Multiple Alu exonizations in 3'UTR of a primate specific isoform of *CYP20A1* creates a potential miRNA sponge

Aniket Bhattacharya[1,2,6], Vineet Jha[3,6], Khushboo Singhal[1,2,6], Mahar Fatima[4], Dayanidhi Singh[1,2], Gaura Chaturvedi[1,2], Dhwani Dholakia[1,2], Rintu Kutum[1,2], Rajesh Pandey[1], Trygve E. Bakken[5], Pankaj Seth[4], Beena Pillai[1,2], Mitali Mukerji[1,2,*]

[1]Genomics and Molecular Medicine, CSIR-Institute of Genomics and Integrative Biology (CSIR-IGIB), Mathura Road, Delhi – 110025, India

[2]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad - 201002, India

[3]Persistent LABS, Persistent Systems Ltd., Pune, Maharashtra – 411004, India

[4]Department of Molecular and Cellular Neuroscience, Neurovirology Section, National Brain Research Centre (NBRC), Manesar, Haryana – 122051, India

[5]Allen Institute for Brain Science, Seattle, WA 98101, USA

[6]Equal contribution

*Corresponding author: mitali@igib.res.in

Keywords: Alu, Alu exonization, *CYP20A1*, miRNA sponge, miRNA recognition elements (MRE), 3'UTR, orphan gene

Running title: Gene sub-functionalization through Alu exonization

Abstract

Background: Alu repeats contribute to phylogenetic novelties in conserved regulatory networks in primates. Exaptation of Alus in transcript isoforms could nucleate large-scale mRNA-miRNA interactions and modulate cellular outcomes.

Result: Using a functional genomics approach, we report a transcript isoform of an orphan gene, *CYP20A1* (*CYP20A1_Alu-LT*) that arise through exonization of 23 Alus in 3'UTR and is expressed in higher primates. *CYP20A1_Alu-LT*, confirmed by 3'RACE, is an outlier in length (9kb) and is expressed in multiple cell lines. Using publicly available datasets, we demonstrate its presence in single nucleus RNA-seq of 15928 human cortical neurons (including rosehip neurons). miRanda predicts ~4700 miRNA recognition elements (MREs; with threshold< -25kcal/mol) for ~1000 miRNAs, which have primarily originated within the 3'UTR-Alus post exonization. *CYP20A1_Alu-LT* could be a potential multi-miRNA sponge as it harbours ≥10 MREs for 140 miRNAs and has cytosolic localization. In order to test this further, we explored whether expression of *CYP20A1_Alu-LT* correlates with genome wide mRNAs harboring similar MRE targets. We carried out RNAseq with conjoint miRNA-seq analysis in primary human neurons as we observed *CYP20A1_Alu-LT* to be downregulated during heat shock response and upregulated in HIV1-Tat treatment. *CYP20A1_Alu-LT* expression was positively correlated with 380 genes that were significantly downregulated in heat shock and upregulated in Tat and harboured MREs for a set of nine expressed miRNAs that were also enriched in *CYP20A1_Alu-LT*. The enrichment of MREs in the 380 genes were significant compared to random sets of expressed (p=4.716e-12) as well as differentially expressed genes (p=8.134e-12). Gene ontology suggested involvement of these genes in neuronal development and hemostasis pathways.

Conclusion: Our study suggests a potential role for *CYP20A1_Alu-LT* as miRNA sponge due to significant enrichment of MREs within Alus in a transcript isoform specific manner. This highlights a novel component of Alu-miRNA mediated transcriptional modulation that could govern specific physiological outcomes in higher primates.

Introduction

Nearly half of the human genome is occupied by transposable elements (TEs) (1). These have been shown to fine-tune conserved gene regulatory networks in a lineage specific manner (2–4) . Depending upon the context, they contribute to gene expression divergence through large scale transcriptional rewiring (3,5–7). Primate specific Alu retrotransposons, which occupy ~11% of the human genome, are one of the major players in this process (1,8). These provide non-canonical transcription factor binding sites and other regulatory sites that govern epigenetic modifications as well as provide cryptic splice sites that lead to alternative splicing or differential mRNA stability (5,9–15). Alu-derived exons exhibit lineage specificity with high transcript inclusion levels and have relatively higher rates of evolution (16–18).

Nearly 14% of the human transcripts contain at least one exonized Alu (19). Exonization is frequently reported in genes that have arisen *de novo* in primates, with most of the events in 3'UTRs (19,20). Such exonized Alus can increase the regulatory possibilities for a transcript, in a spatio-temporal manner, through antisense, miRNAs, Alu-MREs, A-to-I RNA editing, alternative splicing and enhancers. We have earlier shown how a crosstalk between these events could govern transcript isoform dynamics and modulate cellular outcomes (19,21). Besides, Alus provide substrates for other regulatory events such as gain of poly-A sites, AU-rich motifs and miRNA recognition elements (MREs) that can result in alternative polyadenylation, mRNA decay or translation stalling; and formation of specific secondary structures (10,22–26).

In our earlier study on 3177 Alu-exonized genes, we have reported co-occurrence of *cis*- Alu antisense (SAGE datasets) and A-to-I RNA editing (dbEST) marks at the level of single Alu exons (RefSeq) in 319 genes (19). Amongst these genes, during mapping of lineage specific events, we observed that a transcript isoform of *CYP20A1* has acquired an unusually long 3' UTR through exonization of 23 Alus. We report here a unique biological role of this Alu-exonized transcript isoform (*CYP20A1_Alu-LT),* through its extensive characterization. miRNA prediction analysis reveals that its 3'UTR contains predicted target sites for ~1000 miRNAs. We report for the first time that *CYP20A1_Alu-LT* could potentially function as a multi-miRNA sponge that has originated primarily from Alu elements. We demonstrate its regulatory potential in the presence of miRNAs in primary neurons through RNA-seq analysis in two different stress conditions where this transcript has opposite expression. *CYP20A1_Alu-LT* expression correlates with a set of 380 genes that share cognate MREs. These genes are majorly involved in neuro-inflammatory processes, suggesting that the synergistic role of an ensemble of miRNAs function could be modulated by *CYP20A1_Alu-LT.* This can add to the growing repertoire of evolution of lineage specific regulatory functions from transposable elements in the human transcriptome.

Results

CYP20A1 contains a unique 3'UTR with Alu-driven divergence

Literature mining revealed that out of 319 genes with Alu exonization, 91 genes map to apoptosis and nearly 75% of them cluster around three discrete hubs: cell cycle-DNA damage response (p53 hub; 31 genes), mitochondrial events (mito hub; 22 genes) and proteostasis (ubi hub; 15 genes) **(Supplementary Information S1 and Table S1)**. As the majority of these exonization events occur in the 3'UTRs of transcripts, which can modulate miRNA regulatory networks, we focused on identifying specific events in the 3'UTRs of these genes.

In the mitochondrial hub, we found a transcript isoform of *CYP20A1* gene (referred to as *CYP20A1_Alu-LT* hereafter) that has an 8.93kb long 3'UTR, 65% of which is derived from the exonization of 23 Alus **(Figure 1a)**. Since this transcript has an unusual density of Alus across the length of the UTR, we characterized it further for regulatory potential (**Supplementary information, S1**). Amongst the nine transcripts of human *CYP20A1* annotated in NCBI, experimental evidence (i.e., support from RNAseq reads or microarray probe signals) is available only for *CYP20A1_Alu-LT* (NM_177538.2), the longest isoform (10.94 kb). It is an outlier in terms of its 3'UTR length as it occupies the 85th position in the genome-wide length distribution of 3'UTRs **(Figure 1b)**. Such extended 3'UTRs are a rare event. Even among Alu-exonized genes (i.e., genes with Alu exonization events documented in one or more of their transcript isoforms), less than 3% have UTRs longer than 6kb **(Supplementary Figure S1)**. The length and enrichment do not seem to correlate with the density of exonized Alus (r = 0.25) and average of 5.42 exonization events/ 3'UTR in the transcriptome. The Alus in *CYP20A1_Alu-LT* belong to subfamilies of different evolutionary ages, suggesting that their insertion could have happened over a period, though we have not tested this hypothesis in this study.

Genomic region proximal to *CYP20A1_Alu-LT* 3'UTR is relatively well conserved

The coding region of *CYP20A1_Alu-LT* is remarkably well conserved among vertebrates, both at the sequence level as well as the length of the mature protein **(Table 1)**. The chimpanzee, macaque and mouse *CYP20A1* code for the conserved 462-470aa protein as in humans although their annotated transcript orthologs range between 1-3kb. Multiple sequence alignment across vertebrates reveals a strong conservation at both the N and the C terminals **(Supplementary Figure S2)**; of the first 100aa, 62 are completely conserved while 18 contain lineage specific substitutions with residues that have similar functional groups. This is also corroborated by the minimal evolutionary divergence across vertebrate CYP20A1 proteins **(Figure 1c)** and a strong purifying selection in CDS (Ka/Ks ~0.2 in mammals and <0.1 in non-mammalian vertebrates) **(Table 1)**.

On the other hand, 3'UTR extension in *CYP20A1_Alu-LT* seems to be majorly contributed by the primate specific insertion of Alus (65% of this UTR consists of Alu sequence). Its orthologs in mouse, rat and zebrafish are extremely short (within 1kb). In mouse, we observe a sparse presence

of two B1 SINEs, one each of simple repeat and low complexity repeat whereas the zebrafish 3'UTR lacks repeats altogether. The longest annotated *CYP20A1* transcripts for mouse (NM_030013.3), rat (NM_199401.1) and zebrafish (NM_213332.2) are 2.27, 2.03 and 1.79kb, respectively. The 5'UTR appears to be well conserved across the primate lineage (except lemur and proboscis monkey); however, the divergence in the 3'UTR, as evident from Jukes Cantor measure, increases as we move from the great apes to rhesus macaque and is primarily contributed by the Alus, with the breakpoints mostly coinciding with an Alu insertion **(Figure 1d)**. It shows maximum divergence from mouse that were treated as a non-primate evolutionary out-group. To control for the length difference between the 5' and 3'UTRs, we also checked for conservation in the 10kb region upstream of the first transcription start site (TSS) of *CYP20A1* and 10kb downstream of transcription end site (TES) and found it to be conserved among the higher primates, except for some New World monkeys **(Supplementary Figure S3)**. Taken together, these observations suggest that insertion of exonized Alus might have contributed to the specific divergence of this 3'UTR, in a genomic region that is otherwise conserved, at least among the higher primates.

