Research Article

Origin of a novel *CYP20A1* lncRNA through 23 Alu exaptations in the human lineage creates a potential multi-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], 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, email: mitali@igib.res.in

Keywords: Alu, *CYP20A1*, lncRNA, miRNA sponge, neofunctionalization, orphan gene

Running title: Gene subfunctionalization through Alu exaptation: birth of a novel lncRNA from a conserved gene

Abstract

Alu repeats contribute to lineage specific novelties in conserved transcriptional regulatory networks. We report for the first time the origin of a multi-miRNA human specific sponge through exaptation of 23 Alu repeats that forms a novel principal isoform of *CYP20A1* gene with a 9kb 3'UTR. This 3'UTR, confirmed by RACE, is an outlier in terms of its length, with expression in multiple cell lines including brain as evidenced from single nucleus RNA-seq data of ~16000 human cortical neurons. It has diverged from its parent gene through exon skipping into a novel lncRNA. Its uniqueness in humans was validated by its presence in rosehip neurons and absence in closely related primate species through RNA-seq datasets. Strikingly, prediction by miRanda revealed ~4700 MREs for ~1000 different miRNAs, majorly in Alu repeats, in this 3'UTR. Permutations on 1000 random sets suggest their creation is non-random and post Alu exaptation. We hypothesise this lncRNA is an miRNA sponge as it has cytosolic localization and harbors ≥10 MREs for 140 miRNAs (threshold > -25kcal/mol). Under experimental conditions where *CYP20A1* displays differential expression, we probed the expression of the miRNAs that map to this 3'UTR and their cognate targets through small RNA and mRNA-seq, respectively. We observed correlated expression of our lncRNA and a set of 380 genes with downregulation in heat shock and upregulation in HIV1-Tat treatment in primary neurons. GO analyses suggest the involvement of this sponge lncRNA in modulation of processes linked to neuronal development and hemostasis pathways.

Introduction

Transposable elements (TEs), comprising half of th**e** human genome, fine-tune conserved gene regulatory networks in a lineage specific manner (Lander et al. 2001; Chen et al. 2017; Trizzino et al. 2017; Wang et al. 2017). Depending upon the genomic context, TEs contribute to gene expression divergence through large scale transcriptional rewiring (Lynch et al. 2011; Rebollo et al. 2012; Trizzino et al. 2017). These could be through providing non-canonical TFBS (Polak and Domany 2006), epigenetic modifications (Xie et al. 2009; Lynch et al. 2011; Bakshi et al. 2016; Tristán-Flores et al. 2018)), alternate splice isoforms (Sorek et al. 2002; Payer et al. 2019) and differential mRNA stability (An et al. 2004). Primate specific Alu retrotransposons, which occupy 11% of the human genome, are major players in this process (Lander et al. 2001; Grover et al. 2005). Alus can bring about structural variations (Payer et al. 2017) via indels and exhibit a non-random, gene-biased distribution (Lander et al. 2001; Grover et al. 2003; Versteeg et al. 2003; Grover et al. 2004; Xing et al. 2009). Alus harbour cryptic splice sites which potentiate their inclusion into splice isoforms, some of which are ubiquitously expressed. Alu derived exons exhibit lineage specificity, high transcript inclusion levels as well as have much higher rates of evolution (Lin et al. 2008; Sorek 2009; Shen et al. 2011). Exonization is more frequent in genes that have arisen *de novo* in primates, with a strong preference for 3'UTRs (Toll-Riera et al. 2009; Mandal et al. 2013). ~14% of the human transcriptome contains exonised Alus, predominantly in the principal isoforms. These provide substrate for different regulatory events through gain of poly (A) sites (Roy-Engel et al. 2005; Lee et al. 2008), formation of specific secondary structures (Sobczak and Krzyzosiak 2002), AU-rich elements and miRNA recognition elements (MREs) leading to alternative polyadenylation, mRNA decay or interfering with translation (An et al. 2004; Häsler and Strub 2006; Häsler et al. 2007).

Among the TEs, Alu contains the maximum number of MREs (Spengler et al. 2014). A few cases of their involvement in modulating gene expression as well as modifying existing miRNA regulatory networks in a lineage specific manner have been reported . For example, miR-1285-1 is processed from an Alu and predominantly targets exonized Alus (Spengler et al. 2014) . In another instance, 3'UTRs of *MDM2* and *MDM4* harbouring Alu-MREs for primate specific miR-661, have been shown to provide an additional layer of regulation in the conserved p53 pathway in certain primate species (Hoffman et al. 2014). We have also shown that Alu-MREs can alter transcript isoform dynamics during heat shock response and some of these sites are still evolving in humans (Pandey et al. 2016).

Although these are majorly non-canonical, Alu-MREs can be responsive in conditions where miRNA levels or targeting efficiency is altered. For example, in patient derived cells from two different cancerous states, it has been shown that elevated levels of Pol III transcribed, free Alu RNA can sequester miR-566 which correlates with disease progression (Di Ruocco et al. 2018).

Exonized Alus can increase the regulatory possibilities for a transcript by engaging antisense as well as A-to-I RNA editing in a spatio-temporal manner. Since 3'UTR is the hotspot for Alu exonization, miRNAs would be the most common antisense encountered. The transcript isoform dynamics i.e., the relative proportion of exonized versus non-exonized isoforms, depends on the levels of editing as well as antisense RNA expression including the miRNAs and has been shown to be relevant in conditions such as viral and heat shock response (Mandal et al. 2013; Pandey et al. 2016). From our earlier study on 3177 Alu-exonized genes, we had observed the co-occurrence of cis Alu antisense and A-to-I editing at the level of single Alu exons in 319 genes. Enrichment analysis revealed that genes related to apoptosis and lysosomal processes were overrepresented in this set (Mandal et al. 2013). While mapping these 319 genes onto the apoptosis pathway, a serendipitous observation led us to an orphan gene *CYP20A1* that has acquired an unusually long 3'UTR due to exaptation of 23 different Alus and contains target sites for ~1000 different miRNAs. We propose it functions as an Alu-derived, endogenous multi-miRNA sponge to orchestrate systemic changes in associated miRNA regulatory networks.

