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).

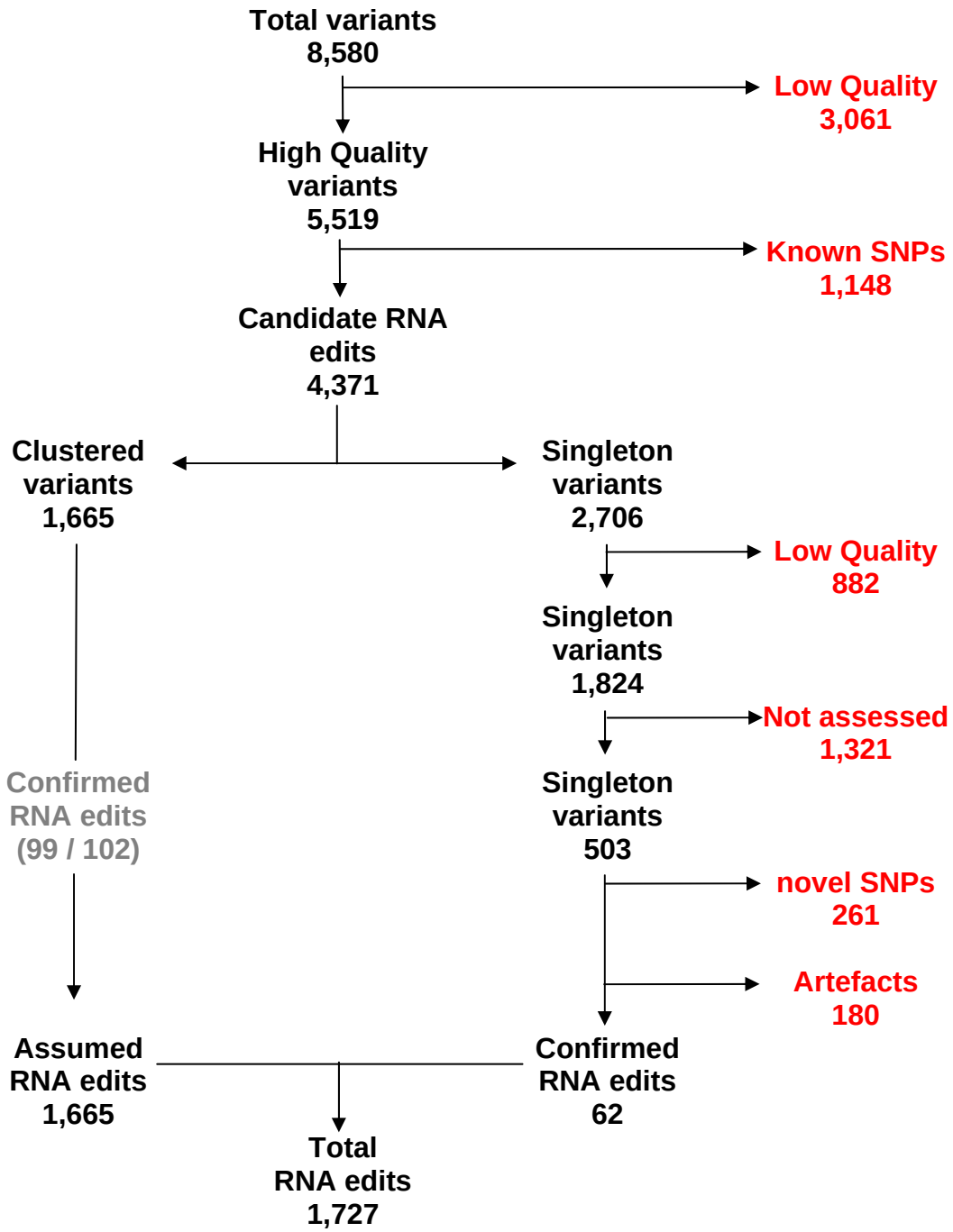

Figure 4-1 Summary of the identification of 1,727 novel A>I RNA edits. Variants are separated into those from sequences with 3 or more variants (clustered), and those from sequences with less than 3 variants (singletons). Variants in red were rejected according to various criteria. Variants in grey

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.