Since this UTR seems to have appeared relatively late in the primate evolution, we tested if it carries variations that can differentiate modern human populations. Among the 23 SNPs in *CYP20A1* 3'UTR (16 within Alus), 11 have average heterozygosity scores >0.2, some of them as high as 0.48. We analyzed the data from 1000 Genomes Phase I and found significant differentiation for seven of these SNPs with global F_{ST} values ranging between 0.2-0.4 **(Supplementary Table S2)**. Interestingly, we also found a GWAS SNP (rs11888559, C/T, T=0.237/1187) in this UTR, which is associated with height in Filipino women (27) with a global F_{ST} of 0.36, rs11888559 differentiates the east Asian (CHB) and European (CEU) from the ancient African (YRI) population. It also exhibits high derived allele frequencies (DAFs) in these populations (0.81 and 0.95 in CHB and CEU, respectively). We also found another SNP rs7577078, within Alu, with high DAFs in all the three populations.

Characterization of *CYP20A1_Alu-LT*

We next investigated whether the full-length transcript containing this 3'UTR is actually transcribed. As two-third of this 3'UTR comprise repetitive sequences, it was slightly challenging to capture the full-length transcript in expression arrays or map it uniquely from sequencing reads. Moreover, there are differences in annotations regarding the full-length 3'UTR-containing isoform in various genomic portals **(Supplementary information, S1)**. Therefore, we designed eleven pairs of primers spanning the entire length of the transcript and experimentally confirmed the expression of *CYP20A1_Alu-LT*. We validated three of its amplicons by Sanger sequencing to negate spurious amplification from other Alu-rich loci in the genome (**Figure 2a**, Supplementary information, S1 for detailed method).

We observed variable expression of *CYP20A1_Alu-LT* in the six cell lines that we had initially tested **(Figure 2a)**. Since we had used cancerous cell lines, its expression could potentially be attributed to the aberrant transcriptional profiles in cancer (28). To delineate if *CYP20A1_Alu-LT*

expression is due to the cancerous state of the cells, we compared its expression in a neuroblastoma cell line (SK-N-SH) with those in primary neuron, glia (astrocyte) and neural progenitor cells (NPC). Neuroblastoma shares features with both mature neurons and NPCs, but is distinct from glia and we found that *CYP20A1_Alu-LT* expression differs significantly only between glia and SK-N-SH but not in neurons or NPCs **(Figure 2b)**. This suggests that our observations in the cancerous cell lines are unlikely to be artifactual.

We selected MCF-7, a breast adenocarcinoma cell line, for some of our subsequent experiments as it has been extensively used for drug screening and studying the effect of xenobiotics on different CYP family genes (29,30). The copy number of *CYP20A1* is not altered in this line (2n=2) (31). We performed 3'RACE to determine the exact transcription termination site for *CYP20A1_Alu-LT*. This was followed by nested PCR and amplicon sequencing to confirm *CYP20A1_Alu-LT* as a RNA transcript **(Figure 2c, Supplementary information, S3)**. Our findings are confirmed in TargetScan (release 7.2) which builds on the longest Gencode 3'UTR and reports even longer, 12.85kb UTR (ENST000000356079.4).

Exon skipping differentiates *CYP20A1_Alu-LT* from the protein coding isoforms

We observed that the expression of *CYP20A1_Alu-LT* is low although CYP20A1 protein is relatively abundant **(Supplementary Figure S4)**, suggesting that other isoforms may contribute to protein levels. When we compare *CYP20A1_Alu-LT* with the shorter 3'UTR containing isoforms, we observe a skipping of the sixth exon in this transcript. Using primers encompassing the sixth exon, we could distinguish between the transcripts; the larger isoform (i.e., *CYP20A1_Alu-LT*) corresponds to the 196bps amplicon and shorter ones to 277bps that also shows relatively higher expression **(Figure 2d)**.

In order to assess the relative contribution of different isoforms to the overall expression of *CYP20A1*, we used publicly available RNA-seq data from 15928 single nuclei derived from the different layers of the human cerebral cortex (32). NM_177538 (*CYP20A1_Alu-LT*) is expressed in 75% of the nuclei whereas all the other RefSeq isoforms are found in <1% (cut-off CPM≥50). There are 7038, 5134 and 1841 single nuclei in which NM_177538 (but no other isoform) is expressed with ≥10, 50 and 100 reads, respectively **(Supplementary Table S3)**. Interestingly, it is expressed in rosehip neurons - a highly specialized cell type in humans **(Supplementary Table S4)** (33).

Although the long 3'UTR transcript is annotated as the principal isoform, its expression level did not correlate with CYP20A1 protein which is relatively abundant in MCF-7 cells **(Supplementary Figure S4)**. Thus, we performed *in silico* translation of all *CYP20A1* isoforms in six-frames and compared them to the annotated human CYP20A1 protein. The two short 3'UTR isoforms matched – one perfectly and another with an additional amino acid stretch **(Figure 2d)**, but the *CYP20A1_Alu-LT* goes out of frame in the sixth and seventh exon and BLAST analysis of the human proteome does not report any hits with the truncated 24 amino acid peptide. Taken together, these data suggest that the *CYP20A1_Alu-LT* is unlikely to be coding for CYP20A1 protein and

may represent a novel non-coding transcript isoform originating from the same locus. This may be an example of evolutionary sub-functionalization of a gene into two different classes of transcripts that might have evolved for different functions **(Figure 2e)** (34).

CYP20A1_Alu-LT expression in non-human primates

Among the non-human primates, we did not find any annotated transcripts beyond 3kb from this locus. Our preliminary analyses of expression data, from public databases, of chimpanzee and macaque (prefrontal cortex, CD4+ T cells) did not yield any reads mapping to this 9kb 3'UTR. Subsequently, we checked for *CYP20A1_Alu-LT* expression in the reference transcriptomes of non-human primates (http://www.nhprtr.org) (35). Total RNA reads derived from 157 libraries of 14 non-human primate species show consistent mapping patterns on *CYP20A1* 3'UTR. Mapping is higher in the neighboring coding exons, but the pattern is consistent across different tissues and the number of reads comparable, with a slightly higher expression in kidney and lungs. In chimpanzee, reads are evenly distributed across the length of the entire 3'UTR; however, distribution is patchy in the other Old world monkeys (with peaks mostly in the non-repeat regions). Expression is minimal in New world monkeys (marmoset, squirrel monkey) and completely absent in lemur, although the adjoining coding exons show comparable expression, suggesting that *CYP20A1_Alu-LT* is expressed in the higher primates **(Supplementary Figure S5)**.

CYP20A1_Alu-LT 3'UTR as an evolving miRNA regulatory hub

Based on our earlier knowledge of 3'UTR exonized Alus providing novel miRNA binding sites, we explored whether the Alu-rich 3'UTR of *CYP20A1_Alu-LT* is also targeted by miRNAs (21). A query in miRTarBase (release 6.0) (36) revealed that *CYP20A1_Alu-LT* 3'UTR had predicted target sites for 169 miRNAs (supported by microarray/ sequencing data), of which 46 were listed as *functional miRNAs* in FuncMir (miRDB) **(Supplementary Table S5)**. Interestingly, ~50% of these are either primate-specific or human-specific miRNAs (microRNAviewer) (37). The occurrence of target sites for human-specific miRNAs in this recently evolved UTR made us carry out further in-depth analysis of miRNA recognition elements (MREs).

Since our 3'UTR has diverged across the vertebrate phylogeny, we did not consider algorithms that employ evolutionary conservation as prediction criteria. Many algorithms which predict target sites based on seed sequence matches also seem to have limitations as the length and position of the seed sequence is variable amongst miRNAs (38,39). In order to reduce false MRE prediction in non-conserved regions, we used miRanda that employs a two-step strategy: sequence complementarity, followed by thermodynamic stability of the predicted miRNA-mRNA duplex (40). Using stringent cut-off criteria, we obtained a total of 4742 MREs for 994 miRNAs, 4500 of which overlap with Alus (4382 MREs, if a conservative estimate of >50% overlap is considered) **(Supplementary Table S6)**. The MREs overlapping with Alu elements are considered as Alu-MREs.

These 4742 MREs span the entire length of the 3'UTR along with several high density pockets in Alu regions **(Figure 3a)**. The 23 exonized Alus mainly belong to Alu S and J family and are from 13 different subfamilies - AluSx, AluSp, AluSc, AluSz6, AluSq2, AluSx3, AluSc8, AluSx1, AluSz, AluSg, AluJo, AluJb and AluJr. Their presence into the 3'UTR of *CYP20A1_Alu-LT* in 5' to 3' direction is represented in Figure 3a from top to bottom of the circos plot. The 994 miRNAs were grouped on the basis of numbers of MREs present in the 3'UTR of *CYP20A1_Alu-LT*. MREs are grouped in range of 1-5, 6-10, 11-20, 21-43, for group 1 (G1), group 2 (G2), group 3 (G3) and group 4 (G4), respectively. The total numbers of miRNAs in each group are 702, 178, 92, 22 for G1, G2, G3 and G4, respectively. Only 2% of total miRNAs have MREs more than 20 (Group G4), whereas ~70% of the miRNAs were in the group G1 with ≤ 5 MREs. The miRNAs present in G4 are shown in figure 3a, with their number of MREs in 3'UTR written in parantheses. In the circos the presence of binding sites in each Alu with the MREs is plotted. Non-Alu region has been placed into one group. As this evident, majority of the sites map to Alus with only ~5% of miRNA binding sites in non-Alu regions**(Figure 3a)**.

