Figure 3-1 Analysis of Human Cerebral cortex nucleic acid preparations. **A**. Electrophoresis of human cerebral cortex total RNA (lanes 1 and 2 in duplicate), and poly-$(A)^+$ purified mRNA (lane 3). Bands corresponding to the 28S and 18S ribosomal RNA subunits are indicated. **B**. Electrophoresis of *EcoRI* digested cDNA from a random sample of 10 cDNA clones (lanes 2 to 11). Lanes 1 and 12 contain a 10kb DNA ladder with bands corresponding to 0.5kb, 1kb, 2kb and 4kb. Images provided by Cytomyx Ltd (Cambridge, UK).

3.2.2 Evaluation of the cDNA library

The sequence composition of the cDNA library was evaluated by sequencing 384 cDNA clones and aligning them to the human genome reference sequence using BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat). For each cDNA clone, the alignment with the highest BLAT score was viewed in the genome browser. The clones were then categorized according to how their alignment to the genome corresponded with known genes (Table 3-1). If a sequence matched more than one category, the category nearest the top of the table took priority.

Category	Description
Exonic (spliced)	The clone aligned to the spliced exons of a known gene.
Exonic (unspliced)	The clone aligned to a single exon of a known gene but cannot be confirmed as spliced
Intronic	The clone aligned to an intron of a known gene
Intergenic	The clone did not align to any known gene
Mitochondrial	The clone aligned to the mitochondrial genome
Failed	The clone failed due to trace quality, the clone had no insert or the clone did not align to the genome.

Table 3-1 Categorisation of cDNA clone sequences based on their alignment to the human genome using BLAT.

76% (292 / 384) clones could be aligned to the genome using BLAT (Table 3-2). Of the aligning clones, only 19% (56 / 292) were exonic (spliced). 14% (41 / 292) of clones were exonic (unspliced), 32% (93 / 292) clones were derived from intronic sequences and 21% (61 / 292) clones were derived from intergenic regions of the genome. Clones aligning to the mitochondrial genome accounted for 14% (41 / 292) of the sequences. Exonic / spliced sequences are the only class of sequence for which alignment to the genome provides direct evidence that they are derived from spliced mRNAs. In principle, all of the remaining 81% of sequences could result from contaminating genomic DNA.

	Clones	Bases	%
Exonic (spliced)	56	28024	19
Exonic (unspliced)	41	18337	14
Mitochondrial	41	20111	14
Intronic	93	45929	32
Intergenic	61	29936	21
Failed	92	-	-
Total sequence	384	142337	100

Table 3-2 Evaluation of the sequence composition of a human brain cDNA library.

The evaluation of the cDNA library indicated potential contamination with genomic DNA. However, the human genome is composed of approximately 2% coding sequence, 20% intronic sequence and 78% intergenic sequence. By contrast, the cDNA library contained 33% coding sequence, 32% intronic sequence and 21% intergenic sequence with the remaining 14% mitochondrial sequences. This indicated that the cDNA library was at least partially enriched in transcribed RNAs. Moreover, there was no guarantee that a cDNA library from another source would give better results. Therefore, it was decided to pursue further experiments with this cDNA library.

3.2.3 Sequencing of 10,000 clones from a human brain cDNA library

The initial sequencing target of this survey was to analyse 1Mb of coding RNA sequence. To compensate for the exonic (spliced) cDNA content of

approximately 25%, the number of clones sequenced from the library was increased four fold. With an average of 400 bases of high quality sequence per clone it was estimated that 10,000 clones would give a total of 4Mb sequence including the desired 1Mb of coding sequence for our analysis.

In total, 9,341 clones comprising 4,982,043bp cDNA sequence were successfully sequenced from the cDNA library. Of this sequence, 15.6% (780,979bp / 4,982,043bp) was masked as cloning vector sequence using the cross_match algorithm (see Methods), leaving 4,201,064bp cDNA sequence.

3.2.4 Automated alignment of 9,341 cDNA clones to the human genome reference sequence

All 9,341 cDNA clones were aligned to the human genome reference sequence (NCBIv34) using BLAT. This program was used in preference to other alignment programs such as BLAST or SSAHA because it is faster and because it can more accurately align spliced cDNA sequences. BLAST and SSAHA produce a separate alignment for each exon of a spliced cDNA sequence, and bases at the ends of an exon may appear in more than one alignment. In contrast, BLAT combines the alignments of individual exons to give a single alignment in which each base of the cDNA sequence is used only once, and in which individual exons are correctly aligned by comparison with splice site consensus sequences (Kent, 2002).

Incorrectly aligned cDNA clone sequences could give rise to erroneous sequence variants which appear to be candidate RNA edits. Therefore, for

each cDNA clone, the BLAT score and the percentage identity score of the two highest scoring alignments to the genome were identified, and used to identify cDNA clones that were incorrectly aligned to the genome.

First, to remove sequences which were incorrectly aligned because of a poor quality sequence trace, or because the target sequence was not present in the genome database, any cDNA clone with a top scoring BLAT alignment of less than 95% was rejected. This relatively low percentage score allowed for the fact that a heavily edited RNA would have a reduced identity to the genome. A cut off of 95% allowed for a 500bp clone to be edited at up to 25 bases in an otherwise perfect alignment.

Second, to remove sequences which aligned with a similar score to more than one region of the genome, the scores of the top two alignments were compared. Any top scoring BLAT alignment that also had a higher percentage score than the second best BLAT alignment was deemed correct. If the second BLAT alignment had a higher percentage score than the first alignment, it was considered potentially ambiguous. In these cases, the product of the BLAT score and percentage score was calculated for the top two alignments. The value obtained for the second alignment was then expressed as a percentage of the value obtained for the top alignment. If this value was greater than 95%, the alignments were considered ambiguous, and the cDNA clone sequence was rejected. If the value was less than 95% similar, the top hit was judged to be better than the second and was accepted.

In total, 92% (8,552 / 9,341) clones comprising 3,787,472bp aligned to the genome (Figure 3-2). Of these, 97% (8,328 / 8,552) clones comprising 3,715,067bp sequence aligned unambiguously to the human genome reference sequences (Figure 3-2). The 1,013 sequences failing to align to the genome were composed of 789 clones which failed to align to the genome at all, 65 clones aligned to the genome with less than 95% identity, and 159 clones for which the top alignment could not be clearly distinguished from lower scoring alignments.

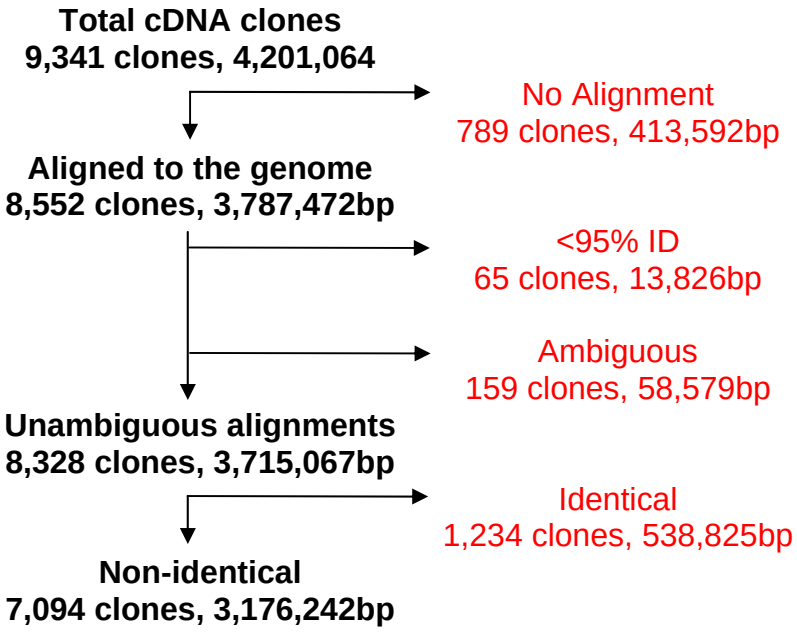


Figure 3-2 Processing of cDNA clone sequence data. The values in red indicate sequences that were rejected for various criteria. Ambiguous sequences are those which aligned to two regions of the genome with similar BLAT scores and percentage scores (defined in the text). Values in black show the remaining good quality cDNA clone sequences at each stage of the analysis.

3.2.4.1 *Investigation of clones failing to align to the genome or failing to align unambiguously*

To investigate the causes of sequences failing to align to the genome, and sequences rejected because of incorrect alignment, examples of each failed category were examined by manual BLAT and BLASTN alignment to the genome. 20 / 789 sequences that failed to align to the genome at all were investigated more closely. The majority (15 / 20) were completely masked as vector sequences, and therefore contained no cDNA insert. One sequence was aligned to 35bp of the mitochondrial genome using BLASTN and was beneath the limits of detection of the BLAT program. The remaining four clone sequences did not align to any sequence in the database. Their sequence traces were of poor quality following mono-nucleotide repeats.

20 out of 65 clones that were rejected because they aligned to the genome with less than 95% identity were looked at in more detail. Most (16 / 20) were due to poor quality sequence traces, and higher quality alignments could not be detected using BLASTN. Three sequences aligned to clones of human chromosome sequences. These clone sequences are not represented in the 'golden path' sequence and therefore were not detected in our BLAT analysis. The remaining sequence had a good quality sequence trace, but could not be aligned to any sequence with BLASTN. The best BLAT alignment was 499 bases long and contained 23 mismatches from A in the genome to G in the clone sequence and only one other (T to G) sequence variant. This pattern of variation was best explained by extensive A to I type RNA editing of the cDNA

clone. This sequence was the first putative novel RNA edited sequence to be identified. A technique was later developed to recover all potentially heavily edited sequences from the cDNA library (see Results, Chapter 4).

20 out of 159 clone sequences rejected because their best and second best alignments to the genome had similar BLAT scores were studied in more detail. 12 of the 20 sequences aligned to more than one region of the same chromosome with identical or near identical scores, and another sequence aligned to two different chromosomes with identical scores. Two sequences aligned to the mitochondrial genome and a region on chromosome 1 with near identical scores. Another two sequences were aligned ambiguously to a gene and a pseudogene. Finally, three sequences were entirely derived from LINE elements and aligned to more than 50 sites in the genome with identical scores. All 159 ambiguous alignments to the genome were removed from subsequent analyses.

Overall, the measures applied to identify ambiguously aligned clones were successful and resulted in 224 being rejected. Apart from the novel heavily edited sequences (which were subsequently recovered) none of the sequences were rejected incorrectly. It is, however, likely that a small number of incorrectly aligned sequences will have been missed because they fell within the acceptable identity scores or similarity scores and have been included in the subsequent analyses

3.2.4.2 *Identification of identical clones and non-identical overlapping clones*

The initial evaluation of the cDNA library indicated that around 1% of the exonic (spliced) clone sequences were overlapping. Identical overlapping clones can in principle derive from a single clone which was amplified in the synthesis of the cDNA library. As artefacts of the cloning process they required removal from our analyses. Non-identical overlapping clones can occur when a highly expressed transcript is cloned and sequenced multiple times. These clones would not be expected to be identical in sequence and were retained as they provided potentially useful biological information.

In principle, ‘identical’ cDNA clones should align to the genome with the same starting site. In practice ‘identical’ clones may align to the genome with slightly different starting positions. The vector masking program can produce subtly different results at the cDNA insert site so that the apparent first base of the insert can vary. Furthermore, different sequence traces of the same clone can produce different results depending on the start of good quality sequence.

To distinguish between identical and non-identical overlapping clones, a comparison of the start position of cDNA clone alignments was performed. All 8,328 unambiguous cDNA clone alignments were sorted by chromosome and start position. The start of each alignment was compared with the start position of the previous clone aligned to that chromosome, and the distance between the two was recorded (Figure 3-3).

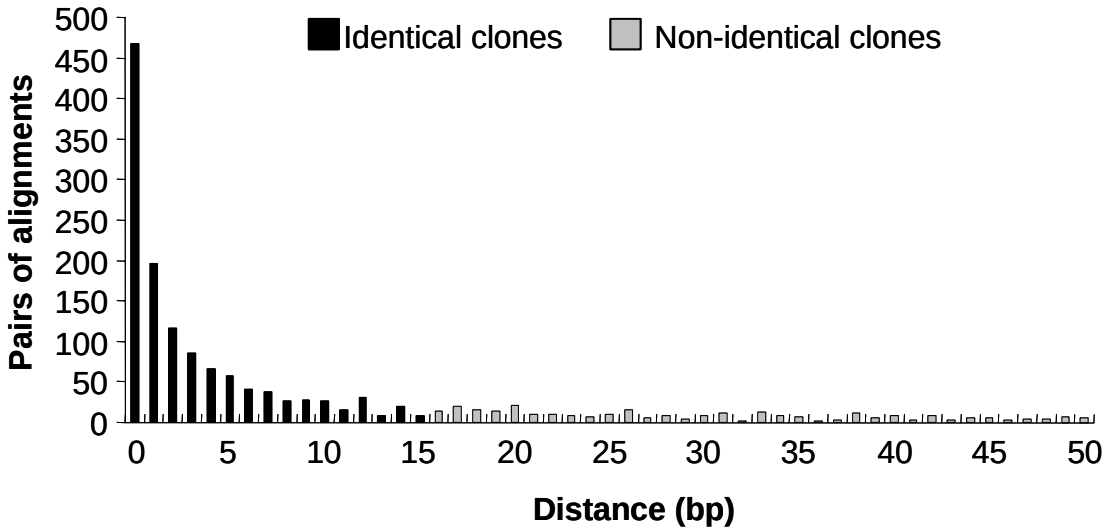


Figure 3-3 Discrimination of identical and non-identical overlapping clones. The number of bases separating the start positions of overlapping alignments was used to discriminate 'identical' from 'non-identical' overlapping cDNA clones. Pairs of alignments with start positions separated by 15 or less bases were deemed to be 'identical' (black bars). Pairs of alignments with start positions separated by more than 15bp were deemed to be non-identical 'overlapping' clones (grey bars).

468 pairs of alignments start at exactly the same position in the genome. These and other alignments which are separated by only a few bases are seen frequently and represent identical clones. Pairs of alignments that are separated by greater distances are less frequent and represent non-identical 'overlapping' clones. The alignments and sequence traces of all pairs of alignments separated by 15 bp were examined. Of nine pairs of alignments, all were non-identical, overlapping clones from the mitochondrial genome. Therefore, a separation of 15 bp was chosen to distinguish between identical

clones (less than 15 bp) and non-identical overlapping clones (greater than or equal to 15 bp) (Figure 3-3). 15% (1,234 / 8,328) clones were classed as identical and removed from subsequent analyses. The remaining 7,094 unambiguously aligned non-identical cDNA clones (3,176,242 bp) were used in the subsequent analyses (Figure 3-2).

The amount of overlap between non-identical 'overlapping' clones was calculated using the custom Perl program *identify_overlapping_clones.pl* (see Methods). Alignments were sorted by chromosome and then by their start position along that chromosome. Moving along a chromosome one alignment at a time, each alignment was compared to all overlapping alignments preceding it on that chromosome. For each clone, the number of bases that were also present in a preceding clone was counted as overlapping. 11% (780 / 7,094) non-identical clones contained a total of 221,429 bp of overlapping sequence.

3.2.5 Evaluation of cDNA library composition by the genomic distribution of cDNA clones

The cDNA clones were next classified according to their origin in the human genome. 95.4% (6,768 / 7,094) cDNA clones, comprising 3,058,468bp non-overlapping cDNA sequence, were derived from the nuclear chromosomes (Table 3-3). A further 0.3% (24 / 7,094) clones, comprising 7,026bp, were derived from the ribosomal DNA repeat sequence. Given that the ribosomal RNA is typically the major component of total RNA, with mRNAs making up only 2-3%, this indicated efficient purification of poly-adenylated RNAs away

from ribosomal RNA in the preparation of this library. The remaining 4.3% (302 / 7,094) clones (110,748bp) were from the mitochondrial genome.

Chromosome	Chromosome length (bp)	Clones	Total bases	Overlapping bases
1	246,127,941	598	268,935	9,731
2	243,615,958	510	235,704	5,168
3	199,344,050	448	206,591	6,533
4	191,731,959	283	131,972	5,069
5	181,034,922	334	156,483	3,191
6	170,914,576	298	137,779	2,533
7	158,545,518	370	171,966	4,895
8	146,308,819	306	140,290	4,096
9	136,372,045	274	123,700	6,384
10	135,037,215	279	130,692	2,330
11	134,482,954	406	182,860	19,622
12	132,078,379	389	173,406	7,417
13	113,042,980	137	61,574	337
14	105,311,216	212	97,086	4,300
15	100,256,656	247	116,491	1,953
16	90,041,932	258	111,998	2,567
17	81,860,266	310	134,638	3,305
18	76,115,139	133	59,227	8,018
19	63,811,651	348	139,076	6,095
20	63,741,868	179	76,775	7,069
21	46,976,097	82	37,323	3,534
22	49,396,972	134	57,226	1,922
X	153,692,391	219	99,863	6,511
Y	50,286,555	14	6,813	660
All chromosomes	3,070,128,059	6,768	3,058,468	123,240
Mitochondrial	16,571	302	110,748	95,254
rRNA	42,999	24	7,026	2,925
Total	**3,070,187,629**	**7,094**	**3,176,242**	**221,419**

Table 3-3 Genome-wide distribution of cDNA clones.

3.2.5.1 *Distribution of cDNA clones aligning to the nuclear chromosomes*

The proportion of cDNA clones from each chromosome was calculated. Overall, the proportion of cDNA sequence derived from each chromosome was similar to the proportion of the genome sequence on that chromosome

(Figure 3-4A). However, the proportion of the cDNA library derived from chromosomes 4 and 13 was only 70% of the proportion of the genome contained on these chromosomes. This ratio fell to 60% for chromosomes 13 and X and 10% for chromosome Y. Conversely, the proportion of cDNA clone sequences from chromosome 19 is more than twice the proportion of the genome on this chromosome. The proportion of the cDNA library derived from each chromosome was next compared with the proportion of all EnsEMBL known genes on that chromosome (Figure 3-4B). The chromosomal distribution of cDNA clone sequences showed a closer correlation with transcribed sequence than total genome sequence. For example, the relatively small amounts of cDNA library sequence from chromosomes 13, X and Y (Figure 3-4A) can be explained by a relatively low proportion of known genes on these chromosomes (Figure 3-4B).

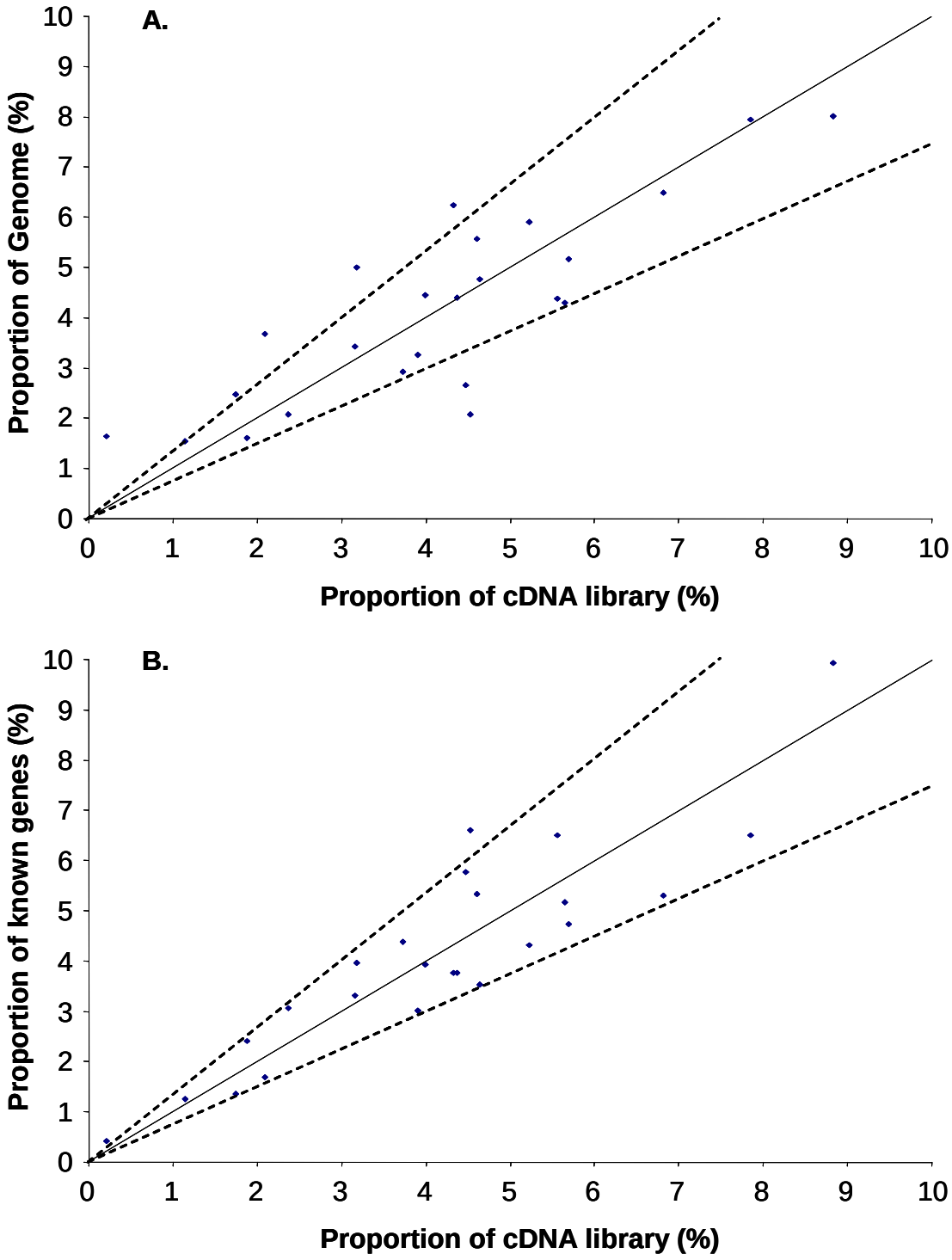


Figure 3-4 Comparison of the proportion of cDNA clones derived from each chromosome with **A**. the proportion of the human genome on each chromosome, and **B**. the proportion of all known genes on each chromosome. **A**. The solid black line indicates an equal proportion from the cDNA library

and the genome. The dashed lines indicate the boundary of values for which the proportion of cDNA library and the proportion of the genome are within 75%. Similarly for **B**.

3.2.5.2 *Distribution of cDNA clones aligning to the mitochondrial genome*

The human mitochondrial genome is a circular DNA from which both strands are transcribed as single molecules. These primary transcripts are then processed into mature transcripts by nuclease cleavage. In total, the mitochondrial genome contains two ribosomal RNAs, 22 tRNA genes and 13 protein coding genes which are poly-adenylated. The genome is extremely gene rich so there is very little intergenic sequence and the genes do not contain introns. Mitochondrial genome replication and transcription is regulated by an intergenic sequence called the D-Loop.

In total, 302 / 7,094 non-identical clones were found to align to the mitochondrial genome. These comprised 110,748 bases of sequence, which overlapped considerably. The total amount of unique sequence was 15,494 bases. As the mitochondrial genome is 16,571bp this corresponds to 93.5% coverage of the mitochondrial genome. To visualise the alignments of clones to the mitochondrion in more detail, they were displayed as a custom track in the UCSC human genome browser (Figure 3-5). Consistent with this, clones span most of the mitochondrial genome and are unspliced. However, the majority of clones cluster into groups corresponding with the known mitochondrial genes and very few clones (<5%) overlap more than one gene. This strongly suggests that most clones are derived from mature transcripts

rather than from precursor transcript, and that the cDNA library is not heavily contaminated with mitochondrial DNA. Only a small number of clones align to the D-Loop of the mitochondrion, consistent with this being intergenic sequence.

