3 SEQUENCING AND EVALUATION OF A HUMAN BRAIN cDNA LIBRARY

3.1 INTRODUCTION

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

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

In this chapter, over 3Mb cDNA sequenced at random from a human brain cDNA library was aligned to the human genome reference sequence. As these alignments were subsequently used to identify novel RNA edits (see Chapter 4), it was important to ascertain whether they were representative of the transcriptome of human brain cells. Therefore, the alignments of cDNA clones to the genome were used to evaluate the quality of the cDNA library with respect to contamination by genomic DNA. The composition of the cDNA library was evaluated by annotation of known genes.

3.2 RESULTS

3.2.1 Construction of a human brain cDNA library

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

Tissue sections were removed from the cerebral cortex of a male donor, whose cause of death was congestive cardiac failure. The brain tissue had been frozen with a post-mortem delay of 9 hours, and was classified as normal from its appearance under the microscope. Total RNA was analysed by denaturing agarose gel electrophoresis. The 28S and 18S ribosomal RNAs were clearly visible indicating that the RNA was reasonably intact. Poly-(A)$^+$ RNA was isolated by two rounds of purification on an oligo-(dT)-cellulose column. Analysis by agarose gel electrophoresis indicated that the majority of the ribosomal RNA was removed (Figure 3-1A). For the purposes of identifying SNPs between the tissue donor and the human genome reference sequence, genomic DNA was isolated from tissue adjacent to that used in the preparation of RNA.

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

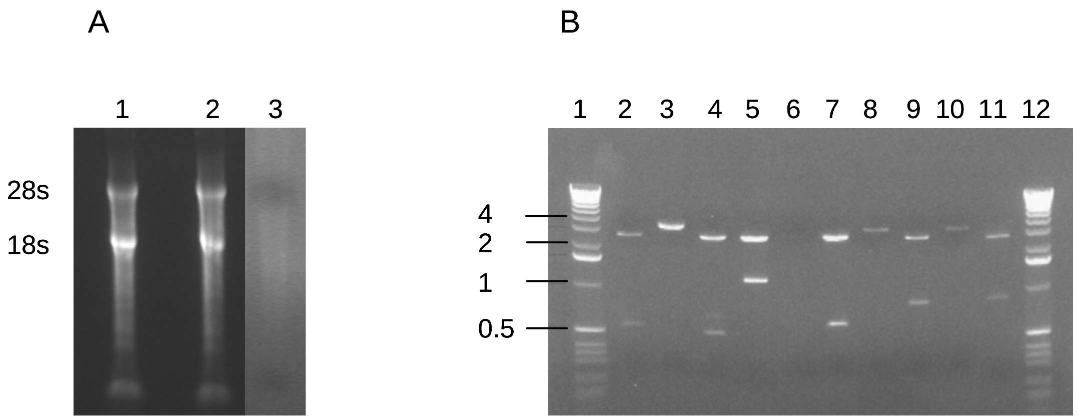

Figure 3-1 Analysis of Human Cerebral cortex nucleic acid preparations. **A**. Electrophoresis of human cerebral cortex total RNA (lanes 1 and 2 in duplicate), and poly-(A)$^+$ purified mRNA (lane 3). Bands corresponding to the 28S and 18S ribosomal RNA subunits are indicated. **B**. Electrophoresis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

3.2.2 Evaluation of the cDNA library

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

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

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

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

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

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

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

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

The initial sequencing target of this survey was to analyse 1Mb of coding RNA sequence. To compensate for the exonic (spliced) cDNA content of approximately 25%, the number of clones sequenced from the library was increased four fold. With an average of 400 bases of high quality sequence per clone it was estimated that 10,000 clones would give a total of 4Mb sequence including the desired 1Mb of coding sequence for our analysis.

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

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

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

Incorrectly aligned cDNA clone sequences could give rise to erroneous sequence variants which appear to be candidate RNA edits. Therefore, for each cDNA clone, the BLAT score and the percentage identity score of the two highest scoring alignments to the genome were identified, and used to identify cDNA clones that were incorrectly aligned to the genome.