It is plausible that the accumulation of so many Alu-MREs (miRNA binding sites present in Alu elements) in this 3'UTR has been due to the retrotransposition or recombination between Alus in pre-existing target sites. To test this possibility, we carried analysis on 1000 random sets of 23 Alus from the genome with matched length, composition and subfamily.We did not observe a similar distribution of Alu-MREs - only 0.5 and 4.2% of these random sets had MREs ≥4742 and 4500, respectively **(Figure 3b)**. All the 23 Alus on *CYP20A1_Alu-LT* UTR have diverged from the consensus sequences of their respective subfamilies; some of these substituted bases might have aided the creation of MREs within this UTR. Next, we queried for the distribution of MREs present in CYP-Alus in 2000 randomly picked Alus of each subfamily from the genome. We found MREs in CYP-Alus as outliers in all 13 subfamilies. However, when a similar distribution is plotted only for 3'UTR-Alus, these lie above the median (but within the distribution) for eight (out of 13) subfamilies viz., AluJo, Sc, Sc8, Sg, Sx, Sx1, Sx3 (1 out of 3) and Sz (**Supplementary information, S4**). Taken together, this data suggests that there are certain subsets of 3'UTR-Alus, at least for the 13 subfamilies analyzed, that have accumulated MREs. This suggests that the chance of Alus having retrotransposed into 3'UTRs with pre-existing MREs is extremely low and these MREs might have been created within Alus post exaptation. An alternative, albeit less likely, possibility is greater retention of MRE sites within 3'UTR Alus than elsewhere in the genome. However, even among the 3'UTRs, the propensity of Alu elements with high MRE content to occur in tandem in a single UTR is low. Taken together, there is a possibility that accumulation of MREs could potentiate its function as miRNA sponge for a regulatory network **(Figure 3c)**.

CYP20A1_Alu-LT isoform has the potential to function as a miRNA sponge

To determine if *CYP20A1_Alu-LT* can be a potential miRNA sponge, we characterized this 3'UTR further using bioinformatics and experimental approaches. Since enrichment of Alu repeats in long RNAs could drive their nuclear localization, we first checked for the localization of the *CYP20A1_Alu-LT* isoform (41). Using RT-qPCR in both nuclear and cytosolic fractions, we found

that it is predominantly localized to the cytosol - a feature observed in most sponges **(Figure 4a)**. A sponge RNA also typically contains 4 to 10 low binding energy MREs for a particular miRNA that are separated by a few nucleotides and is generally devoid of destabilizing RNA elements. In *CYP20A1_Alu-LT*, using a stringent cut-off for MRE prediction (binding energy≤ -25kcal/mol), we observed miRNAs with as many as 43 MREs and binding energy as low as -47kcal/mol. Out of the 994 miRNAs, 140 have ≥10 MREs and are distributed across the length of the UTR **(Figure 3a, Supplementary information, S4)**.

To screen for MREs that would efficiently dock miRNA without degrading the *CYP20A1_Alu-LT* transcript, we next checked for the presence of bulge within the MREs for the 23 prioritized miRNAs **(Table 2)** using miRanda with default parameters. We used twin criteria – a complete match in 6-mer (2nt-7nt) seed site and presence of mismatch or insertion at 9nt-12nt position. 6-mer sites with wobble base pairing were also retained as two wobble-pairs were maximally present in some of the MREs. We found five such sites for miR- 6724-5p, two each for miR-1254, miR-4767 and miR-3620-5p and one each for miR-941, miR-4446-3p, miR- 296-3p, miR-619-5p, miR-6842-3p and miR-1226-5p **(Table 3)**. At all these sites, we observed insertion in *CYP20A1_Alu-LT*, which suggests the possibility of a bulge formation in the transcript. This can potentially prevent the transcript from miRNA directed degradation and increase its efficiency to sequester miRNA molecules.

Potential sponge activity of *CYP20A1_Alu-LT* in primary neurons in response to heat shock and HIV1-Tat

To probe if the alteration in *CYP20A1_Alu-LT* level could affect expression of transcripts containing cognate MREs, we looked for conditions where it is likely to be altered. In these conditions the miRNA that targets these MREs should also be expressed. We anticipated that in conditions where there is a higher expression of the potential 'sponge' (*CYP20A1_Alu-LT*), the abundant MREs would sequester the miRNAs. This potentially would relieve its other cognate targets resulting in higher expression of those genes. Whereas in conditions where *CYP20A1_Alu-LT* is downregulated, the miRNA would be free to bind its cognate targets, thereby potentially reducing their expression **(Figure 3c)**.

We first queried for the expression of the 994 miRNAs having potential MREs in *CYP20A1_Alu-LT* from miRNA expression profiles available in public datasets. These experiments, mostly microarray based, showed low concordance across replicates and high variability across experiments **(Supplementary information, S1)**. So we tested this experimentally in MCF-7 and primary neurons. Since primary neurons preferentially express longer 3'UTRs, we hypothesised that it would be a good model to study miRNA-mediated regulation events (42,43). We carried out small RNA-seq and using a cut-off of at least 10 MREs on *CYP20A1_Alu-LT* 3'UTR and TPM value of 50, we obtained a set of 21 and 9 miRNAs in MCF-7 and neurons, respectively, of which seven were common to both **(Table 2)**.

Since *CYP20A1* has been identified as a candidate from a set of Alu exonized genes that map to apoptosis, we asked if this would respond to triggers that induce cell death. HIV1-Tat is a potent neurotoxin that kills ~50% more neurons compared to the vehicle control **(Figure 4b)**. Upon treating primary human neurons with HIV1 full length Tat protein, followed by 6 hours recovery, *CYP20A1_Alu-LT* was found to be significantly upregulated (1.65 fold). However, progenitor cells, which are immune to Tat **(Figure 4b)**, did not show any such trend **(Figure 4c)**. *CYP20A1_Alu-LT*'s 3'UTR also carries 17 potential binding sites for HSF1, 14 of them within Alus, which show positional conservation in agreement with previous study (44). This suggests that this transcript may also be amenable to antisense-mediated downregulation during heat shock response as demonstrated by an earlier work from our lab (44). We found *CYP20A1_Alu-LT* to be significantly downregulated (2.68 folds) in primary neurons upon heat shock, followed by 1hr recovery **(Figure 4d)**.

In order to query the expression of the other cognate targets of the nine prioritized miRNAs in these two conditions, we performed strand sepcific RNA-seq of primary neurons after these treatments. The expression of *CYP20A1_Alu-LT* in RNA-seq showed similar patterns of expression as observed in RT-qPCR, significantly downregulated by 2.68 folds (log2FC=-1.42) upon heat shock (HS) recovery and 1.21 folds upregulated (log2FC=0.28) during Tat response. The latter, however, did not cross our stringent statistical significance threshold. Out of the 3876 genes differentially expressed in HS or Tat, 380 exhibit positively correlated expression patterns as *CYP20A1_Alu-LT* **(Figure 5a, Supplementary Table S7)**. All of these 380 genes contain at least one MRE for one or more of the nine prioritized miRNAs and the majority of their MREs are canonical and not Alu-derived. On the contrary, *CYP20A1_Alu-LT* contains a total of 116 MREs for all these nine miRNAs combined **(Supplementary Table S7)**. There is a significant enrichment of MRE sites (for our nine prioritized miRNAs) in this set of 380 genes compared to all expressed genes (complete transcriptome; FPKM>2) or those with a significant differential expression (FPKM>2; FDR<0.05) (**Table 4**; Kolmogorov-Smirnov test). The MRE distribution in expressed genes and significantly differentially expressed genes were not significantly different. Similar results were obtained by comparing the median distribution of MREs in the 3 sets using Mann-Whitney U test (**Supplementary Figure S6**, **Table 4**). The abundance of MREs (for our 9 prioritized miRNAs), plotted against the FPKM values in significantly differentially expressed genes, exhibit a normal distribution (Shapiro-Wilk test, p<2.2e-16). Genes with high MRE counts have lower expression values and vice-versa (**Supplementary Figure S7**). To further validate the enrichment of MREs in the 380 genes, we performed Monte-Carlo simulation using one million random sets of 380 genes. The MRE densities in the 380 genes with expression pattern correlated to *CYP20A1*_Alu-LT, were outliers when plotted with the distributions derived from random sets (p-value 9.99999e-07), for all the miRNAs except miR-5096 (**Figure 5b)**. The 9 miRNAs studied had nearly similar distribution of MREs across genes (**Supplementary Figure S8**). Taken together these data suggest that the set of 380 genes represents potential cognate targets whose expression levels can be modulated by *CYP20A1*_Alu-LT through competing for the miRNAs targeting them.

According to Gene Ontology analysis for these 380 genes set using Toppfun (FDR <0.05) the top five processes were hemostasis (28 genes), axon guidance (25 genes), neutrophil degranulation (23 genes), platelet signaling, activation and aggregation (18 genes) and ECM organization (18 genes). Other processes include mRNA processing and mitochondria translation, metabolism, amino acid and nucleotide synthesis and antigen presentation **(Supplementary Table and Figure S9)**. Pathways from the Topfunn enrichment were manually grouped into major biological processes. For the nine miRNAs, analysis of target genes in the set of 380 genes and their pathways revealed blood coagulation to be the major biological process targeted by almost all miRNAs followed by neuron development **(Figure 5c).** This implies that *CYP20A1_Alu-LT* might be important in maintaining the homeostasis and fine tuning the neurological pathways.

