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.

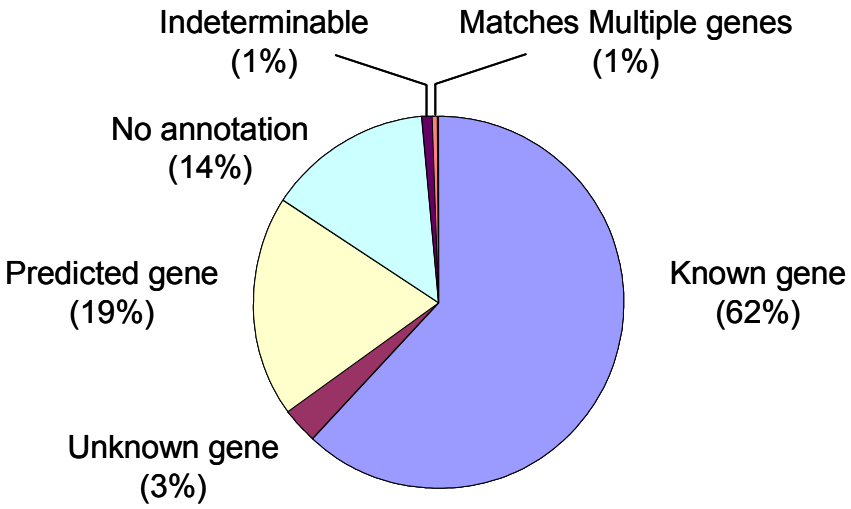


Figure 5-1 Breakdown of RNA edits by gene class. Annotation of edited cDNA clone sequences was derived from the annotation of all cDNA library sequences (see Chapter 3).

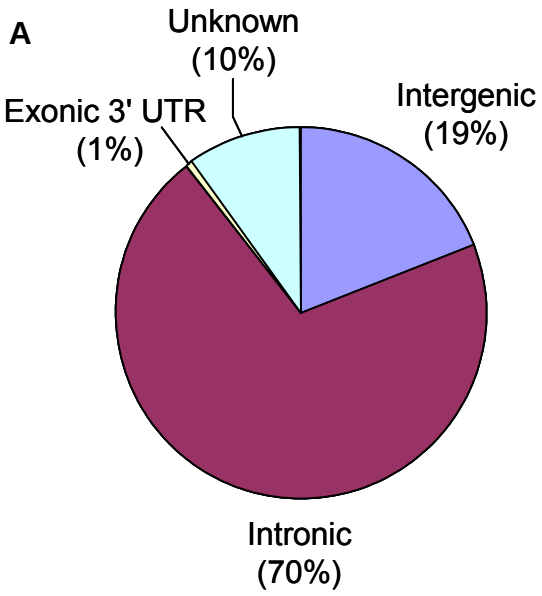
Of transcripts for which there was evidence of editing, 91% (167 / 183) were found in the cDNA library as a single edited clone. The most frequently edited transcript from the library was FRMD4 from which three non-overlapping edited clones and four non-overlapping unedited clones were sequenced. All 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.

