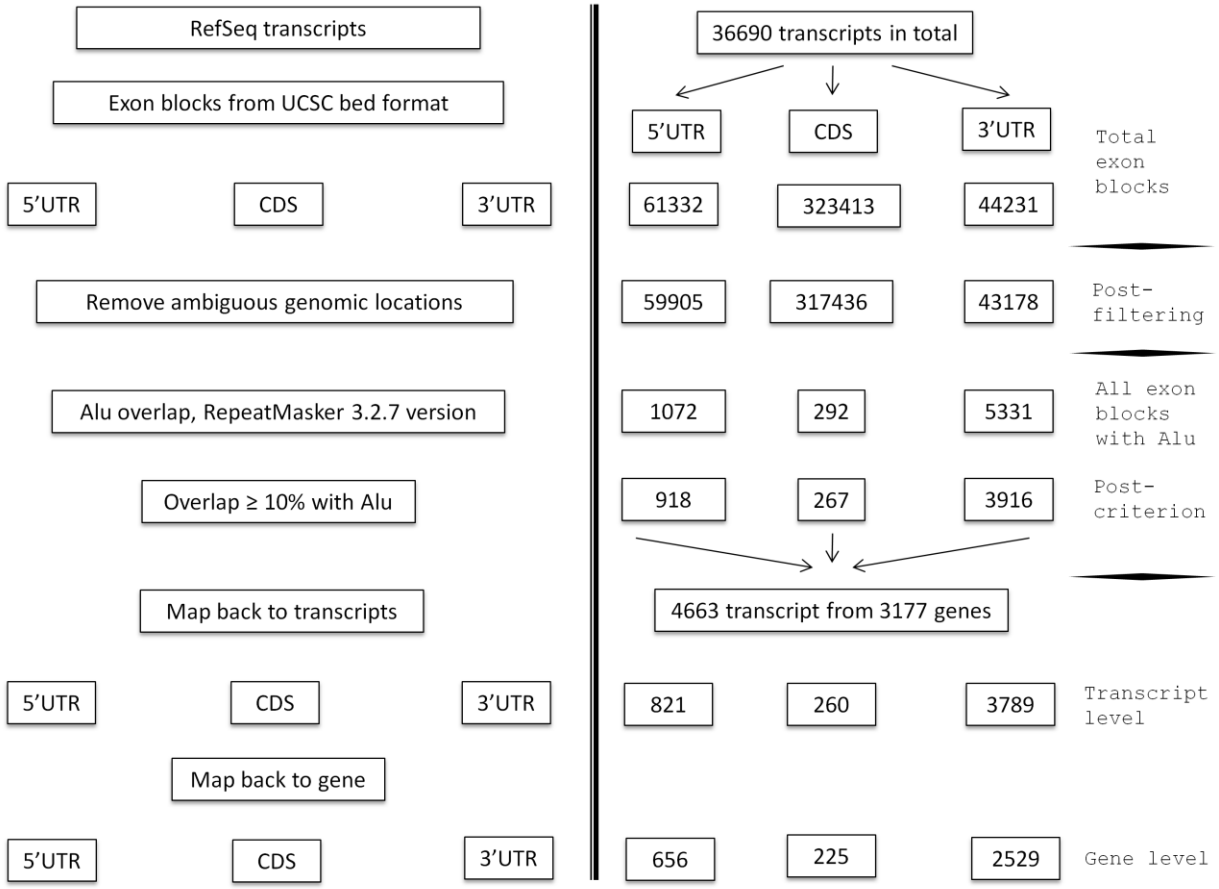
Detailed Methodology of Alu mediated Exonization, A→I editing and antisense events:

Identification of Alu exonization events in the transcriptome

Data-mining of the RefSeq database (release 45, Jan 2011) which comprehensively catalogs mature mRNA sequences and annotations with respect to the UTR regions and CDS start-end positions was carried out to identify Alu exonization events in the entire transcriptome **(Supplementary Figure 1)**. Using the UCSC Table Browser (for genome build hg18), exon block alignments for 5'UTR, 3'UTR and CDS regions in BED (browser extensible data) format were exported to the Galaxy framework of tools. Alignments from alternate assemblies (HapMap regions) and unplaced contigs (*chr*_random*) were removed from each of the exonic classes. The start-end positions of each alignment blocks were then overlapped with Alu element coordinates from RepeatMasker (version 3.2.7) data using the *Coverage* Tool in Galaxy. In order to avoid retaining blocks with only minimal Alu content a threshold of ≥10% overlap of Alu elements within the exon blocks was set. After all the above filtering processes, the exonised Alus were mapped back to the gene through the mRNA accession numbers. The number of transcripts (and related genes) in each category, that is 5'UTR, 3'UTR and CDS, that contained Alu-in exons were documented individually.