Discussion

Exonized Alus create a unique 3'UTR in *CYP20A1_Alu-LT*

In this study, we report a transposable element (TE)-derived putative miRNA sponge from a transcript isoform of *CYP20A1* gene. The uniqueness of this isoform is that it is derived from 23 Alus that are exonized into its 9kb-long 3'UTR (against the transcriptomic average of 5.42/ 3'UTR). Although there are transcripts with even longer 3'UTRs, none of them have matching repeat content. After 23, the second highest number of exonized Alus present in a transcript's 3'UTR is 17. Second, all these 23 Alus bear A-to-I RNA editing marks and have *cis*- antisense mapping to them, suggesting that their presence can modulate this RNA's expression in a spatio-temporal manner (19,21). *CYP20A1_Alu-LT* isoform seems to have been neo-functionalized from a protein coding locus and its expression is detected in the higher primates. This could result in inclusion of the *CYP20A1* gene into new regulatory networks. Interestingly, our study in primary human neurons suggests that differential expression of this RNA could modulate expression of multiple modules of a regulatory network and synchronize specific outcomes in response to environmental cues.

We highlight both coding as well as non-coding transcripts from the *CYP20A1* locus. While the protein is highly conserved across the vertebrate phylogeny, we observe a novel transcript with a skipped exon that results in an out-of-frame CDS. It is not only expressed in 75% of single nuclei derived from different layers of the human cerebral cortex but also in rosehip neurons - a highly specialized cell type in humans. This is further corroborated by its presence exclusively in the higher primates. Further, the RNA isoform has a significantly longer UTR with multiple features that suggests it could be a hub of post-transcriptional regulatory events.

CYP20A1_Alu-LT: a potential miRNA sponge

CYP20A1_Alu-LT having ≥10 MREs for as many as 140 miRNAs stands out from naturally occurring sponges that have been so far have been mechanistically characterized. Most of them contain MREs for a single or a few related miRNA species (45–47). Among TEs, Alus contains the maximum number of MREs (48). We also observed that out of the 4742 total MREs more than

80% are within Alu. Involvement of Alu in modulating gene expression as well as modifying existing miRNA regulatory networks in a lineage specific manner has been highlighted. Our earlier work has reported the functional significance of MREs within 3'UTR Alus in fine tuning the p53 regulatory network during stress response (21). It has also been reported that, miR-1285-1 is processed from an Alu and predominantly targets exonized Alus (48) and primate specific miR-661 targets Alu-MREs in3'UTRs of *MDM2* and *MDM4* (49). Besides elevated levels of free Alu RNA can sequester miR-566 which correlates with cancer progression (50). Multiple exonized Alus in *CYP20A1_Alu-LT* 3'UTR can facilitate secondary structure that can additionally alter bioavailability for MREs. Future studies on secondary structure simulations would allow us to assess the availability and accessibility of these MREs. However, the presence of MREs for miRNAs raises the exciting possibility that it can exert a systemic effect through titrating the cellular levels of many miRNAs simultaneously.

_CYP20A1_Alu-LT_ sponge activity could modulate mRNA-miRNA networks in neuro-coagulopathy

We observe that expression of *CYP20A1_Alu-LT* in human primary neurons is inducible in response to HIV1-Tat and decreased during HS response. Expression of competing endogenous (ce)RNAs, such as sponge RNAs, is also tightly regulated and often specific to tissue, development stage or stress conditions (51,52). Since sponge RNA could titrate miRISC complexes, their expression could correlate with expression of mRNA having shared targets (53,54) **(Figure 3c)**. We have not yet looked at its turnover rates, nevertheless we make similar observations in a set of 380 genes which correlate with the expression pattern of *CYP20A1_Alu-LT* i.e., downregulated during heat-shock and upregulated upon Tat treatment. This set also shows a significant, non random enrichment of MREs for the 9 prioritized miRNAs, suggesting that it potentially represents cognate target genes whose levels can be competitively modulated by *CYP20A1*_Alu-LT by titrating the miRNAs that target them. The binding of these miRNAs and downstream effects on target expression levels is currently under experimental investigation in our lab.

The 380 genes map to processes that are involved in blood coagulation and neuronal pathways. Blood coagulation factors have been reported to affect pathophysiology of CNS via coagulation protein mediated signal transduction (55). This process is being linked to several neurodegenerative diseases including multiple sclerosis, cancer of the CNS, addiction and mental health (55). Although the exact biological role of *CYP20A1_Alu-LT* remains to be mechanistically elucidated, yet enrichment of coagulation pathways in gene set showing correlated expression with this transcript suggests that it may be involved in fine-tuning inflammatory responses in neuro-coagulopathy, a possibility for future studies.

Exposure to HIV1-Tat is known to cause axonal damage, loss of blood brain barrier integrity, changes in neurite outgrowth, etc. These are mediated by astrocyte activation, inflammatory cytokine expression, inducing mitochondrial injury and rearrangement of microtubules (ref). The

set of 380 genes which correlate with the expression pattern of *CYP20A1_Alu-LT* were also enriched in similar pathways like axon guidance, hemostasis, platelet activation and aggregation, ECM organization, regulation of actin cytoskeleton, antigen presentation, Golgi to ER transport and mitochondrial translation. In the light of our observations, it is possible that the changes observed upon Tat exposure could partly be mediated and synergised by the sponging effect of *CYP20A1_Alu-LT*. Upon activation by Tat, the sponge could titrate out the miRNA that target the 380 genes and hence modulate all the pathways simultaneously. Validation of *CYP20A1_Alu-LT* in the context of neuronal damage might shed further light on this plausible involvement. It could also play a role in normal neuronal functions through fine tuning expression in NPCs during neurogenesis, neuronal migration during differentiation, etc.

During the course of this study, we noticed reads mapping beyond the longest 3'UTR annotated in UCSC **(Supplementary Figure S10)** raising the possibility that isoforms with even longer 3'UTRs may be transcribed from this locus. Beyond its role as miRNA sponge, this UTR may have a role in sequestering RNA binding proteins and deplete their cellular reserves, thereby indirectly affecting other genes (56). Besides all the 23 Alus contains A to I editing marks, which could further contribute to phylogenetic novelties especially in the brain (57,58). These events, though dynamic, could further disrupt or create new miRNA binding sites from existing MREs, thereby increasing the regulatory repertoire. Other possibilities such as independent transcription events from this UTR or additional polyadenylation sites also cannot be ruled out. The impact of these events in case of *CYP20A1_Alu-LT,* remains a possibility for future investigation.

Conclusion

In this study, we postulate a novel dimension of its regulatory potential - that of creation of a miRNA sponge through Alu exaptations in the 3'UTR regions. *CYP20A1* provides an interesting model for studying Alu derived novel transcripts that can function as ceRNAs and co-regulate multiple genes in a network or cellular process. Thus, the addition of a lineage specific sponge could be a top-up on existing networks that modulate intermediate phenotypes such as neuro-coagulation. These could act as regulatory switches and in response to biological cues rapidly release or sequester miRNAs to govern specific cellular outcome.

Materials and Methods

Data mining

Characterization of a novel transcript isoform of *CYP20A1*

Extensive annotation of different transcript isoforms of *CYP20A1* was carried out using Ensembl, NCBI and UCSC. Details are provided in **Supplementary information S1**. For comparison of the length of 3'UTR of *CYP20A1_Alu-LT* with other 3'UTRs at genome-wide scale, the coordinates for human transcripts (NM and XM IDs) were downloaded from NCBI RefSeq version 74 (hg38). For every gene, only the longest 3'UTR was considered. The summary statistics for size distribution were calculated using R scripts. DNA sequence conservation across different species

was checked with UCSC genome browser using multiple alignment across 20 species generated by multiz (59). Both gaps as well as unaligned sequences were treated as 'missing' data. For protein conservation, CYP20A1 protein sequences from different species were taken from the top hits obtained in NCBI pBLAST by using the human protein as a reference. Multiple sequence alignment was performed using Clustal Omega (O 1.2.2). As described in Gautam et al. 2015 Ka/Ks ratio was calculated (see Supplementary information, S1 for details) (60).

CYP20A1_Alu-LT expression in non-human primates

We used publicly available chimp and macaque RNA-seq datasets from GEO [GSM1432846, 55, 65 (SRR1510158, 167, 177); GSM2265102, 4, 6 (SRR4012405, 08, 09, 13)]. Reads were mapped to both human and chimp/macaque 3'UTR to increase fidelity and mapping on housekeeping genes like *ACTB, GAPDH* and *EIF4A2* was also checked to control for data quality and mapping parameters. To query more expression datasets, we took advantage of the sequence differences in this transcript due to skipping of sixth exon. We performed BLAST against human datasets in SRA using a 289bp sequence reconstructed by joining exon 5 and 7. The hits were reconfirmed by alignment of reads to the 3'UTR.

RNA-Seq Reads from non-human primate reference transcriptome mapped on hg19 were exported as UCSC genome browser tracks. We additionally incorporated the stranded RNA-seq data generated as a part of this study to compare the expression level of this transcript between human and other non-human primates.

miRNA target prediction in *CYP20A1_Alu-LT*

miRNA target sites (MREs) on *CYP20A1* 3'UTR were predicted using miRanda (version 3.3a) (36), with the parameters set as follows: score threshold(-sc): 100, gap opening penalty(-go): -8, gap extension penalty(-ge): -2, binding energy(-en): -25kcal/mol., 'strict' (i.e., G:U pairs and gaps were not tolerated in the seed region). miRanda uses miRBase (which contains ~2500 miRNAs) for annotation. For bulge analysis, target prediction for 23 miRNAs on *CYP20A1* 3'UTR was performed using miRanda offline version 3.2a with default parameters (gap opening penalty= -8, gap extension= -2, score threshold= 50, energy threshold= -20kcal/mol, scaling parameter= 4).