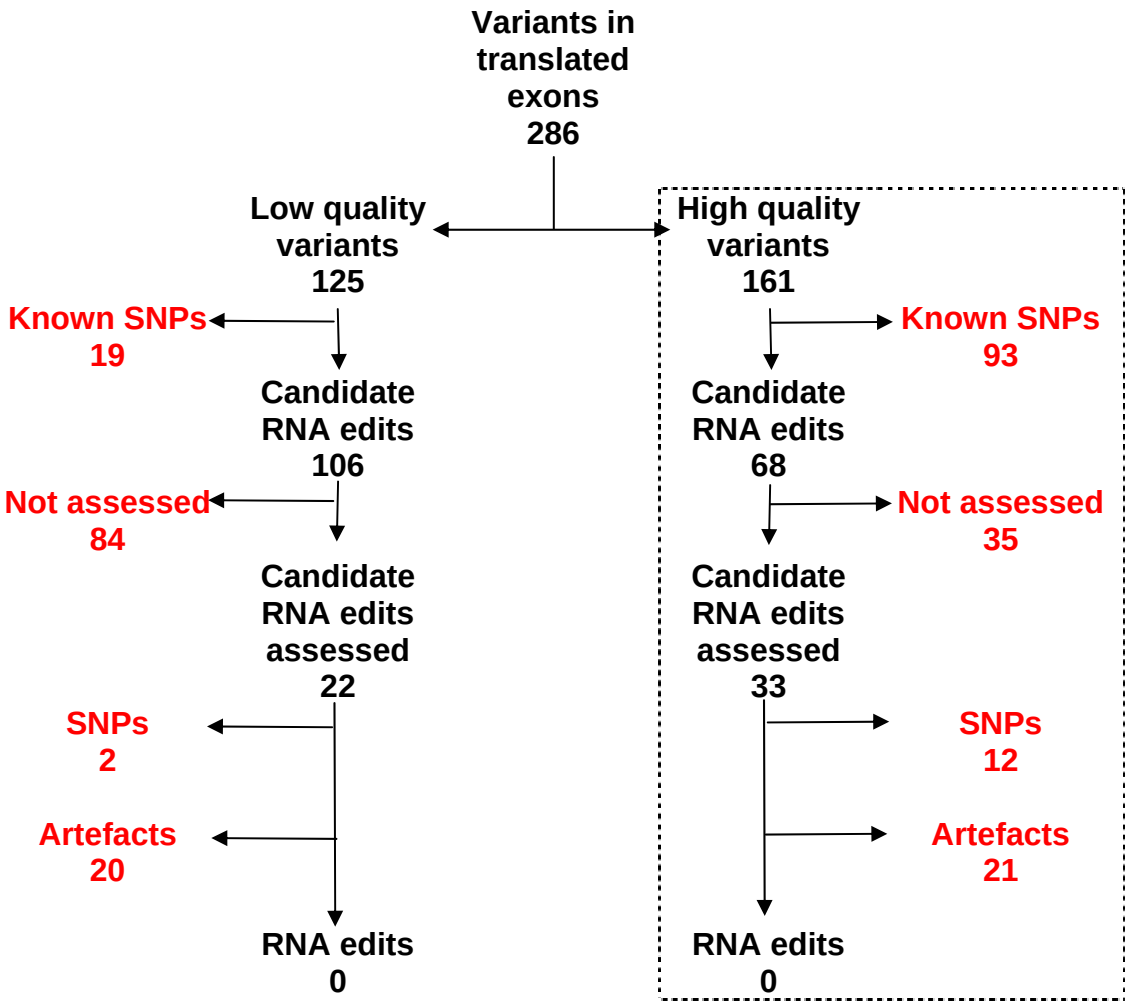


Figure 5-3 Summary of the analysis of the subset of 286 variants from translated exon sequence. Variants were classified as high quality or low quality (see Methods). Variants listed in red were rejected for various criteria Values in black show the remaining candidate RNA edits at each stage of the 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.