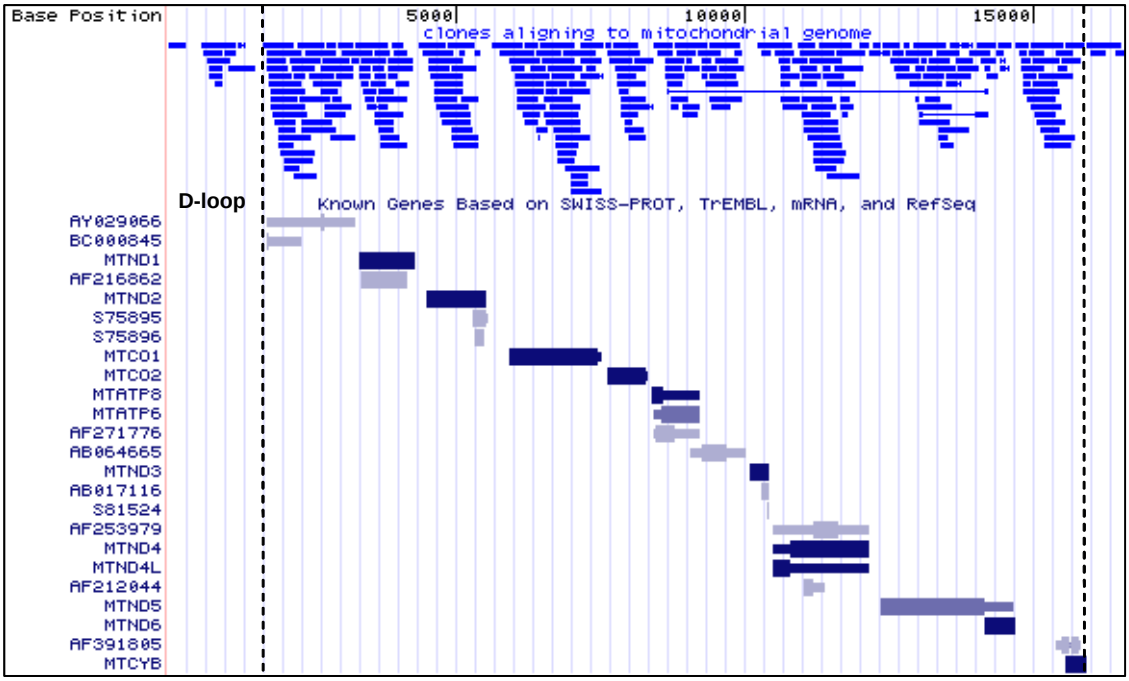


Figure 3-5 Mitochondrial cDNA clones. Mitochondrial DNA is a circular molecule. The extreme left and right of the display represent the same point on this molecule. The upper blue bars represent clone sequences. The lower bars represent the known mitochondrial genes for comparison. Dashed lines indicate the boundaries of the regulatory D-loop region.

3.2.6 Evaluation of cDNA library by annotation of known genes

Detailed annotation of cDNA clones aligning to the nuclear chromosomes was required to provide context for any novel RNA edits. For each clone, the EnsEMBL database was searched for evidence of unambiguous alignment to

a known or predicted gene. For clones derived from known genes, the alignment was related to the positions of boundaries between introns and exons and between coding and non-coding regions. To annotate all 7,094 clones an automated method was developed based on the EnsEMBL database, and the associated EnsEMBL API.

3.2.6.1 *Evaluation of the gene content of the cDNA library*

The custom Perl program *annotate_exons.pl* (see Methods) was used to identify cDNA clones which aligned unambiguously with the exon structure of overlapping genes from the EnsEMBL database. The program first searched for overlap with EnsEMBL known genes (constructed from alignments of cDNAs or proteins to the genome) or EnsEMBL novel genes (constructed from alignments of spliced ESTs to the genome). If no overlap was found, then the program searched for overlap with EnsEMBL gene predictions (constructed using gene prediction programs such as Genescan). Clones were then classified according to Table 3-4. In total, 87% (5,892 / 6,768) of cDNA clones overlapped with an EnsEMBL gene. These included 70% (4,760 / 6,768) known genes, 13% (910 / 6,768) predicted genes and 3% (222 / 6,768) novel genes (Table 3-4). Only 2% (141 / 6,768) cDNA clones could not be unambiguously annotated because they matched multiple genes. The remaining 11% (735 / 6,768) clones did not overlap any annotation in the EnsEMBL database and were classed as intergenic.

Classification	Description	Clones
Intergenic	The cDNA clone does not overlap with any gene.	735
Indeterminable	The cDNA clone alignment is unspliced, and overlaps more than one gene.	82
Matches multiple genes	The cDNA clone alignment is spliced but matches exons from different genes.	59
Known gene	The cDNA clone alignment is spliced and matches the exon structure of a single known gene.	4760
Novel gene	The cDNA clone alignment is spliced and matches the exon structure of a single novel gene.	222
Predicted gene	The cDNA clone alignment is spliced and matches the exon structure of a single predicted gene.	910

Table 3-4 Classification of cDNA clones according to overlap with gene annotation in the EnsEMBL genome database.

For each cDNA clone aligning unambiguously to an EnsEMBL gene, the gene number and gene description was retrieved. By counting the number of times each gene was sequenced, a list of the most highly represented transcripts in the cDNA library was constructed (Table 3-5). As expected from a brain cDNA library, most of the frequently sequenced genes had ‘housekeeping’ functions

such as Actin (9 clones) and Glyceraldehyde 3-phosphate dehydrogenase (8 clones), and neuronal functions such as Myelin basic protein (24 clones) and Synaptosomal protein (12 clones). However, by far the most frequently detected gene in the library was ENSG00000185316 (32 clones). BLASTN alignment of the minimum genomic DNA sequence containing all 32 cDNA clone sequences against the NCBI non-redundant database identified a gene with no protein product, metastasis associated lung adenocarcinoma transcript 1 (MALAT-1)(Ji et al., 2003). This transcript was reported to be significantly associated with metastasis in NSCLC patients (Ji et al., 2003), but there is currently no information about its function in normal cells.

To evaluate further the non-coding RNA content of the cDNA library, the genomic coordinates of all cDNA clones were compared with the genomic coordinates of a list of known RNA genes. Nine cDNA clones overlapped with a non-coding RNA gene. Three were small nucleolar RNAs (snoRNAs), three were small nuclear RNAs (snRNAs), one micro RNA (miRNA), one 28S ribosomal RNA related transcript and one mitochondrial derived pseudogene. In all cases the cDNA clone extended beyond the genomic coordinates of the fully processed non-coding RNA, suggesting that the cDNA clone was derived from an unprocessed transcript.

Gene ID	Length (bp)	Description	Clones
ENSG00000185316	167	MALAT-1	32
ENSG00000151507	38224	Myelin basic protein	24
ENSG00000080824	58630	Heat shock protein HSP 90-alpha	12
ENSG00000132639	88588	Synaptosomal-associated protein 25	12
ENSG00000142192	290270	Amyloid beta A4 protein precursor	12
ENSG00000187391	1436238	Atrophin-1 interacting protein 1	11
ENSG00000075624	3445	Actin, cytoplasmic 1	9
ENSG00000087460	71450	Guanine nucleotide-binding protein G(S), alpha subunit	9
ENSG00000179915	1107923	Neurexin 1-alpha precursor	9
ENSG00000018625	27905	Sodium/potassium-transporting ATPase alpha-2 chain precursor	8
ENSG00000087258	166052	Guanine nucleotide-binding protein G(O), alpha subunit 1	8
ENSG00000111640	3852	Glyceraldehyde 3-phosphate dehydrogenase	8
ENSG00000123560	15791	Myelin proteolipid protein	8
ENSG00000081853	174192	Protocadherin gamma C5 precursor	7
ENSG00000092964	80195	Dihydropyrimidinase related protein-2	7
ENSG00000109472	119386	Carboxypeptidase H precursor	7
ENSG00000123416	3610	Tubulin alpha-1 chain	7
ENSG00000127603	405671	Microtubule-actin crosslinking factor 1, isoform 4	7
ENSG00000131711	102036	Microtubule-associated protein 1B	7
ENSG00000139720	170836	Nuclear receptor co-repressor 2	7
ENSG00000142599	465067	arginine-glutamic acid dipeptide (RE) repeats	7

Table 3-5 The 20 most commonly sequenced genes in the cDNA library.

3.2.6.2 *Evaluation of cDNA library composition by sequence class*

For each cDNA clone aligning unambiguously to a known gene, the amount of translated, 5' untranslated and 3' untranslated exonic sequence was calculated using the custom Perl program *genomic_coding_script.pl* (see Methods). Although the cDNA library is enriched in gene sequences (Table 3-4), only 33% of sequences were derived from exons (Figure 3-6). The majority of the cDNA library was intronic (54%), and intergenic (13%).

The presence of intronic and intergenic sequences raised the possibility that the cDNA library was contaminated with genomic DNA. However, comparison with the composition of genomic DNA (78% intergenic, 20% intronic and 2% exonic) indicates that the cDNA library was highly enriched in intronic and exonic sequence.

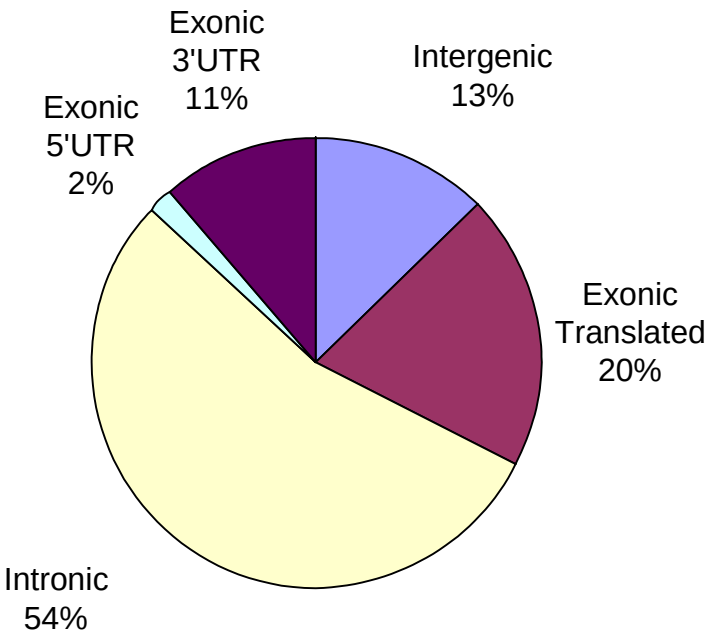


Figure 3-6 Sequence class composition of the cDNA library.

Instead, these results suggest that the cDNA library was composed of a mixture of fully processed, partially processed and unprocessed transcripts along with an unknown amount of contaminating genomic DNA. If 17% of sequences in the cDNA library were derived from genomic DNA (which is approximately 78% intergenic), this would explain the observed 13% intergenic sequence content of the cDNA library. However, this is likely to be an over-estimate of the genomic DNA content of the cDNA library. A small number of intergenic sequences (2%) were spliced when aligned to the genome and therefore represent processed transcripts. As the cDNA library clearly contains unprocessed transcripts from annotated genes, it is likely that a proportion of the un-annotated intergenic sequences are also from unprocessed transcripts. The subsequent identification of RNA editing of these sequences (see Results, Chapter 4) provides further evidence that at least a proportion (and possibly all) of the intergenic sequences were transcribed. Contamination of the library with genomic DNA is therefore likely to be much less than 17%.

3.3 DISCUSSION

3.3.1 Choice of experimental strategy for a survey of RNA editing

In order to investigate the genome wide patterns of RNA editing, randomly selected cDNA clones from a randomly primed human brain cDNA library were sequenced and aligned to the human genome reference sequence.

These alignments subsequently formed the basis of a search for novel sequence variants, and ultimately novel RNA edits. This approach was chosen because it offered an unselective insight into the targets and patterns of RNA editing in human cells.

An alternative would have been to use a targeted RT-PCR based sequencing approach to analyse sequences with similarity to known RNA editing substrates. Candidate RNA edits could be identified from homologues of RNA editing substrates in humans, and orthologues of RNA editing substrates from other organisms. These could be extended to include whole gene families, or genes with related function for which a common mechanism of regulation by RNA editing seems reasonable. Whilst this type of approach might be expected to yield more edits, and perhaps novel coding edits, it would be biased towards variants with the characteristics of known RNA edits.

Another source of candidate RNA edits, from multiple tissue types, would be from alignments of EST sequences to the human genome reference sequence. Indeed this approach was successfully employed in a recent systematic search for A>I edits in human tissues (Levanon et al., 2004). Sequence variants identified from EST alignments would include sequence trace errors, unforeseen artefacts relating to cDNA library construction, SNPs, and RNA edits. Although frequent editing events and dominant patterns of RNA editing would be readily detectable, infrequent RNA editing events would be extremely difficult to separate from other sources of sequence variation. When this thesis was started, very few EST sequence traces were available

from sequence trace repositories, and therefore there was no information about the quality of sequence traces. Furthermore, there is no matching genomic DNA sequence with which to compare EST sequences and identify SNPs. In contrast, the cDNA clone sequencing approach used in this thesis allowed sequencing artefacts to be identified from sequence traces, and SNPs to be identified by reference to matching genomic DNA. This allowed an untargeted evaluation of infrequent as well as frequent editing events.

3.3.2 Choice of tissue for a survey of RNA editing

Previous data indicated that levels of RNA editing may be highest in the brain. Therefore the cDNA library used for this survey was constructed from RNA derived from human cerebral cortex. As this is heterogenous tissue, the library is not representative of a single cell type, but of the constituent cell types including nerve cells, astrocytes, oligodendrocytes, endothelial cells and microglia. Consequently, transcripts that are edited in only one cell type would be diluted by unedited transcripts from other cell types and the chances of detecting rare transcripts would be reduced. An alternative would have been to analyse RNA editing in a tissue type that is more homogeneous, for example muscle, which consists predominantly of only one cell type. However, there is less evidence for RNA editing in these tissues than in brain. Alternatively, we could have examined RNA editing in a cell line in which the cells are clonal and therefore represent a single cell type. However, cultured cells are known to undergo extensive genetic and transcriptional alteration, so the transcriptomes of cultured cells may differ widely from the *in vivo* transcriptomes of the cells from which they were derived.

3.3.3 Extent to which the cDNA library is representative of the human brain transcriptome

For an unselective survey of RNA editing, it was important that the cDNA library was representative of the transcriptome of normal human brain cells. Therefore, several measures were taken to minimise the impact of experimental artefacts. The tissue sample from which the RNA was extracted was obtained with minimum delay following death and was judged to be 'normal' in appearance. Synthesis of cDNA was performed using random primers. This prevented a bias towards the 3' end of transcripts that would have resulted from the use of oligo-dT as a primer. To prevent distortion of the cDNA library composition from altered growth rates of bacterial clones, the cDNA library was subject to only one round of amplification performed in semi-solid media which allows for uniform colony growth.

Total RNA was poly (A)+ RNA purified prior to cDNA synthesis. This was necessary to remove ribosomal RNA from the cDNA library, but would also result in the exclusion of other transcripts that are not poly-adenylated. This includes the majority of RNA Pol I and Pol II transcripts including rRNA, tRNA, snRNA, snoRNAs and miRNAs, which are potential RNA editing substrates. Several of these classes of RNA undergo modification (eg pseudouridylation and o-methylation), and both tRNAs and miRNAs are known targets of A > I RNA editing (Maas et al., 1999, Luciano et al., 2004). Despite selection for poly-adenylated transcripts, non poly-adenylated transcripts are represented in the cDNA library, albeit at reduced levels compared to the original brain

tissue. 24 cDNA clones were derived from the rRNA repeat and a further nine sequences overlap non-coding RNAs including three snRNAs, three snoRNAs and one miRNA.

3.3.4 Sequence class composition of the cDNA library

The initial evaluation of the cDNA library indicated that only 19% of cDNA clones were exonic (spliced). This raised the possibility that the library was heavily contaminated with genomic DNA. However, this concern was influenced by the preconception that a high quality cDNA library should be composed almost completely of sequences derived from fully processed mRNAs (i.e. nearly 100% exonic (spliced)). In fact, the composition of the cDNA library is consistent with derivation almost entirely from poly-adenylated transcripts which have undergone varying degrees of splicing. Several lines of evidence support this hypothesis. 1) The distribution of cDNA clone sequences by chromosome correlates with gene density rather than the DNA content of the chromosome (Figure 3-4). 2) cDNA clones derived from the mitochondrial genome align with the boundaries of genes implying that they originate from processed mitochondrial transcripts rather than contaminating mitochondrial DNA (Figure 3-5). 3) The cDNA library is enriched in intron and exon sequences compared with the estimated composition of total genomic DNA. 4) A small proportion (1.5%) of intergenic cDNA sequence is processed and therefore must be transcribed. 5) Subsequent experiments showed that intergenic sequences are subject to RNA editing, and therefore that they must be transcribed (see Results, Chapter 4).

The cDNA library contained a large number of mitochondrial cDNA sequences. Although the mitochondrial genome is much smaller than the nuclear genome, there are on average 1,000 mitochondria per cell, and each mitochondrion may contain several molecules of DNA. As a result, mitochondrial DNA contributes significantly to total cellular DNA. Furthermore, whereas only a fraction of the nuclear genome is transcribed the entire mitochondrial genome is transcribed, so the contribution of mitochondrial RNA to total RNA is even higher.

Overall, the most abundant transcripts from the nuclear genome were derived from house-keeping genes and genes involved in neuronal function. However, the most highly represented transcript in the library is from a region on chromosome 11 corresponding with a putative non-coding RNA (MALAT-1). Many other cDNA clones were from unannotated regions and are therefore putative novel transcripts. This is consistent with recent observations that the transcriptional output of the genome is far higher than can be accounted for by known protein coding genes (Kapranov et al., 2002).

In conclusion, these results indicate that the cDNA library is derived from human cerebral cortex RNA and contains a low level of contamination of genomic DNA. Over 3Mb unique cDNA sequence was aligned unambiguously to nuclear genome, sufficient for an extensive search for sequence variants and novel RNA edits.

4 IDENTIFICATION OF NOVEL RNA EDITS IN HUMAN BRAIN

4.1 INTRODUCTION

All previously reported RNA edits in humans are small changes to the nucleotide sequence by base substitution or nucleotide insertions or deletions. In principle, these are detectable as differences between the alignments of cDNA clone sequences and the human genome reference sequence. In the previous chapter, general features of the sequences obtained from a human brain cDNA library were evaluated. The results indicated that most sequences were derived from transcripts and were suitable for the detection of RNA edits. In this chapter, identification, confirmation and initial characterisation of RNA edits present in these cDNA sequences is described.

4.2 RESULTS

4.2.1 Computational detection of high quality candidate RNA edits from human brain cDNA

Most of the available software for analysing sequence variants deal with one sequence at a time, and require manual inspection of sequence traces. They are primarily designed for comparing two DNA sequences, and incorporate sophisticated methods to distinguish heterozygous sequence variants from sequence trace errors. In contrast, detection of sequence variants between cDNA clones and genomic DNA is relatively simple. Because both the cDNA

clone sequence and the human genome reference sequence represent a single allele, all sequence variants will be homozygous. Therefore, assuming that the sequence traces being compared are of high quality, variants can be detected by comparing the letters of the two aligned sequences. In this study, variants were detected computationally from the sequence alignments generated by BLAT.

The custom Perl program used previously to identify the best alignment of each cDNA clone to the genome reference sequence was modified to compare the two sequences, and record variants (see Methods). For each variant detected, the variant type and location in the genome was reported. In total, 8,580 variants were identified from 6,768 cDNA clones (Figure 4-1).

To rule out sequence variants that were due to sequence trace errors, sequence trace quality was evaluated using 'q-scores'. These are automatically generated by the *Phred* base calling algorithm (see Methods). To assess the quality of each sequence variant identified from the cDNA sequence alignments, q-scores corresponding to each variant base and the five flanking nucleotides on either side in the cDNA clone sequence were identified. Initially, the cut-offs used to identify high quality variants were taken from a method to identify SNPs from overlapping genome sequence reads, in which a variant was deemed to be 'high quality' if it had a quality score of 20 or over, and the five bases either side had quality scores of 15 or over (Altshuler et al., 2000, Mullikin et al., 2000). Using this threshold, 64% (5,519 / 8,580) variants were classified as high quality (Figure 4-1).

To rule out known SNPs, the dbSNP database (dbSNP, http://www.ncbi.nlm.nih.gov/SNP/) was queried with the genomic coordinates of each variant. In total 21% (1,148 / 5,519) high quality variants were known SNPs. This is equivalent to one difference every 3,300bp for the whole cDNA library. This is approaching half of the expected number of SNPs, based on the estimate that differences in nucleotide sequence occur every 1,331bp when two chromosomes of similar ethnicity are compared (Sachidanandam et al., 2001). The remaining 4,371 high quality variants were candidate RNA edits (Figure 4-1).

4.2.2 Extensive A > I RNA edits but no other class of RNA edits are present in human brain cDNA

The 4,371 candidate RNA edits were next subject to experimental evaluation. To discriminate RNA edits from other causes of sequence variation (including novel SNPs and sequence artefacts) genomic DNA from the individual from whom the cDNA library was constructed was analysed and compared to cDNA clone sequences. Since there were a large number of potential edits which would have required extensive PCR based genomic DNA and cDNA sequencing for complete assessment, we implemented a parsimonious, two stage evaluation of these variants. First, the cDNA library was searched for putative multiply edited transcripts from sequences containing more than three sequence variants. Second, a subset of the candidate RNA edits from sequences containing only one or two variants (subsequently referred to as singleton variants) were evaluated (Figure 4-1).

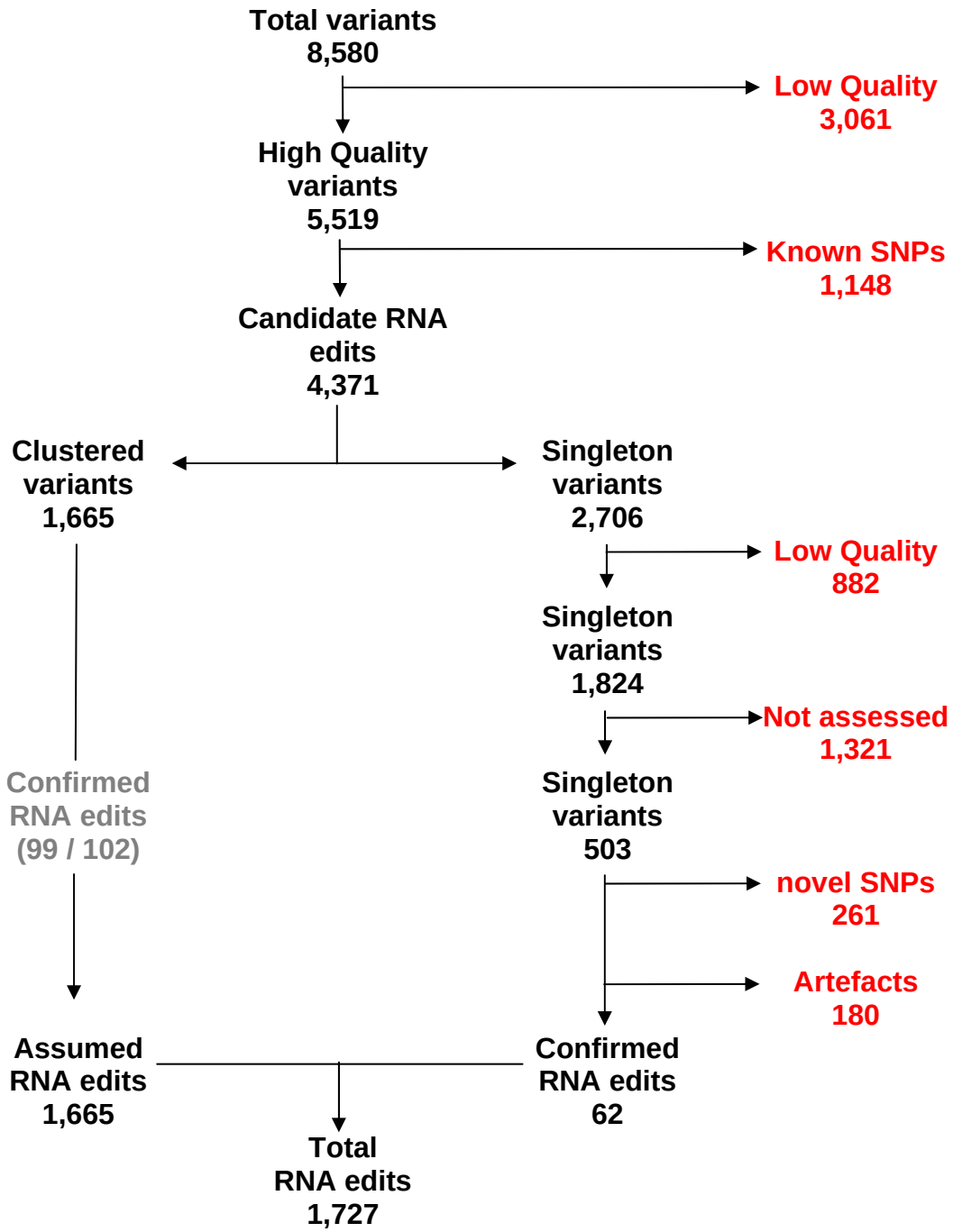


Figure 4-1 Summary of the identification of 1,727 novel A>I RNA edits. Variants are separated into those from sequences with 3 or more variants (clustered), and those from sequences with less than 3 variants (singletons). Variants in red were rejected according to various criteria. Variants in grey

indicate a partial analysis of clustered variants. Variants in black show the remaining candidate RNA edits at each stage of the analysis.