MRE Enrichment Analysis

The abundance of MREs for our 9 prioritized miRNAs in the 380 *CYP20A1*_Alu-LT co-regulated genes (listed in **Supplementary table S7;** correspond to 399 unique transcripts) was compared to that in all expressed genes (cut-off: FPKM>2) and in genes with significant differential expression (FPKM>2, FDR<0.05) in our RNA-seq datasets. The MRE counts for each of the individual nine miRNAs were added to obtain a single value (total MREs) against each transcript 3'UTR, followed by statistical tests for the median (Mann-Whitney U-test) as well as for the distribution (Kolmogorov Smirnov test). p values less than 0.05 were considered statistically significant. Expressed genes were also binned by MRE count - individually for each miRNA - to evaluate if

the presence of MRE sites was correlated to their expression. The MRE density in 380 genes was compared to the distribution in one million random subsets (of 380 genes each) using Monte Carlo simulation. All of these analyses (averaged expression values, MRE prediction) were done at the transcript level and subsequently mapped to the gene.

Experimental

Expression analysis of *CYP20A1_Alu-LT* across diverse cell lines

RNA isolation, cDNA synthesis and RT-qPCR

Total RNA was isolated using TRIzol (Ambion, Cat. No. 15596-026) as per manufacturer's protocol and its integrity was checked on 1% agarose gel followed by Nanodrop quantification (ND1000, Nanodrop technologies, USA). cDNA was prepared from oligo(dT)-primed DNase-treated RNA (Invitrogen, Cat. No. AM1907) and SuperScript III RT (Invitrogen, Cat. No. 18080-044). RNA template was digested from the cDNA using 2 units of E. coli RNaseH (Invitrogen, Cat. No. 18021071). Primers were designed using Primer3 (version 4.0.0) and were synthesized by Sigma **(Supplementary information, S2)**. To ensure there was no spurious amplification, we designed two pairs of overlapping primers both on the 5' as well as 3'ends of our transcript of interest and included 'minus-RT' controls in every reaction. Additionally, we sequenced three amplicons (1, 5 and 10) to check the specificity of amplification **(Supplementary information, S1)**. BLASTN (NCBI; 2.4.0+) against the corresponding *in silico* predicted amplicons had revealed >95% sequence identity with an average query cover of 90%; BLAT against the whole genome (hg38) gave *CYP20A1* as the top hit in every case. RT-qPCR was performed using 2X SYBR Green I master mix (Kapa Biosystems, Cat. No. KK3605) and the reaction was carried out in Roche LightCycler 480 (USA) **(Supplementary information, S1)** Melting curves were confirmed to contain a single peak and the fold change was calculated by $^{\Delta\Delta}$Ct method. MIQE guidelines were followed for data analysis.

3'RACE for mapping the full length transcripts

cDNA for 3'RACE was prepared using RLM-RACE kit (Ambion, Cat. No. AM1700) with 1µg MCF-7 total RNA as per the manufacturer's recommendation. Nested PCR was performed with FP10 and an internal primer using the amplicon produced by FP9 and external primer. The product of this nested PCR was electrophoresed on 2% agarose gel and four major bands were observed, which were gel eluted using Qiaquick gel extraction kit (Qiagen, Cat. No. 28704) and subsequently sequenced. Details of the results are provided in the **Supplementary information, S3**.

Cell culture studies

MCF-7 cell line was procured from National Centre for Cell Sciences, (Pune, India) and cultured in GlutaMax-DMEM high glucose (4.5gm/l) (Gibco, Cat. No. 10569044) supplemented with 10% heat inactivated FBS (Gibco, Cat. No. 10082147), HEPES (Gibco, Cat. No. 11560496) and 1X antibiotic-antimycotic (Gibco, Cat. No. 15240096). The culture was maintained at 70-80%

confluency at 37°C, 5% CO^2. Cell line lineage was confirmed by STR profiling and cells were routinely screened for any contamination **(Supplementary information, S1)**.

Primary human neuron and astrocyte cultures comply with the guidelines approved by the Institutional Human Ethics Committee of NBRC as well as the Stem Cell and Research Committee of the Indian Council of Medical Research (ICMR) (Fatima et al. 2017). Briefly, neural progenitor cells (NPCs) derived from the telencephalon region of a 10-15 week old aborted foetus were isolated, suspended into single cells and plated on poly-D-lysine (Sigma, Cat. No. P7886) coated flasks. The cells were maintained in neurobasal media (Gibco, Cat. No. 21103049) containing N2 supplement (Gibco, Cat. No. 17502048), Neural Survival Factor 1 (Lonza, Cat. No. CC-4323), 25ng/ml bovine fibroblast growth factor (bFGF) (Sigma, Cat. No. F0291), 20ng/ml human epidermal growth factor (hEGF) (Sigma, Cat. No.E9644) and allowed to proliferate over one or two passages. The stemness of NPCs was functionally assayed by - i) formation of neurospheres, and ii) ability to differentiate into neurons or astrocytes. Additionally, NPCs were also checked for the presence of specific markers like Nestin. For commitment to the neuronal lineage, NPCs were starved of bFGF and EGF; with 10ng/ml each of PDGF (Sigma, Cat. No. P3326) and BDNF (Sigma, Cat. No. B3795) added to the media cocktail. Differentiation of NPCs to astrocytes required Minimum Essential Medium (MEM) (Sigma, Cat. No. M0268-10x) supplemented with 10% FBS. The process of neuronal differentiation completes in exactly 21 days; our experiments were completed within a week post-differentiation. Differentiated cultures of primary neurons and astrocytes were also checked for specific markers by immunostaining to determine the efficiency of the differentiation process **(Supplementary Figure S11)**.

Nuclear -cytosolic localization of *CYP20A1_Alu-LT*

Nuclear and cytosolic RNA were isolated using PARIS kit (Ambion, Cat. No. AM1921) as per manufacturer's protocol. Briefly, nearly 10 million cells were resuspended in a fractionation buffer, incubated on ice and centrifuged at 4°C to separate the nuclear and cytosolic fractions. The nuclear pellet was additionally treated with a cell disruption buffer before mixing with the 2X lysis/binding solution and absolute ethanol and passing through a column. The RNA was subsequently eluted in a hot elution buffer; quantified using Nanodrop and its integrity was checked on 1% agarose gel. Nuclear RNA contains an additional hnRNA band above the 28S rRNA band and is usually of lower yield than cytosolic RNA. RT-qPCR was done as described earlier using gene specific primers from 5'UTR and 3'UTR region.

Induction of Stress

Cells were gently washed once with 1X PBS and fresh media was replenished before treatment for accurate quantification of stress response. Heat shock was given at 45°C (±0.2) for 30 min in a water bath. Subsequent to the treatment, cells were transferred to 37°C/5% CO^2 for recovery and harvested after 1hr, 3hrs and 24hrs. For Tat treatment, full-length lyophilized recombinant HIV1 Tat protein was purchased from ImmunoDX, LLC (Woburn, Massachusetts, USA) and reconstituted in saline. The dosage for treatment was determined by drawing a 'kill curve' using

graded dose of Tat on neurons **(Supplementary Figure S12)**. Treatment was performed for 6hrs with 100ng/ml Tat and cells were either harvested just after the treatment or allowed to recover at 37°C and 5% CO^2 for another 6hrs prior to harvesting.

TUNEL Assay

The assay was performed with in situ Cell Death Detection kit, TMR red (Millipore Sigma, Cat. No. 12156792910). Nearly 20,000 cells were seeded per well (on coverslips) in 12-well plates. Post Tat treatment, cells were washed once with 1X PBS and fixed with 4% PFA, followed by three washes with 1X PBS, permeabilization and blocking with 4% BSA containing 0.5% Triton-X 100, incubation with TdT for 1hr in the dark and three washes with 1X PBS. Coverslips were then mounted on clean glass slides using hardset mounting media containing DAPI (Vectashield, Cat. No. H-1500). Six to eight random fields were imaged for each experimental group using AxioImager, Z1 microscope (Carl Zeiss, Germany). Fixed cells treated with 2 units of DNaseI (for 10 minutes at RT, followed by the addition of EDTA to stop the reaction) were used as a positive control in this experiment. TUNEL positive nuclei were scored using ImageJ software (NIH, USA). Minimum of 1000 cells were scored for each replicate.

RNA sequencing and small RNA sequencing

Detailed methods for library preparation for RNA-seq and small RNA-seq are provided in supplementary information (S1). Briefly, libraries for RNA-seq were made using 500ng of total RNA per sample and three biological replicates were taken per experimental condition. Libraries were prepared following Illumina's TruSeq stranded total RNA protocol. The final libraries were pooled, diluted and denatured to a final concentration of 8pM. Clusters were generated using TruSeq PE cluster kit V3-cBot-HS on cBot system, followed by paired end sequencing on HiSeq2000 using TruSeq SBS kit V3-HS (200 cycles). Libraries for small RNA sequencing were prepared using Illumina's TruSeq small RNA library preparation kit from 1µg total RNA. The libraries were normalized to 2nM, denatured and subjected to cluster generation on cBot using TruSeq SR cluster kit v3-cBOT-HS. Single read sequencing was performed on HiSeq2000 using TruSeq SBS kit v3-HS (50 cycles).

For RNA sequencing analysis, Fastq files were checked using FASTQC and overall Q score was >20 with no adapter contamination. Overrepresented sequences were not removed. Reads were mapped on hg38 using Tophat, followed by isoform quantification (Cufflinks) and collation (Cuffmerge). Overall read mapping rate was between 59-86.3% and concordant pair alignment ranged between 53.1 and 81.1%. Cuffdiff was used to calculate the differential expression (D.E.; calculated for each experimental condition against untreated). Summary of sample- wise RNAseq data is provided in **supplementary information (S3)**.

