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).

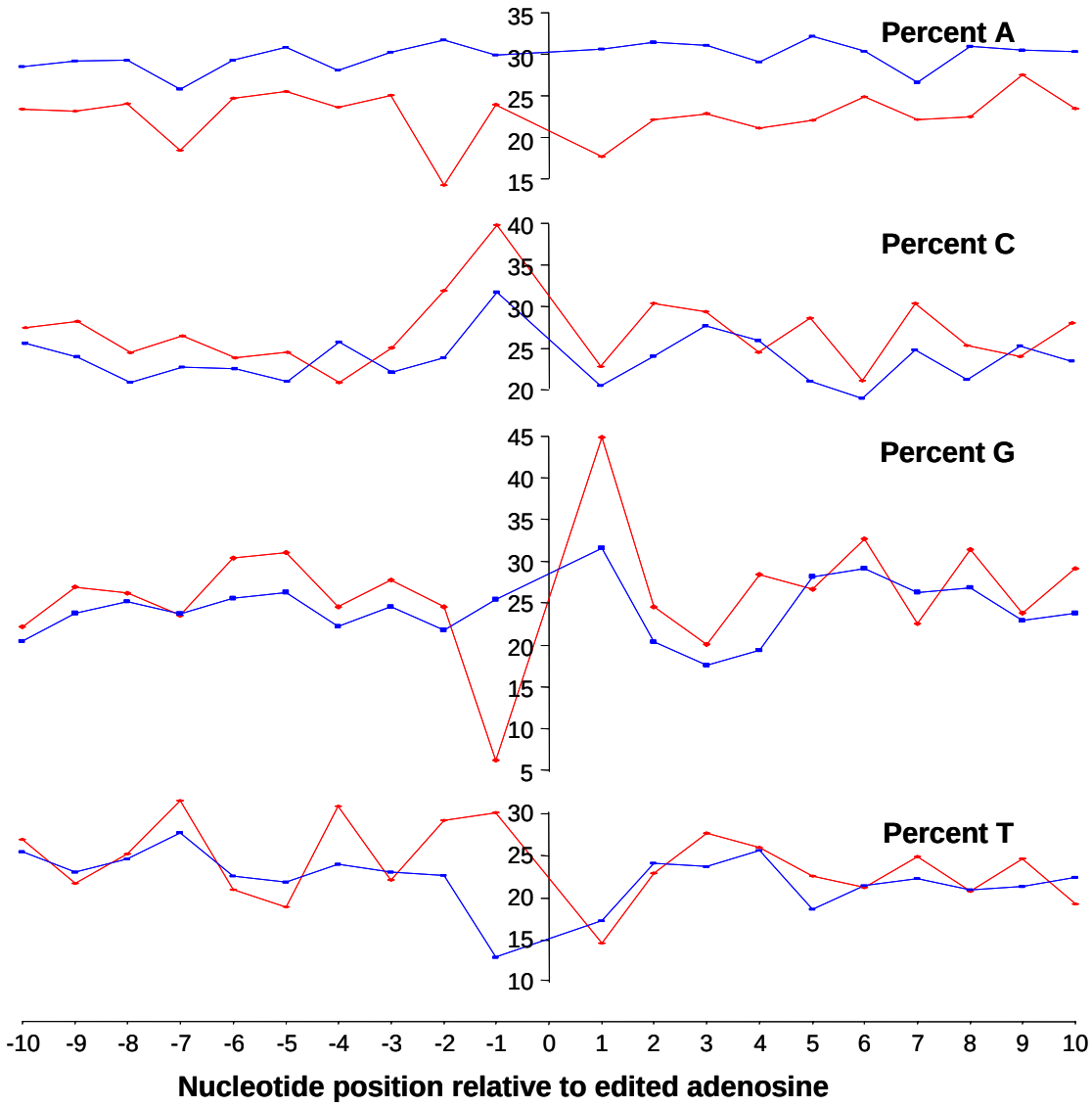

Figure 6-1 Sequence context of adenosines in edited Alu sequences. The sequence context of all edited adenosines and all unedited adenosines from all edited Alu sequences was compared. For each of the ten bases either side of edited adenosines (red lines) and unedited adenosines (blue lines) the proportion 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.