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

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

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

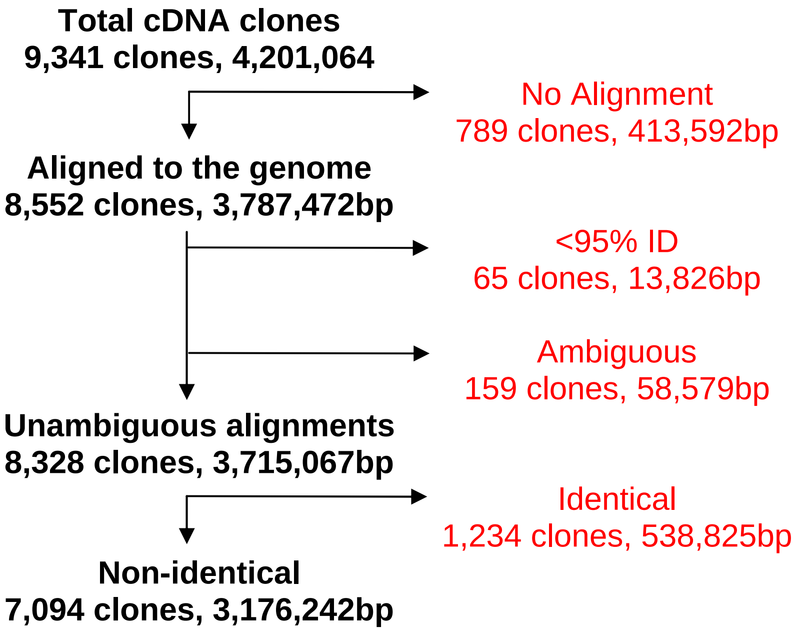

Figure 3-2 Processing of cDNA clone sequence data. The values in red indicate sequences that were rejected for various criteria. Ambiguous sequences are those which aligned to two regions of the genome with similar BLAT scores and percentage scores (defined in the text). Values in black show the remaining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

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

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

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

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

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

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

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

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

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

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

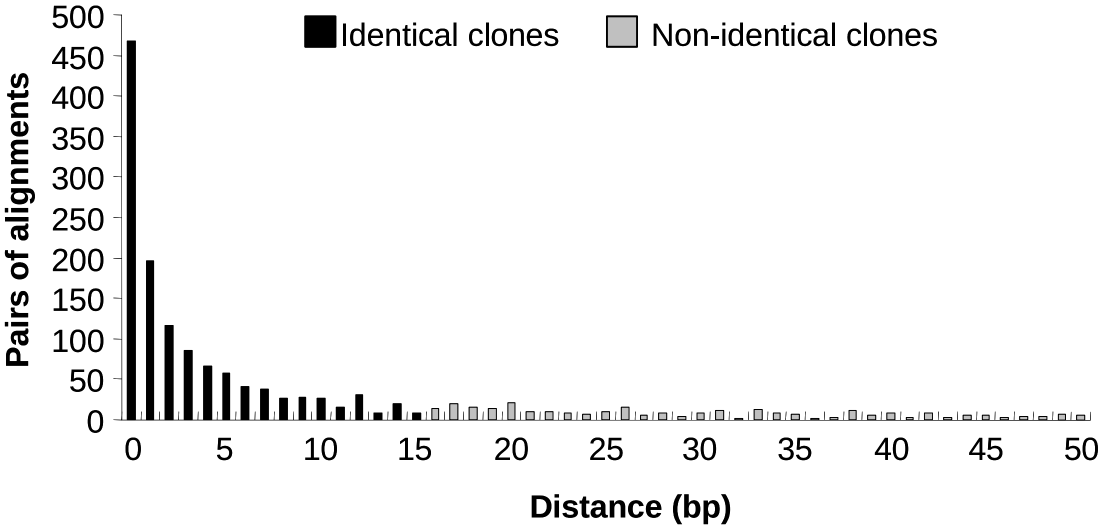

Figure 3-3 Discrimination of identical and non-identical overlapping clones. The number of bases separating the start positions of overlapping alignments was used to discriminate 'identical' from 'non-identical' overlapping cDNA clones. Pairs of alignments with start positions separated by 15 or less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

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

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

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

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

The cDNA clones were next classified according to their origin in the human genome. 95.4% (6,768 / 7,094) cDNA clones, comprising 3,058,468bp non-overlapping cDNA sequence, were derived from the nuclear chromosomes (Table 3-3). A further 0.3% (24 / 7,094) clones, comprising 7,026bp, were derived from the ribosomal DNA repeat sequence. Given that the ribosomal RNA is typically the major component of total RNA, with mRNAs making up only 2-3%, this indicated efficient purification of poly-adenylated RNAs away from ribosomal RNA in the preparation of this library. The remaining 4.3% (302 / 7,094) clones (110,748bp) were from the mitochondrial genome.