For the small RNA sequencing, the data were quality checked using FastQC (version 0.11.2), followed by adapter trimming by cutadapt (version 1.18) and reads were not discarded. As expected, around 95% of the adapter trimming events happened at the 3'end of the reads. Filtering

based on length and quality were carried out by cutadapt; Q30 reads with sequence length >15 but < 35nt were retained for mapping. Nearly 80% of the reads were retained after these filtering steps. Size distribution of the reads and k-mer position were (21-25, 28-32) and 26-28 respectively. Subsequently, these reads were mapped onto hg38 using Bowtie2. On average, 61% of the reads were uniquely mapped. miRDeep2 was run to obtain the read counts as TPM. Summary of sample-wise small RNAseq data is provided in supplementary information (S3).

Figure Legends

Figure 1: *CYP20A1* contains a unique 3'UTR with Alu-driven divergence. (a) UCSC tracks representing the four transcript isoforms of *CYP20A1* with varying 3'UTR length. Only isoform 1 (NM_177538) contains the full length 8932bp 3'UTR (*CYP20A1_Alu-LT*). The RepeatMasker track shows this 3'UTR harbours 23 Alu repeats from different subfamilies. (b) Genome-wide analysis of length distribution of 3'UTR reveals *CYP20A1_Alu-LT* to be an outlier. Mean and median 3'UTR lengths were 1553bp and 1007bp, respectively. (c) Cladogram of CYP20A1 protein sequence divergence among different classes of vertebrates. At the protein level this gene seems to have diverged minimally. Values within the brackets represent branch length (unit: substitutions per site) (d) DNA level conservation analysis of 5'UTR and 3'UTR among 20 mammals reveals that 5'UTR is well conserved among all primate lineages, suggesting that divergence is unique to 3'UTR. Repeat masker track shows the position of Alu elements in the UTR region. (also see **Supplementary Figure S3**).

Figure 2: *CYP20A1_Alu-LT* is expressed and may be a long non-coding RNA. (a) A schematic representation of the primers designed on the *CYP20A1_Alu-LT* to encompass 5'UTR and full length 3'UTR. To check for full length expression of transcript, cDNA from multiple cell lines of different tissue origin was used for amplification. Lanes-a-g are from Hela-S3, A549, HeLa, HEK293, MCF7, SK-N-SH and gDNA(positive control) respectively. Representative gel images of this isoform expression, via amplification from the starting of 5'UTR and the end of 3'UTR, are shown by primer pairs 1 and 10, respectively. Amplicons (1, 5 and 10) were also confirmed by Sanger sequencing. (b) RT-qPCR for *CYP20A1_Alu-LT* expression in cancerous and non cancerous cell types of neuronal origin. Fold change was calculated with respect to SK-N-SH, after normalization with the geometric mean of expression values from *β-actin*, *GAPDH* and *18S rRNA*. The error bars represent the SD of three biological replicates and the average of three technical replicates were taken for each biological replicate (** → $p<0.01$; Student's t test). (c) 3'RACE confirms the expression of the full length transcript. The schematic depicts the oligo(dT) (attached to a tag sequence) primed reverse transcription, followed by nested PCR. The amplification products corresponding to the bands below 900bp and above 700bp mapped to *CYP20A1_Alu-LT* 3'UTR, suggesting that the full length transcript is expressed in untreated MCF-7 cells (n=3). (d) Differentiating the *CYP20A1_Alu-LTR* transcript from other isoforms. The schematic in figure 1a highlights the skipped exon 6 and the position of flanking primers on shared exons in green color. The presence of at least two different types of transcripts was confirmed. 277bp amplicon corresponds to isoform(s) that contain exon 6, but have shorter 3'UTRs (isoforms 2 and 3 in **figure 2e**) and 196bp amplicon corresponds to the long-3'UTR isoform (isoform 1). None of the six translation frames of the long 3'UTR isoform match with the annotated protein. The amino acids marked in red are common to both isoform 2 and 3, blue exclusive to isoform 3 and green represents the sequence from isoform 1. (e) Schematic representation summarizing the differences between *CYP20A1* transcript isoform 1 (*CYP20A1_Alu-LT*) and isoforms 2 and 3.

Figure 3: *CYP20A1_Alu-LT* putative lncRNA may act as a miRNA sponge. (a) Circos plot representing the MREs for the 994 miRNAs on *CYP20A1_Alu-LT* 3'UTR. miRNAs are grouped on the basis of the number of MREs. 23 Alus in this 3'UTR contribute to 65% of its length and

are distributed throughout the UTR. Only 11% of miRNAs have MREs >10 (92 and 22 in G3 and G4, respectively). (b) Distribution of MREs for these 994 miRNAs on 1000 random sets of 23 length and subfamily matched Alu repeats. Only 6 sets contain MREs in range of 4701-4800 suggesting this is a non random phenomenon and MREs are created post Alu exaptations. Highlighted in green are sets with more than 4500 MREs. (c) Proposed model to demonstrate the effect of potential sponge activity of *CYP20A1_Alu-LT*. In the condition where it is highly expressed, it will recruit multiple miRISC complexes which could relieve the repression of cognate targets leading to their translation; whereas in case of its reduced expression, those miRISC complexes remain free to load on the cognate targets and affect translational repression or promote mRNA degradation. *CYP20A1_Alu-LT* has the potential to sponge multiple miRNAs at the same time thereby regulating a large repertoire of transcripts.

Figure 4: Features of *CYP20A1_Alu-LT* for being a potential sponge RNA. (a) Cytosolic localisation of *CYP20A1_Alu-LT* confirmed by RT-qPCR. Fold change was calculated with respect to total RNA, after internal normalization using the primers against spiked-in control. The error bars represent the SD of four independent experiments and the average of two technical replicates was used for each experiment. Quality controls for assessing the purity of cytosolic (*GAPDH*) and nuclear (*MALAT1*) fractions are also shown. The RT-qPCR data were analyzed in accordance with the MIQE guidelines (Bustin et al. 2009) **(Supplementary information, S3)**. (b) Late apoptotic cells in primary neurons and NPCs in response to HIV1-Tat treatment were scored by the number of TUNEL positive nuclei. Tat is neurotoxic and kills ~50% more neurons compared to the vehicle control (VC i.e., saline), whereas the difference is not statistically significant for NPCs (p-values 0.04 and 0.21 for primary neurons and NPCs, respectively, for Student's t-test assuming equal variance). The data represents the mean and SD of three independent experiments and >1000 nuclei were scored per condition for each experiment. (c and d) Expression of *CYP20A1_Alu-LT* in response to HIV1-Tat (c) and heat shock (d) treatment was assessed by RT-qPCR using both 5' and 3'UTR primers. The 3'UTR was found to be upregulated following 6hrs recovery after Tat treatment in neurons (p value = 0.035 *p value <0.05, Student's t-test), but not in NPCs (p value = 0.348) (c). It was also strongly downregulated in neurons (p value = 0.031) immediately after heat shock (HS+1hr recovery). This difference was not significant during recovery (p value = 0.310; HS+3hrs recovery) (d). In both these cases, the 5'UTR primer exhibits the same trend as the 3'UTR but does not qualify the statistical significance cut-off of p< 0.05. Fold change was calculated with respect to saline (vehicle) treatment, after internal normalization with the geometric mean of *GAPDH*, *ACTB* and *18S rRNA* in (c) and with respect to control (no heat shock treatment) cells, after internal normalization with the geomean of *GAPDH* and *ACTB* (d). The error bars represent the SD of three independent experiments and the average of 2-3 technical replicates was taken for each experiment.

Figure 5: Fold change (Log2FC values) of 380 genes.(a) Figure represents log2FC of a set of 380 genes upregulated in response to Tat treatment (red) and downregulated during heat shock recovery (green) in primary neurons, resonating with the trend exhibited by *CYP20A1_Alu-LT*. All the transcripts contain one or more MREs for the 9 miRNAs that can be potentially titrated by

sponge activity of *CYP20A1_Alu-LT* in neurons. These represent potential cognate targets whose expression can be regulated by *CYP20A1_Alu-LT* perturbation. Genes are plotted in order as **Supplementary table S7**. (b) Enrichment of MRE sites in the 380 gene-set compared to one million random sets of equal number of genes (Monte-Carlo simulations, p =9.99999e-07) (c) The heat map represents the top 5 biological processes targeted by each miRNA from pathway enrichment of 380 genes. Scale 0-5 is an arbitrary scale where 5 being the most targeted process.

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Funding: This work was supported by the Council of Scientific and Industrial Research grant MLP-901 to MM. Financial support from CSIR in the form of fellowships to AB and KS are acknowledged. VJ was supported by Persistent Systems LTD. GC and DD were supported by fellowships from the University Grants Commission (UGC) and Department of Biotechnology (DBT), respectively. MF and PS acknowledge the support of the facilities provided through the Distributed Information Centre at NBRC, Manesar, under the Biotechnology Information System Network (BTISNET) grant, DBT, India. MF was supported by a fellowship from CSIR, New Delhi and PS was partially supported by research grants from DBT and NBRC core funds.

Acknowledgements: The authors acknowledge Chitra Mohindar Singh Singal, NBRC for her help with the TUNEL assay, Parashar Dhapola for his help with data visualization in the initial phase of this work, Madiha Haider for data visualisation of miRNA-pathway enrichment, Dr. Amit Chaurasia for Jukes Cantor divergence analysis, Dr. Rakesh Dey for providing reagents and many fruitful discussions and Drs. Gaurav Ahuja and Debarka Sengupta (IIIT Delhi) for their help with the MRE-enrichment analyses. The authors would also like to acknowledge Mr. Raghunandanan MV and Mr. Amit Khulve at IT division CSIR- IGIB for their constant help and support for data upload in GEO.