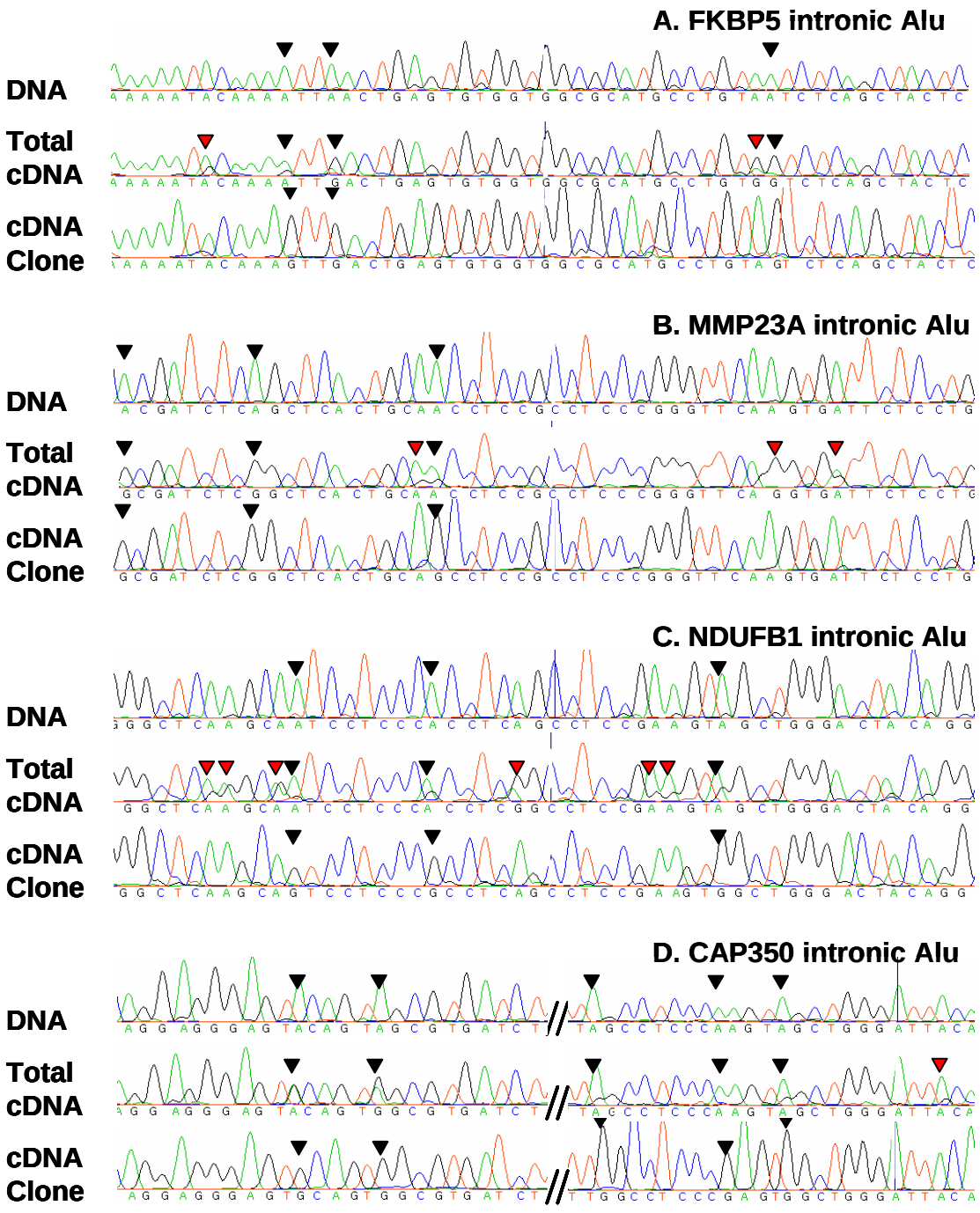

Figure 4-2 Confirmation of RNA editing of heavily edited sequences. cDNA clones containing three or more variants of the same type were evaluated by sequencing of PCR products from genomic DNA, and Total cDNA. Sequence traces are shown for four of the 11 / 12 sequences that were confirmed to be edited.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

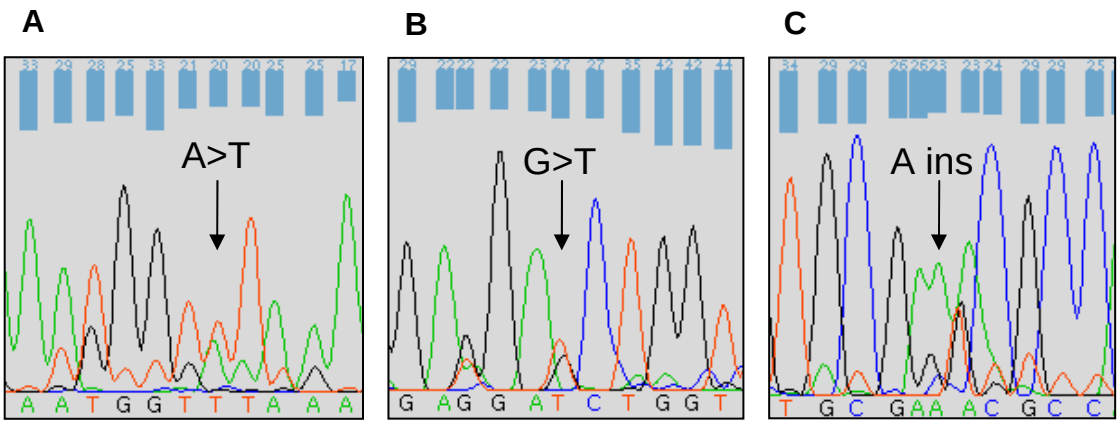

Figure 4-3 Examples of variants identified incorrectly by automated detection. The traces are base called using *Phred*, quality scores are shown above each

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.