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

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

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

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

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

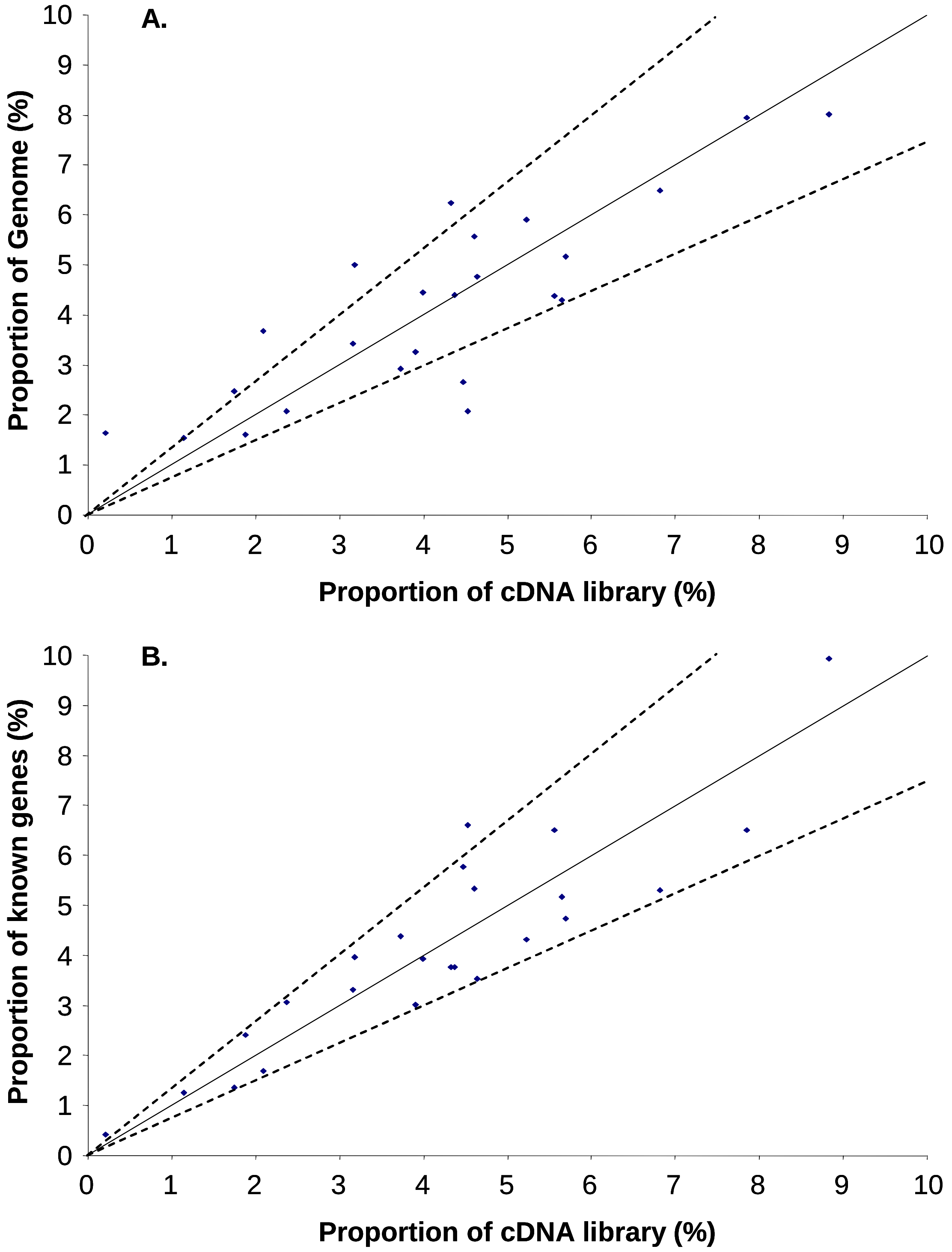

Figure 3-4 Comparison of the proportion of cDNA clones derived from each chromosome with **A**. the proportion of the human genome on each chromosome, and **B**. the proportion of all known genes on each chromosome. **A**. The solid black line indicates an equal proportion from the cDNA library