Supplementary Figure 1. Schema of analyses for identification and mapping of Alus to Exonized transcript isoforms from RefSeq database

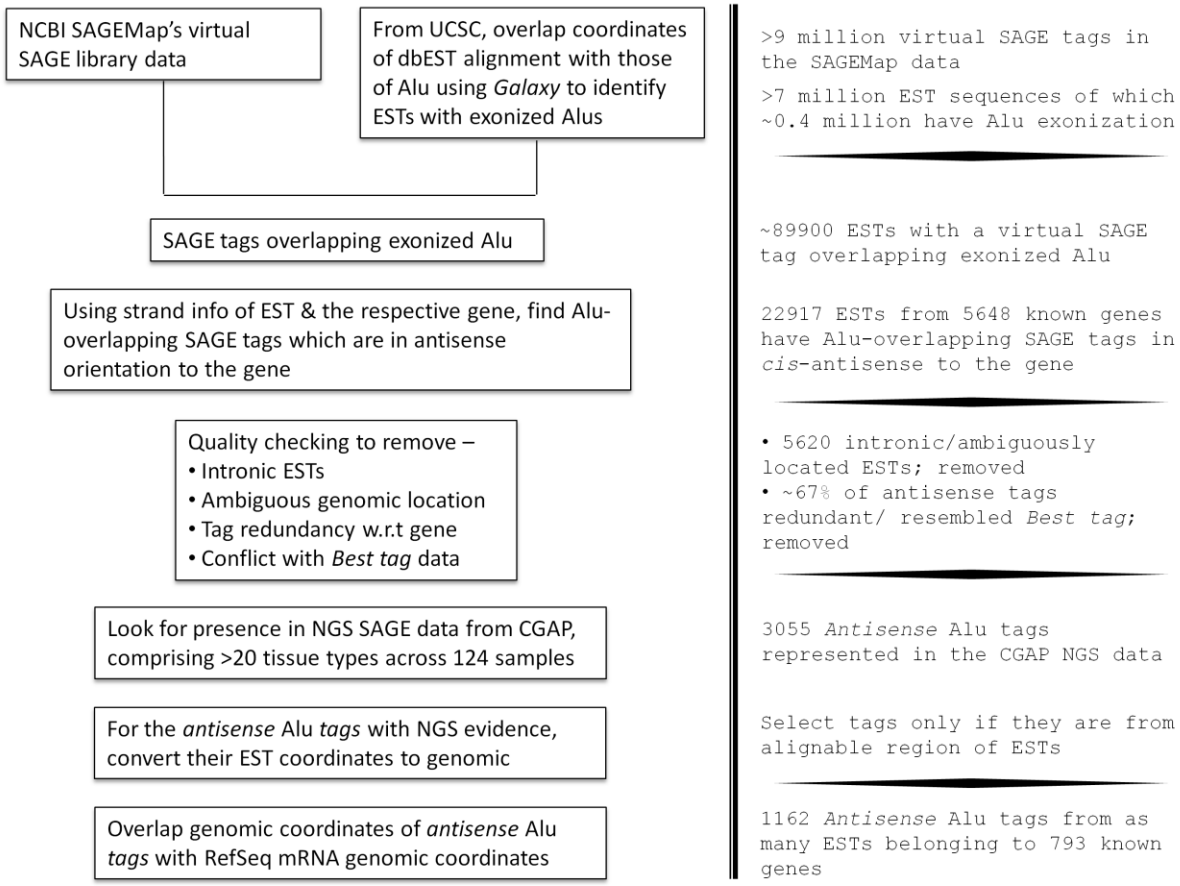
The strategy followed from RefSeq genes to exon-level information for Alu exonization harboured within 5'UTR, CDS or 3'UTR has been depicted. Using datasets from UCSC TableBrowser and coordinate overlay using Galaxy tools, specific counts for Alu-harboring exons were determined with respect to the CDS or UTR positions at the transcriptome-wide level. The panel on the left represents the steps involved and the right panel represents the numbers obtained in each of these steps.