4.2.2.1 *RNA editing of nuclear transcripts containing multiple variants*

To search for potentially heavily edited sequences, the cDNA library was searched for sequences with three or more variants of the same type, where the total number of variants of this type constituted over 75% of the total number of variants in the sequence. These criteria were designed to detect transcripts that contained multiple edits of the same type. In total, 256 sequences (comprising 1,665 variant bases) were identified which contained three or more high quality variants. In all cases the variants were A > G or T > C. The most variants seen in a single cDNA clone sequence was 28 A > G changes.

A random sample of 12 out of 256 cDNA clone sequences containing three or more A > G or T > C changes were experimentally verified. Sequence analysis of genomic DNA from the individual from whom the library was constructed demonstrated that none of the A > G / T > C variants observed in these 12 sequences were SNPs. In order to confirm the variants as RNA edits, RT-PCR sequences from total brain RNA (subsequently referred to as total cDNA sequences) were analysed. In these experiments, RNA editing was confirmed by the presence of the edited nucleotide in the total cDNA sequence but not in the matching genomic DNA sequence and by a decrease in the genomically encoded nucleotide in the total cDNA sequence compared to the matching genomic DNA sequence. The decrease in the genomically

encoded nucleotide was measured relative to an unedited nucleotide of the same type in the adjacent sequence trace, in order to rule out variability between the two sequence traces being compared. For each variant, sequencing was performed in duplicate and in sense and anti-sense orientation.

In total, 97% (99 / 102) of variants (from 11 out of 12 sequences) were confirmed as RNA edits by sequencing of total cDNA (Figure 4-2). In some cases RNA editing was very subtle (for example Figure 4-2D, CAP350 intronic Alu sequence). A possible explanation for this is that only a small proportion of the transcripts were edited in the total RNA sample, and that the sequenced cDNA clone was derived from the minority edited population. This may also explain the one sequence for which RNA editing could not be reproduced. Since almost all variants (99 out of 102 from the 12 sequences) in this class of sequence appeared to be RNA edits, all 1,665 A > G or T > C variants from all 256 sequences were classified as RNA edits and included in the subsequent analyses without further confirmation (Figure 4-1). However, it should be noted that a small proportion of these 1,665 presumed A > I edits (an estimated 3%) may not be correct.

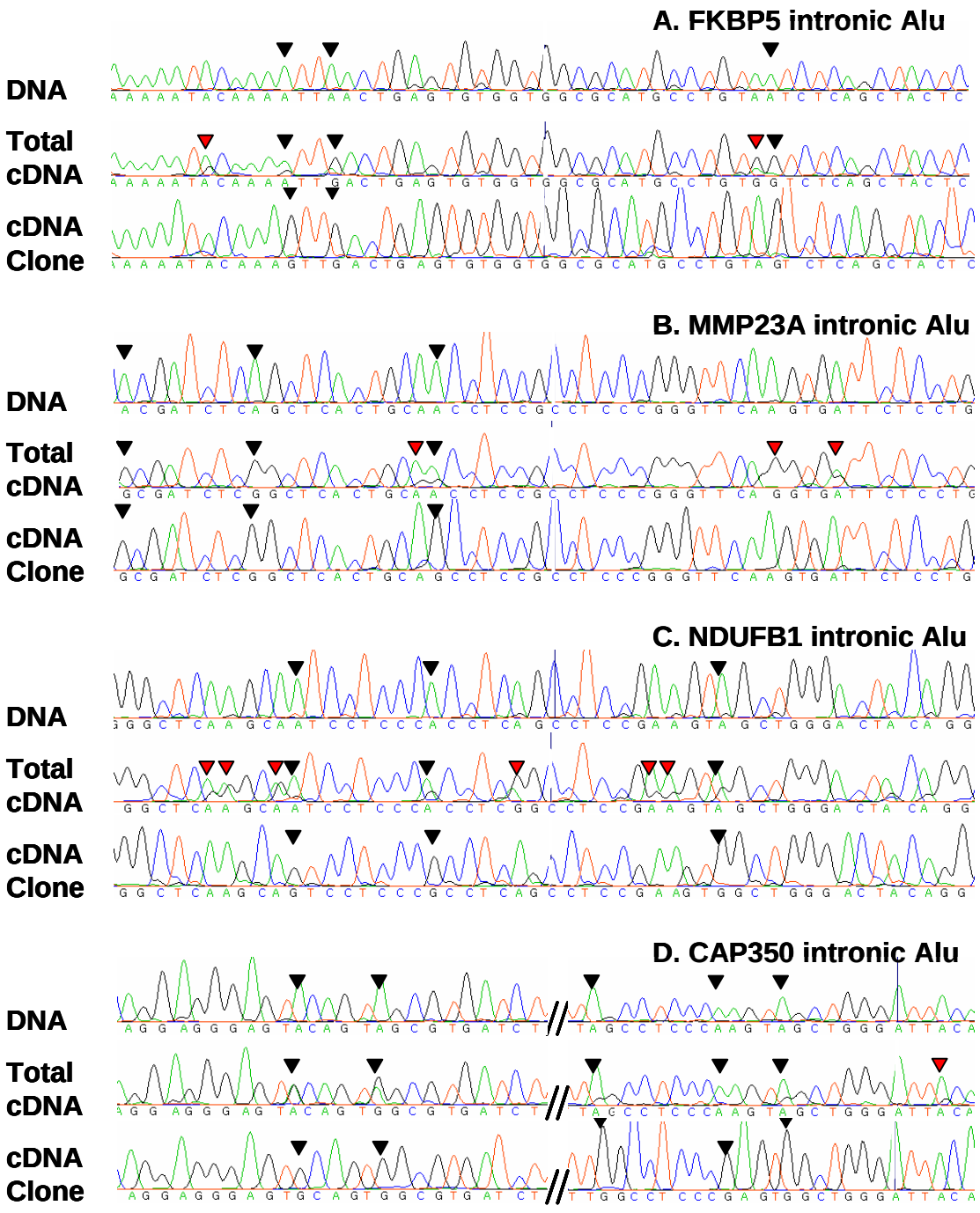


Figure 4-2 Confirmation of RNA editing of heavily edited sequences. cDNA clones containing three or more variants of the same type were evaluated by sequencing of PCR products from genomic DNA, and Total cDNA. Sequence traces are shown for four of the 11 / 12 sequences that were confirmed to be edited. Black arrows indicate the position of RNA edits identified in the original

cDNA clone sequences. Red arrows indicate the position of additional sites of RNA editing identified from total cDNA.

4.2.2.2 *RNA editing of nuclear transcripts with low frequency sequence variants*

Next, an evaluation of singleton variants was performed. In order to identify the highest quality candidate RNA edits for sequencing, more stringent quality scoring criteria were established. 120 'high quality' sequence variants were selected at random from the 2,706 singleton variants, and their sequence traces were examined manually. Although 93 variants were found to be genuine variants, 27 variants appeared to be sequence trace artefacts. These artefacts include mis-called substitution variants following mononucleotide or dinucleotide repeats (Figure 4-3A), mis-called substitutions variants due to low intensity G peaks (possibly resulting from a problem with sequencing reagents) (Figure 4-3B), and mis-called insertion variants where sequence traces were incorrectly processed so that peaks were 'split' (Figure 4-3C).

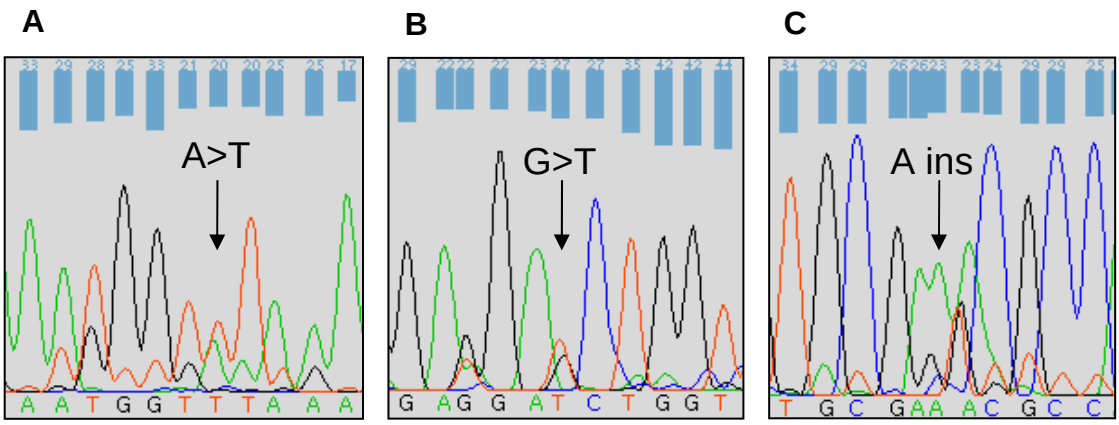


Figure 4-3 Examples of variants identified incorrectly by automated detection. The traces are base called using *Phred*, quality scores are shown above each

peak. **A**, a false positive A>T variant. **B**, a false positive G>T variant in a sequence with low intensity G-peaks. **C**, a false positive A-insertion variant caused by splitting of an A peak into two A peaks.

The quality scores of variant nucleotides and the five bases either side were determined for all 93 correctly called variants and compared to all 27 mis-called variants. By trial and error, new quality score criteria were established which led to the rejection of as many of the sequence trace artefacts but as few of the genuine variants as possible. Under these criteria, variants with a quality score of 30 or more with two preceding bases of quality score 30 or more and a following base of score 20 were classed as high quality. Re-analysis of the test set of 120 randomly selected sequence variants resulted in rejection of 85% (23 / 27) of the sequence artefacts, whereas only 19% (18 / 93) of the correctly identified variants were rejected. Applying these criteria to the full set of 2,706 singleton variants resulted in 882 variants being classified as low quality, leaving 1,824 high quality singleton variants (Figure 4-1).

Next, a subset of the 1,824 singleton variants was evaluated. 503 variants (from 374 different PCR fragments) were successfully amplified from genomic DNA and, if the variants were shown not to be SNPs, were evaluated by sequencing of total brain cDNA. Of 185 A > G / T > C variants in these experiments, 62 variants (from 41 sequences) were confirmed as RNA edits (Figure 4-1). Of 285 other base substitution variants and 33 insertion / deletion variants all were either SNPs or artefacts.

The 1,665 edits from the first stage of evaluation were combined with the 62 confirmed edits from the second stage of evaluation (Figure 4-1). In total 1,727 edits of which 9% (161 / 1,727) were directly confirmed by sequencing of total cDNA were included in the analyses described below. Because only 503 of the 1,824 potential edits that were present in sequences with fewer than three variants were evaluated, A > I edits which occur in such sequences are underrepresented in the final 1,727. However, evaluation of the remaining 1,321 / 1,824 potential edits by sequencing would have increased the total number of A>I edits by less than 10%. Moreover, subsequent analyses indicate that A > I edits from sequences with fewer than three variants show similar patterns to A > I edits from multiply edited sequences, and therefore are likely be the product of the same editing activity responsible for multiply edited sequences.

4.2.3 A > G / T > C variants are all likely A > I edits

During synthesis of the cDNA library, cDNA sequences were randomly cloned. Sense and anti-sense variants have therefore been combined in analyses to this point. All A > G / T > C edits are assumed to be A > I (A > G) rather than T > C. To test this assumption, all novel RNA edits from cDNA clones aligning to known genes were reoriented according to the transcribed strand. Of 180 edited cDNA clones from known genes, 96% (173 / 180) were confirmed to be A > G edited when oriented to the known gene. All seven remaining sequences aligned to regions of the genome which for which there is EST evidence of transcription of both strands. Therefore it seems likely that all novel edits are truly A > I.

4.2.4 RNA editing is absent from mitochondrial transcripts in human brain

1,055 cDNA clone sequences aligned to the mitochondrial genome reference sequence. A total of 230 high quality sequence variants were identified from 60 unique variant positions in the mitochondrial genome. At each unique variant position, the number of overlapping clones with the variant allele, and the number of overlapping clones with the reference allele was used to calculate the frequency of the variant allele within the cDNA library.

14 / 60 variants were present in more than one cDNA clone, and were selected as candidate novel RNA edits. Twelve of these variants were successfully evaluated by PCR and sequencing the genomic DNA of the individual from which the cDNA library was made (Table 4-1). 10 / 12 variants were detectable in DNA and therefore were polymorphisms in the mitochondrial DNA sequence. The remaining two variants were not detectable in genomic DNA, but were not confirmed in cDNA. In both cases, the number of clones containing the variant allele was vastly outnumbered by those containing the reference allele (two clones containing the variant allele compared to 115 with the reference allele and 2 clones containing the variant allele compared to 125 with the reference allele). Therefore if these variants did arise through RNA editing, the frequency of editing would be extremely low (less than 1 in 57 transcripts).

A further 4 / 60 variants were identified from a single cDNA clone, with one or zero clones containing the reference allele (Table 4-1). All were detectable from genomic DNA and were therefore polymorphisms. The remaining 42 / 60 variants were from a single cDNA clone, with more than 1 (and up to 129) cDNA clones containing the reference allele. These variants were likely to be a cloning or sequencing artefact and were not evaluated further. Overall, in these analyses no examples of RNA editing of mitochondrial transcripts were identified.

	Gene	Variant	Variant clones	Total clones	Frequency
1	**16s rRNA**	A > G	33	54	61%
2	**ND2**	A > G	30	37	81%
3	**CYTB**	A > G	30	37	81%
4	**16s rRNA**	C del	24	40	60%
5	**ATP8**	A > G	13	20	65%
6	**ND5**	T > C	12	18	67%
7	**ATP8**	G > A	12	19	63%
8	**ND5**	G > A	11	17	65%
9	**ND1**	C > T	8	8	100%
10	**12s rRNA**	T > C	3	4	75%
11	**16s rRNA**	A del	2	117	2%
12	**16s rRNA**	G > A	2	127	2%
13	**D-Loop**	A > G	1	1	100%
14	**D-Loop**	G > A	1	1	100%
15	**D-Loop**	T > C	1	2	50%
16	**D-Loop**	T > C	1	2	50%

Table 4-1 List of evaluated sequence variants in mitochondrial cDNA clone sequences. 14 / 60 variants from mitochondrial cDNA sequences were evaluated by sequencing from genomic DNA including 12 variants identified in more than one cDNA clone, and 4 variants identified from transcripts with only one variant clone.

4.2.5 The estimated frequency of RNA editing in the human brain

These results strongly indicate that A > I RNA editing is the predominant form of RNA editing in human brain. A > I editing occurs at approximately 1 in 1,700 nucleotides (600 bases / Mb) in the cDNA library. In contrast, none of the 318 / 1183 RNA edits of other categories that were evaluated were found to be RNA edits. To estimate frequency of non A > I RNA editing in human brain, the probability of obtaining these results if a proportion of the 1183 variants was actually an RNA edit was calculated (Table 4-2). These data suggest that it is highly unlikely that more than 20 of the 1,183 variants identified from the 3.06Mb sequence are actually RNA edits (Table 4-2). It is therefore unlikely that there are more than 7 non A > I edits / Mb, in contrast to approximately 600 A > I edits / Mb, in RNA from human brain.

p(edit)	**1 / 1183**	**5 / 1183**	**10 / 1183**	**15 / 1183**	**20 / 1183**
p(0 / 318)	76%	26%	6.7%	1.7%	0.4%

Table 4-2 Estimation of the frequency of non A > I RNA editing in the human brain. In total 318 / 1183 non A > I variants from 3.06Mb cDNA were evaluated by RT-PCR and sequencing and shown not to be RNA edits. The probability of sampling 318 and finding no RNA edits (p (0 / 318)) was evaluated for several hypothetical frequencies of RNA editing (p(edit)) using the calculation $p(0 / 318) = (1-p(edit))^{318}$. These values indicate a low probability (less than 1%) of there being twenty RNA edits in 3.06Mb.

4.3 DISCUSSION

4.3.1 Classes of RNA editing in the human brain

In this survey, there is strong evidence for widespread A > I editing, but no evidence for other classes of RNA editing in human brain RNA. This result does not rule out the possibility that other types of RNA edit occur in the brain at a very low frequency, or in a restricted sub-set of cells (and therefore were not sampled in this survey). Neither does it exclude the possibility that abundant RNA editing by other mechanisms exists in other tissues. The results are, however, consistent with the known expression patterns of RNA editing enzymes. The A > I editing enzymes ADAR1 and ADAR2 are expressed widely in the brain, and are likely to be the enzymes responsible for the A > I edits identified in this survey. In contrast, expression of the only confirmed RNA cytidine deaminase, APOBEC-1 (which catalyses C > U RNA editing), is confined to the small intestine of humans, whilst the related candidate cytidine deaminase APOBEC-2 is expressed only in heart and skeletal muscle (Liao et al., 1999). There are currently no known enzymes capable of catalysing other classes of nucleotide substitutions in human brain RNA.

Despite extensive sequence analysis, no RNA editing of mitochondrial cDNAs was observed. This could be due to absence of dsRNA formation in mitochondrial transcripts, or because potential substrates are physically isolated from RNA editing enzymes in the cytoplasm and nucleus.

4.3.2 Frequency of RNA editing in the human brain

A > I editing occurs at a frequency of approximately 1 in 1,700bp in the RNA sample used for this survey. This is almost ten-fold more than previous estimates of 1 in 17,000 nucleotides (Paul and Bass, 1998). This may be a slight overestimate, as a small proportion of assumed A > I edits identified from sequences with more than three A > G or T > C changes may be incorrect. Conversely, the number of A > I edits from sequences with one or two variants may be slightly underestimated as not all candidate A > I edits were examined. On balance, 1 in 1,700 is likely to be a reasonable estimate of the frequency of A > I editing in the human brain.

The reason for the discrepancy between this and previous estimates is unclear, but may be due to differences in the RNA samples evaluated. The sample used in this survey represents the steady-state poly-(A)+ RNA population of human brain cells. In contrast, the RNA sample used in the previous study was derived from rat brain (Paul and Bass, 1998). In Chapter 5 we show that the majority of the RNA edits are in Alu repeats in introns. As the Alu repeat is primate specific, the number of potential A > I editing sites in rat may be smaller than in human RNA. Furthermore, if our analysis had been performed on more completely processed RNA, for example cytoplasmic RNA, the number of RNA edits would have been smaller.

In conclusion, we have demonstrated that A > I editing of transcripts is the predominant RNA editing activity in human brain. The initial characterisation of these edited sequences forms the next chapter of this thesis.

5 THE CHARACTERISTICS OF A > I EDITED TRANSCRIPTS FROM HUMAN BRAIN

5.1 INTRODUCTION

A small number of A > I RNA edits in human brain transcripts are known to be in translated exon sequences. These include A > I edits in the serotonin receptor transcript and various glutamate receptor transcripts (Bass, 2002). A larger number of A > I edits have been identified in untranslated sequences including introns, 3' untranslated exons and 5' untranslated exons of transcripts from human brain (Morse et al., 2002). However, the overall patterns of A > I RNA edits in different classes of sequence from human brain transcripts is unknown.

The known A > I RNA editing substrates are associated with the formation of dsRNA. In the case of A > I edits in coding sequence, dsRNA is commonly formed between the edited exon and complementary sequence in an adjacent intron. A > I edits in non-coding sequence are commonly found in high copy repeat sequences such as Alus which are predicted to form dsRNA by base-pairing with inverted copies in the same transcript (Morse et al., 2002).

The analysis of sequence variants from 3.1Mb human brain cDNA library sequence led to the discovery of 1,727 novel A > I RNA edits. In this chapter, the genome in the vicinity of these edits was analysed in order to characterise the targets of A > I editing in human brain, and the potential involvement of dsRNA formation.

5.2 RESULTS

5.2.1 A > I RNA editing targets a wide variety of human brain transcripts

In order to identify the transcripts that are subject to A > I editing, the novel A > I edited sequences were compared with the EnsEMBL annotation of the cDNA clones from which they were identified (Figure 5-1). 62% (183 / 297) of sequences were from known genes, 20% (58 / 297) from predicted genes, 3% (9 / 297) from novel genes and 1% (4 / 297) overlapped with more than one gene and therefore could not be clearly identified. The remaining 14% (43 / 297) of sequences were from regions of no annotation, probably representing novel or poorly defined transcripts.

There was no obvious association of RNA editing with any one gene or family of genes. To search for association of RNA editing with gene function, the gene ontologies associated with edited and unedited sequences were compared using GOstat (http://gostat.wehi.edu.au/). However this did not reveal any statistically significant over-representation or under-representation of any function associated with edited genes.

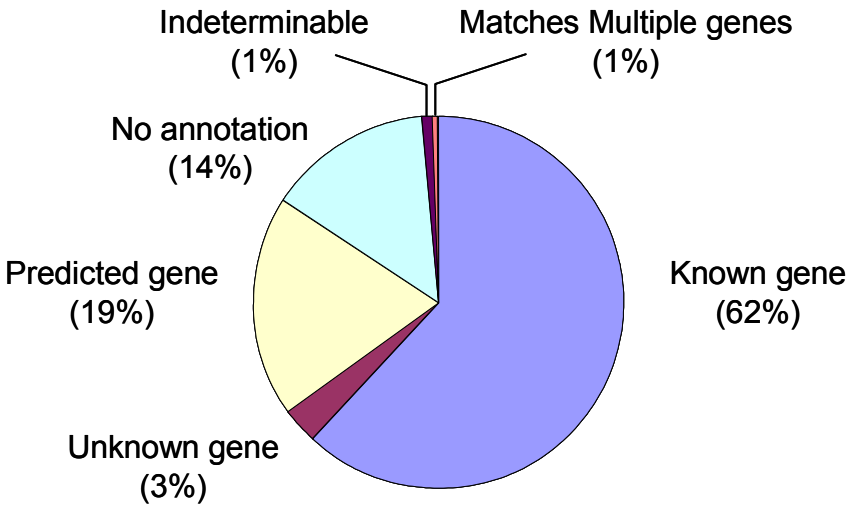


Figure 5-1 Breakdown of RNA edits by gene class. Annotation of edited cDNA clone sequences was derived from the annotation of all cDNA library sequences (see Chapter 3).

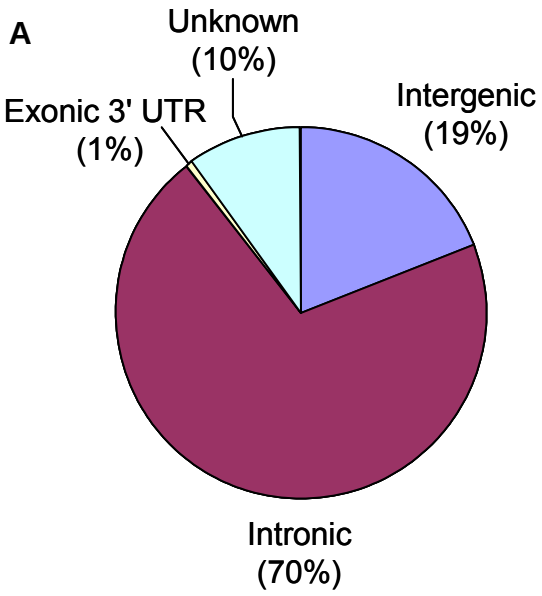
Of transcripts for which there was evidence of editing, 91% (167 / 183) were found in the cDNA library as a single edited clone. The most frequently edited transcript from the library was FRMD4 from which three non-overlapping edited clones and four non-overlapping unedited clones were sequenced. All seven clones from this gene were from the large (approx 0.5Mb) first intron. None of the 20 most abundant transcripts in the cDNA library (see Chapter 3, Table 3-5) were found to be edited. Potential reasons why these sequences are unedited are discussed below.

5.2.2 A > I RNA editing is predominantly in non-coding RNA

Novel RNA edits were next compared with the class of sequence from which they were derived (Figure 5-2A). All RNA edits were in non-coding RNA

sequence, including 70% (1,214 / 1,727) from intronic RNA and 19% (333 / 1,727) were in intergenic transcripts. None of these intergenic edited sequences could be identified by comparison with a database of all known non-coding RNA genes. Only 1% (9 / 1,727) of edits were in 3' untranslated exons and none were found in 5' untranslated exons or in translated exon sequences.

A



B

	Edits/Mb
Intergenic	962
Exonic translated	0
Intronic	816
Exonic 5'UTR	0
Exonic 3'UTR	29
unknown	545
Total	566

Figure 5-2 Distribution of A > I RNA edits by sequence class. **A**. the sequence class distribution of A > I edits. **B**. The frequency of A > I editing in each class of sequence.

RNA editing did not occur at an equal frequency in all classes of non-coding sequence (Figure 5-2B). The most frequently edited class of sequence was intergenic (962 edits per Mb) with a similar, but slightly lower frequency of RNA editing in intronic sequences (816 edits per Mb). RNA editing of 3'

untranslated exons was much less frequent, and no RNA edits were identified in 5'UTR or translated exons.