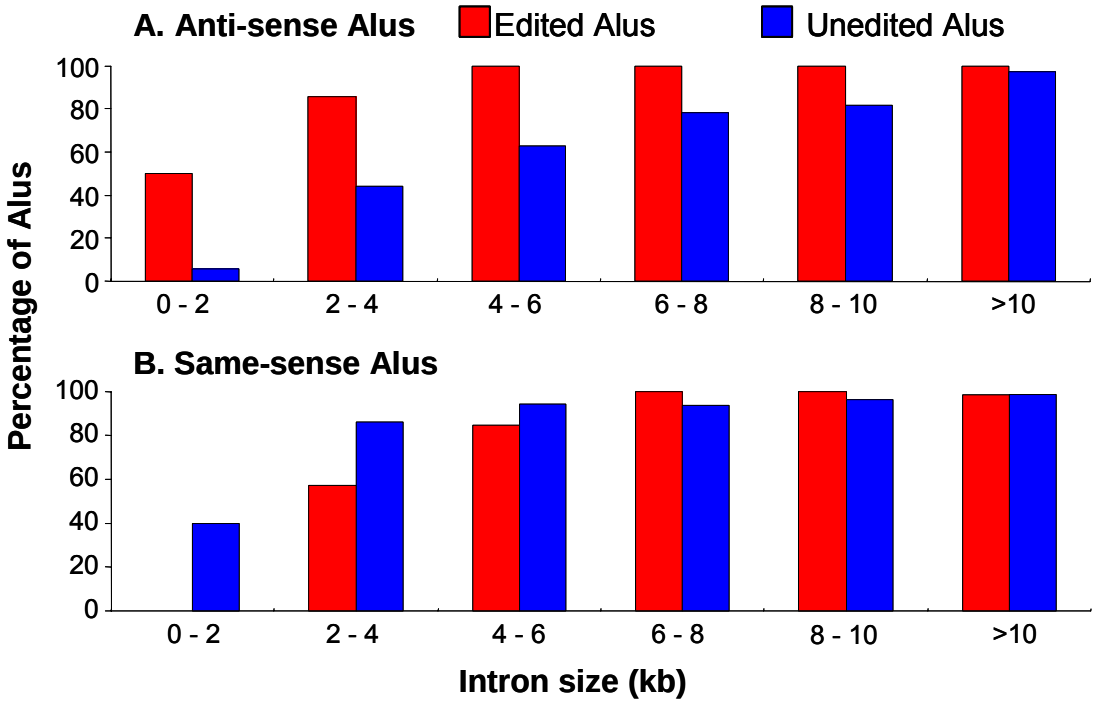


Figure 5-4 Proportion of edited and unedited Alus with additional Alus in the same intron. All Alus aligning to the introns of known genes, and for which ≥80% of the genomic extent of the Alu was sequenced were included in this analysis. The proportion of edited Alus (red bars) and the proportion of 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.

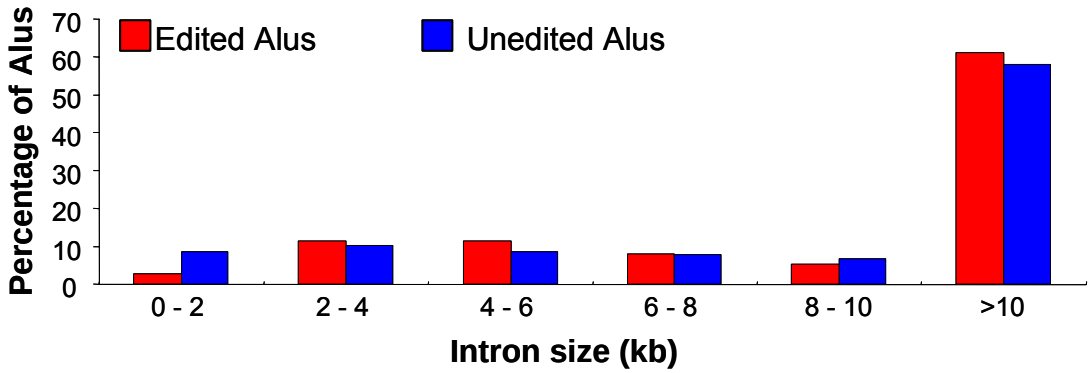


Figure 5-5 Proportion of edited and unedited Alus from introns of different sizes. Intron sizes were recorded for all Alus aligning to the introns of known genes, and for which ≥80% of the genomic extent of the Alu was sequenced. The proportion of edited alus (red bars) and unedited alus (blue bars)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