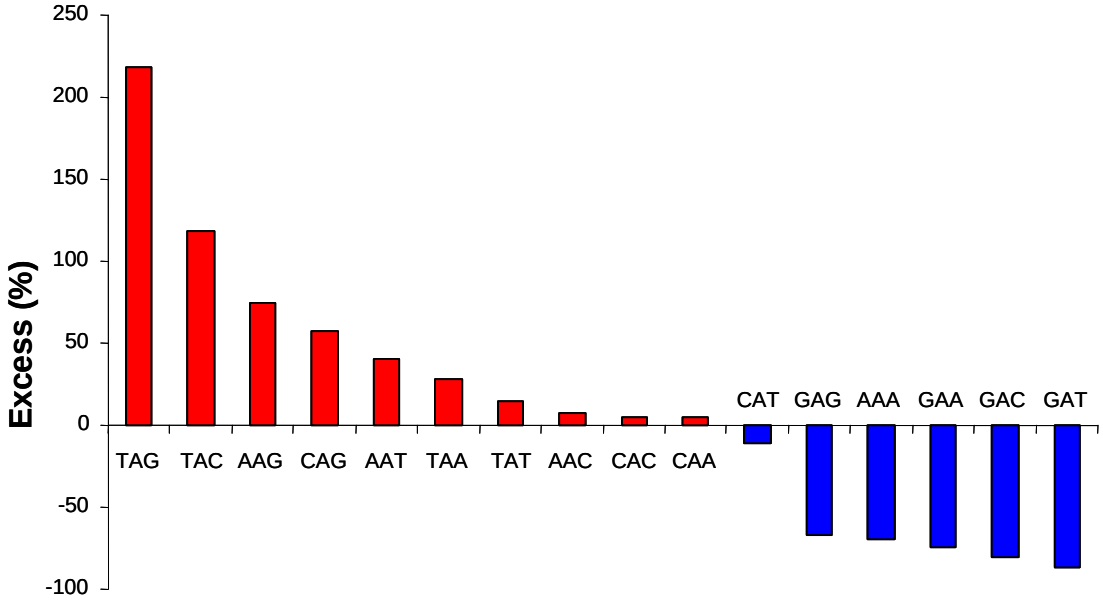

Figure 6-2 Tri-nucleotide sequence context of adenosines in edited Alu sequences. For each tri-nucleotide sequence centred on an adenosine, the number of edited and unedited adenosines present in that sequence context from cDNA clone sequences was determined. The percentage excess of edited adenosines 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).

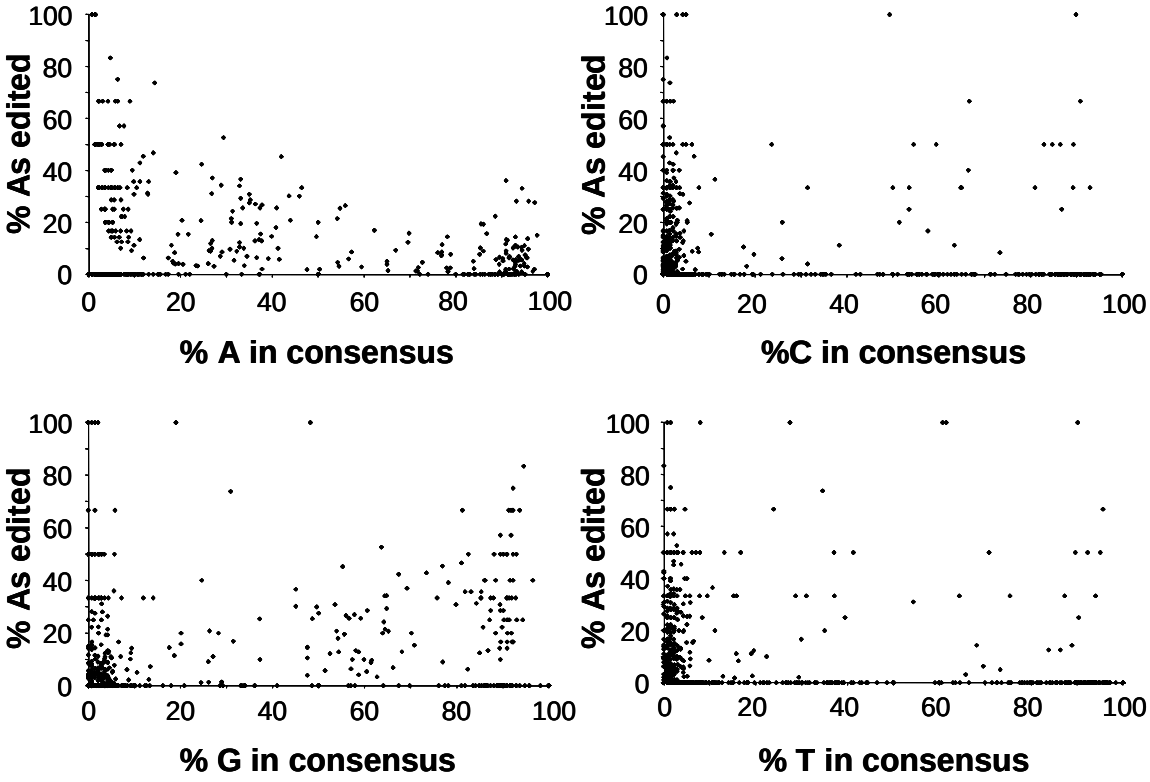

Figure 6-3 Effect of sequence composition on the likelihood of RNA editing. A multiple alignment of all edited Alu sequences was prepared using CLUSTALW. At each position in the alignment, the proportion of edited adenosines was calculated from the number of sequenced edited adenosines and the total 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).