5.2.3 RNA editing of translated exons is a rare event in human brain

The cDNA library contained 541,777bp of translated exon sequence. Initially, variants from translated exon sequence were evaluated using the lower quality score threshold (see Methods).This allowed us to include as many potential RNA edits as possible in our subsequent analyses. In total, 286 sequence variants were detected (one per 1.9kb) using the lower quality threshold. 125 of these variants failed the higher quality score threshold. 19 out of these 125 were known SNPs, leaving 106 potentially novel variants, 22 of which were successfully evaluated further. 9% (2 / 22) were novel SNPs, and the remaining 91% (20 / 22) were artefacts. None were RNA edits (Figure 5-3, low quality variants). As variants passing only the lower quality score threshold were enriched in sequence artefacts, no further assessment of variants from this category was performed.

161 out of 286 translated sequence variants passed the higher quality score threshold (one per 3.3kb). 93 were known SNPs leaving 68 potentially novel variants. 33 of these 68 variants were evaluated and shown either to be either SNPs or artefacts (Figure 5-3, high quality variants). There were 17 potential non-synonymous coding variants present in the set of 68. Of these, 13 were successfully sequenced as part of these analyses and shown not to be RNA edits.

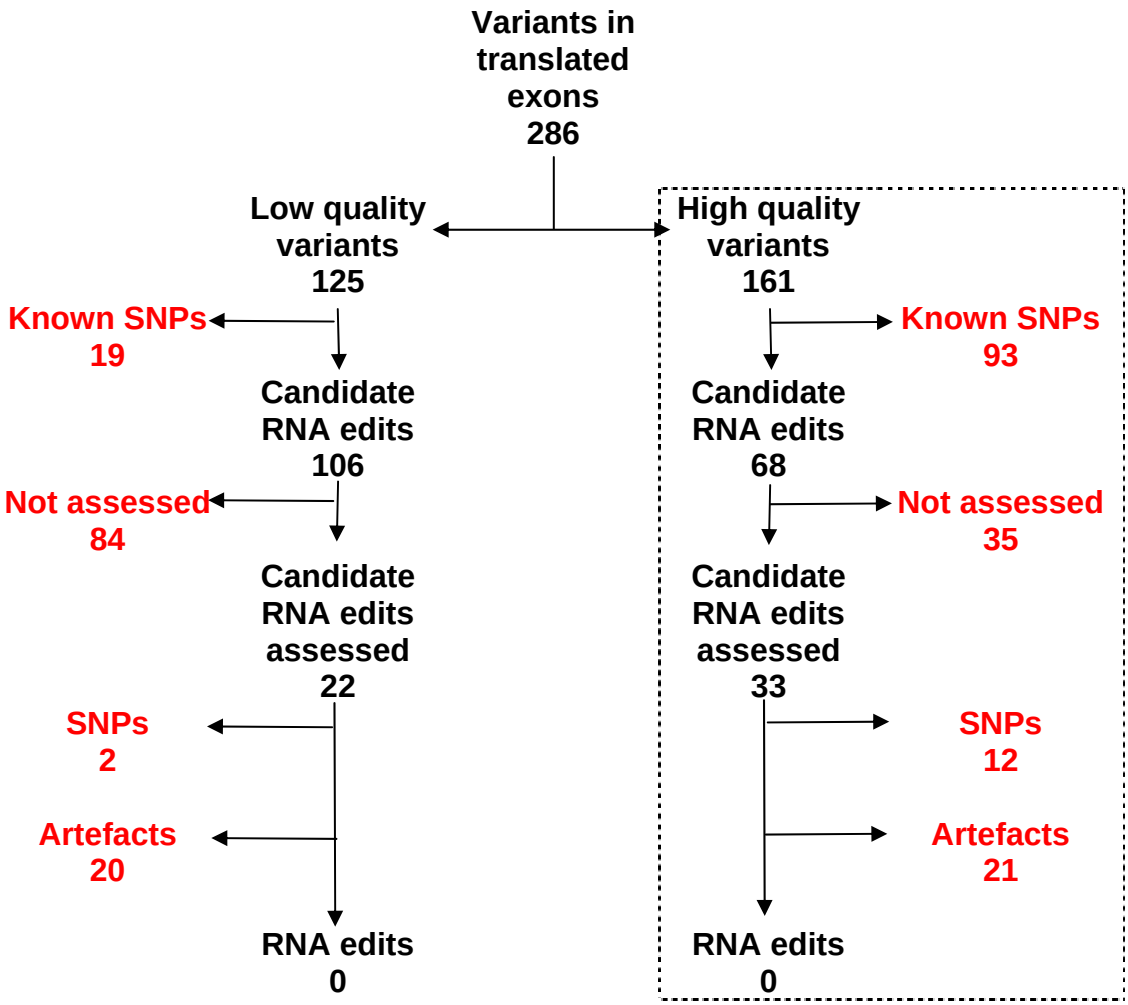


Figure 5-3 Summary of the analysis of the subset of 286 variants from translated exon sequence. Variants were classified as high quality or low quality (see Methods). Variants listed in red were rejected for various criteria Values in black show the remaining candidate RNA edits at each stage of the analysis. The high quality variants formed part of the evaluation of 503 variants from sequences with less than 3 variants described in Chapter 4 (indicated by dashed line). Low quality variants were evaluated in additional experiments.

Although only 167 out of 286 variants from the 541,777bp translated exon sequence were directly investigated and categorised, none out of 55 that were not previously known SNPs turned out to be RNA edits. This suggests that very few of the remaining 119 are likely to be edits and therefore that the total number of edits in the 541,777bp translated exon sequence is very small. To confirm the presence of edited coding sequences in the RNA sample used for this survey, A > I editing of the Q / R site and R / G of the Glutamate Receptor B subunit transcript in total cDNA was successfully demonstrated (data not shown).

5.2.4 A > I RNA editing is associated with Alu repeat sequences

Many of the previously reported RNA edits in non-coding RNA from human brain were in high copy repeat sequences, and were predicted to form dsRNA with inverted copies in the same transcript (Morse et al., 2002).
Therefore, the repeat content of the edited sequences identified in this survey was determined (Table 5-1).

98% (1693 / 1727) A > I RNA edits were in high copy number repeats. The majority, 89% (1548 / 1727), were in Alu repeats which also showed more edits per base sequenced than other repeat classes (Table 5-1). The frequency of editing in Alus (4559 edits / Mb) is almost ten fold greater than the frequency of A > I editing in simple repeats (519 edits / Mb), the second most frequently edited class of repeats.

Repeat	Bases sequenced	Repeats sequenced	Repeats edited	Edits	Edits/Mb
SINE/Alu (All)	339546	2151	302	1548	4559
AluJ	83801	519	79	367	4379
AluS	196178	1197	164	900	4588
AluY	45628	283	43	231	5063
FLAM	9256	99	8	23	2485
FRAM	3114	34	8	27	8671
Alu (MISC)	1569	19	0	0	0
SINE/MIR	49704	455	1	5	101
LINE/L1	269044	1258	18	116	431
LINE/L2	71420	456	0	0	0
SIMPLE	21191	497	6	11	519
LOW COMPLEXITY	18502	471	0	0	0
DNA	54155	398	2	6	111
LTR	103375	505	4	7	68
Other Repeats	10743	69	0	0	0
Other Sequences	2111380	11041	20	35	17

Table 5-1 Distribution of RNA edits by repeat class and subclass.

Amongst the subfamilies of Alus, the number of edits per base analysed did not differ markedly. Three-fold greater numbers of edits were observed in Free Right Arm Monomers (FRAMs) than in Free Left Arm Monomers

(FLAMs). However, subsequent analyses showed no evidence for comparable differences in the number of A > I edits in the FRAM or FLAM components of complete Alus (see Chapter 6). There was considerable variation in the extent of editing of individual Alus in the cDNA library. The Alu with the greatest number of edits had 20 edits from 529 bases sequenced.

Of the other classes of repeats, simple repeats and LINE / L1 repeats were most frequently edited. Although a lower proportion of LINE / L1 repeats were edited, they included the most heavily edited sequence in the cDNA library, containing 28 edits in 568 bases. A small number of RNA edits were not obviously in highly repetitive sequences (Other Sequences in Table 5-1).

5.2.5 The presence of an anti-sense repeat in the same transcript increases the likelihood of RNA editing of Alu sequences

To investigate the role of dsRNA formation in the editing of sequences identified in this survey, custom Perl programs were used to analyse the human genome for the presence or absence of same-sense and anti-sense Alu sequences in the same introns as edited and unedited Alus (see Methods).

Although novel A > I edits were found in several classes of repeat and non-repeat sequences, the majority (90%) were in Alu sequences. The following analyses were therefore simplified by primarily restricting them to Alu repeats. However, there is no reason to believe that the patterns identified do not apply to other classes of repeat sequence. The analysis was further restricted to Alu

sequences which were from known genes. This allowed transcript boundaries, and intron / exon boundaries to be accurately determined in the analysis of the genome sequence flanking Alu repeats. Finally, to avoid wrongly classifying repeat sequences as unedited because of insufficient sequencing, only Alus for which more than 80% of the genomic extent of the repeat was sequenced were used in the analysis. 38% (115 / 302) edited Alus and 22% (411 / 1849) unedited Alus satisfied all of these requirements and were included in the following analyses.

Overall, edited Alus are more likely to have an anti-sense repeat in the same transcript than unedited Alus (Figure 5-4A). For example 50% (2 / 4) edited Alus compared to 6% (2 / 35) unedited Alus from introns of less than 2kb have an anti-sense repeat in the same intron (χ^2 = 4.77, $p \leq 0.05$). In total, 97% (111 / 115) edited Alus had an inverted copy in the same intron, whereas 78% (322 / 411) unedited Alus had an inverted copy in the same intron (χ^2 = 20.4, $p \leq 0.001$). This was not due to a difference in the overall density of repeats flanking edited sequences as there was little difference between the proportion of edited and unedited Alus with a same-sense copy in the same intron (88%, and 91% respectively, Figure 5-4B). The results confirm that A > I RNA editing of the sequences identified in this survey is associated with the presence of an inverted sequence in the same transcript. This is consistent with dsRNA formation through intra-molecular base-pairing between the two repeats.

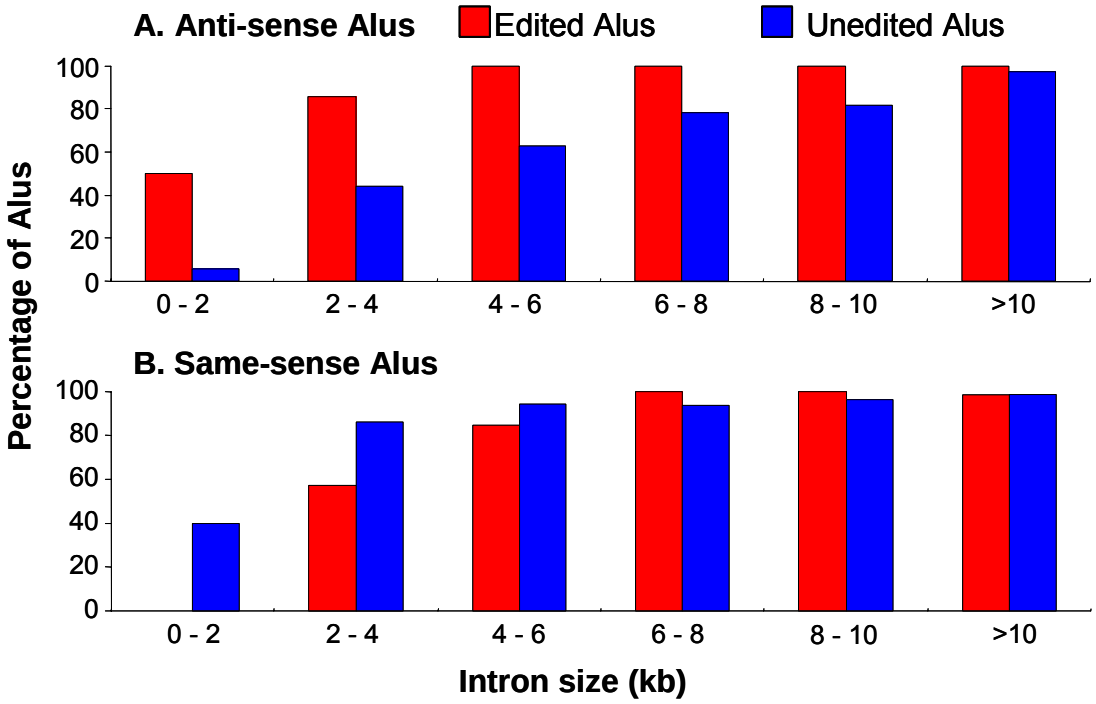


Figure 5-4 Proportion of edited and unedited Alus with additional Alus in the same intron. All Alus aligning to the introns of known genes, and for which ≥80% of the genomic extent of the Alu was sequenced were included in this analysis. The proportion of edited Alus (red bars) and the proportion of unedited Alus (blue bars) having an anti-sense Alu (**A**) or a same-sense Alu (**B**) in the same intron is shown for different intron sizes.

5.2.6 The presence of an anti-sense Alu in the same intron increases the likelihood of RNA editing

To investigate whether the presence of an inverted copy of an Alu in the same intron (as opposed to an adjacent intron) influences A > I RNA editing of Alu sequences, the sizes of introns containing edited and unedited Alus was compared (Figure 5-5). In general, edited and unedited Alus are found with similar frequency in introns of different sizes. For example, 11% (13 / 115)

edited Alus and 10% (42 / 411) unedited Alus are found in introns of 2 to 4 kb in length. However, edited Alus are found less frequently than unedited Alus in introns smaller than 2kb. Only 3% (6 / 189) of all edited Alus from the introns of a known gene compared to 9% (35 / 411) of unedited Alus from the intron of a known gene (and for which greater than 80% of the genomic extent of the Alu was sequenced) are in introns smaller than 2kb ($\chi^2 = 5.16$, $p \leq 0.025$). If RNA editing occurred preferentially at Alus with an inverted copy nearby in the same transcript, but not necessarily in the same intron, the presence of an inverted copy in the same intron would not be important, and we would have observed an equal number of edited and unedited Alus in introns of all sizes. Instead, RNA editing of Alus in small introns is rare. Presumably, this is because introns shorter than 2kb have less space to accommodate multiple Alus and so are less likely to contain inverted copies which have the potential to form dsRNA. This result suggests that RNA editing occurs preferentially at Alus that are enriched in inverted copies *in the same intron*, rather than nearby in the same transcript.

Although having an inverted copy in the same intron clearly increases the likelihood of editing, it is not always required. There were four edited Alus that did not have an anti-sense copy of a repeat in the same intron. All of these sequences were situated in a small intron (<5kb), all were close to an intron / exon boundary (<1kb), and all were close to an anti-sense repeat in an adjacent intron (<2kb). This suggests that infrequently, RNA editing may take place between closely placed Alus in adjacent introns.

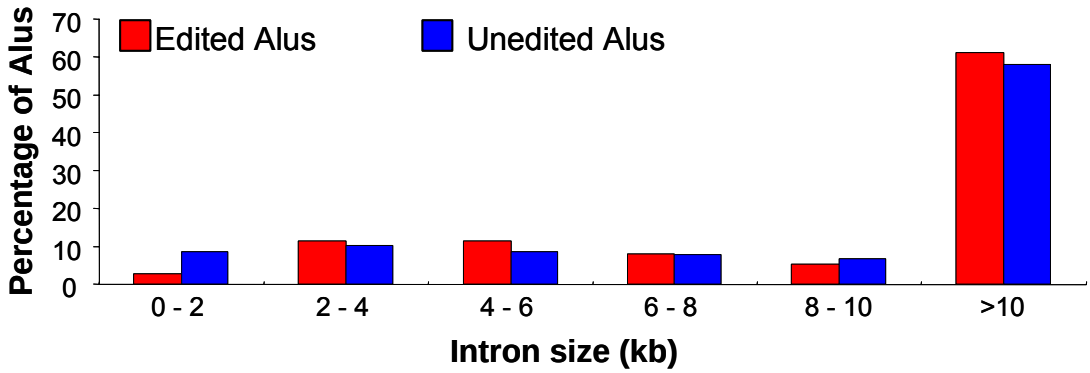


Figure 5-5 Proportion of edited and unedited Alus from introns of different sizes. Intron sizes were recorded for all Alus aligning to the introns of known genes, and for which ≥80% of the genomic extent of the Alu was sequenced. The proportion of edited alus (red bars) and unedited alus (blue bars) from different intron sizes is compared.

5.2.7 The proximity of inverted Alu sequence influences the likelihood of RNA editing

The effect of the proximity of an inverted Alu repeat within the same intron upon the likelihood of an Alu being edited was studied. Custom Perl programs were used to calculate the proportion of edited and unedited Alu sequences with an anti-sense Alu within 0-1kb in the same intron. The results were then broken down according to the size of the intron from which edited or unedited Alus were derived (Figure 5-6A and 5-6B).

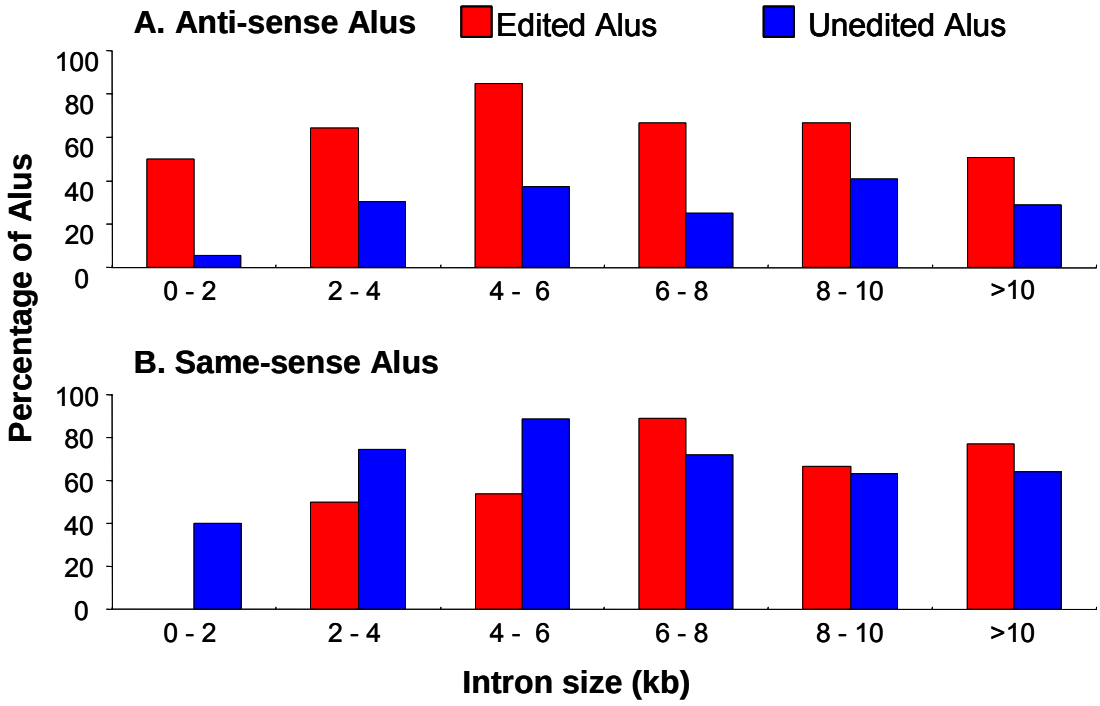


Figure 5-6 Proportion of edited and unedited Alus with additional Alus within 0 to 1 kb in the same intron. The Alu sequences included in this analysis are the same as in Figure 5-4. The proportion of edited Alus (red bars) and unedited Alus (blue bars) with an anti-sense Alu (**A**) or a same-sense Alu (**B**) within 1kb in the same intron is shown for different intron sizes.

Overall, edited Alus are more likely than unedited Alus to have an inverted Alu within 1kb in the same intron. For example 50% (2 / 4) edited Alus compared to 6% (2 / 35) unedited Alus from introns smaller than 2kb have an inverted Alu within 1kb (χ^2 = 4.77, p ≤ 0.05). Even in large introns, edited Alus are more likely than unedited Alus to have an inverted copy within 1kb. For example, although all (69 / 69) edited Alus and nearly all (97%, 232 / 239) unedited Alus from introns larger than 10kb have an anti-sense copy in the same intron (Figure 5-4A, >10kb), only 29% (69 / 239) unedited Alus

compared to 51% (35 / 69) edited Alus have an anti-sense copy within 1kb (χ^2 = 11.4, $p \leq 0.001$) (Figure 5-6A, >10kb). Therefore, this effect is not simply a consequence of the preference for RNA editing of Alus with an anti-sense copy in the same intron. The effect is not attributable to a high density of Alu repeats in general in the vicinity of edited Alus, as there is little difference between the proportion of edited and unedited Alus with a same-sense copy within 1kb in the same intron (Figure 5-6B). Instead, the effect is best explained by preferential editing of dsRNAs formed by *closely spaced* inverted Alus in the same intron.

To investigate further the effect of proximity of inverted copies on RNA editing, the proportion of edited and unedited Alus at different distances from the nearest anti-sense Alu in the same intron was calculated (Figure 5-7A). Overall, edited Alus are more frequently close to an inverted copy within the same intron than unedited Alus. The effect is most marked at shorter distances, with 58% (67 / 115) of all edited Alus compared to only 27% (112 / 411) of all unedited Alus having an inverted copy within 1kb (χ^2 = 38.49, $p \leq 0.001$). Conversely, no association with likelihood of Alu editing is observed for proximity of same-sense Alus (Figure 5-7B). Consistent with previous results, A > I editing is most strongly associated with the presence of an inverted repeat within 2kb in the same intron. Fewer edited than unedited Alus are more than 2kb from the nearest inverted copy. For example, 32% (132 / 411) unedited Alus compared with only 5% (6 / 115) edited Alus are more than 5kb from the nearest inverted copy in the same intron.

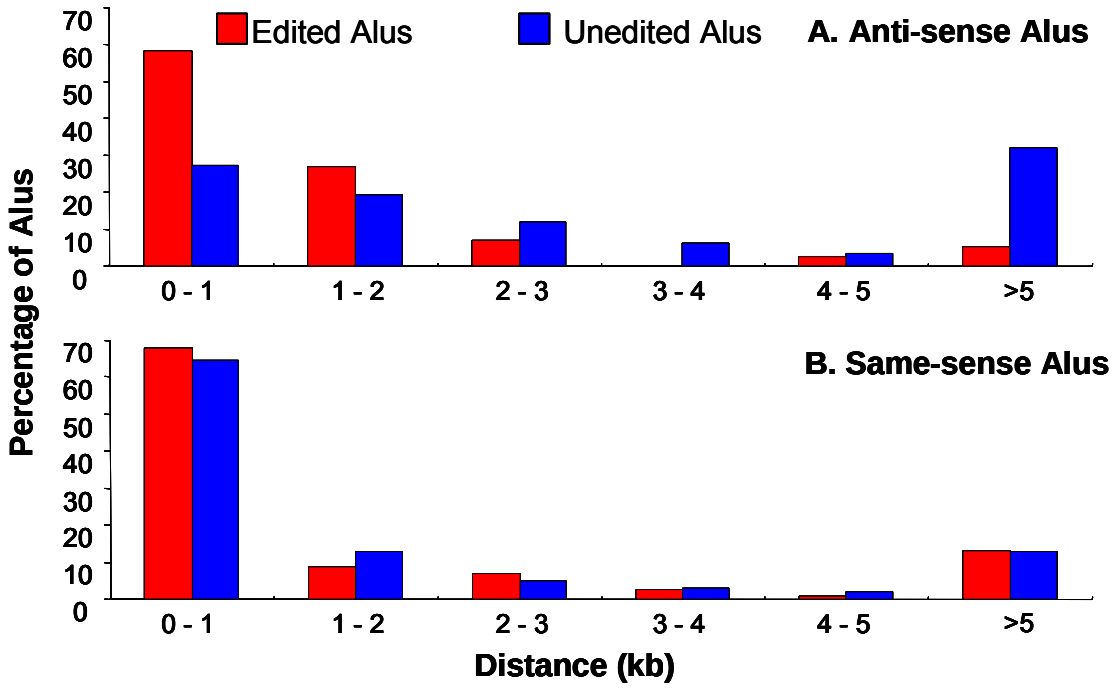


Figure 5-7 Distance from edited and unedited Alus to the nearest Alu in the same intron. The Alu sequences included in this analysis are the same as in Figure 5-4. The proportion of edited Alus (red bars) and unedited Alus (blue bars) at different distances from the nearest anti-sense Alu (**A**) or same-sense Alu (**B**) is shown.

5.2.8 The amount of inverted Alu sequence is associated with the likelihood of RNA editing

To investigate whether the amount of inverted Alu copy sequence in the vicinity of an Alu influences the likelihood of A > I editing, the amount of Alu sequence flanking all edited and unedited Alus was determined. Edited Alus have more inverted Alu copies than unedited Alus at all distances up to 10kb (Figure 5-8A). For example, the average amount of anti-sense Alu sequence within 4 - 5kb flanking edited Alus is 99 bp / kb, compared with 50 bp / kb at

the equivalent distance flanking unedited Alus. However, the effect is strongest within 1kb of the Alu where the average amount of flanking anti-sense Alu sequence in the vicinity of edited Alus (64 bp / kb), is greater than three-fold more than the average amount of anti-sense sequence in the vicinity of unedited Alus (20 bp / kb).