Identification of Alu in the antisense transcriptome

There are no high-throughput experimental platforms that can detect antisense transcripts, that too from the repetitive sequences. For instance, microarray platforms have defined probes primarily designed to query the sense transcripts and that too mapping to unique regions. Unbiased approaches such as RNA sequencing can provide information on antisense transcripts provided one can develop methodologies both experimental as well as computational to differentiate sense from antisense. We sought to harvest the potential of SAGE (Serial analysis of Gene expression) methodology to determine the contribution of Alu elements to natural antisense transcription **(Supplementary Figure 2)**. To carry this out comprehensively we started with database of all possible SAGE tags i.e. NCBI SAGEMap's virtual SAGE library data (long SAGE, 17bp sequences). This dataset contains 9 million virtual SAGE tags generated from *in silico* digestion of transcripts by *NlaIII* from the 3' end. The transcripts profiled in the virtual SAGE library are derived from heterogeneous sources like mRNA, cDNA and ESTs. The majority of the SAGE tags are from ESTs which though not full length, present the most exhaustive resource mirroring the expression repertoire of diverse tissues. Therefore we selected dbEST, containing 7 million EST sequences for identifying not only Alu antisense but also A→I editing events, the later being described in the following section. Using the UCSC Table Browser we obtained the genomic alignment blocks for ESTs following which we used the Galaxy framework as described for exonised transcripts, to overlap with Alu element coordinates from RepeatMasker and identified all the ESTs that had Alu exonization. Then using the start-end coordinates of SAGE tags from the virtual SAGE Map data and the positional information of Alu element within the ESTs, we identified all the SAGE tags that overlap with the exonized Alus. We used the strand information of EST alignment and the corresponding gene to identify the antisense transcripts and filter out the virtual SAGE tags that are in *cis*-antisense orientation to the gene and overlap with Alu. We wanted to restrict our analysis to only those antisense events that target mature mRNA and thus we filtered out the intronic ESTs and ESTs with ambiguous genomic location (i.e. those that either aligns to multiple places onto the genome, or onto alternate assemblies or unplaced contigs). SAGE tag analysis is dependent on the assignment of specific tag sequences as the *Best Tag* representing a given gene. Annotation of *Best Tag* is available as a resource in NCI-CGAP (Cancer Genome Anatomy Project) FTP. As an additional quality filter criterion we removed all those antisense tags which matched to known *Best Tag.*

We also ensured that all possible Alu overlapping antisense tags, not conflicting with *Best Tag*, are represented only once for any given gene i.e. we made the list of filtered tags non-redundant gene-wise. Therefore, through the above exercise of extensive series of filtering criteria we identified a set of virtual SAGE tags in the transcripts that were potentially derived from Alu and *cis*-antisense to the genes in the transcriptome. We then queried for these Virtual SAGE Tags for their actual presence in experimental datasets. For this we used two Next-generation sequencing based SAGE datasets (GSE1902 and GSE15314), a part of the CGAP, available from NCBI GEO. These datasets have information on >20 different tissue types across 124 samples. By pooling the SAGE tag sequences from these two datasets and comparing with the filtered set of virtual SAGE tags we identified all Alu antisense events harbouring Alu sequences that have experimental evidence. We refer these as Alu antisense.

The Alu antisense that had been identified from the ESTs and had experimental evidence were then anchored to the exons to localise these events in the transcripts with respect to 5'UTR, CDS and 3'UTR regions. The EST-based coordinates were first mapped to RefSeq transcripts coordinates through the UCSC Table Browser resource which has genomic coordinates for both ESTs and RefSeq transcripts. Since Alu antisense events were identified in ESTs, they had been annotated for gene information using UniGene database. In order to ensure that the Alu antisense events on the ESTs were from alignable portions we anchored the antisense event for only those genes which were present in the RefSeq database. That is we left out all those antisense events where the EST stretch had no alignment information available. Using the EST genomic alignment block data we next converted the EST-based Alu antisense start-end positions into genomic coordinates. These were then overlapped with Alu-containing exon blocks from 5'UTR, CDS and 3'UTR. Through this exercise we could anchor and find the preference for Alu antisense events with respect to positions within the mRNA.