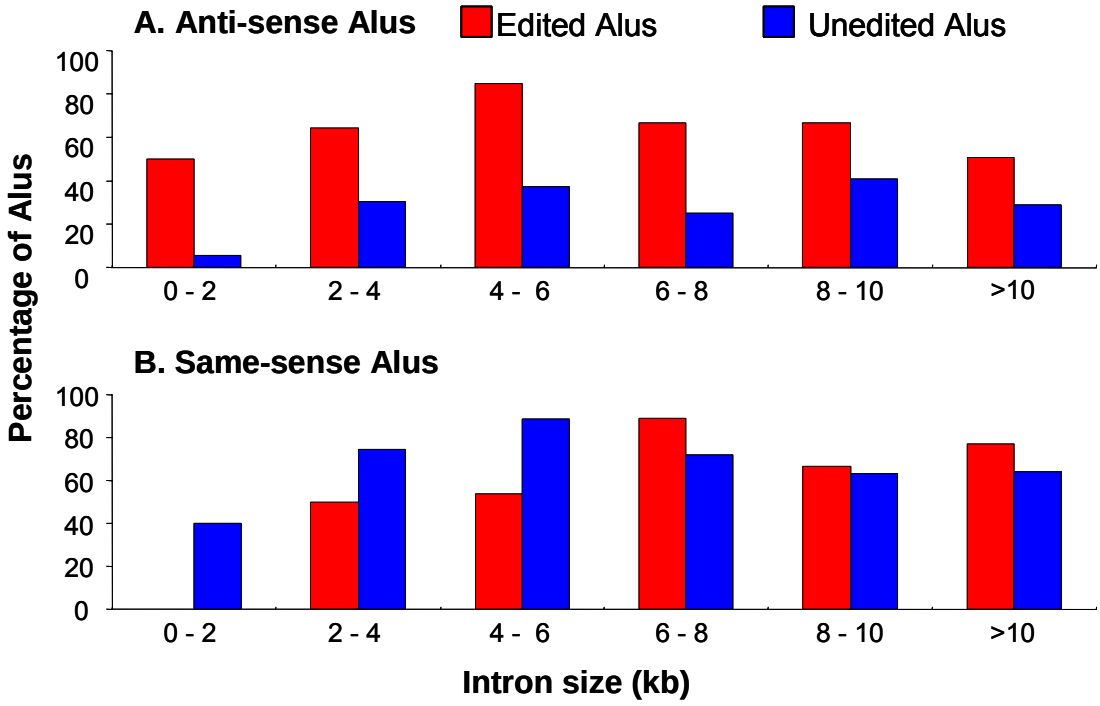


Figure 5-6 Proportion of edited and unedited Alus with additional Alus within 0 to 1 kb in the same intron. The Alu sequences included in this analysis are the same as in Figure 5-4. The proportion of edited Alus (red bars) and unedited Alus (blue bars) with an anti-sense Alu (**A**) or a same-sense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

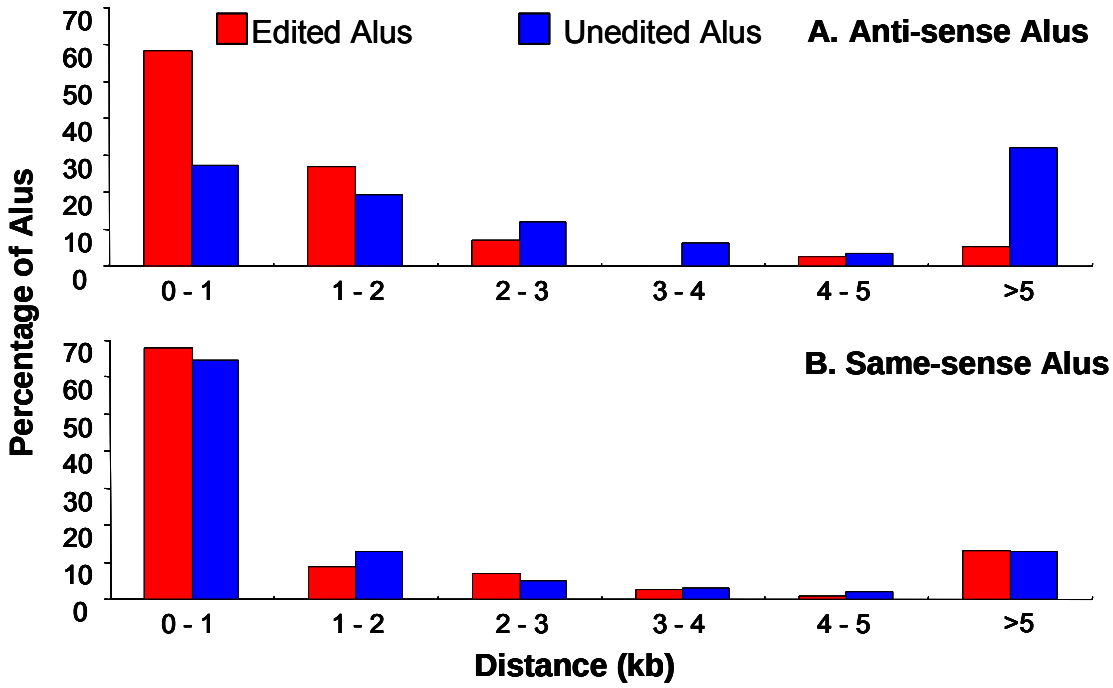


Figure 5-7 Distance from edited and unedited Alus to the nearest Alu in the same intron. The Alu sequences included in this analysis are the same as in Figure 5-4. The proportion of edited Alus (red bars) and unedited Alus (blue bars) at different distances from the nearest anti-sense Alu (**A**) or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

