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°C overnight. Protein was precipitated and removed by addition of 8ml NaCl (5M) and centrifugation (3000rpm, 4°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°C overnight (17 hours) then placed at 4°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°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°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.

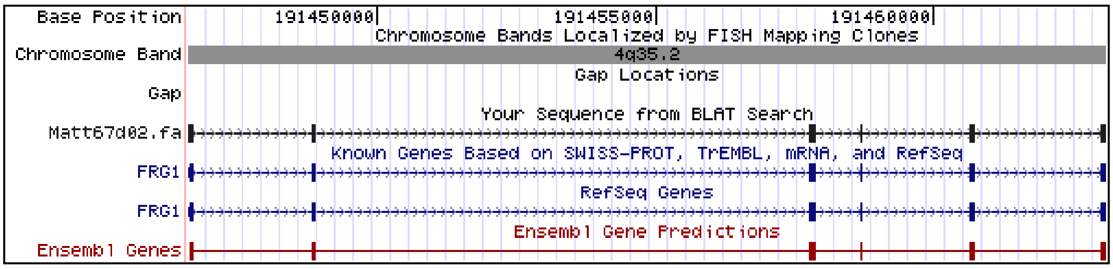
seq_context.pl: For every edited and unedited adenosine from all cDNA clones returns the 10bp of sequence from either side of that adenosine

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,gaatctggtggacagtaacaaactt
tggtgaaatttcaggaaccatagccattgaaatggataagggaacctatatacatgcactcgacaatggtctttttaccctgggagctccacaca
aagaag,ttgatgagggccctagtcctccagagcagtttacggctgtcaaattatctgattccag,aattgccctgaagtctggctatggaaaat
atcttggtataaattcagatggacttgttgttgg,cgttcagatgcaattggaccangagaacaatgggaaccagtctttcaaaatg,[22]
ccggcttcagcctctccgcgcagaagtctcccggagccatggcctagtattcttatgtgaagtctaccaagcttgtgctcaagggaacc,agtaa
gaagaaaaagagcaaagataagaagagaaaaagagaagaagatgaagaaacccagcttgatat,tgttg,gaatctggtgaacagtaacaaactt
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.