Availability of data and materials:

(1) The raw data for small RNA and mRNA sequencing generated in this study have been submitted to GEO (GSE132447).

(2) dbGap link for the human temporal cortex (MTG) raw sequence reads https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001790.v1.p1

Author's contribution: MM and AB designed the study and co-wrote the manuscript along with KS and RP. AB performed conservation analyses, miRNA target prediction, cell culture and molecular biology experiments and helped in RNA seq data analysis. VJ analyzed mRNA and

small RNA seq data, ran miRNA target prediction on miRanda and helped in data visualization. KS performed molecular biology experiments, ran miRNA bulge analysis with DD and contributed to improving data visualization along with RK. MF carried out primary human neuron and NPC culture under the supervision of PS. DS performed some of the cellular assays. GC prepared NGS libraries and carried out sequencing, assisted by AB and KS. TEB analyzed the single nuclei RNA-seq data. BP contributed reagents, helped in troubleshooting experiments and provided critical inputs. MM supervised the overall study.

Conflict of interest: All authors have read and approved the final version of the manuscript. Authors declare no conflict of interest.

References:

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860–921.
2. Chen H, Chen L, Wu Y, Shen H, Yang G, Deng C. The Exonization and Functionalization of an Alu-J Element in the Protein Coding Region of Glycoprotein Hormone Alpha Gene Represent a Novel Mechanism to the Evolution of Hemochorial Placentation in Primates. Mol Biol Evol. 2017 Dec 1;34(12):3216–31.
3. Trizzino M, Park Y, Holsbach-Beltrame M, Aracena K, Mika K, Caliskan M, et al. Transposable elements are the primary source of novelty in primate gene regulation. Genome Res. 2017;27(10):1623–33.
4. Wang L, Rishishwar L, Mariño-Ramírez L, Jordan IK. Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. Nucleic Acids Res. 2017 Mar 17;45(5):2318–28.
5. Lynch VJ, Leclerc RD, May G, Wagner GP. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat Genet. 2011 Sep 25;43(11):1154–9.
6. Rebollo R, Romanish MT, Mager DL. Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. Annual Review of Genetics. 2012;46(1):21–42.
7. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 2016 Mar 4;351(6277):1083–7.
8. Grover D, Kannan K, Brahmachari SK, Mukerji M. ALU-ring elements in the primate genomes. Genetica. 2005 Jul;124(2–3):273–89.
9. Sorek R, Ast G, Graur D. Alu-containing exons are alternatively spliced. Genome Res. 2002 Jul;12(7):1060–7.
10. An HJ, Lee D, Lee KH, Bhak J. The association of Alu repeats with the generation of potential AU-rich elements (ARE) at 3' untranslated regions. BMC Genomics. 2004 Dec 21;5(1):97.
11. Polak P, Domany E. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. BMC Genomics. 2006 Jun 1;7:133.
12. Xie H, Wang M, Bonaldo M de F, Smith C, Rajaram V, Goldman S, et al. High-throughput sequence-based epigenomic analysis of Alu repeats in human cerebellum. Nucleic Acids Res. 2009 Jul;37(13):4331–40.
13. Bakshi A, Herke SW, Batzer MA, Kim J. DNA methylation variation of human-specific Alu repeats. Epigenetics. 2016 Feb 18;11(2):163–73.
14. Tristán-Flores FE, Guzmán P, Ortega-Kermedy MS, Cruz-Torres G, de la Rocha C, Silva-Martínez GA, et al. Liver X Receptor–Binding DNA Motif Associated With Atherosclerosis-Specific DNA

Methylation Profiles of Alu Elements and Neighboring CpG Islands. J Am Heart Assoc [Internet]. 2018 Jan 31 [cited 2019 Apr 6];7(3). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5850253/
15. Payer LM, Steranka JP, Yang WR, Kryatova M, Medabalimi S, Ardeljan D, et al. Structural variants caused by Alu insertions are associated with risks for many human diseases. Proc Natl Acad Sci U S A. 2017 May 16;114(20):E3984–92.
16. Lin L, Shen S, Tye A, Cai JJ, Jiang P, Davidson BL, et al. Diverse splicing patterns of exonized Alu elements in human tissues. PLoS Genet. 2008 Oct 17;4(10):e1000225.
17. Sorek R. When new exons are born. Heredity. 2009 Oct;103(4):279–80.
18. Shen S, Lin L, Cai JJ, Jiang P, Kenkel EJ, Stroik MR, et al. Widespread establishment and regulatory impact of Alu exons in human genes. PNAS. 2011 Feb 15;108(7):2837–42.
19. Mandal AK, Pandey R, Jha V, Mukerji M. Transcriptome-wide expansion of non-coding regulatory switches: evidence from co-occurrence of Alu exonization, antisense and editing. Nucleic Acids Res. 2013 Feb 1;41(4):2121–37.
20. Toll-Riera M, Castelo R, Bellora N, Albà MM. Evolution of primate orphan proteins. Biochem Soc Trans. 2009 Aug;37(Pt 4):778–82.
21. Pandey R, Bhattacharya A, Bhardwaj V, Jha V, Mandal AK, Mukerji M. Alu-miRNA interactions modulate transcript isoform diversity in stress response and reveal signatures of positive selection. Sci Rep. 2016 02;6:32348.
22. Sobczak K, Krzyzosiak WJ. Structural determinants of BRCA1 translational regulation. J Biol Chem. 2002 May 10;277(19):17349–58.
23. Roy-Engel AM, El-Sawy M, Farooq L, Odom GL, Perepelitsa-Belancio V, Bruch H, et al. Human retroelements may introduce intragenic polyadenylation signals. Cytogenet Genome Res. 2005;110(1–4):365–71.
24. Häsler J, Strub K. Alu elements as regulators of gene expression. Nucleic Acids Res. 2006;34(19):5491–7.
25. Häsler J, Samuelsson T, Strub K. Useful “junk”: Alu RNAs in the human transcriptome. Cell Mol Life Sci. 2007 Jul;64(14):1793–800.
26. Lee JY, Ji Z, Tian B. Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3’-end of genes. Nucleic Acids Res. 2008 Oct;36(17):5581–90.
27. Croteau-Chonka DC, Marvelle AF, Lange EM, Lee NR, Adair LS, Lange LA, et al. Genome-Wide Association Study of Anthropometric Traits and Evidence of Interactions With Age and Study Year in Filipino Women. Obesity. 2011 May;19(5):1019–27.
28. Suzuki A, Makinoshima H, Wakaguri H, Esumi H, Sugano S, Kohno T, et al. Aberrant transcriptional regulations in cancers: genome, transcriptome and epigenome analysis of lung adenocarcinoma cell lines. Nucleic Acids Res. 2014 Dec 16;42(22):13557–72.
29. Coumoul X, Diry M, Robillot C, Barouki R. Differential Regulation of Cytochrome P450 1A1 and 1B1 by a Combination of Dioxin and Pesticides in the Breast Tumor Cell Line MCF-7. Cancer Res. 2001 May 15;61(10):3942–8.
30. Ptak A, Ludewig G, Rak A, Nadolna W, Bochenek M, Gregoraszczuk EL. Induction of cytochrome P450 1A1 in MCF-7 human breast cancer cells by 4-chlorobiphenyl (PCB3) and the effects of its hydroxylated metabolites on cellular apoptosis. Environ Int. 2010 Nov;36(8):935–41.
31. Pan S-T, Xue D, Li Z-L, Zhou Z-W, He Z-X, Yang Y, et al. Computational Identification of the Paralogs and Orthologs of Human Cytochrome P450 Superfamily and the Implication in Drug Discovery. International Journal of Molecular Sciences. 2016 Jun 28;17(7):1020.
32. Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, Graybuck LT, et al. Conserved cell types with divergent features in human versus mouse cortex. Nature. 2019 Sep;573(7772):61–8.
33. Boldog E, Bakken TE, Hodge RD, Novotny M, Aevermann BD, Baka J, et al. Transcriptomic and morphophysiological evidence for a specialized human cortical GABAergic cell type. Nat Neurosci. 2018;21(9):1185–95.
34. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes