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

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

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

In total, 302 / 7,094 non-identical clones were found to align to the mitochondrial genome. These comprised 110,748 bases of sequence, which overlapped considerably. The total amount of unique sequence was 15,494 bases. As the mitochondrial genome is 16,571bp this corresponds to 93.5% coverage of the mitochondrial genome. To visualise the alignments of clones to the mitochondrion in more detail, they were displayed as a custom track in the UCSC human genome browser (Figure 3-5). Consistent with this, clones span most of the mitochondrial genome and are unspliced. However, the majority of clones cluster into groups corresponding with the known mitochondrial genes and very few clones (<5%) overlap more than one gene. This strongly suggests that most clones are derived from mature transcripts rather than from precursor transcript, and that the cDNA library is not heavily contaminated with mitochondrial DNA. Only a small number of clones align to the D-Loop of the mitochondrion, consistent with this being intergenic sequence.

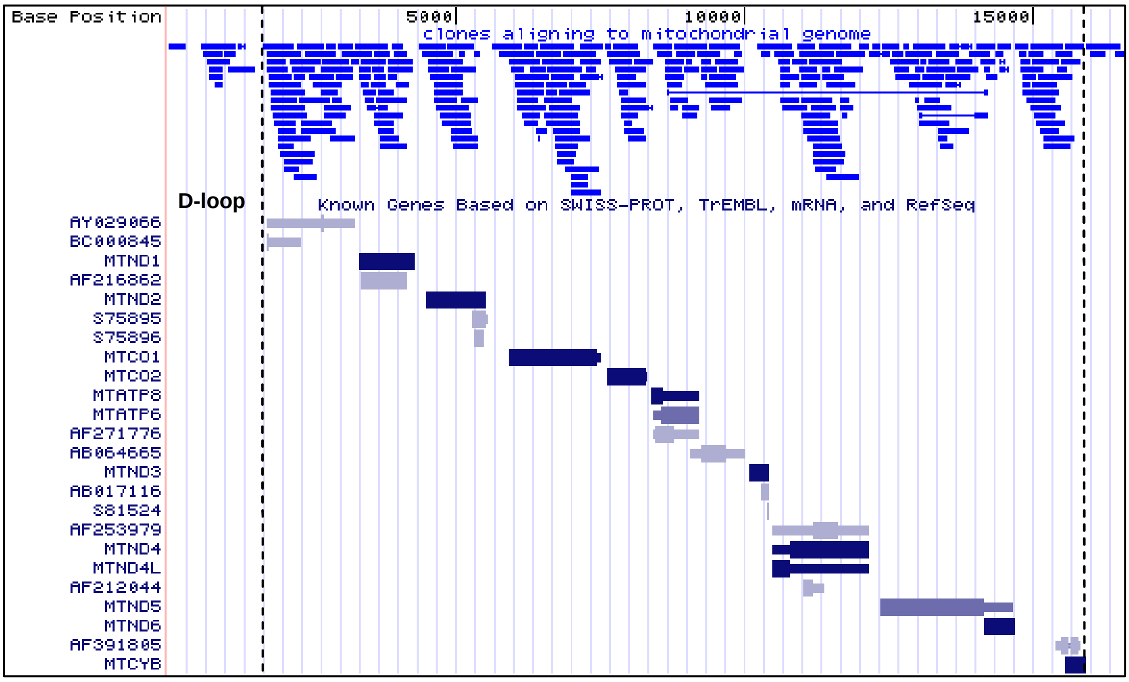

Figure 3-5 Mitochondrial cDNA clones. Mitochondrial DNA is a circular molecule. The extreme left and right of the display represent the same point on this molecule. The upper blue bars represent clone sequences. The lower bars represent the known mitochondrial genes for comparison. Dashed lines indicate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

3.2.6 Evaluation of cDNA library by annotation of known genes

Detailed annotation of cDNA clones aligning to the nuclear chromosomes was required to provide context for any novel RNA edits. For each clone, the EnsEMBL database was searched for evidence of unambiguous alignment to a known or predicted gene. For clones derived from known genes, the alignment was related to the positions of boundaries between introns and exons and between coding and non-coding regions. To annotate all 7,094 clones an automated method was developed based on the EnsEMBL database, and the associated EnsEMBL API.