Supplementary Figure 2. Schema of analyses for identification of Alu antisense events and mapping to RefSeq coordinates

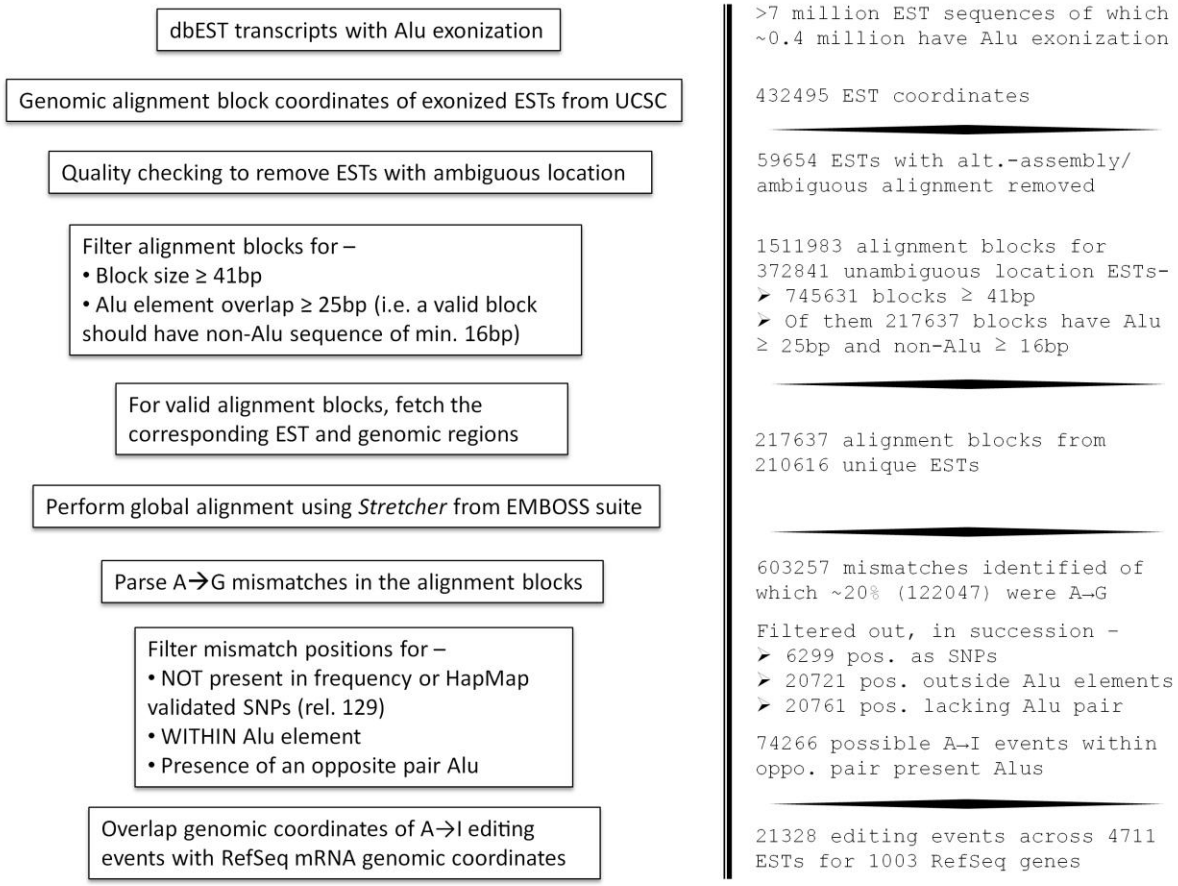
Alu antisense evidence was drawn by first determining computationally possible Alu antisense events (virtual tags) and then probing their evidence in Solexa datasets for experimental support. After quality checks and ensuring non-conflict with genic information, genomic coordinates for Alu antisense events were mapped onto RefSeq coordinates for exonization with respect to CDS and UTR positions. The panel on the left represents the steps involved and the right panel represents the numbers obtained in each of these steps.