Interestingly, a similar effect, of lesser magnitude, is observed for same-sense Alus (Figure 5-8B). For example, the average same-sense Alu sequence content within 4 - 5kb flanking edited Alus is 96bp / kb compared with 71bp / kb for unedited Alus.

For both edited and unedited Alus, there is a decrease in the quantity of anti-sense Alus and an increase in the quantity of same-sense Alus at close proximity. Whilst the average amount of anti-sense Alu sequence within 0 – 1kb flanking unedited Alus (18bp / kb) is one third of that between 3 and 4kb (54bp / kb, Figure 5-8A), conversely, the average amount of same-sense Alu sequence within 0 – 1kb flanking unedited Alus (137bp / kb) is nearly twice that in the flanking sequence within 3 to 4kb (75bp / kb). It has previously been reported that genome-wide, there is an over-representation of same-sense Alus, and an under-representation of anti-sense Alus in close proximity to Alu repeats (Stenger et al., 2001). The over-representation of same–sense Alus is thought to arise through insertion of multiple Alu repeats in sequences which satisfy the local sequence preferences of Alu insertion (i.e. AT rich sequences), whilst the under-representation of anti-sense Alus is thought to

relate to toxic effects, perhaps genome instability associated with closely spaced inverted repeats.

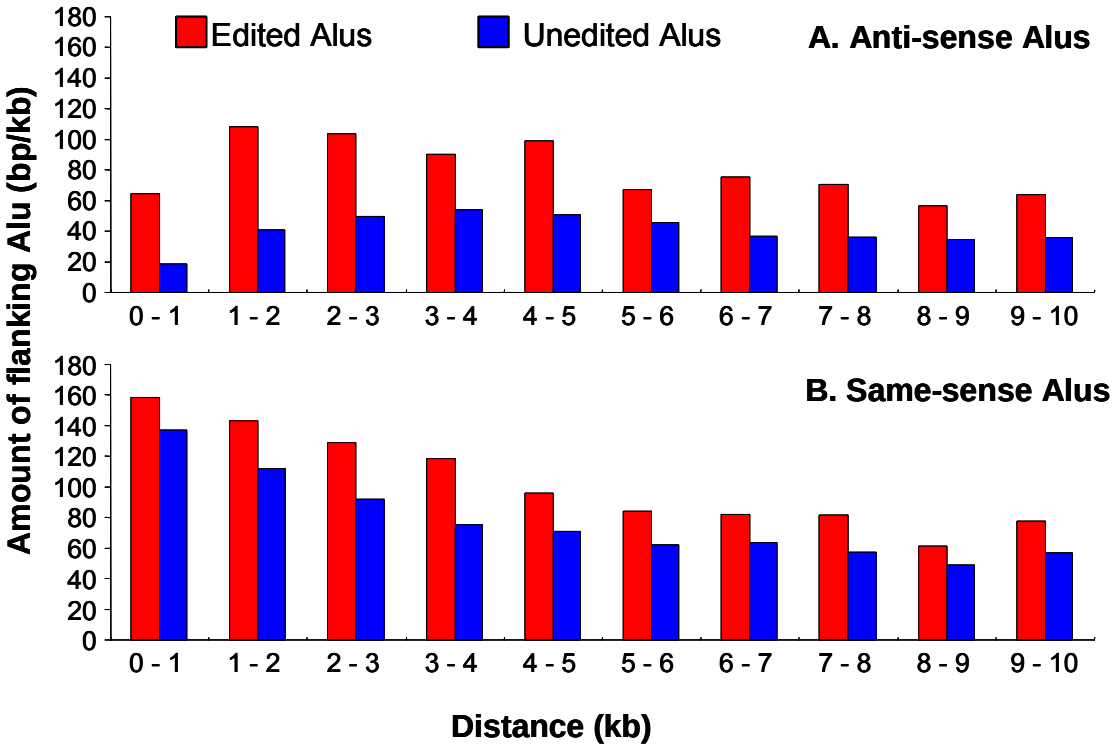


Figure 5-8 Amount of flanking Alu sequence at different distances from edited and unedited Alus. All Alus for which ≥80% of the genomic extent of the Alu was sequenced were included in this analysis. For each Alu, the amount of flanking Alu sequence in the opposite orientation (**A**) or same orientation (**B**) in successive 1kb windows was recorded. For each distance, the flanking Alu sequences in the 1kb window 5' and 3' of the reference Alu were combined. The data presented is the average amount of Alu sequence flanking all edited Alus (red bars) or unedited Alus (blue bars).

5.2.9 The orientation of Alus with respect to transcription has no impact on RNA editing

The Alu repeat is asymmetrical, consisting of a FLAM monomer, a FRAM monomer and a poly-(A) tail (see Introduction Figure 1-1). Therefore, as components of other RNAs, Alus can be transcribed in the forward orientation (with a poly-A tail) or reverse orientation (with a leading poly-T sequence). To investigate potential differences in A > I RNA editing of forward and reverse Alu sequences, Alus were oriented with respect to the transcribed strand, and the number of edited Alus transcribed in the forward orientation and reverse orientation was compared.

In total, 20% (53 / 265) of Alus transcribed in the forward orientation, and 24% (67 / 283) of Alus transcribed in the anti-sense orientation were edited. Therefore, there is no strong preference for editing of Alus in a particular orientation (χ^2 = 0.69, p ≤ 1). This result is consistent with the formation of dsRNA between inverted Alu repeats, and with both strands of the dsRNA being edited.

5.2.10 The orientation of Alus with respect to each other has no impact on RNA editing

An Alu can potentially form dsRNA with inverted copies positioned either 3' or 5' in the flanking transcript. For each Alu, this results in two possible RNA duplexes. If dsRNA is formed between a forward Alu and a reverse Alu 3' in the same transcript (or between a reverse Alu and a forward Alu 5' in the same transcript), the poly-(A) tail of the forward Alu and the poly-T tail of the

reverse Alu will base pair towards the loop of the RNA hairpin (Figure 5-9, Tails in). Conversely, if dsRNA is formed between a reverse Alu and a forward Alu 3' in the same transcript (or between a forward Alu and a reverse Alu 5' in the same transcript), the poly-A tail of the forward Alu and the poly-T tail of the reverse Alu will be at the base of the RNA hairpin (Figure 5-9, Tails out).

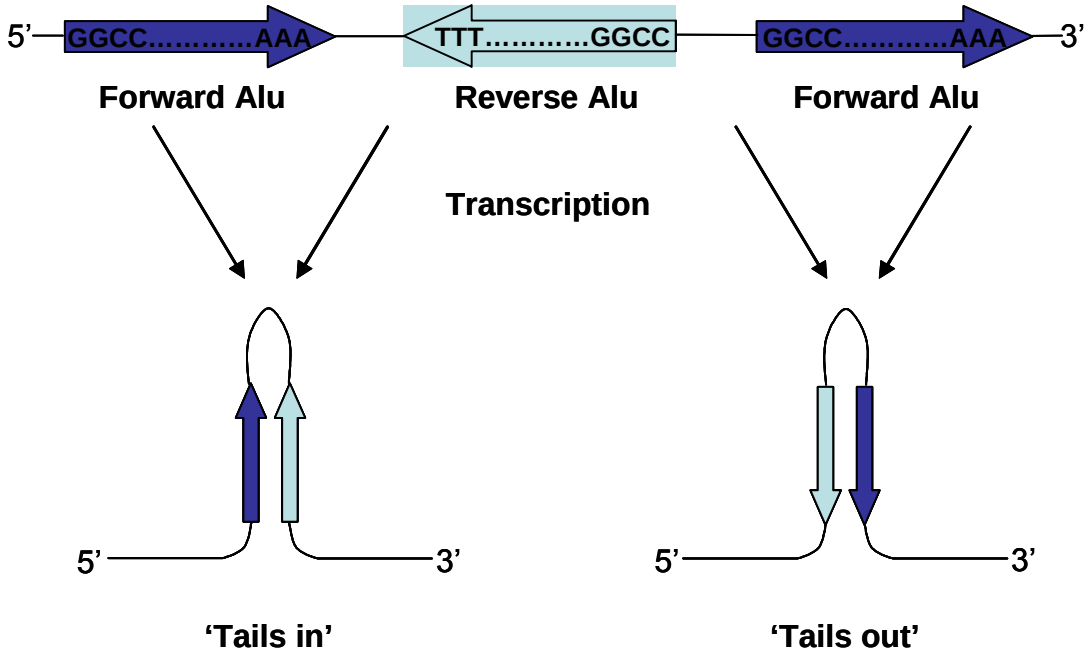


Figure 5-9 Orientation of Alu sequences with respect to each other. Alus may be transcribed in the forward (dark blue arrows), or reverse (light blue arrows) orientation. Arrowheads indicate the position of the poly-A tail (forward Alus) or leading poly-T sequence (reverse Alus). A pair of inverted repeats may be transcribed in the 'tails in' or 'tails out' conformation.

To investigate the effect of the orientation of Alus within hairpins on A > I RNA editing, the amount of flanking Alu sequence in a 'tails-in' orientation (Figure 5-10A), and in a 'tails-out' orientation (Figure 5-10B) was calculated for edited

and unedited Alus. No clear difference in the amount of tails-out or tails-in anti-sense sequence flanking edited compared to unedited Alus was observed. For example, within 1kb of edited Alus there is an average of 87bp / kb anti-sense Alu sequence in the 'tails-in' orientation, and similarly there is an average of 101bp / kb anti-sense Alu sequence in the 'tails-out' orientation. This suggests that 'tails-in' and 'tails-out' hairpins are edited with similar efficiency.

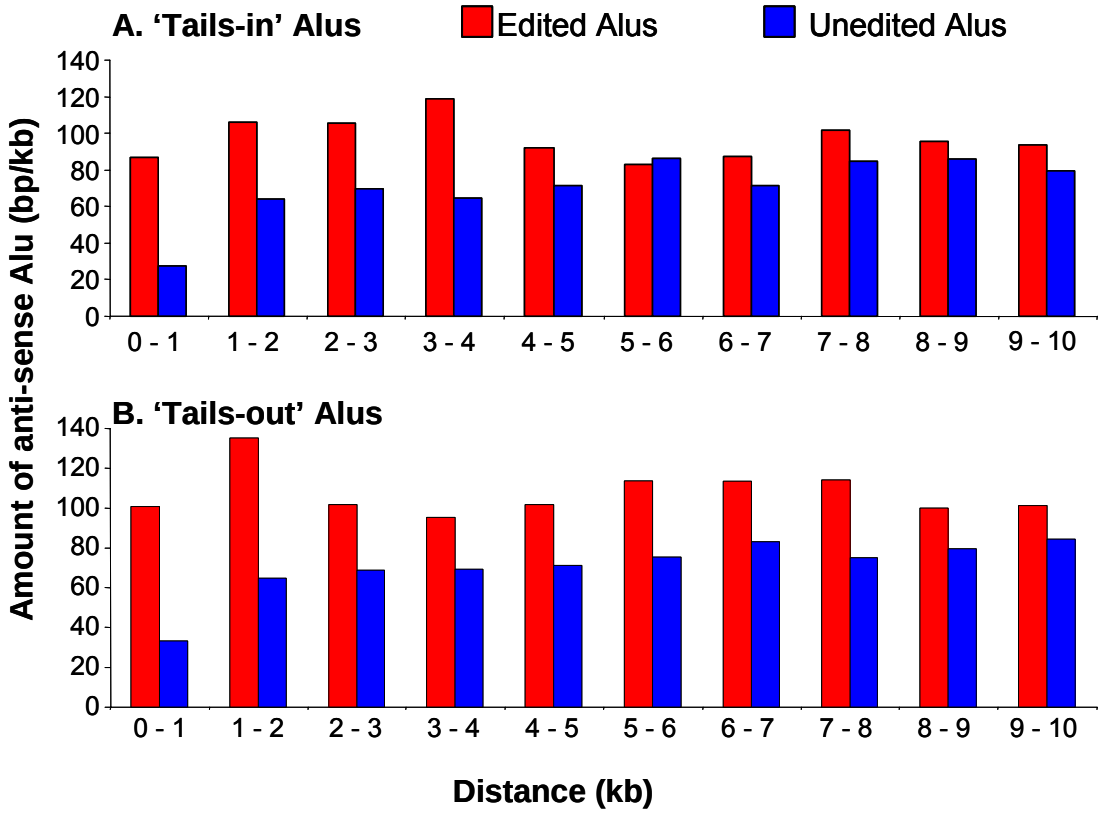


Figure 5-10 Amount of anti-sense Alu sequence at different distances from edited and unedited Alus in 'Tails-Out' orientation (**A**), or 'Tails-in' orientation (**B**). 'Tails-in' anti-sense Alus are all reverse Alus 3' in the same transcript as forward Alus, and all forward Alus 5' in the same transcript as reverse Alus. 'Tails-out' anti-sense Alus are all forward Alus 3' in the same transcript as

reverse Alus, and all reverse Alus 5' in the same transcript as forward Alus. All Alus from known genes from which ≥80% of the genomic extent of the Alu was sequenced were included in this analysis. For each Alu, the amount of flanking Alu sequence in successive 1kb windows was recorded. For each distance, the flanking Alu sequences in the 1kb window 5' and 3' of the reference Alu were combined. The data presented is the average amount of Alu sequence flanking all edited Alus (red bars) or unedited Alus (blue bars).

5.2.11 Further analysis of Alus that have an inverted repeat in the same intron but are apparently unedited

Although the vast majority of edited cDNA clone sequences were Alu repeats, the cDNA library contained many more unedited Alus (1,849) than edited Alus (302) (Table 5-1). Many apparently unedited Alus have an inverted copy in the same intron, and might therefore be predicted to undergo A > I RNA editing. It is possible that these sequences are actually weakly edited in the cell, but by chance we cloned unedited rather than edited transcripts. For example, some transcripts containing inverted Alus may be weakly expressed in a sub-set of brain cells in which A > I RNA editing occurs, but overwhelmingly expressed in another sub-set of brain cells in which A > I RNA editing is absent. The total RNA population of such a transcript would contain predominantly unedited molecules, and these would be more likely than edited molecules to be sampled by the random cDNA cloning and sequencing approach used in this survey.

To investigate the possibility that apparently unedited Alus (from cDNA clone sequencing) with an inverted copy within 2kb are actually edited, 63 unedited Alus with an inverted copy within 2kb were amplified by RT-PCR from human brain total RNA, sequenced, and compared to the matching genomic DNA sequence. 54% (34 / 63 including 11 with an inverted copy in the same intron) were, as expected unedited. However, the remaining 46% (29 / 63 including 13 with an inverted copy in the same intron) did show evidence of editing. These results suggest that the presence of an inverted Alu within 2kb is not sufficient for RNA editing. The results also indicate that a small proportion of the Alus classed as unedited in the earlier analyses are actually edited. Therefore, the differences demonstrated between unedited and edited sequences are likely to be underestimated.

5.2.12 The genome wide distribution of inverted Alus within 2kb in the same intron

To estimate the genome wide prevalence of potential RNA editing substrates formed by inverted Alu repeats, a search was performed of all transcripts from all known EnsEMBL genes. For each transcript, the number of pairs of inverted Alu sequences within 10kb in the same intron, with at least 50bp of complementary sequence, was recorded. Of 25,662 transcripts from known genes that were evaluated, 63% (16,249 / 25,662), have at least one intron containing a pair of inverted Alus, and therefore are potential RNA editing substrates. This includes 844 transcripts with more than 100 pairs of intronic inverted Alus. The remaining 9,413 transcripts contained no pairs of inverted Alus within an intron. These comprised 2,660 transcripts with no introns and

6,753 with no inverted repeats despite having at least one intron and up to 25 Alus in the transcripts. Transcripts without an intronic Alu hairpin included olfactory receptors (which are intron-less) and many housekeeping genes such as actin and tubulin. Housekeeping genes are compact, with an average intron size of 2kb compared to the genome wide average of 5kb. The median intron size of housekeeping genes is 600bp which would be insufficient for accommodating two Alus. These intronic characteristics may underlie the absence of RNA editing of any of the most frequently sequenced transcripts from the cDNA library (see section 5.2.1).

To identify potential RNA editing substrates involving translated exons, the dataset of transcripts described above was searched for intronic Alu repeats with an inverted copy in an adjacent exon. In total 236 potential Alu hairpins involving translated exon sequences were identified. However, there was no obvious enrichment of any gene or group of genes.

5.2.13 The role of dsRNA formation in non-Alu edited sequences.

Although most of the observed RNA edits were in Alu sequences, there were 145 edits in 31 edited sequences from other repeats (Table 5-1). The majority of these sequences were LINE / L1 repeats which accounted for 116 edits from 18 sequences. Unfortunately, because of the relatively small amount of data from edited LINE sequences, it was not possible to repeat the detailed analyses performed for Alu sequences. However, 57% (8 / 14) of edited LINE / L1 repeats compared with only 15% (152 / 995) unedited LINE / L1 repeats contained an inverted LINE / L1 copy which overlapped by at least 50bp, and was within 5kb in the flanking sequence (χ^2 = 18.14, $p < 0.001$). These data

suggest that as with RNA editing of Alu sequences, LINE / L1 editing is influenced by the presence of a nearby inverted copy.

Although a similar amount of LINE / L1 (270kb) and Alu (340kb) repeat sequence was obtained from the cDNA library, only 1% (18 / 1258) LINE / L1 repeats compared with 14% (302 / 2,151) Alu repeats were edited (Table 5-1). The lower frequency of editing of LINE / L1 repeats may simply be a consequence of a lower likelihood of nearby inverted copies that would be available for dsRNA formation. Consistent with this, 71% (1,305 / 1,837) Alus have an inverted Alu within 5kb which overlaps by at least 50bp, and therefore may form dsRNA. Conversely, only 11% (104 / 1010) LINE / L1 repeats have an inverted LINE / L1 within 5kb that overlaps by at least 50bp (χ^2 = 961, $p < 0.001$).

The only repeat class in which there clearly did not seem to be a relationship between the likelihood of A > I RNA editing and the presence of a nearby inverted copy was simple repeats. All six of these sequences were TA dinucleotide repeats. These can form dsRNA molecules internally and therefore the presence of an inverted copy in the flanking sequence is not required for the formation of dsRNA.

20 edited sequences were not from high copy number repeats (other sequences, Table 5-1). On further inspection, 18 of these were from cDNA clones containing high copy number repeats and are therefore likely to be close to the dsRNAs formed through these repeats. Two of the sequences

were not close to any high copy repeat sequence. However BLAST analysis of the flanking sequence revealed inverted repeats within 1kb (one sequence forming a predicted duplex of approximately 35bp, the other a duplex of approximately 100bp). Therefore, these sequences are likely to form dsRNAs and to be substrates for ADAR editing.

5.3 DISCUSSION

5.3.1 Sequence class composition of RNA editing substrates

The novel A > I RNA edits identified in this survey were confined to untranslated RNA sequence, including introns, 3' untranslated exons and intergenic RNAs. Although the majority of edited sequences were from introns, the most heavily edited sequences identified were from intergenic regions of the genome (962 edits per Mb compared with 816 edits per Mb in intronic sequences). The reason for the higher frequency of editing in intergenic sequence is unclear.

RNA editing of untranslated exons is less frequent than editing of either intronic or intergenic classes of non-coding sequence. As discussed below, the majority of RNA editing is associated with repeat sequences, particularly Alus, where pairs of inverted repeats are predicted to underlie formation of dsRNA. The Alu sequence content of 5' UTR (2% Alu), and 3' UTR (5% Alu) is less than that of introns (13% Alu). Therefore, pairs of inverted repeats would be expected to occur less frequently in untranslated exons than in

introns. Furthermore, the average 5' UTR (300bp) and 3' UTR (770bp) are shorter than the average intron (3,365bp) and therefore may be unable to harbour a pair of inverted Alus (300bp each). Finally, unlike introns, 5' and 3' UTRs are retained in the mature mRNA. The presence of dsRNA in mature mRNA, may be subject to additional selective pressures, for example by impacting on polyadenylation, translation or stability of mRNA.

Despite sequencing 167 out of 286 variants from 541,777bp coding cDNA sequence, no novel coding RNA edits were confirmed, indicating that the frequency of A > I editing in coding sequence is low compared to that in non coding sequence. However, this analysis of RNA edits in coding sequence was not exhaustive, and does not rule out the existence of novel A > I or other types of RNA edits in coding exons of human brain transcripts. Further analysis would be necessary to evaluate the number of RNA editing sites in coding sequence, and to completely catalogue the coding RNA edits of the human brain transcriptome.

5.3.2 Association of RNA editing with repeat sequences

RNA editing is strongly associated with the presence of repeat elements, especially Alus. Consistent with previous observations, this appears to be a consequence of dsRNA formation between inverted repeats in the same transcript (Morse et al., 2002). Although Alu subfamilies vary substantially in their genomic copy number, there seems to be little difference in the frequency of editing of these subfamilies. This would suggest that members of Alu subfamilies do not discriminate between each other in the formation of

double stranded mRNA i.e. that a member of one subfamily is as likely to form dsRNA and be edited with a member of its own subfamily as with a member of another subfamily.

5.3.3 The role of dsRNA formation in RNA editing

The analysis of the finished human genome sequence in the vicinity of edited Alu sequences confirms that the potential for dsRNA formation is associated with whether or not a sequence is edited. The likelihood of a sequence being edited is increased in proportion to the amount and proximity of inverted copy sequence (which can potentially serve as a partner in dsRNA formation) with the strongest effects observed when the two copies are within 2kb of each other.

The likelihood of a sequence being edited also appears to be dependent upon the two inverted copies being within the same intron. Thus edited Alus are observed less frequently than unedited Alus within small introns (<2kb), presumably because of the preference for an inverted copy within the restricted space. These data suggest that inverted copies of a sequence can form dsRNA and become edited if they are within the same loop (lariat) of RNA that is removed during RNA splicing, but are much less likely to do so if they are in different loops.

The preference for a pair of inverted repeats in the same intron may add to the reasons why A > I RNA editing in untranslated exons is less frequent than in introns or transcripts of intergenic sequences. Alus in untranslated exons

are separated from inverted Alu repeats in the neighbouring intron by an intron / exon boundary. This may have the same negative effect on A > I RNA editing as the presence of an exon between a pair of inverted Alus in adjacent introns.

The presence of inverted copies at distances greater than 2kb appears to have less influence on the likelihood of an Alu being edited. Nevertheless, the frequency of inverted Alu repeats up to 10kb distant is higher for edited sequences than unedited sequences. Although this may in part be due to a direct biological interaction between two distant inverted copies to form dsRNA, the effect (although less marked) is observed for same-sense sequences as well. These longer distance associations of repeat copy density with likelihood of editing may be a reflection of the existence of large Alu rich genomic domains. Edited Alus are more likely to be in Alu rich domains because this will be associated with a higher frequency of Alus in close proximity.

If the likelihood of editing is increased by the proximity of inverted sequence copies, it is conceivable that proximity of same-sense copies might reduce the likelihood of editing, perhaps by competing for nearby inverted copies in the formation of dsRNA. The results suggest, however, that the presence of a same-sense Alu in the vicinity is not associated with a decrease in the likelihood of editing (except in small introns, where they occupy the space that might be taken by an inverted copy). Indeed, there is a slightly higher frequency of same-sense Alus at all distances up to 10kb from edited

sequences compared to unedited sequences. These are perhaps due to the existence of large Alu rich domains in which both sense and anti-sense Alus are more common. Indeed, there is known to be widespread variation in Alu repeat density. For example, a 100kb region of chromosome 7q11 has an Alu repeat sequence content in excess of 56%, whereas each of the human homeobox gene clusters contains a region of around 100kb of less than 2% interspersed repeat sequence (Lander et al., 2001).

5.3.4 Edited Alus with no inverted copy in the same intron

There are, however, edited Alus for which no inverted copy within the same intron can currently be identified. Some of these may be due to anomalies in gene annotation. Alternatively, double stranded mRNA formation with independent mRNA molecules such as anti-sense transcripts, double stranded mRNA formation with an inverted copy in an adjacent intron before the splicing machinery separates the two copies, or conceivably an editing process which does not rely on double stranded mRNA, may be responsible.