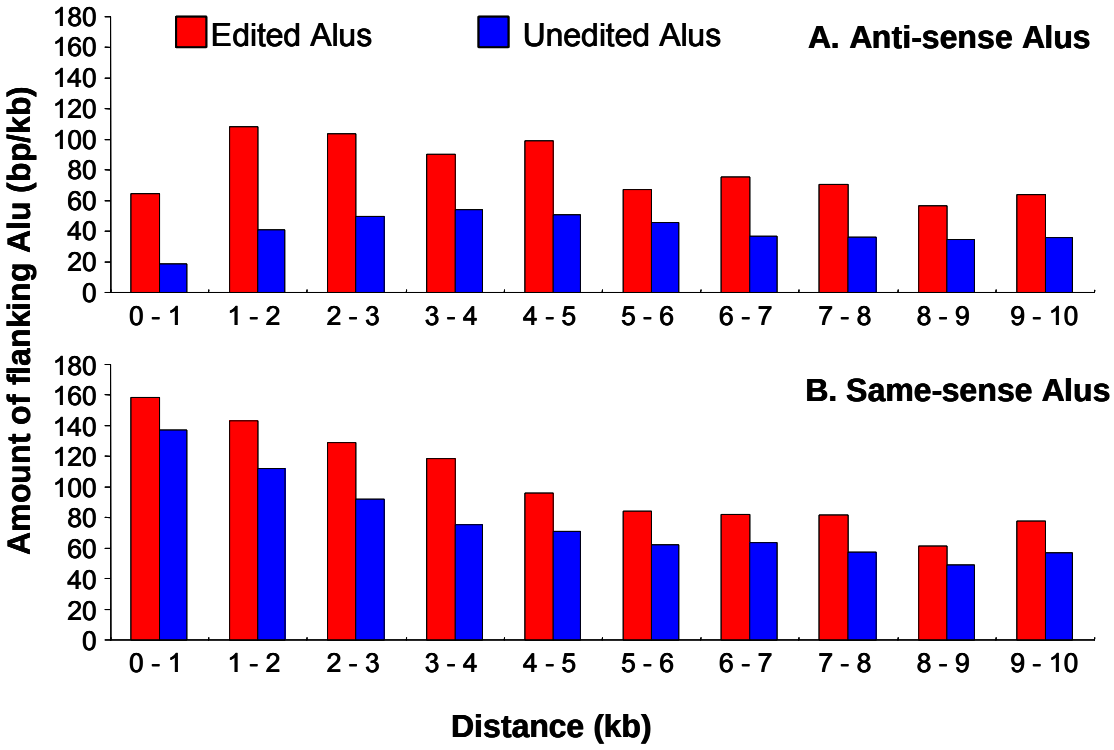


Figure 5-8 Amount of flanking Alu sequence at different distances from edited and unedited Alus. All Alus for which ≥80% of the genomic extent of the Alu was sequenced were included in this analysis. For each Alu, the amount of flanking Alu sequence in the opposite orientation (**A**) or same orientation (**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).