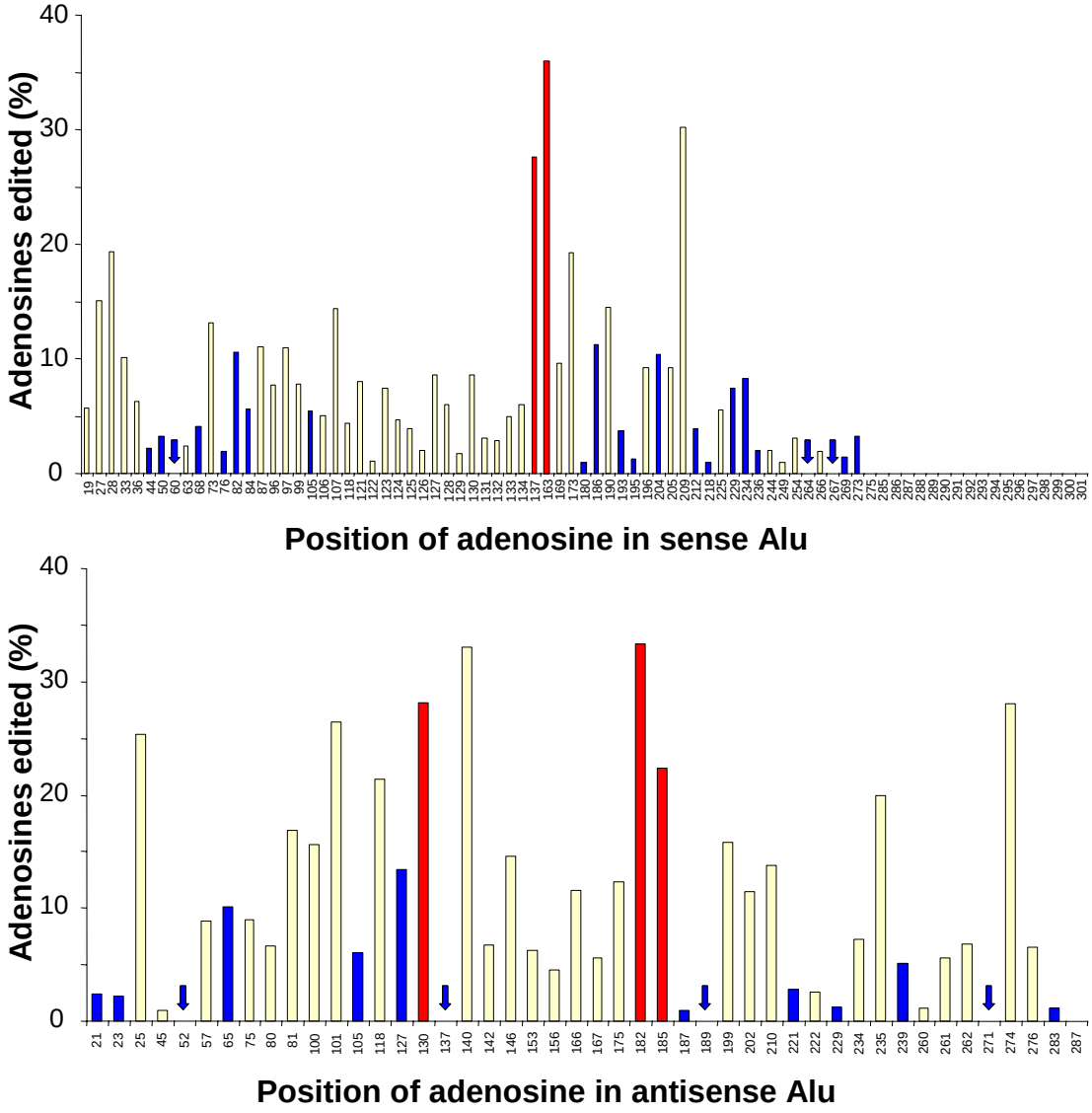

Figure 6-4 Frequency of editing at adenosines in edited sense and anti-sense Alus. For each adenosine in the consensus sequence, the proportion of adenosines which were edited was calculated from all sequenced adenosines. All adenosines within TAG tri-nucleotides (red bars), and GAX (X = A,C,G or T)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