5.3.5 Unedited Alus with an inverted copy in the same intron

Some Alus are not edited to a detectable extent even if there is an inverted repeat within 2kb in the same intron. This suggests that, in addition to the presence of a nearby inverted copy within the same intron, other factors influence the likelihood of editing. One of these may simply be whether a transcript is predominantly expressed in a cell type(s) that has low levels of editing. Previous data show that the extent of A > I RNA editing is highly

variable between tissues (Paul and Bass, 1998). Brain is a heterogeneous tissue composed of several constituent cell types including nerve cells, astrocytes, oligodendrocytes, endothelial cells and microglia. Therefore, unedited Alus with an inverted copy in the same intron may simply be part of transcripts that are expressed exclusively in cells with no editing activity, (and similarly fully edited transcripts may be expressed only in cells with high editing activity).

5.3.6 RNA editing of non-Alu repeat sequences

The most commonly edited repeats are Alus. A much smaller proportion of MIRs, LINEs and other repeats are edited. The lower frequency of editing of repeats other than Alus may simply be a consequence of lower genome copy number and hence lower likelihood of nearby inverted copies that would be available for dsRNA formation. For example, the full length LINE / L1 repeat is approximately 6.1kb, and therefore approximately twenty times the length of a full length Alu sequence (approximately 300bp). Therefore, despite LINE / L1 sequences occupying a higher proportion of the genome than Alu sequences, the effective genome copy number of LINE / L1 repeats is much lower than that of Alus. Furthermore, LINE / L1 repeats are underrepresented in gene rich regions of the genome, whereas Alu sequences are enriched in gene rich regions. As a result, the difference in copy number between the two classes of repeats will be even greater in transcribed regions of the genome. For example, only 10% of LINE / L1 repeats compared with 71% of Alu repeats have an overlapping inverted copy within 5kb in the same transcript.

Overall, the data presented in this chapter is consistent with a model in which the likelihood of A>I editing is largely dependent on the likelihood of dsRNA formation. This in turn is predominantly determined by the proximity and amount of inverted copy sequence, particularly in the same intron. By implication, the results also indicate that most edited dsRNAs are formed by intramolecular RNA base pairing. Although other sources of dsRNA cannot be ruled out (for example through base pairing of independent sense and anti-sense transcripts), the very low frequency of edited Alus without an inverted copy in the close vicinity suggests that these only account for a small fraction of edited Alus (although possibly more of other classes of repeat).

These observations are broadly consistent with previous reports of A > I edited transcripts identified by cloning of inosine-containing transcripts from human (Morse et al., 2002), and by computational analysis of human ESTs and cDNAs (Levanon et al., 2004, Kim et al., 2004), in which editing was found predominantly in transcribed Alus in non-coding sequence, and was associated with dsRNA formation.

6 THE ROLE OF LOCAL SEQUENCE EFFECTS IN RNA EDITING

6.1 INTRODUCTION

The ADAR RNA editing enzymes bind to their substrates predominantly through their dsRNA binding domains. The dsRNA binding domain has a general affinity for RNA duplexes, so dsRNA formed between inverted Alu sequences and dsRNA formed between inverted LINE/L1 sequences are both targets for RNA editing. However, within these dsRNAs, preferences for adenosines in certain sequence contexts have been previously demonstrated. In the case of RNA editing by ADAR2, this is at least partly attributable to binding selectivity of the dsRNA binding domain (Stephens et al., 2004).

In vitro analyses of RNA editing of synthetic dsRNAs indicate that A > I editing by Xenopus ADAR1 takes place preferentially at adenosines that are immediately 3' to U = A > C > G, but with no preference for the nucleotide immediately 3' of the adenosine (Polson and Bass, 1994). Human ADAR2 A > I editing occurs preferentially at adenosines immediately 3' to U = A > C = G, and immediately 5' to U = G > C = A (Lehmann and Bass, 2000). Analyses of a small number of edited adenosines in ADAR2 itself were broadly concordant with these patterns (Dawson et al., 2004).

Further *in vitro* experiments indicate that base-pairing of adenosines within dsRNA also influences the likelihood of RNA editing. Adenosines at A:C mismatches are more efficiently edited than adenosines at A:U matches or other mismatches (Wong et al., 2001).

The large number of novel RNA edits identified in this survey enabled a more in-depth analysis of sequence preferences and base-pairing preferences than has previously been possible from the relatively small number of known substrates or from synthetic dsRNAs.

6.2 RESULTS

6.2.1 Local sequence preferences A > I RNA editing

The role of local sequence context in RNA editing was addressed by selecting edited Alu sequences, identifying the bases at positions up to 10bp 5' and up to 10 bp 3' of edited adenosines and comparing these to the bases up to 10bp 5' and up to 10bp 3' of unedited adenosines. The results show that there is a marked deficit of G at the 5' position to an edited A. There is a compensatory increase of U (and to a lesser extent C) (Figure 6-1). There is also an excess of G at the 3' position to an edited A with minor compensatory fluctuations of the other bases. At all positions 5' and 3' to the edited adenosine, edited bases show fewer adenosines than unedited bases. This seems to be attributable mainly to complete absence of editing of the FRAM associated poly-(A) tail of Alus (see Figure 6-4).

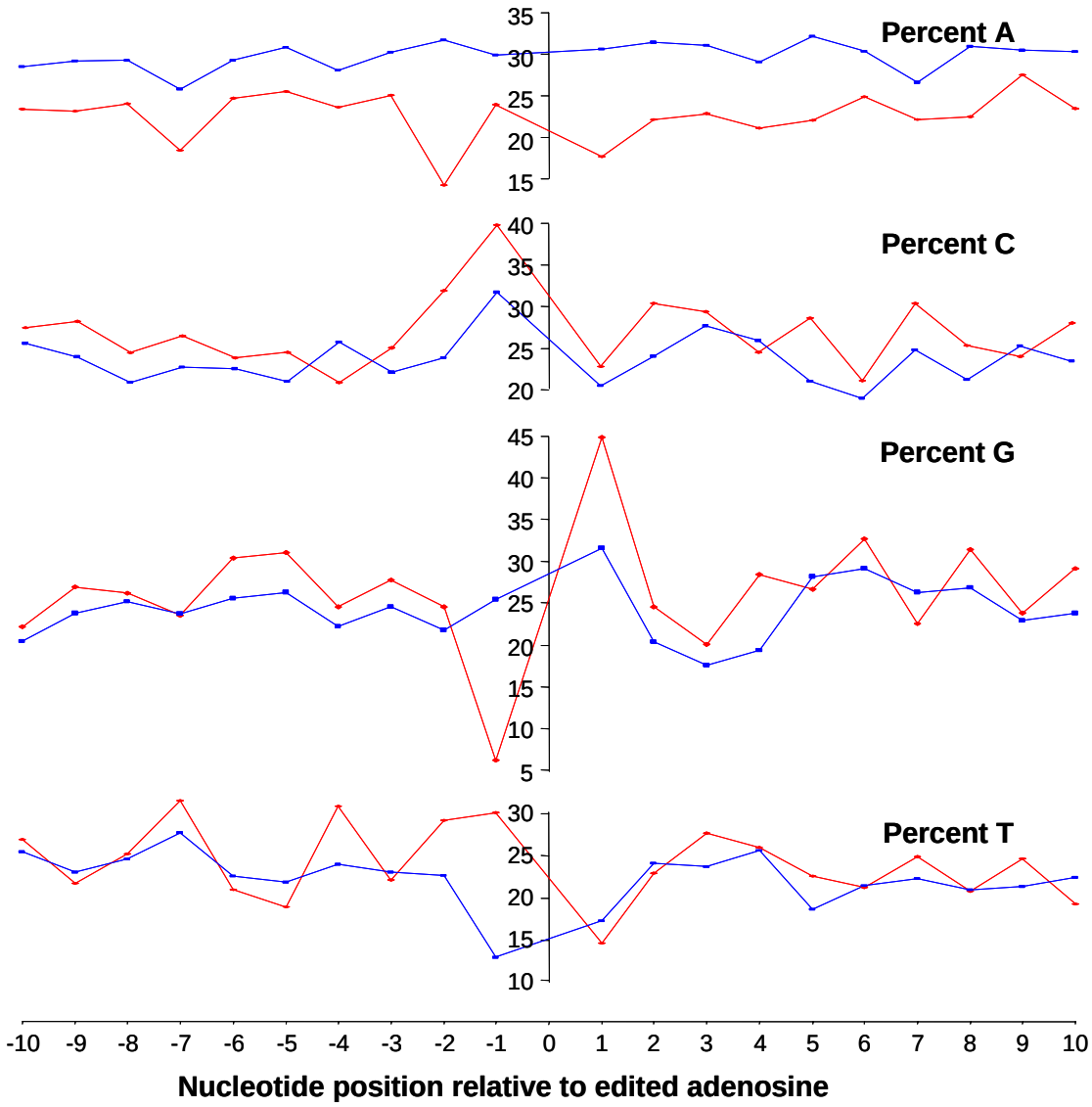


Figure 6-1 Sequence context of adenosines in edited Alu sequences. The sequence context of all edited adenosines and all unedited adenosines from all edited Alu sequences was compared. For each of the ten bases either side of edited adenosines (red lines) and unedited adenosines (blue lines) the proportion of adenosines with A, C, G or T at that position was calculated.

To further investigate the local sequence preferences of A > I editing, the tri-nucleotide composition of all edited and unedited adenosines was compared

(Figure 6-2). Consistent with the previous analysis, A > I editing was found to occur preferentially at TAG tri-nucleotides, whilst editing at any tri-nucleotide with a guanine at the 5' position was under-edited.

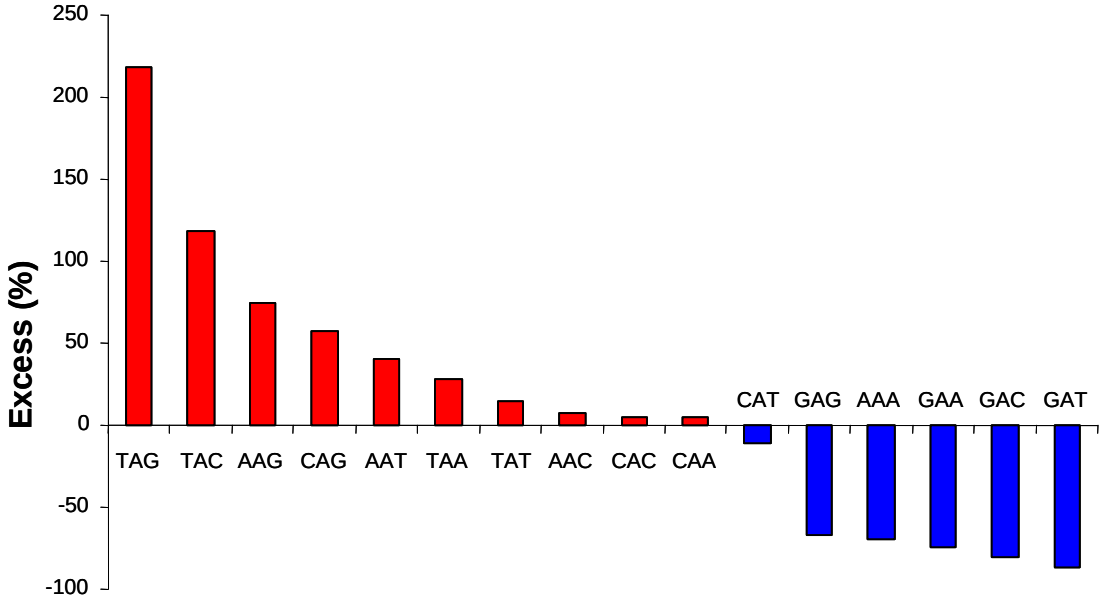


Figure 6-2 Tri-nucleotide sequence context of adenosines in edited Alu sequences. For each tri-nucleotide sequence centred on an adenosine, the number of edited and unedited adenosines present in that sequence context from cDNA clone sequences was determined. The percentage excess of edited adenosines in each tri-nucleotide was calculated. Tri-nucleotide sequences that are over-represented (red bars) or underrepresented (blue bars) at edited adenosines are indicated. The analysis was performed on DNA sequences. U replaces T in the equivalent RNA sequences.

6.2.2 BLAST alignment of inverted Alus indicates base-pairing preferences for A > I RNA editing

To investigate how the position of adenosines within matches or mismatches in dsRNA effects the likelihood of RNA editing, hypothetical dsRNA molecules were formed by BLAST alignments between edited Alus and the nearest inverted repeat copy. Mismatches and matches in each hypothetical dsRNA molecule were identified, and by superimposing the observed edits, the likelihood of A > I editing at each class of mismatch and match was assessed (Table 6-1).

Match / Mismatch	**Subset of Alus**	**Total bp**	**Edits**	**Edited %**
A:U	All Alus	5839	465	8
Matches		581	44	8
A:G	All Alus	217	13	6
Mismatches	Alus with one inverted copy	23	0	0
A:C	All Alus	1166	249	21
Mismatches	Alus with one inverted copy	113	24	21
A:A	All Alus	264	11	4
Mismatches	Alus with one inverted copy	24	1	4
Total	All Alus	25363	465	1.8
Matches	Alus with one inverted copy	2400	44	1.8
Total	All Alus	8368	273	3.3
Mismatches	Alus with one inverted copy	769	25	3.1

Table 6-1 A > I editing at different RNA base pairings. Each edited Alu was BLAST aligned to the nearest inverted Alu copy in the same transcript to form a hypothetical dsRNA molecule. The number of adenosines that are matched

(A:U) and mismatched (A:A, A:C, A:G) and the numbers of each class of match/mismatch that are edited was calculated. The calculations were performed for all edited Alus (all Alus) and separately for the subset which have only a single inverted copy in the same intron (Alus with one inverted copy). The results were from 159 alignments and 738 RNA edits (all Alus), and from 14 alignments and 69 RNA edits (Alus with one inverted copy in the same intron).

The results indicate that A > I editing at an A:C mismatch (which will generate an I:C matched base pair) is more likely than editing at other types of base pair (Table 6-1, all Alus). For example, 21% (249 / 1,166) A:C mismatches are edited, whereas 8% (465 / 5,839) A:U matches are edited (χ^2 = 190, p < 0.001).

Our previous results indicate that Alus are more likely to be edited if the nearest inverted copy is in the same intron rather than in an adjacent intron. If there is only one inverted Alu in the same intron as an edited Alu, this is most likely to be the copy with which dsRNA is formed *in vivo*. Using this subset of Alus (although smaller) is therefore probably a more accurate simulation of the *in vivo* situation. Analysis of this subset (Table 6-1, Alus with one inverted copy), similar to the analysis of all Alus, showed that A > I editing at A:C mismatches is more likely than editing at other mismatches or at A:U matches.

6.2.3 Alu multiple sequence alignments indicate base-pairing preferences for A > I RNA editing

To further investigate whether edited adenosines were likely to be at matches or mismatches within dsRNA, ClustalW was used to create multiple alignments of all edited sense Alus and all edited anti-sense Alus from the cDNA library. At each position in the multiple alignments, the proportion of edited adenosines was compared to the proportion of each nucleotide at that position. The scatter graphs (Figure 6-3) show that a high proportion of adenosines at a particular position in the alignment (which would be uridine in the anti-sense strand forming A:U matches in dsRNA) is correlated with a low frequency of editing, whilst a high proportion of guanosines at a particular position in the alignment (which would be cytidine in the anti-sense strand forming A:C mismatches in dsRNA) is correlated with a high frequency of editing (Figure 6-3, %G in consensus).

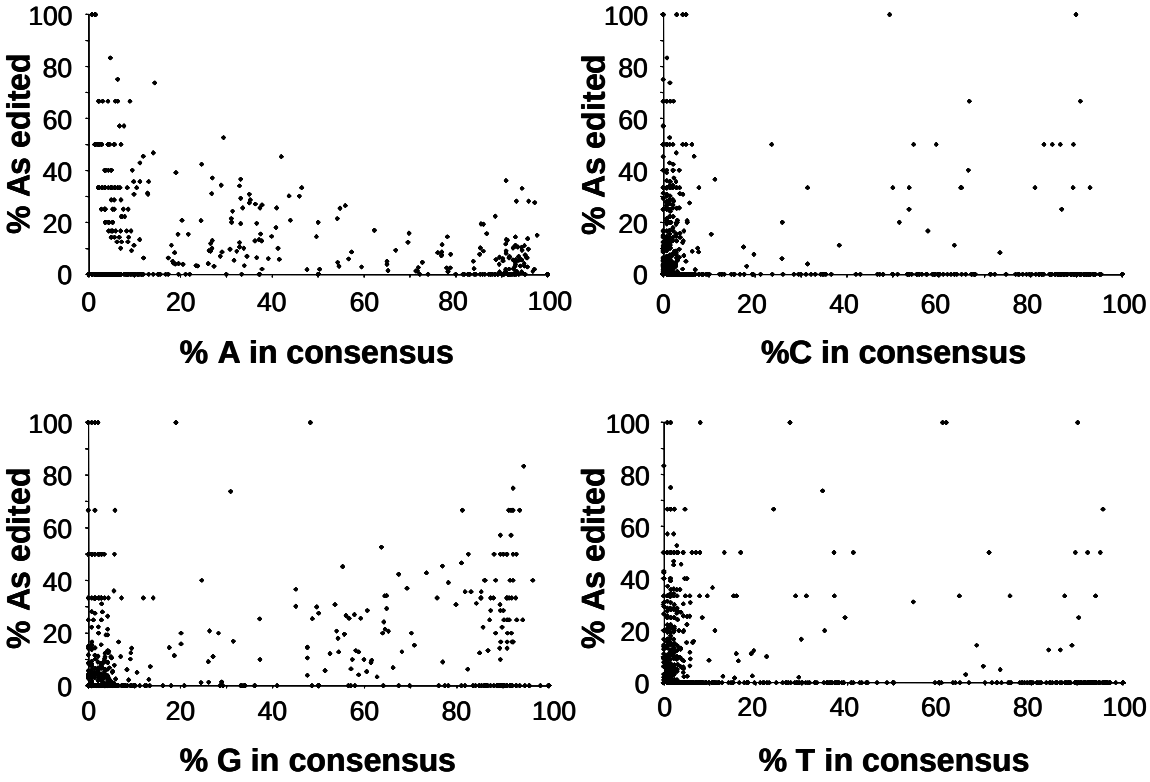


Figure 6-3 Effect of sequence composition on the likelihood of RNA editing. A multiple alignment of all edited Alu sequences was prepared using CLUSTALW. At each position in the alignment, the proportion of edited adenosines was calculated from the number of sequenced edited adenosines and the total number of sequenced adenosines. The sequence composition at each position was calculated from all Alus. For each position in the alignment, the proportion of edited adenosines is compared to the proportion of A, C, G or T at that position in the consensus.

Finally, the effect of RNA editing on base pairing was evaluated using the alignments of all edited Alus to all other edited Alus. The average nucleotide composition at 1,539 edited adenosines from 301 multiply aligned Alus was determined. The results indicate that 57% of editing reactions create a

mismatch (I:U) from a match (A:U), 28% create a match (I:C) from a mismatch (A:C) and 15% create a mismatch from a mismatch. Therefore, on balance, the effect of A > I editing would be predicted to increase the number of mismatches in Alu dsRNAs. This is consistent with the previous analyses.

6.2.4 A > I RNA editing results in a marginal decrease in base pairing in predicted dsRNA

The results of these analyses indicate that A>I editing may result in matching base pairs being formed from mismatched base pairs (A:C > I:C), mismatches being formed from matches (A:U > I:U) and mismatches from mismatches (A:A > I:A and A:G > I:G). Therefore, the overall effect of RNA editing on the balance of matched base pairing in hypothetical dsRNA molecules was further investigated.

The effect of RNA editing on base pairing in dsRNAs was evaluated from the BLAST alignments of all edited Alus to their nearest inverted copy (Table 6-1, all Alus). In these simulations, 63% (465 / 738) A > I edits convert A:U matches to I:U mismatches, 34% (249 / 738) convert A:C mismatches to I:C matches, and 3% (24 / 738) convert A:A or A:G mismatches to I:A or I:G mismatches respectively. The overall effect is a net increase of 216 mismatches. Taking into account all matches and mismatches in the alignments, A > I editing results in a net increase in mismatches of approximately 2.6% (from 8,368 to 8,584) resulting, on balance, in an additional 0.6% (216 / 33,731) of bases in dsRNAs becoming mismatched after editing. Since these analyses evaluate editing of only one strand of RNA

in the double stranded molecule, and A > I editing targets both strands, it is likely that the number of additional mismatched base pairs is twice this estimate, i.e. 1.2%. It should be noted, however, that in a minority of individual simulated dsRNA molecules there was on balance an apparent increase in matches (data not shown).

Next, the effect of RNA editing on hypothetical dsRNA molecules formed by BLAST alignment of repeats that have only a single inverted copy within the same intron was examined (Table 6-1, Alus with one inverted copy). Of 69 A > I edits in this set, 64% (44 / 69) convert A:U matches to I:U mismatches, 35% (24 / 69) convert A:C mismatches to I:C matches, and 1% (1 / 69) converts an A:A mismatches to an I:A mismatch. Following editing, there is a 2.5% (from 796 to 816) increase in mismatches resulting, on balance, in an additional 0.6% of bases (20 out of 3,196) becoming mismatched after editing (1.2% taking into account both strands). However, one out of the 14 dsRNA molecules included in this analysis still would appear slightly better matched after editing (six matches to mismatches and seven mismatches to matches, data not shown).

6.2.5 Distribution of A > I editing sites in the Alu consensus sequence

To search for patterns in the distribution of A > I edits in Alu sequences, the multiple sequence alignments of edited sense and anti-sense Alu sequence were used to derive a consensus sequence. At every adenosine in the consensus sequence, the frequency of RNA editing was determined (Figure 6-4).

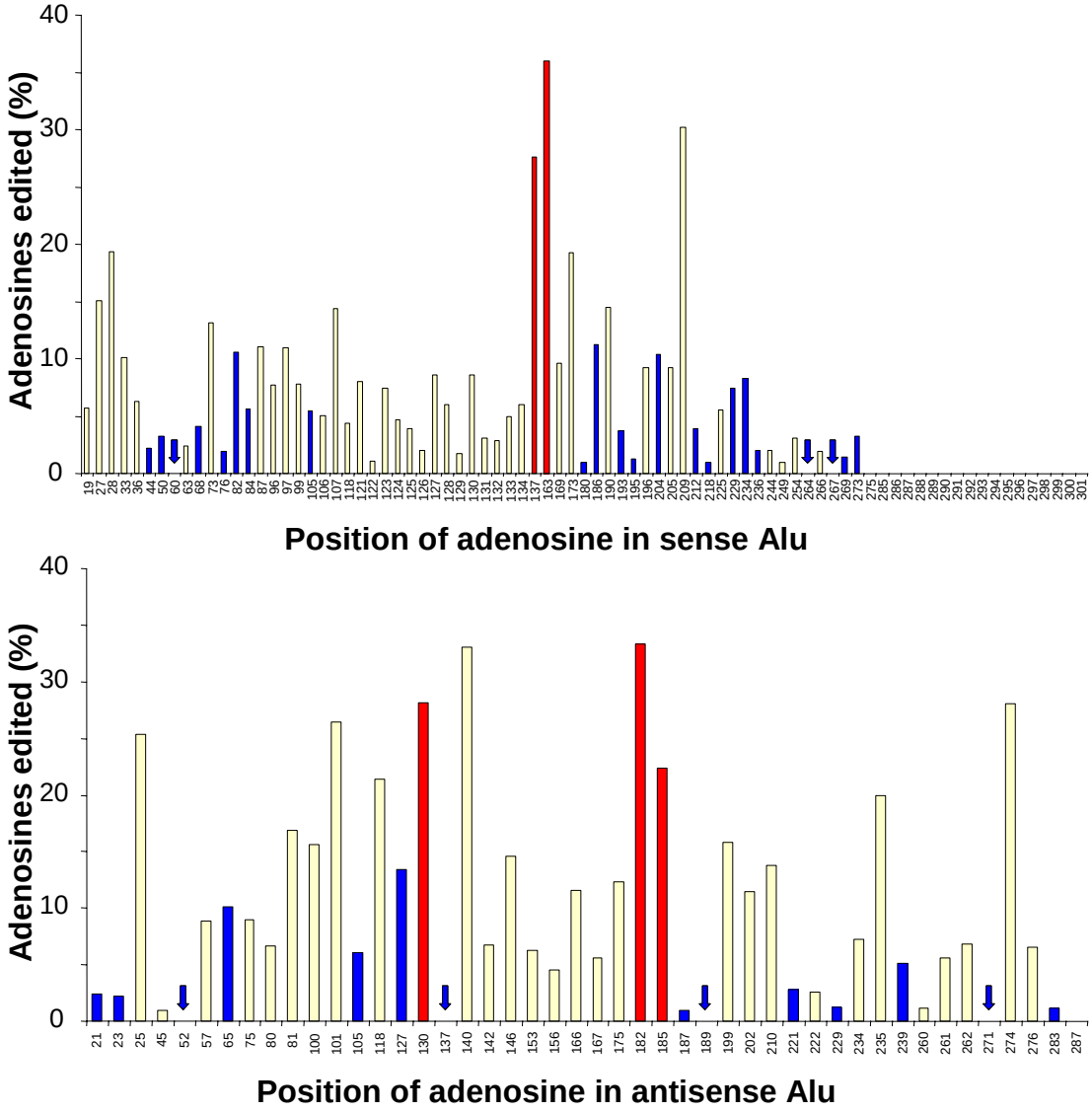


Figure 6-4 Frequency of editing at adenosines in edited sense and anti-sense Alus. For each adenosine in the consensus sequence, the proportion of adenosines which were edited was calculated from all sequenced adenosines. All adenosines within TAG tri-nucleotides (red bars), and GAX (X = A,C,G or T) tri-nucleotides (blue bars or blue arrows where editing is absent) are highlighted.