Identification of Alu editing in the transcriptome

As described in the earlier section we selected dbEST for profiling A→I editing within Alu repeats. The same set of Alu exonized ESTs that was used for detecting antisense transcripts

were used in this analysis **(Supplementary Figure 3)**. The editing sites were identified through alignment of EST stretches with corresponding genomic regions through A→G mismatches. For this the alignment block coordinates for Alu exonized ESTs were retrieved. These were then quality filtered for ambiguous location i.e. alignments on multiple places of genome or onto alternate assemblies or onto unplaced contigs. We set criteria for the block size and the ratio of Alu versus non-Alu sequence in these alignment blocks. We set a criteria that a block should have a minimum block size of ≥41bp comprising 16bp of non-Alu sequence (4^{16} > Genome size, the probability of finding a 16bp sequence stretch more than once exceeds the genome size) and ≥25bp of Alu sequence. These block coordinates of these filtered sequences were then fetched from the paired genomic sequence and EST stretches. Using the *Stretcher* program from the EMBOSS suite we performed global alignment and parsed the output for mismatch positions. Since we were interested in A→I editing, for subsequent steps we profiled the A→G mismatch positions only. All mismatch positions were noted in genomic coordinates to aid downstream analysis steps. These A→G positions containing stretches were then successively filtered through a series of quality checks. These checks included that these positions are not SNPs (as revealed from frequency information in dbSNP) or HapMap validated SNPs (release 129) and must reside in Alu sequence. Alu editing is commonly reported in transcripts if there is another oppositely oriented Alu as the editing enzyme is specific for double stranded RNA. Therefore as additional criteria we also looked for an oppositely oriented Alu proximal to the edited sequence. Post-filtering, the set of A→G mismatch positions obtained finally were termed as possible A→I editing events within Alu elements.

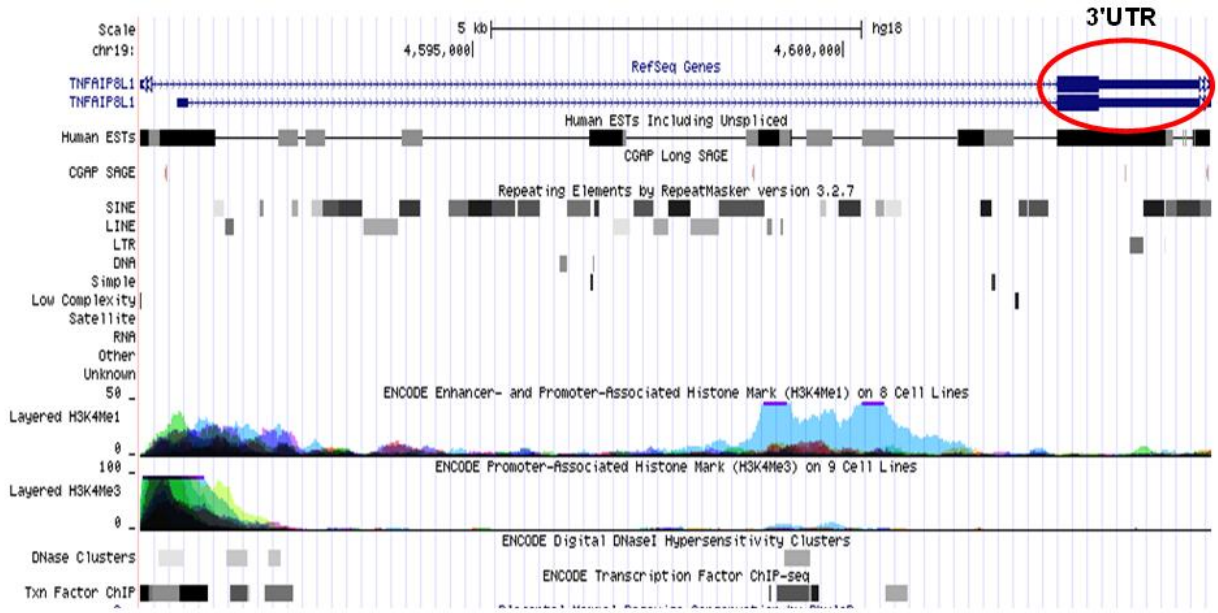
As the A→I editing events were studied initially in dbEST, as in the case of antisense, the preference of these events for 5'UTR, CDS or 3'UTR within exons were then further mapped as described earlier. To achieve this we overlapped the genomic coordinates of the possible Alu editing events with those of the Alu-containing exon blocks from 5'UTR, CDS and 3'UTR. By mapping editing positions onto genomic coordinates, we could, as in Alu antisense, study positional preference for these events within RefSeq mRNAs.