by complementary, degenerative mutations. Genetics. 1999 Apr;151(4):1531–45.
35. Pipes L, Li S, Bozinoski M, Palermo R, Peng X, Blood P, et al. The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. Nucleic Acids Research. 2013 Jan 1;41(D1):D906–14.
36. Hsu S-D, Lin F-M, Wu W-Y, Liang C, Huang W-C, Chan W-L, et al. miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res. 2011 Jan;39(Database issue):D163-169.
37. Kiezun A, Artzi S, Modai S, Volk N, Isakov O, Shomron N. miRviewer: a multispecies microRNA homologous viewer. BMC Research Notes. 2012;5(1):92.
38. Ellwanger DC, Büttner FA, Mewes H-W, Stümpflen V. The sufficient minimal set of miRNA seed types. Bioinformatics. 2011 May 15;27(10):1346–50.
39. Mullany LE, Herrick JS, Wolff RK, Slattery ML. MicroRNA Seed Region Length Impact on Target Messenger RNA Expression and Survival in Colorectal Cancer. PLoS One [Internet]. 2016 Apr 28 [cited 2019 Apr 2];11(4). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4849741/
40. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. Genome Biol. 2003;5(1):R1.
41. Lubelsky Y, Ulitsky I. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. Nature. 2018 Mar 1;555(7694):107–11.
42. Miura P, Shenker S, Andreu-Agullo C, Westholm JO, Lai EC. Widespread and extensive lengthening of 3' UTRs in the mammalian brain. Genome Res. 2013 May;23(5):812–25.
43. Tushev G, Glock C, Heumüller M, Biever A, Jovanovic M, Schuman EM. Alternative 3′ UTRs Modify the Localization, Regulatory Potential, Stability, and Plasticity of mRNAs in Neuronal Compartments. Neuron. 2018 May 2;98(3):495-511.e6.
44. Pandey R, Mandal AK, Jha V, Mukerji M. Heat shock factor binding in Alu repeats expands its involvement in stress through an antisense mechanism. Genome Biology. 2011 Nov 23;12(11):R117.
45. Franco-Zorrilla JM, Valli A, Todesco M, Mateos I, Puga MI, Rubio-Somoza I, et al. Target mimicry provides a new mechanism for regulation of microRNA activity. Nature Genetics. 2007 Aug;39(8):1033–7.
46. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 2010 Jun;465(7301):1033–8.
47. Hu H, Liu J-M, Hu Z, Jiang X, Yang X, Li J, et al. Recently Evolved Tumor Suppressor Transcript TP73-AS1 Functions as Sponge of Human-Specific miR-941. Mol Biol Evol. 2018 01;35(5):1063–77.
48. Spengler RM, Oakley CK, Davidson BL. Functional microRNAs and target sites are created by lineage-specific transposition. Hum Mol Genet. 2014 Apr 1;23(7):1783–93.
49. Hoffman Y, Pilpel Y, Oren M. microRNAs and Alu elements in the p53-Mdm2-Mdm4 regulatory network. J Mol Cell Biol. 2014 Jun;6(3):192–7.
50. Di Ruocco F, Basso V, Rivoire M, Mehlen P, Ambati J, De Falco S, et al. Alu RNA accumulation induces epithelial-to-mesenchymal transition by modulating miR-566 and is associated with cancer progression. Oncogene. 2018 Feb;37(5):627–37.
51. Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, Chinappi M, et al. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell. 2011 Oct 14;147(2):358–69.
52. Tay Y, Rinn J, Pandolfi PP. The multilayered complexity of ceRNA crosstalk and competition. Nature. 2014 Jan;505(7483):344–52.
53. Ebert MS, Sharp PA. MicroRNA sponges: Progress and possibilities. RNA. 2010 Nov;16(11):2043–50.
54. Ebert MS, Neilson JR, Sharp PA. MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. Nat Methods. 2007 Sep;4(9):721–6.

55. De Luca C, Virtuoso A, Maggio N, Papa M. Neuro-Coagulopathy: Blood Coagulation Factors in Central Nervous System Diseases. Int J Mol Sci [Internet]. 2017 Oct 12 [cited 2019 Apr 2];18(10). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5666810/
56. Kelley DR, Hendrickson DG, Tenen D, Rinn JL. Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. Genome Biol. 2014 Dec 3;15(12):537.
57. Paz-Yaacov N, Levanon EY, Nevo E, Kinar Y, Harmelin A, Jacob-Hirsch J, et al. Adenosine-to-inosine RNA editing shapes transcriptome diversity in primates. PNAS. 2010 Jul 6;107(27):12174–9.
58. Rosenthal JJC, Seeburg PH. A-to-I RNA editing: effects on proteins key to neural excitability. Neuron. 2012 May 10;74(3):432–9.
59. Blanchette M. Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. Genome Research. 2004 Apr 1;14(4):708–15.
60. Gautam P, Chaurasia A, Bhattacharya A, Grover R, Indian Genome Variation Consortium, Mukerji M, et al. Population diversity and adaptive evolution in keratinization genes: impact of environment in shaping skin phenotypes. Mol Biol Evol. 2015 Mar;32(3):555–73.

Tables

Organism	Scientific name	Tax ID	% Identity	Query cover	Protein ID	Length (aa)	Ka/Ks	P-Value (Fisher)

Chimpanzee	*Pan troglodytes*	9598	99	100	XP_516042.2	462	0.711	0.499867
Bonobo	*Pan paniscus*	9597	99	100	XP_003820821.1	462	0.497	0.266782
Gorilla	*Gorillla gorilla*	9595	82	100	XP_004033127.1	417	1.014	1
Orangutan	*Pongo abelii*	9601	99	100	XP_002812812.1	462	0.242	0.0078154
Rhesus macaque	*Macaca mulatta*	9544	96	100	EHH21600.1	470	0.370	0.0042967
House mouse	*Mus musculus*	10090	83	100	NP_084289.1	462	0.216	4.77E-35
Rat	*Rattus norvegicus*	10116	82	100	NP_955433.1	462	0.194	1.19E-41
Cow	*Bos taurus*	9913	91	100	NP_001015644.1	462	0.163	1.28E-27
Grey wolf	*Canis lupus familiaris*	9615	91	96	XP_003434295.2	613	0.184	4.38E-29
Chicken	*Gallus gallus*	9031	74	99	XP_426572.2	463	0.064	1.58E-207
Anole lizard	*Anolis carolinensis*	28377	76	99	XP_003223588.2	557	0.080	2.87E-172
Xenopus	*Xenopus (Silurana) tropicalis*	8364	73	99	NP_001039140.1	463	0.054	0
Zebra fish	*Danio rerio*	7955	64	99	NP_998497.1	462	0.080	0
Drosophila	*Drosophila melanogaster*	7227	23	97	NP_573003.2	495	0.490	8.73E-100
Sea urchin	*Strongylocentrotus purpuratus*	7668	32	99	XP_792896.2	475	0.320	0
Arabidopsis	*Arabidopsis thaliana*	3702	22	94	BAA28539.1	500	0.559	1.08E-49

Table 1: CYP20A1 protein sequence conservation across vertebrates.CYP20A1 protein sequences among different vertebrate classes were compared using NCBI pBLAST. Drosophila, sea urchin and Arabidopsis were used as the evolutionary outgroups. All the pairwise comparisons are done with respect to the 462aa human CYP20A1 protein (NP_803882) and a query cover of ~95% was obtained in each case. CYP20A1 is well conserved among vertebrates. Except for the three great apes (chimpanzee, bonobo and gorilla), all the Ka/Ks comparisons are significant (Fisher's exact test; $p<0.01$; marked in red). Lesser the value of Ka/Ks, the more stringent is the negative selection operative on the protein i.e., fewer non-synonymous substitutions are tolerated in it.

miRNA	**Expression(TPM)**	

	MCF-7	Pr. Neurons	MREs with binding energy ≤ -25Kcal/mol
miR-941	119639.3	950.33	10
miR-3677-3p	2892.33	51.67	12
miR-1304-3p	1922.5	80.33	10
miR-4446-3p	1839.33	-	13
miR-296-3p	1406.5	70.33	10
miR-1254	1235	-	10
miR-6724-5p	330.17	-	20
miR-619-5p	193.17	89	26
miR-1908-3p	191.67	-	16
miR-3944-3p	158.67	-	14
miR-6842-3p	158.17	175.67	10
miR-4767	129.83		18
miR-5096	98.5	72.33	14
miR-7703	97.17	-	13
miR-939-5p	81.83	-	19
miR-3620-5p	81.33	-	21
miR-1226-5p	81.17	-	18
miR-1915-5p	78.5	-	14
miR-6732-3p	57.5	-	13
miR-1273g-3p	56.67	-	11
miR-4707-3p	53	-	10
miR-668-3p	-	175.33	10
miR-370-3p	-	244.33	14

Table 2: List of 23 prioritized miRNAs.

miRNAs were prioritized based on their expression level (≥50 TPM), number of MREs ≥10 with binding energy ≤ 25kcal/mol.

* expression values <50 TPM have not been represented.

miRNA					Features of MREs

	Total no. of MREs on *CYP20A1_Alu-Lt* 3'UTR	**Average of Overall complementarity of miRNA with *CYP20A1_Alu-Lt* 3'UTR**	**MRES with seed match(2-7 nt) including wobble**	**Number of MREs with mismatch or insertion at bulge position**	**Forward score**	**Binding energy**	**Alignment length**
miR-941	29	59.37	2	1	146	-22.05	21
miR-3677-3p	57	54.14	3				
miR-1304-3p	28	59.57	12				
miR-4446-3p	34	60.56	3	1	120	-27.05	22
miR-296-3p	42	54.76	5	1	146	-25.23	25
miR-1254	45	59.72	18	2	126,128	-22.05, -23.75	25,23
miR-6724-5p	105	44.18	19	5	99,99,103,107, 126	-21.14, -23.87, -26.94, -22.16, -23.6	22,22,22,21, 18
miR-619-5p	45	67.17	27	1	115	-21.76	20
miR-1908-3p	43	52.60	1				
miR-3944-3p	64	54.82	23				
miR-6842-3p	39	59.20	11	1	104	-24.14	29
miR-4767	72	54.71	9	2	101,124	-25.3, -23.93	27, 16
miR-5096	22	65.80	8				
miR-7703	51	57.20	4				
miR-939-5p	63	51.32	9				
miR-3620-5p	77	53.71	32	2	108,130	-21.24, -22.89	22,20
miR-1226-5p	73	56.95	6	1	124	-26.76	29
miR-1915-5p	40	54.88	1				
miR-6732-3p	35	60.37	0				
miR-1273g-3p	30	67.30	12				
miR-4707-3p	30	58.03	6				
miR-668-3p	38	52.86	12				
miR-370-3p	56	56.57	2				

Table 3: Features of MREs with seed site match and presence of bulge.

Test	Kolmogorov-Smirnov Test		Mann-Whitney U-test	
Gene Set	380 genes	Expressed genes (FPKM>2)	380 genes	Expressed genes (FPKM>2)
Expressed genes (FPKM>2)	5.102e-07		4.716e-12	
Differentially expressed genes (FPKM>2; FDR<0.05)	4.311e-07	0.9925	8.134e-12	0.7505

Table 4: 380 genes (expression correlated with *CYP20A1*_Alu-LT) are significantly enriched in MREs compared to all expressed genes and those with significant differential expression