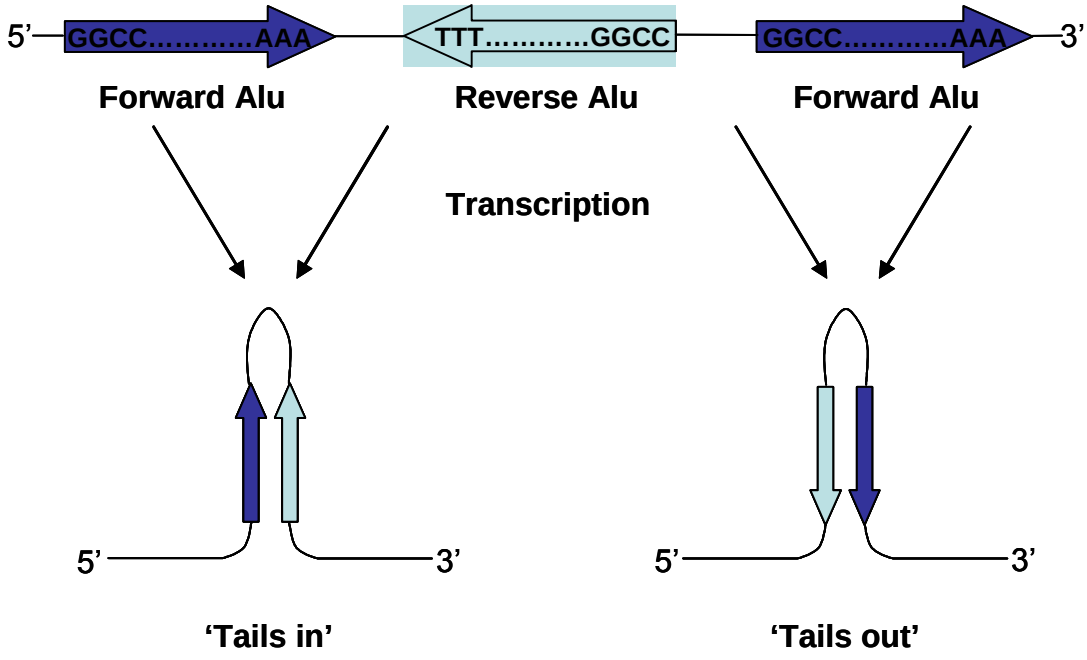


Figure 5-9 Orientation of Alu sequences with respect to each other. Alus may be transcribed in the forward (dark blue arrows), or reverse (light blue arrows) orientation. Arrowheads indicate the position of the poly-A tail (forward Alus) or leading poly-T sequence (reverse Alus). A pair of inverted repeats 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.

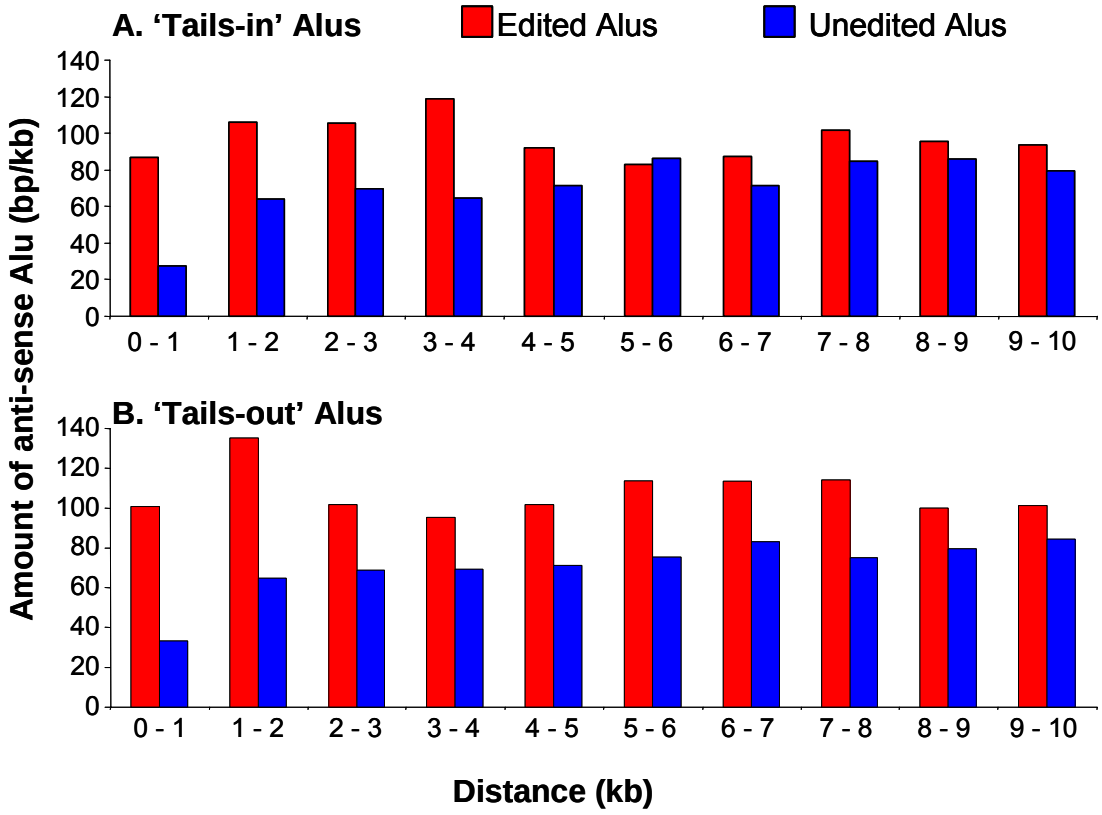


Figure 5-10 Amount of anti-sense Alu sequence at different distances from edited and unedited Alus in 'Tails-Out' orientation (**A**), or 'Tails-in' orientation (**B**). 'Tails-in' anti-sense Alus are all reverse Alus 3' in the same transcript as forward Alus, and all forward Alus 5' in the same transcript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