Overall, 9 % (774 / 8,893) adenosines from 149 aligned edited sense Alu sequences and 12% (706 / 6,057) adenosines from 152 aligned edited anti-sense sequences were edited. The sense Alu consensus sequence contains more adenosines than the anti-sense consensus sequence (86 and 46 respectively). However, the 23 adenosines in the FRAM associated poly-A tail of the sense Alu were devoid of editing, and account for the small difference in editing between the sense and anti-sense consensus sequences (excluding the poly-(A) tail, 11% (774 / 7,644) adenosines from sense Alus were edited).

With the exception of the FRAM associated poly-(A) tail of the sense Alu, edited adenosines are widely distributed along both the sense and anti-sense Alu consensus sequences. The frequency of editing at individual adenosines varies substantially, but generally can be explained by the local sequence context and base-pairing preferences determined above. For example, two of the most frequently edited adenosines in the sense Alu consensus, and three of the most frequently edited adenosines in the anti-sense Alu consensus are at preferentially edited TAG tri-nucleotides (Figure 6-4 red bars). Conversely, many of the least edited adenosines are in GAX tri-nucleotides (Figure 6-4 blue bars).

It was previously shown that FRAM monomers were more frequently edited than FLAM monomers (see Chapter 5, Table 5-1). From these analyses, 9% (637 / 6,082) adenosines in FLAM and 10% (843 / 7,388) adenosines in FRAM components of Alu sequences were edited. There is therefore no

evidence of differential editing of the FLAM and FRAM derived components of complete Alus.

6.3 DISCUSSION

6.3.1 Local sequence preferences of Alu A > I editing

The results indicate that at the immediately 5' position to an edited adenosine there is a relative deficit of guanine and a compensatory increase in uridine (thymidine) and cytidine, and at the immediately 3' position to an edited adenosine there is a relative excess of guanosine with compensatory decrease of all other nucleotides, mainly adenosine. These results are corroborated by two recent analyses of A > I editing of Alu sequences in which similar sequence preferences were observed (Levanon et al., 2004, Kim et al., 2004). Analysis of the tri-nucleotide sequence preferences of A > I editing indicate an over-representation of UAG and an under-representation of all GAX tri-nucleotides at edited adenosines compared with unedited adenosines. These results are consistent with the 5' and 3' neighbouring nucleotide preferences, and in agreement with similar analyses by others (Kim et al., 2004).

The 5' neighbour preferences of edited adenosines identified in these analyses are consistent with the previously reported patterns associated with ADAR1 and ADAR2 editing. The 3' neighbour preference matches the observed preferences of ADAR2, but not of ADAR1 for which no 3' preference was observed (Polson and Bass, 1994). This may reflect a predominant role of ADAR2 in editing of brain mRNA. Alternatively, ADAR1 may have *in vivo* A

> I editing sequence preferences that were not detected by the previous *in vitro* analyses. ADAR1 and ADAR2 are both expressed in the brain (O'Connell et al., 1995, Melcher et al., 1996a), and have overlapping specificities (Lehmann and Bass, 2000). Therefore the edited Alu sequences in these analyses may represent the combined output of A > I editing by both ADAR1 and ADAR2.

6.3.2 Distribution of A > I edits in the Alu consensus sequence

A > I editing does not occur uniformly at all adenosines in the forward or reverse Alu consensus sequences. Instead, there are some positions at which editing is overrepresented, and others at which editing is underrepresented. Generally these positions are consistent with the sequence preferences or base-pairing preferences established in these analyses. However, there appears to be negligible A > I editing of the FRAM associated poly-(A) tail. It is possible that the high degree of variation in the lengths of Alu poly-(A) tails results in only a small proportion of adenosines within the Alu poly-(A) tail being matched in RNA duplexes. Furthermore, A:U base pairs are less stable than G:C base pairs, such that extended poly-(A):poly-(U) duplexes may be less stable substrates of ADARs. In contrast to the FRAM poly-(A) tail which is at the ends of the duplex, the FLAM poly-(A) sequence is internal and clamped by more stable dsRNA either side. There is evidence of A > I editing (although weakly) at all positions of the internal FLAM associated sequence.

These results are in general agreement with other analyses of the positions of A > I editing sites within Alus (Levanon et al., 2004, Kim et al., 2004). Both

report the hotspots of A > I editing (for example at adenosines 137 and 163 in the sense Alu), and under-editing at several GAX trinucleotides, as well as virtual absence of editing of the Alu poly-(A) tail.

6.3.3 Base-pairing preferences of Alu A > I editing

To evaluate base pairing preferences of A > I RNA editing, dsRNA molecules were simulated by BLAST alignment of edited Alus to the nearest inverted Alu copy in the same transcript. BLAST is not generally regarded as an algorithm for RNA structural prediction. However, comparison with MFOLD (which is an RNA secondary structure prediction algorithm), revealed that predicted base-pairing of edited adenosines was identical using the two methods. Therefore BLAST was considered suitable for these analyses as it allowed a more rapid and easily interpretable analysis of all edited Alu sequences than was possible using MFOLD, with no apparent loss in the accuracy of the predictions.

For the BLAST simulations, it was assumed that the dsRNA which was the *in vivo* substrate for A > I editing enzymes was formed between the edited Alu and the closest inverted copy. Although this assumption is unlikely to be correct for all sequences, the results of Chapter 5 indicate that it is often likely to be the case. The advantage of invoking this assumption is that it allows use of most available information. A second series of BLAST analyses were performed on a subset of edited Alus with only a single inverted copy in the same intron. Whilst these represent a fraction of the available information, the

results indicate that these are likely to be more accurate simulations of the *in vivo* substrate.

DsRNA formed between a sequence and an inverted copy usually includes a number of unpaired bases. In addition to the BLAST simulations of dsRNA, alignments of all edited Alus to all other edited Alus were used to investigate whether editing is equally likely at mismatches and matches. In these analyses, the hypothetical dsRNA molecules generated are dependent on the parameters used to generate the alignments and are unlikely to completely replicate the biological conditions present *in vivo*. Moreover, the results only provide information on editing of one strand of the dsRNA molecule. Editing on the other strand (probably at an equivalent rate) is likely, but cannot be evaluated from the data generated in this survey. Although each of these simulations has its deficiencies, their results are very similar and taken together they probably provide a realistic representation of dsRNA formation.

The likelihood of editing at A:C mismatches in dsRNA appears to be higher than at A:G or A:A mismatches or at A:U matches. Since an A:C mismatch is converted into an I:C base pair by A > I editing, the enzymatic configuration of the editing machinery seems to favour the creation of fully matched dsRNA. These observations are consistent with previous *in vitro* experiments which indicate that editing at A:C mismatches is more efficient than at A:U matches or other mismatches (Wong et al., 2001).

6.3.4 The overall effect of A > I editing on base-pairing in dsRNA

Although adenosines at A:C mismatches are more efficiently edited than adenosines at A:U matches, the frequency of A:U matches in most RNA duplexes formed by inverted copies is much higher than the number of A:C mismatches. Therefore, despite the higher likelihood of editing at A:C mismatches, the overall effect of RNA editing may be to increase the number of mismatches in dsRNA molecules, albeit by a relatively modest amount (in edited sequences, an additional 1-2% of base pairs become mismatched after editing). This appears to be the prediction of all three types of analysis. The role and functional consequences of this are considered in the General Discussion.

The conclusions of this chapter are broadly concordant with those from a recent study of the base pairing preferences of A > I edits within Alus (Levanon et al., 2004), in which A > I edits were found more frequently than expected at A:C mismatches, but were predominantly at A:U matches.

7 GENERAL DISCUSSION

In this thesis, a survey of the types and targets of RNA editing in the human brain is presented. Approximately 1 in 1,700 nucleotides in the human brain RNA sample used were subject to A > I editing. By contrast, RNA editing by mechanisms other than A > I is a rare event in the human brain. The majority of A > I edits are in transcribed intronic and intergenic Alu repeats, and are associated with dsRNA formation with inverted Alus in the same transcript. Within edited Alu sequences, A > I editing occurs preferentially at adenosines with a deficit of guanine at the immediately 5' adjacent nucleotide, and an increase in guanine at the immediately 3' adjacent nucleotide. Editing is also more efficient at A:C mismatches than at other mismatches or A:U matches in simulations of dsRNA. The results suggest that the effect of A > I editing is to increase the number of mismatches in dsRNA molecules, albeit by a relatively modest amount (in edited sequences, an additional 1-2% of base pairs become mismatched after editing).

7.1 FUTURE CHALLENGES

We cannot currently rule out the existence of non A > I RNA edits in human brain RNA. The scarcity of such edits means that evaluation of additional sequence variants from a more extensive survey of the type described in this thesis or by a targeted approach such as RT-PCR product sequencing will be necessary for their identification. A more extensive survey would allow the frequency with which such RNA edits occur in the human brain to be determined more accurately. It is also possible that human brain transcripts harbour additional coding RNA edits. A more exhaustive investigation directed

at coding sequences is warranted to detect rare, functionally important coding edits. This could be achieved by sequencing from cDNA clones derived from cytoplasmic RNA, or by sequencing RT-PCR products designed to amplify specifically from coding sequences. Experimental analysis of the exonic Alu sequences with an inverted copy in an adjacent intron identified in this survey may also reveal novel coding A > I edits.

This survey was performed on poly (A)+ RNA. Further work is required to investigate the extent to which non-coding unadenylated RNAs are subject to RNA editing. The function of many non-coding RNAs is dependent on base pairing and local dsRNA structures, and may plausibly be regulated by RNA editing. The presence of known A > I edits in tRNAs (Maas et al., 1999) and miRNAs (Luciano et al., 2004) are further indications that a survey of non-coding RNAs is warranted.

Several analyses indicate that A > I editing varies widely between different tissues (Paul and Bass, 1998, Levanon et al., 2004, Kim et al., 2004). It will be interesting to carry out a more exhaustive analysis of the patterns of A > I editing in different tissues, and to look for correlation with the expression levels of the different ADAR editing enzymes in these tissues. More extensive evaluation of the patterns of A > I editing, and ADAR expression in diseased tissues is also warranted, as aberrant A > I editing has previously been linked with tumour progression in gliomas (Maas et al., 2001b), and a number of neurological disorders including amyotrophic lateral sclerosis (ALS) and epilepsy (Kawahara et al., 2004, Kortenbruck et al., 2001). As C > U RNA

editing of ApoB mRNA occurs specifically expressed in the small intestine (Teng et al., 1993), it is possible that additional tissue specific RNA editing activities may exist. This could be assessed by performing a survey of RNA editing in other tissues similar to the one described in this thesis.

In our analyses of A > I editing from total cDNA, we found that the extent of A > I editing varied between different adenosines in the same transcript. This suggests that within the total population of transcripts, individual molecules are differently edited. Cloning and sequencing of multiple individual cDNAs from the same transcript will be required to better understand the patterns of A > I editing at the level of individual RNA molecules.

Currently, the extent to which each of the ADAR editing enzymes contributes to the pattern of A > I edits observed in this survey is unknown. One way of investigating this further would be to use RNA interference to selectively down-regulate ADAR1 or ADAR2 in cultured cells in order to investigate the contribution of each enzyme to the pattern of A > I edits identified in Alu sequences from this survey. This type of analysis may also help elucidate the functional consequences of Alu A > I editing.

7.2 THE FUNCTION OF A > I EDITING

The functions of RNA editing in mammals are still being investigated. On the basis of previously reported evidence a small number of edits alter the coding sequence and activities of certain proteins. An additional small number have direct effects on mRNA splicing, by altering transcript sequence at consensus

splice sites. However, the function of the large majority of RNA edits, which are within intronic or intergenic high copy number repeats, is not known. One possibility is that they have no function at all. They may simply be the collateral damage of an enzyme system which uses dsRNA as a template and which therefore generates large numbers of edits of high copy number repeat elements. According to this hypothesis, the important functional consequences for the cell reside in the small number of coding, splice site and other functional edits. This would be a system of remarkable metabolic profligacy since fewer than 1% (and probably fewer than 0.1%) edits would be functional.

Alternatively, editing of intronic and intergenic high copy number repeats may have a function. One possibility is that RNA editing inhibits non-specific cellular responses to dsRNA which are deleterious to cellular function. These potentially include activation of 2',5'-oligoA synthetase / RNaseL resulting in single stranded RNA degradation, activation of the dsRNA dependent Protein kinase (PKR) resulting in suppression of protein synthesis and activation of the interferon response leading to apoptosis (Kumar and Carmichael, 1998).

Another possibility is that A > I editing prevents gene silencing via the RNAi pathway (Mello and Conte, 2004). It is conceivable that endogenously transcribed dsRNA formed by pairs of inverted Alu repeats are substrates of the dsRNA ribonucleases Dicer, giving rise to Alu derived short interfering RNAs (siRNAs). Given the abundance of Alu sequences in the transcriptome, the number of potential binding sites of Alu siRNAs would be huge and could

have catastrophic effects on the cell. Previous studies in *C. elegans* support the notion that RNA editing abrogates RNAi dependent toxic effects of endogenous dsRNAs (Tonkin and Bass, 2003). An increased number of mismatches generated by editing of dsRNA molecules may limit their deleterious RNAi dependent effects by destabilising the hairpin, by reducing the efficiency of processing (perhaps by retention in the nucleus (Zhang and Carmichael, 2001)), by generating products which are less effective in mediating the effects of RNAi, (for example, by interrupting long, perfectly matched stretches of base pairing) or by other, currently obscure, mechanisms. Our data is broadly consistent with this model, as A > I editing results in an overall increase in the number of mismatches in dsRNA.

An alternative explanation is that dsRNAs formed between inverted Alu repeats are not toxic to the cell, but play a functional role that is regulated by A > I editing. Although closely spaced inverted repeats are apparently toxic to the cell and are underrepresented in the genome (Stenger et al., 2001), our results indicate that nearly 65% of all transcripts have at least one intron with a pair of inverted Alus, and therefore are potential A > I RNA editing substrates. Given that they have accumulated to such a high level in the human genome, it is possible that not all dsRNAs formed by inverted Alu repeats are subject to negative selection.

No function has been ascribed to transcribed inverted repeats in mammals. One possibility, as suggested above, is that they are processed into short RNAs and act in a manner analogous to siRNAs or miRNAs. Rather than

having a toxic effect on cell function, these may be functional molecules which regulate the expression of target transcripts. The role of RNA editing may be to regulate rather than to prevent the entry of Alu derived dsRNA into this pathway.

Interestingly, there were several edited sequences for which, in the simulations, the effect of A > I editing appeared to increase base pairing in dsRNA. This would apparently lead to a small number of dsRNAs becoming more stable and therefore, presumably better substrates for RNAi. Also, A > I RNA editing of a miRNA precursor was recently demonstrated, and predicted to have an effect on the biogenesis and function of the encoded miRNA (Luciano et al., 2004). These results are consistent with a regulatory rather than a preventative role for A > I editing. The use of Alu sequences in such a way may account for their toleration in high abundance in the human genome and in particular their accumulation in gene rich sequences.

There is evidence that A > I RNA editing influences splicing by competing with splicing machinery for RNA at the intron exon junction (Bratt and Ohman, 2003, Flomen et al., 2004), by editing and destroying a branch site adenosine (Beghini et al., 2000) or by creating splice sites (Rueter et al., 1999). In the latter case, a novel splice site is created by ADAR2 editing of an AA dinucleotide in an intronic Alu sequence of its own transcript, to an AG splice site acceptor. In the absence of RNA editing, Alu sequences have been shown to generate splice variants, by virtue of both splicing donor and acceptor consensus sequences within transcribed intronic Alu sequences

(Sorek et al., 2002). The large number of edited intronic Alu sequences identified in this survey includes AA > AI and AG > IG edits. It is therefore possible that regulation of splicing by RNA editing of intronic Alu sequences is widespread. However, none of the edited Alu sequences identified in this survey were spliced, and given that intronic Alu RNA editing substrates are widespread (>60% of all transcripts contain an intronic inverted Alu repeat), it is difficult to envisage specific regulation of splicing through RNA editing.

Whatever the function of A > I editing, it is necessary to account for the observation that the extent of A > I editing and the expression levels of ADAR editing enzymes varies between tissues. It is conceivable that the requirement for RNA editing in a particular tissue is linked to the fate of endogenous dsRNA or the product of dsRNA metabolism in that tissue. For example, cells in which endogenous dsRNA can have deleterious consequences (perhaps by eliciting an RNAi response), may require RNA editing to prevent such a response occurring. Conversely, RNA editing of dsRNA may not be as important in tissues in which endogenous dsRNAs do not have such an effect.

Finally, the association of A > I editing with high copy repeats suggests that A > I editing may function in the biology of retrotransposons. For example, it is possible that A > I editing may lead to the mutation and inactivation of transcribed Alus to prevent their re-insertion into the genome. However, active Alus tend to be transcribed under the control of their own promoters, rather than as components of other transcripts, and therefore would not necessarily be expected to form the types of dsRNA molecules that were found to be

edited in this survey. The potential for single Alu repeats to form dsRNA structures that are substrates for A > I editing is unclear, but seems to be low from our data. If active Alus are subject to modification by A > I editing prior to retrotransposition, evidence for this should be present in the sequence of Alus in the human genome, and may be detectable among other causes of variation such as error prone reverse transcription, and conventional DNA mutation.

There are clearly many interesting unanswered questions regarding the function of RNA editing in human cells. This thesis describes a survey of the patterns of RNA editing in the human brain, and forms a basis for future analyses.

8 REFERENCES

Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J., Baldwin, J., Linton, L. and Lander, E. S. (2000) *Nature,* **407,** 513-6.

Anant, S. and Davidson, N. O. (2003) *Trends Mol Med,* **9,** 147-52.

Anant, S., MacGinnitie, A. J. and Davidson, N. O. (1995) *J Biol Chem,* **270,** 14762-7.

Anant, S., Mukhopadhyay, D., Sankaranand, V., Kennedy, S., Henderson, J. O. and Davidson, N. O. (2001) *Am J Physiol Cell Physiol,* **281,** C1904-16.

Avner, P. and Heard, E. (2001) *Nat Rev Genet,* **2,** 59-67.

Bachellerie, J. P., Cavaille, J. and Huttenhofer, A. (2002) *Biochimie,* **84,** 775-90.

Backus, J. W., Schock, D. and Smith, H. C. (1994) *Biochim Biophys Acta,* **1219,** 1-14.

Bartel, D. P. (2004) *Cell,* **116,** 281-97.

Bass, B. L. (2000) *Cell,* **101,** 235-8.

Bass, B. L. (2002) *Annu Rev Biochem,* **71,** 817-46.

Bass, B. L. and Weintraub, H. (1988) *Cell,* **55,** 1089-98.

Batzer, M. A., Deininger, P. L., Hellmann-Blumberg, U., Jurka, J., Labuda, D., Rubin, C. M., Schmid, C. W., Zietkiewicz, E. and Zuckerkandl, E. (1996) *J Mol Evol,* **42,** 3-6.

Beghini, A., Ripamonti, C. B., Peterlongo, P., Roversi, G., Cairoli, R., Morra, E. and Larizza, L. (2000) *Hum Mol Genet,* **9,** 2297-304.

Begum, N. A., Kinoshita, K., Kakazu, N., Muramatsu, M., Nagaoka, H., Shinkura, R., Biniszkiewicz, D., Boyer, L. A., Jaenisch, R. and Honjo, T. (2004) *Science,* **305,** 1160-3.

Benne, R., Van den Burg, J., Brakenhoff, J. P., Sloof, P., Van Boom, J. H. and Tromp, M. C. (1986) *Cell,* **46,** 819-26.

Bhalla, T., Rosenthal, J. J., Holmgren, M. and Reenan, R. (2004) *Nat Struct Mol Biol*.

Blanc, V., Kennedy, S. and Davidson, N. O. (2003) *J Biol Chem,* **278,** 41198-204.

Blanc, V., Navaratnam, N., Henderson, J. O., Anant, S., Kennedy, S., Jarmuz, A., Scott, J. and Davidson, N. O. (2001) *J Biol Chem,* **276,** 10272-83.

Blum, B., Bakalara, N. and Simpson, L. (1990) *Cell,* **60,** 189-98.

Bock, R. (2000) In *RNA editing*(Ed, B.L, B.) Oxford University Press, Oxford.

Bock, R., Hermann, M. and Kossel, H. (1996) *Embo J,* **15,** 5052-9.

Bock, R. and Koop, H. U. (1997) *Embo J,* **16,** 3282-8.

Bratt, E. and Ohman, M. (2003) *Rna,* **9,** 309-18.

Brenner, S. (1961) *Cold Spring Harb Symp Quant Biol,* **26,** 101-10.

Brouha, B., Schustak, J., Badge, R. M., Lutz-Prigge, S., Farley, A. H., Moran, J. V. and Kazazian, H. H., Jr. (2003) *Proc Natl Acad Sci U S A,* **100,** 5280-5.

Burns, C. M., Chu, H., Rueter, S. M., Hutchinson, L. K., Canton, H., Sanders-Bush, E. and Emeson, R. B. (1997) *Nature,* **387,** 303-8.

Cavaille, J., Buiting, K., Kiefmann, M., Lalande, M., Brannan, C. I., Horsthemke, B., Bachellerie, J. P., Brosius, J. and Huttenhofer, A. (2000) *Proc Natl Acad Sci U S A,* **97,** 14311-6.

Chen, C. X., Cho, D. S., Wang, Q., Lai, F., Carter, K. C. and Nishikura, K. (2000) *Rna,* **6,** 755-67.
Chen, C. Z., Li, L., Lodish, H. F. and Bartel, D. P. (2004) *Science,* **303,** 83-6.
Chen, S. H., Habib, G., Yang, C. Y., Gu, Z. W., Lee, B. R., Weng, S. A., Silberman, S. R., Cai, S. J., Deslypere, J. P., Rosseneu, M. and et al. (1987) *Science,* **238,** 363-6.
Chen, Z., Eggerman, T. L. and Patterson, A. P. (2001) *Biochem J,* **357,** 661-72.
Cheng, Y. W., Visomirski-Robic, L. M. and Gott, J. M. (2001) *Embo J,* **20,** 1405-14.
Cho, D. S., Yang, W., Lee, J. T., Shiekhattar, R., Murray, J. M. and Nishikura, K. (2003) *J Biol Chem,* **278,** 17093-102.
Christiansen, T. and Torkington, N. (2003) *Perl Cookbook,* O'Reilly.
Clement, J. Q., Maiti, S. and Wilkinson, M. F. (2001) *J Biol Chem,* **276,** 16919-30.
Colgan, D. F. and Manley, J. L. (1997) *Genes Dev,* **11,** 2755-66.
Crick, F. H. (1958) *Symp Soc Exp Biol,* **12,** 138-63.
Dance, G. S., Beemiller, P., Yang, Y., Mater, D. V., Mian, I. S. and Smith, H. C. (2001) *Nucleic Acids Res,* **29,** 1772-80.
Dawson, T. R., Sansam, C. L. and Emeson, R. B. (2004) *J Biol Chem,* **279,** 4941-51.
Decatur, W. A. and Fournier, M. J. (2002) *Trends Biochem Sci,* **27,** 344-51.
Deininger, P. L. and Batzer, M. A. (2002) *Genome Res,* **12,** 1455-65.
Desterro, J. M., Keegan, L. P., Lafarga, M., Berciano, M. T., O'Connell, M. and Carmo-Fonseca, M. (2003) *J Cell Sci,* **116,** 1805-18.
Dewannieux, M., Esnault, C. and Heidmann, T. (2003) *Nat Genet,* **35,** 41-8.
Eckmann, C. R., Neunteufl, A., Pfaffstetter, L. and Jantsch, M. F. (2001) *Mol Biol Cell,* **12,** 1911-24.
Eddy, S. R. (2001) *Nat Rev Genet,* **2,** 919-29.
Fatica, A. and Tollervey, D. (2002) *Curr Opin Cell Biol,* **14,** 313-8.
Fatica, A. and Tollervey, D. (2003) *Nat Struct Biol,* **10,** 237-9.
Flomen, R., Knight, J., Sham, P., Kerwin, R. and Makoff, A. (2004) *Nucleic Acids Res,* **32,** 2113-22.
Gallo, A., Keegan, L. P., Ring, G. M. and O'Connell, M. A. (2003) *Embo J,* **22,** 3421-30.
Garriga, G., Lambowitz, A. M., Inoue, T. and Cech, T. R. (1986) *Nature,* **322,** 86-9.
George, C. X. and Samuel, C. E. (1999) *Proc Natl Acad Sci U S A,* **96,** 4621-6.
Gerber, A., Grosjean, H., Melcher, T. and Keller, W. (1998) *Embo J,* **17,** 4780-9.
Gerber, A., O'Connell, M. A. and Keller, W. (1997) *Rna,* **3,** 453-63.
Gerber, A. P. and Keller, W. (1999) *Science,* **286,** 1146-9.
Gesteland, R. F. (1999) *The RNA world,* New York Cold Spring Harbor Laboratory Press 1999.
Giege, P. and Brennicke, A. (1999) *Proc Natl Acad Sci U S A,* **96,** 15324-9.
Gitlin, L. and Andino, R. (2003) *J Virol,* **77,** 7159-65.
Gott, J. M. (2000) In *RNA editing*(Ed, Bass, B. L.) Oxford University Press, Oxford, pp. 20-36.