Supplementary Figure 3. Schema of analyses for identification of Alu editing events in the transcriptome

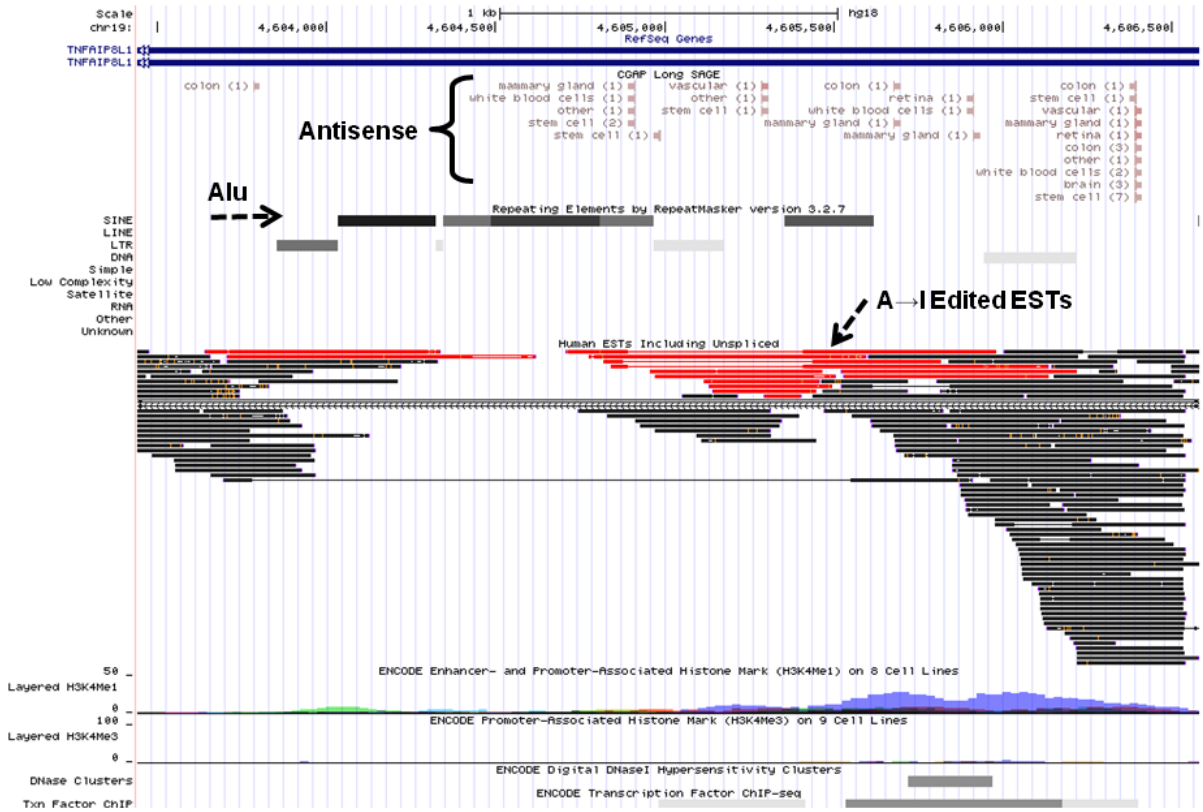
The comprehensive collection of transcripts in dbEST were first selected for Alu Exonization and then using UCSC alignment blocks, EST and corresponding genomic stretches were realigned to infer A→G mismatches. Post-processing for quality criteria, genomic coordinates for the possible A→I editing events were overlapped with RefSeq coordinates for Alu exonization with respect to CDS and UTR positions. The panel on the left represents the steps involved and the right panel represents the numbers obtained in each of these steps.