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

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

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

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

For each cDNA clone aligning unambiguously to an EnsEMBL gene, the gene number and gene description was retrieved. By counting the number of times each gene was sequenced, a list of the most highly represented transcripts in the cDNA library was constructed (Table 3-5). As expected from a brain cDNA library, most of the frequently sequenced genes had ‘housekeeping’ functions such as Actin (9 clones) and Glyceraldehyde 3-phosphate dehydrogenase (8 clones), and neuronal functions such as Myelin basic protein (24 clones) and Synaptosomal protein (12 clones). However, by far the most frequently detected gene in the library was ENSG00000185316 (32 clones). BLASTN alignment of the minimum genomic DNA sequence containing all 32 cDNA clone sequences against the NCBI non-redundant database identified a gene with no protein product, metastasis associated lung adenocarcinoma transcript 1 (MALAT-1)(Ji et al., 2003). This transcript was reported to be significantly associated with metastasis in NSCLC patients (Ji et al., 2003), but there is currently no information about its function in normal cells.

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

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

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

3.2.6.2 *Evaluation of cDNA library composition by sequence class*

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

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

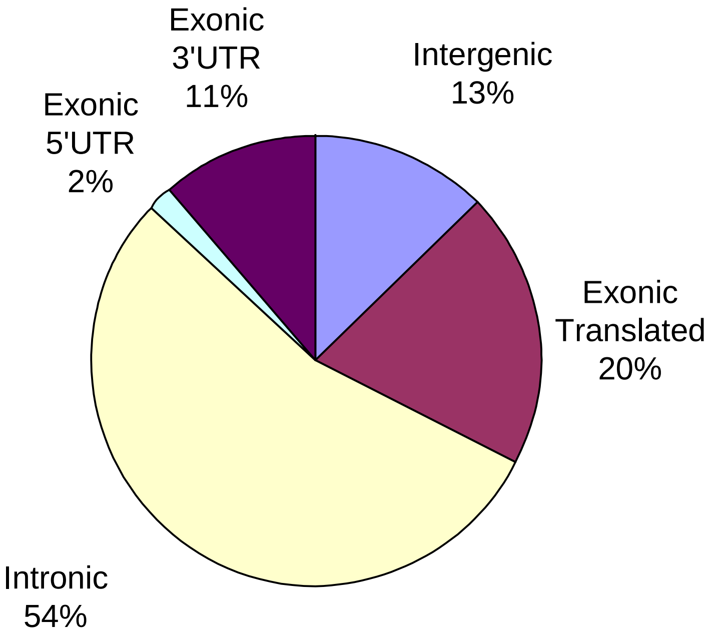

Figure 3-6 Sequence class composition of the cDNA library.

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

3.3 DISCUSSION

3.3.1 Choice of experimental strategy for a survey of RNA editing

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

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

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

Another source of candidate RNA edits, from multiple tissue types, would be from alignments of EST sequences to the human genome reference sequence. Indeed this approach was successfully employed in a recent systematic search for A>I edits in human tissues (Levanon et al., 2004). Sequence variants identified from EST alignments would include sequence trace errors, unforeseen artefacts relating to cDNA library construction, SNPs, and RNA edits. Although frequent editing events and dominant patterns of RNA editing would be readily detectable, infrequent RNA editing events would be extremely difficult to separate from other sources of sequence variation. When this thesis was started, very few EST sequence traces were available from sequence trace repositories, and therefore there was no information about the quality of sequence traces. Furthermore, there is no matching genomic DNA sequence with which to compare EST sequences and identify SNPs. In contrast, the cDNA clone sequencing approach used in this thesis allowed sequencing artefacts to be identified from sequence traces, and SNPs to be identified by reference to matching genomic DNA. This allowed an untargeted evaluation of infrequent as well as frequent editing events.

3.3.2 Choice of tissue for a survey of RNA editing

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

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

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

Total RNA was poly (A)+ RNA purified prior to cDNA synthesis. This was necessary to remove ribosomal RNA from the cDNA library, but would also result in the exclusion of other transcripts that are not poly-adenylated. This includes the majority of RNA Pol I and Pol II transcripts including rRNA, tRNA, snRNA, snoRNAs and miRNAs, which are potential RNA editing substrates. Several of these classes of RNA undergo modification (eg pseudouridylation and o-methylation), and both tRNAs and miRNAs are known targets of A > I RNA editing (Maas et al., 1999, Luciano et al., 2004). Despite selection for poly-adenylated transcripts, non poly-adenylated transcripts are represented in the cDNA library, albeit at reduced levels compared to the original brain tissue. 24 cDNA clones were derived from the rRNA repeat and a further nine sequences overlap non-coding RNAs including three snRNAs, three snoRNAs and one miRNA.

3.3.4 Sequence class composition of the cDNA library

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

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

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

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