Gott, J. M., Visomirski, L. M. and Hunter, J. L. (1993) *J Biol Chem,* **268,** 25483-6.
Greger, I. H., Khatri, L., Kong, X. and Ziff, E. B. (2003) *Neuron,* **40,** 763-74.
Grosjean, H., Auxilien, S., Constantinesco, F., Simon, C., Corda, Y., Becker, H. F., Foiret, D., Morin, A., Jin, Y. X., Fournier, M. and Fourrey, J. L. (1996) *Biochimie,* **78,** 488-501.
Gunning, K. B., Cohn, S. L., Tomlinson, G. E., Strong, L. C. and Huff, V. (1996) *Oncogene,* **13,** 1179-85.
Gurevich, I., Tamir, H., Arango, V., Dwork, A. J., Mann, J. J. and Schmauss, C. (2002) *Neuron,* **34,** 349-56.
Hanrahan, C. J., Palladino, M. J., Ganetzky, B. and Reenan, R. A. (2000) *Genetics,* **155,** 1149-60.
Hartner, J. C., Schmittwolf, C., Kispert, A., Muller, A. M., Higuchi, M. and Seeburg, P. H. (2004) *J Biol Chem,* **279,** 4894-902.
Haussmann, S., Garcin, D. and Kolakofsky, D. (2001) In *RNA Editing*(Ed, Bass, B. L.) Oxford University Press, pp. 139-59.
Herbert, A., Alfken, J., Kim, Y. G., Mian, I. S., Nishikura, K. and Rich, A. (1997) *Proc Natl Acad Sci U S A,* **94,** 8421-6.
Hersberger, M. and Innerarity, T. L. (1998) *J Biol Chem,* **273,** 9435-42.
Hersberger, M., Patarroyo-White, S., Arnold, K. S. and Innerarity, T. L. (1999) *J Biol Chem,* **274,** 34590-7.
Higuchi, M., Maas, S., Single, F. N., Hartner, J., Rozov, A., Burnashev, N., Feldmeyer, D., Sprengel, R. and Seeburg, P. H. (2000) *Nature,* **406,** 78-81.
Higuchi, M., Single, F. N., Kohler, M., Sommer, B., Sprengel, R. and Seeburg, P. H. (1993) *Cell,* **75,** 1361-70.
Hirotsune, S., Yoshida, N., Chen, A., Garrett, L., Sugiyama, F., Takahashi, S., Yagami, K., Wynshaw-Boris, A. and Yoshiki, A. (2003) *Nature,* **423,** 91-6.
Hoagland, M. (2004) *Nature,* **431,** 249.
Honjo, T., Kinoshita, K. and Muramatsu, M. (2002) *Annu Rev Immunol,* **20,** 165-96.
Hoopengardner, B., Bhalla, T., Staber, C. and Reenan, R. (2003) *Science,* **301,** 832-6.
Hopper, A. K. and Phizicky, E. M. (2003) *Genes Dev,* **17,** 162-80.
Hough, R. F. and Bass, B. L. (1997) *Rna,* **3,** 356-70.
Huttenhofer, A., Kiefmann, M., Meier-Ewert, S., O'Brien, J., Lehrach, H., Bachellerie, J. P. and Brosius, J. (2001) *Embo J,* **20,** 2943-53.
Jarmuz, A., Chester, A., Bayliss, J., Gisbourne, J., Dunham, I., Scott, J. and Navaratnam, N. (2002) *Genomics,* **79,** 285-96.
Ji, P., Diederichs, S., Wang, W., Boing, S., Metzger, R., Schneider, P. M., Tidow, N., Brandt, B., Buerger, H., Bulk, E., Thomas, M., Berdel, W. E., Serve, H. and Muller-Tidow, C. (2003) *Oncogene,* **22,** 8031-41.
Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R. and Shoemaker, D. D. (2003) *Science,* **302,** 2141-4.
Joyce, G. F. (2002) *Nature,* **418,** 214-21.
Kallman, A. M., Sahlin, M. and Ohman, M. (2003) *Nucleic Acids Res,* **31,** 4874-81.

Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P. and Gingeras, T. R. (2002) *Science,* **296,** 916-9.
Kawahara, Y., Ito, K., Sun, H., Aizawa, H., Kanazawa, I. and Kwak, S. (2004) *Nature,* **427,** 801.
Kawahara, Y., Ito, K., Sun, H., Kanazawa, I. and Kwak, S. (2003) *Eur J Neurosci,* **18,** 23-33.
Kawasaki, H. and Taira, K. (2004) *Nature,* **431,** 211-7.
Keegan, L. P., Leroy, A., Sproul, D. and O'Connell, M. A. (2004) *Genome Biol,* **5,** 209.
Kent, W. J. (2002) *Genome Res,* **12,** 656-64.
Kikuno, R., Nagase, T., Waki, M. and Ohara, O. (2002) *Nucleic Acids Res,* **30,** 166-8.
Kim, D. D., Kim, T. T., Walsh, T., Kobayashi, Y., Matise, T. C., Buyske, S. and Gabriel, A. (2004) *Genome Res,* **14,** 1719-25.
Kim, J. W., Kim, H. C., Kim, G. M., Yang, J. M., Boeke, J. D. and Nam, K. (2000) *Nucleic Acids Res,* **28,** 3666-73.
Kim, U., Wang, Y., Sanford, T., Zeng, Y. and Nishikura, K. (1994) *Proc Natl Acad Sci U S A,* **91,** 11457-61.
Knight, S. W. and Bass, B. L. (2002) *Mol Cell,* **10,** 809-17.
Kohler, M., Burnashev, N., Sakmann, B. and Seeburg, P. H. (1993) *Neuron,* **10,** 491-500.
Kondo, N., Matsui, E., Kaneko, H., Aoki, M., Kato, Z., Fukao, T., Kasahara, K. and Morimoto, N. (2004) *Clin Exp Allergy,* **34,** 363-8.
Kortenbruck, G., Berger, E., Speckmann, E. J. and Musshoff, U. (2001) *Neurobiol Dis,* **8,** 459-68.
Kugita, M., Yamamoto, Y., Fujikawa, T., Matsumoto, T. and Yoshinaga, K. (2003) *Nucleic Acids Res,* **31,** 2417-23.
Kumar, M. and Carmichael, G. G. (1997) *Proc Natl Acad Sci U S A,* **94,** 3542-7.
Kumar, M. and Carmichael, G. G. (1998) *Microbiol Mol Biol Rev,* **62,** 1415-34.
Lai, F., Chen, C. X., Carter, K. C. and Nishikura, K. (1997) *Mol Cell Biol,* **17,** 2413-24.
Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F.,

Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001) *Nature,* **409,** 860-921.
Lau, P. P., Xiong, W. J., Zhu, H. J., Chen, S. H. and Chan, L. (1991) *J Biol Chem,* **266,** 20550-4.
Lau, P. P., Zhu, H. J., Baldini, A., Charnsangavej, C. and Chan, L. (1994) *Proc Natl Acad Sci U S A,* **91,** 8522-6.
Lehmann, K. A. and Bass, B. L. (1999) *J Mol Biol,* **291,** 1-13.
Lehmann, K. A. and Bass, B. L. (2000) *Biochemistry,* **39,** 12875-84.
Levanon, E. Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z. Y., Shoshan, A., Pollock, S. R., Sztybel, D., Olshansky, M., Rechavi, G. and Jantsch, M. F. (2004) *Nat Biotechnol,* **22,** 1001-5.
Liao, W., Hong, S. H., Chan, B. H., Rudolph, F. B., Clark, S. C. and Chan, L. (1999) *Biochem Biophys Res Commun,* **260,** 398-404.
Lippman, Z. and Martienssen, R. (2004) *Nature,* **431,** 364-70.
Liu, Z., Song, W. and Dong, K. (2004) *Proc Natl Acad Sci U S A,* **101,** 11862-7.
Lomeli, H., Mosbacher, J., Melcher, T., Hoger, T., Geiger, J. R., Kuner, T., Monyer, H., Higuchi, M., Bach, A. and Seeburg, P. H. (1994) *Science,* **266,** 1709-13.
Luciano, D. J., Mirsky, H., Vendetti, N. J. and Maas, S. (2004) *Rna,* **10,** 1174-7.
Ma, E., Gu, X. Q., Wu, X., Xu, T. and Haddad, G. G. (2001) *J Clin Invest,* **107,** 685-93.
Maas, S., Gerber, A. P. and Rich, A. (1999) *Proc Natl Acad Sci U S A,* **96,** 8895-900.
Maas, S., Kim, Y. G. and Rich, A. (2001a) *Mamm Genome,* **12,** 387-93.
Maas, S., Patt, S., Schrey, M. and Rich, A. (2001b) *Proc Natl Acad Sci U S A,* **98,** 14687-92.
Maden, B. E. (1990) *Prog Nucleic Acid Res Mol Biol,* **39,** 241-303.
Medstrand, P., van de Lagemaat, L. N. and Mager, D. L. (2002) *Genome Res,* **12,** 1483-95.
Mehta, A., Kinter, M. T., Sherman, N. E. and Driscoll, D. M. (2000) *Mol Cell Biol,* **20,** 1846-54.
Meister, G. and Tuschl, T. (2004) *Nature,* **431,** 343-9.
Melcher, T., Maas, S., Herb, A., Sprengel, R., Higuchi, M. and Seeburg, P. H. (1996a) *J Biol Chem,* **271,** 31795-8.
Melcher, T., Maas, S., Herb, A., Sprengel, R., Seeburg, P. H. and Higuchi, M. (1996b) *Nature,* **379,** 460-4.
Mello, C. C. and Conte, D. (2004) *Nature,* **431,** 338-342.
Mighell, A. J., Markham, A. F. and Robinson, P. A. (1997) *FEBS Lett,* **417,** 1-5.
Miyamura, Y., Suzuki, T., Kono, M., Inagaki, K., Ito, S., Suzuki, N. and Tomita, Y. (2003) *Am J Hum Genet,* **73,** 693-9.
Morse, D. P., Aruscavage, P. J. and Bass, B. L. (2002) *Proc Natl Acad Sci U S A,* **99,** 7906-11.
Morse, D. P. and Bass, B. L. (1997) *Biochemistry,* **36,** 8429-34.
Morse, D. P. and Bass, B. L. (1999) *Proc Natl Acad Sci U S A,* **96,** 6048-53.

Muddashetty, R., Khanam, T., Kondrashov, A., Bundman, M., Iacoangeli, A., Kremerskothen, J., Duning, K., Barnekow, A., Huttenhofer, A., Tiedge, H. and Brosius, J. (2002) *J Mol Biol,* **321,** 433-45.

Mullikin, J. C., Hunt, S. E., Cole, C. G., Mortimore, B. J., Rice, C. M., Burton, J., Matthews, L. H., Pavitt, R., Plumb, R. W., Sims, S. K., Ainscough, R. M., Attwood, J., Bailey, J. M., Barlow, K., Bruskiewich, R. M., Butcher, P. N., Carter, N. P., Chen, Y., Clee, C. M., Coggill, P. C., Davies, J., Davies, R. M., Dawson, E., Francis, M. D., Joy, A. A., Lamble, R. G., Langford, C. F., Macarthy, J., Mall, V., Moreland, A., Overton-Larty, E. K., Ross, M. T., Smith, L. C., Steward, C. A., Sulston, J. E., Tinsley, E. J., Turney, K. J., Willey, D. L., Wilson, G. D., McMurray, A. A., Dunham, I., Rogers, J. and Bentley, D. R. (2000) *Nature,* **407,** 516-20.

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y. and Honjo, T. (2000) *Cell,* **102,** 553-63.

Muramatsu, M., Sankaranand, V. S., Anant, S., Sugai, M., Kinoshita, K., Davidson, N. O. and Honjo, T. (1999) *J Biol Chem,* **274,** 18470-6.

Muto, T., Muramatsu, M., Taniwaki, M., Kinoshita, K. and Honjo, T. (2000) *Genomics,* **68,** 85-8.

Neuberger, M. S., Harris, R. S., Di Noia, J. and Petersen-Mahrt, S. K. (2003) *Trends Biochem Sci,* **28,** 305-12.

Niswender, C. M., Herrick-Davis, K., Dilley, G. E., Meltzer, H. Y., Overholser, J. C., Stockmeier, C. A., Emeson, R. B. and Sanders-Bush, E. (2001) *Neuropsychopharmacology,* **24,** 478-91.

Novo, F. J., Kruszewski, A., MacDermot, K. D., Goldspink, G. and Gorecki, D. C. (1995) *Nucleic Acids Res,* **23,** 2636-40.

Numata, K., Kanai, A., Saito, R., Kondo, S., Adachi, J., Wilming, L. G., Hume, D. A., Hayashizaki, Y. and Tomita, M. (2003) *Genome Res,* **13,** 1301-6.

Nutt, S. L., Hoo, K. H., Rampersad, V., Deverill, R. M., Elliott, C. E., Fletcher, E. J., Adams, S. L., Korczak, B., Foldes, R. L. and Kamboj, R. K. (1994) *Receptors Channels,* **2,** 315-26.

O'Connell, M. A., Gerber, A. and Keller, W. (1997) *J Biol Chem,* **272,** 473-8.

O'Connell, M. A., Krause, S., Higuchi, M., Hsuan, J. J., Totty, N. F., Jenny, A. and Keller, W. (1995) *Mol Cell Biol,* **15,** 1389-97.

Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., Nagai, K., Kimura, K., Makita, H., Sekine, M., Obayashi, M., Nishi, T., Shibahara, T., Tanaka, T., Ishii, S., Yamamoto, J., Saito, K., Kawai, Y., Isono, Y., Nakamura, Y., Nagahari, K., Murakami, K., Yasuda, T., Iwayanagi, T., Wagatsuma, M., Shiratori, A., Sudo, H., Hosoiri, T., Kaku, Y., Kodaira, H., Kondo, H., Sugawara, M., Takahashi, M., Kanda, K., Yokoi, T., Furuya, T., Kikkawa, E., Omura, Y., Abe, K., Kamihara, K., Katsuta, N., Sato, K., Tanikawa, M., Yamazaki, M., Ninomiya, K., Ishibashi, T., Yamashita, H., Murakawa, K., Fujimori, K., Tanai, H., Kimata, M., Watanabe, M., Hiraoka, S., Chiba, Y., Ishida, S., Ono, Y., Takiguchi, S., Watanabe, S., Yosida, M., Hotuta, T., Kusano, J., Kanehori, K., Takahashi-Fujii, A., Hara, H., Tanase, T. O., Nomura, Y., Togiya, S., Komai, F., Hara, R., Takeuchi, K., Arita, M., Imose, N., Musashino, K., Yuuki, H., Oshima, A., Sasaki, N., Aotsuka, S., Yoshikawa, Y., Matsunawa, H., Ichihara, T., Shiohata, N., Sano, S., Moriya, S., Momiyama, H., Satoh, N., Takami, S., Terashima, Y., Suzuki, O., Nakagawa, S., Senoh, A.,

Mizoguchi, H., Goto, Y., Shimizu, F., Wakebe, H., Hishigaki, H., Watanabe, T., Sugiyama, A., et al. (2004) *Nat Genet,* **36,** 40-5.
Palladino, M. J., Keegan, L. P., O'Connell, M. A. and Reenan, R. A. (2000a) *Rna,* **6,** 1004-18.
Palladino, M. J., Keegan, L. P., O'Connell, M. A. and Reenan, R. A. (2000b) *Cell,* **102,** 437-49.
Panigrahi, A. K., Gygi, S. P., Ernst, N. L., Igo, R. P., Jr., Palazzo, S. S., Schnaufer, A., Weston, D. S., Carmean, N., Salavati, R., Aebersold, R. and Stuart, K. D. (2001) *Mol Cell Biol,* **21,** 380-9.
Patterson, J. B. and Samuel, C. E. (1995) *Mol Cell Biol,* **15,** 5376-88.
Patton, D. E., Silva, T. and Bezanilla, F. (1997) *Neuron,* **19,** 711-22.
Paul, M. S. and Bass, B. L. (1998) *Embo J,* **17,** 1120-7.
Paule, M. R. and White, R. J. (2000) *Nucleic Acids Res,* **28,** 1283-98.
Paupard, M. C., O'Connell, M. A., Gerber, A. P. and Zukin, R. S. (2000) *Neuroscience,* **95,** 869-79.
Pennisi, E. (2003) *Science,* **300,** 1484.
Peters, N. T., Rohrbach, J. A., Zalewski, B. A., Byrkett, C. M. and Vaughn, J. C. (2003) *Rna,* **9,** 698-710.
Polson, A. G. and Bass, B. L. (1994) *Embo J,* **13,** 5701-11.
Polson, A. G., Bass, B. L. and Casey, J. L. (1996) *Nature,* **380,** 454-6.
Poulsen, H., Nilsson, J., Damgaard, C. K., Egebjerg, J. and Kjems, J. (2001) *Mol Cell Biol,* **21,** 7862-71.
Powell, L. M., Wallis, S. C., Pease, R. J., Edwards, Y. H., Knott, T. J. and Scott, J. (1987) *Cell,* **50,** 831-40.
Raitskin, O., Cho, D. S., Sperling, J., Nishikura, K. and Sperling, R. (2001) *Proc Natl Acad Sci U S A,* **98,** 6571-6.
Revy, P., Muto, T., Levy, Y., Geissmann, F., Plebani, A., Sanal, O., Catalan, N., Forveille, M., Dufourcq-Labelouse, R., Gennery, A., Tezcan, I., Ersoy, F., Kayserili, H., Ugazio, A. G., Brousse, N., Muramatsu, M., Notarangelo, L. D., Kinoshita, K., Honjo, T., Fischer, A. and Durandy, A. (2000) *Cell,* **102,** 565-75.
Rosenthal, J. J. and Bezanilla, F. (2002) *Neuron,* **34,** 743-57.
Rueter, S. M., Dawson, T. R. and Emeson, R. B. (1999) *Nature,* **399,** 75-80.
Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S. and Altshuler, D. (2001) *Nature,* **409,** 928-33.
Sansam, C. L., Wells, K. S. and Emeson, R. B. (2003) *Proc Natl Acad Sci U S A,* **100,** 14018-23.
Saunders, L. R. and Barber, G. N. (2003) *Faseb J,* **17,** 961-83.
Scadden, A. D. and Smith, C. W. (2001) *EMBO Rep,* **2,** 1107-11.
Semenov, E. P. and Pak, W. L. (1999) *J Neurochem,* **72,** 66-72.
Shah, R. R., Knott, T. J., Legros, J. E., Navaratnam, N., Greeve, J. C. and Scott, J. (1991) *J Biol Chem,* **266,** 16301-4.

Sharma, P. M., Bowman, M., Madden, S. L., Rauscher, F. J., 3rd and Sukumar, S. (1994) *Genes Dev,* **8,** 720-31.
Shuman, S. (2002) *Nat Rev Mol Cell Biol,* **3,** 619-25.
Sijen, T. and Plasterk, R. H. (2003) *Nature,* **426,** 310-4.
Simpson, L., Aphasizhev, R., Gao, G. and Kang, X. (2004) *Rna,* **10,** 159-70.
Simpson, L., Sbicego, S. and Aphasizhev, R. (2003) *Rna,* **9,** 265-76.
Skuse, G. R., Cappione, A. J., Sowden, M., Metheny, L. J. and Smith, H. C. (1996) *Nucleic Acids Res,* **24,** 478-85.
Smith, C. M. and Steitz, J. A. (1998) *Mol Cell Biol,* **18,** 6897-909.
Smith, L. A., Peixoto, A. A. and Hall, J. C. (1998) *J Neurogenet,* **12,** 227-40.
Sodhi, M. S., Burnet, P. W., Makoff, A. J., Kerwin, R. W. and Harrison, P. J. (2001) *Mol Psychiatry,* **6,** 373-9.
Sommer, B., Kohler, M., Sprengel, R. and Seeburg, P. H. (1991) *Cell,* **67,** 11-9.
Song, W., Liu, Z., Tan, J., Nomura, Y. and Dong, K. (2004) *J Biol Chem,* **279,** 32554-61.
Sorek, R., Ast, G. and Graur, D. (2002) *Genome Res,* **12,** 1060-7.
Stenger, J. E., Lobachev, K. S., Gordenin, D., Darden, T. A., Jurka, J. and Resnick, M. A. (2001) *Genome Res,* **11,** 12-27.
Stephens, O. M., Haudenschild, B. L. and Beal, P. A. (2004) *Chem Biol,* **11,** 1239-50.
Teng, B., Burant, C. F. and Davidson, N. O. (1993) *Science,* **260,** 1816-9.
Tijsterman, M., Ketting, R. F. and Plasterk, R. H. (2002) *Annu Rev Genet,* **36,** 489-519.
Tisdall, J. (2001) *Begining Perl for bioinformatics,* O'Reilly.
Tonkin, L. A. and Bass, B. L. (2003) *Science,* **302,** 1725.
Tonkin, L. A., Saccomanno, L., Morse, D. P., Brodigan, T., Krause, M. and Bass, B. L. (2002) *Embo J,* **21,** 6025-35.
van Leeuwen, F. W., de Kleijn, D. P., van den Hurk, H. H., Neubauer, A., Sonnemans, M. A., Sluijs, J. A., Koycu, S., Ramdjielal, R. D., Salehi, A., Martens, G. J., Grosveld, F. G., Peter, J., Burbach, H. and Hol, E. M. (1998) *Science,* **279,** 242-7.
Visiers, I., Hassan, S. A. and Weinstein, H. (2001) *Protein Eng,* **14,** 409-14.
Walter, P. and Blobel, G. (1982) *Nature,* **299,** 691-8.
Wang, L., Kimble, J. and Wickens, M. (2004a) *Rna,* **10,** 1444-8.
Wang, Q., Miyakoda, M., Yang, W., Khillan, J., Stachura, D. L., Weiss, M. J. and Nishikura, K. (2004b) *J Biol Chem,* **279,** 4952-61.
Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D.,

Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigo, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., et al. (2002) *Nature,* **420,** 520-62.

Wei, C. M., Gershowitz, A. and Moss, B. (1976) *Biochemistry,* **15,** 397-401.

Wolf, J., Gerber, A. P. and Keller, W. (2002) *Embo J,* **21,** 3841-51.

Wong, S. K. and Lazinski, D. W. (2002) *Proc Natl Acad Sci U S A,* **99,** 15118-23.

Wong, S. K., Sato, S. and Lazinski, D. W. (2001) *Rna,* **7,** 846-58.

Yamanaka, S., Poksay, K. S., Arnold, K. S. and Innerarity, T. L. (1997) *Genes Dev,* **11,** 321-33.

Yang, J. H., Luo, X., Nie, Y., Su, Y., Zhao, Q., Kabir, K., Zhang, D. and Rabinovici, R. (2003a) *Immunology,* **109,** 15-23.

Yang, J. H., Nie, Y., Zhao, Q., Su, Y., Pypaert, M., Su, H. and Rabinovici, R. (2003b) *J Biol Chem,* **278,** 45833-42.

Yelin, R., Dahary, D., Sorek, R., Levanon, E. Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R., Nemzer, S., Pinner, E., Walach, S., Bernstein, J., Savitsky, K. and Rotman, G. (2003) *Nat Biotechnol,* **21,** 379-86.

Yusupov, M. M., Yusupova, G. Z., Baucom, A., Lieberman, K., Earnest, T. N., Cate, J. H. and Noller, H. F. (2001) *Science,* **292,** 883-96.

Zeng, Y., Yi, R. and Cullen, B. R. (2003) *Proc Natl Acad Sci U S A,* **100,** 9779-84.

Zhang, H., Yang, B., Pomerantz, R. J., Zhang, C., Arunachalam, S. C. and Gao, L. (2003) *Nature,* **424,** 94-8.