UCSC Genome Browser snapshots for Alu exonization, antisense and editing for a representative gene: *TNFAIP8L1*



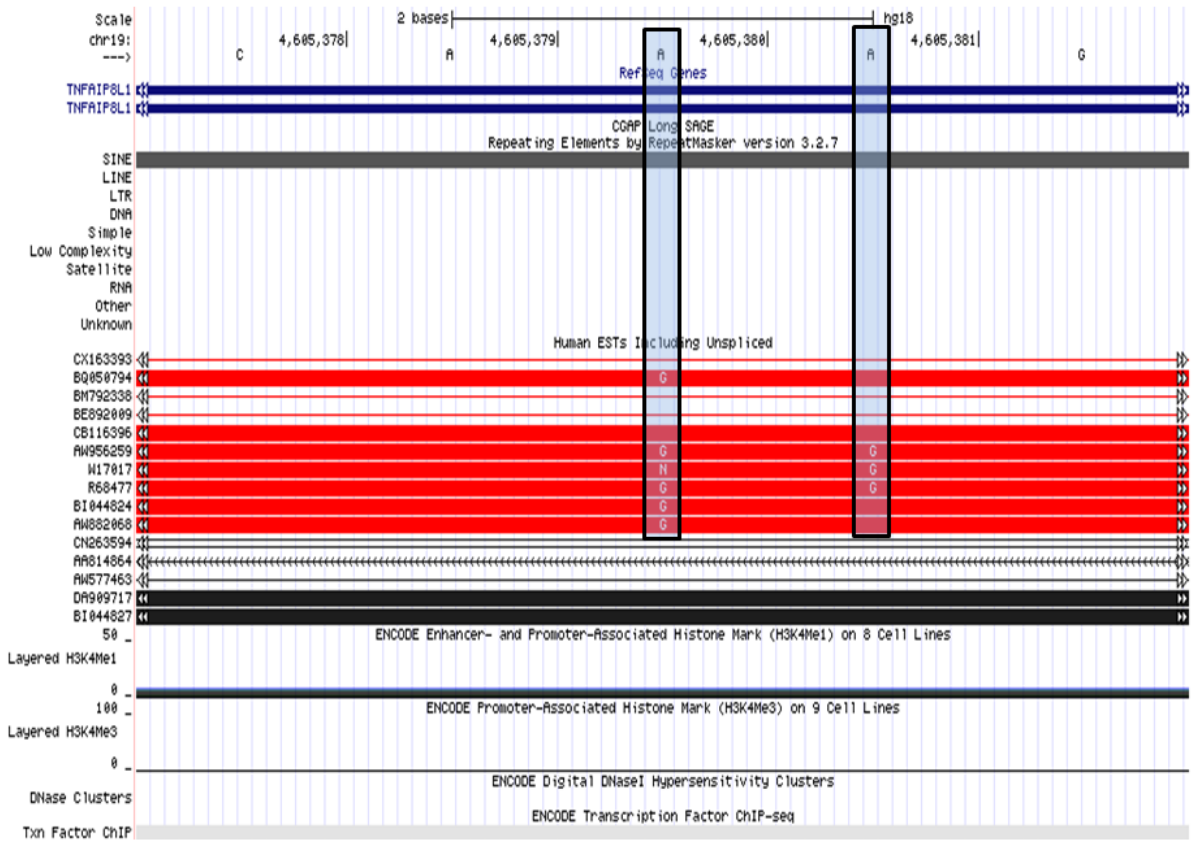
Supplementary Figure 4: Genome Browser snapshot for *TNFAIP8L1*.

In this snapshot, Alu harboring 3'UTR of the gene is encircled in red. The gene is in plus orientation with respect to the genome.



Supplementary Figure 5: 3'UTR of the gene zoomed-in to depict Alu mediated events.

Apart from the exonized Alu visible in the RepeatMasker track, the antisense and A→I edited information are also visible from the CGAP Long SAGE and Human ESTs tracks respectively, in the 3'UTR of the gene. The thin end of the SAGE tag points to the direction of the transcription. Hence, from the CGAP Long SAGE track, antisense tags overlapping Alu can be visualized. In the Human ESTs track, A→I edited ESTs have been marked in red.



Supplementary Figure 6: A→I editing within Alu in 3'UTR visualized.

Two of the A→I editing positions within the exonized Alu in the 3'UTR of the *TNFAIP8L1* gene has been visualized. In the Human ESTs track, the editing position can be seen as an A→G mismatch and the respective Alu in the RepeatMasker track.

Categorization of Exonization length for Pearson Chi-squared test

Class	Exonization length with incidence of both Editing and Antisense (bps)	Exonization length devoid of co-occurrence (bps)
5'UTR	2348	172643
CDS	256	19817
3'UTR	77956	1071978

Using this categorical information, Pearson Chi-squared test was performed in the R statistical package using the *chisq.test* command.