Results

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

Apoptosis entails a cascade of sequential events at the molecular level. We performed an extensive, systematic manual curation and text mining to ascertain where in the apoptotic pathway the list of 319 genes mapped (**supplementary information**). 91 out of 319 genes mapped to apoptosis with 75% non-randomly clustered 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 Table S1**).

Since majority of the events occur in the 3'UTRs of the exonized transcripts which can also modulate miRNA regulatory networks in a lineage specific manner (Hoffman et al. 2014; Pandey et al. 2016), we further focussed on looking at phylogenetic novelties in the 3'UTRs of these genes. Interestingly,

we identified a novel principal isoform of orphan *CYP20A1* gene with a 8.93kb long 3'UTR, 65% of which is contributed by exonization of 23 Alus (**Figure 1a**). The orthologs of this 3'UTR in mouse, rat and zebrafish are extremely short, well within 1kb and there is only a sparse presence of two B1 SINEs, one simple repeat and one low complexity repeat in mouse whereas the zebrafish 3'UTR lacks repeats altogether. Amongst the eight transcripts of human *CYP20A1* annotated in NCBI, experimental evidence is available only for the longest isoform (NM_177538). This 10.94kb RNA not only appears to be unique to humans, but also its 3'UTR is an outlier in terms of size and occupies the 85th position in the genome-wide length distribution (**Figure 1b**). Such extended stretches of UTR are extremely rare as less than 3% of Alu-exonized genes have 3'UTRs greater than 6kb (**Supplementary Figure S1**). 3'UTR length also does not seem to correlate with the density of exonized Alus (r= 0.25) and compared to the genomic average of exonization events of 5.42 per UTR, *CYP20A1* 3'UTR is clearly an outlier with 23 Alus. The longest annotated *CYP20A1* transcripts for mouse (NM_030013.3), rat (NM_199401.1) and zebrafish (NM_213332.2) were 2.27, 2.03 and 1.79kb, respectively. Even among the non-human primates, we did not find any annotated transcripts beyond 3kb from this locus. Taken together, these suggest that the event of *CYP20A1* 3'UTR extension is lineage specific. Interestingly, these Alus belong to different subfamilies which differ in their evolutionary ages, suggesting that their exaptation had happened gradually over a period of time.

Unlike its 3'UTR, the coding region of *CYP20A1* is remarkably well conserved among vertebrates, both at the sequence level as well as in the length of the mature protein (**Table 1**). The chimpanzee, macaque and mouse *CYP20A1* all code for the same 462-470aa protein as humans although their transcript orthologs range between 1-3kb. The protein sequence identity is greater than 60% between zebrafish and human. 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 contain strongly similar functional groups. This is also corroborated by the minimal evolutionary divergence across vertebrate CYP20A1 proteins (**Figure 1c**) and a strong purifying selection operative on the CDS (Ka/Ks ~0.2 in mammals and <0.1 in non-mammalian vertebrates) (**Table 1**). The 3'UTR, however, becomes increasingly divergent as we move from the great apes to rhesus macaque and is maximum for mouse (the non-primate evolutionary out-group) and is maximally contributed by the Alu repeats, as evidenced through Jukes Cantor divergence.

Next, we asked if the patterns of sequence conservation and divergence are also reflected at the DNA level. We observed that though the 5'UTR appears to be well conserved across the primate lineage (except lemur and proboscis monkey), the 3'UTR exhibits divergence right from bonobo, with the breakpoints mostly coinciding with an Alu insertion (**Figure 1d**). 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 TSS of *CYP20A1* and 10kb downstream of transcription end site (TES) and found it to be almost perfectly conserved among the higher primates, except for some New world monkeys (**Supplementary Figure S3**). Hence, a human specific divergence in this 3'UTR seems to have occurred in a genomic context that is *per se* conserved, at least among the higher primates.

CYP20A1 full length transcript may be a lncRNA

Multiple lines of evidence suggesting human-specific divergence of this orphan 3'UTR prompted us to investigate if this is transcribed. As two thirds of the 3'UTR is primarily comprised of repetitive sequences, the full length transcript is not captured adequately in expression arrays nor does it map uniquely from sequencing reads. Moreover, there are disparities regarding the annotation of the full length 3'UTR-containing isoform in different genomic portals (**supplementary information**). Thus instead of using publically available expression data, we confirmed the expression on this long 3'UTR experimentally using eleven pairs of primers that span the entire length of the transcript and also validated three amplicons by sequencing (see **methods** for details). Although highly variable, we observed the transcript of our interest was expressed in all the six cell lines tested (**Figure 2a**). The cell lines tested are cancerous and aberrant transcriptional profiles are reported in cancer (Suzuki et al. 2014), so next we compared its expression in a neuroblastoma cell line (SKNSH) with those in primary neuron, glia (astrocyte) and progenitor cells (NPC) to check if it is expressed due to the cancerous state of the cell. Neuroblastoma shares features with both mature neurons and NPCs, but is distinct from glia. *CYP20A1* expression differs significantly only between glia and SKNSH (**Figure 2b**) but not in neurons or NPCs, suggesting it exhibits tissue specificity and is unlikely to be an artifact of the cancerous state of the cell.

We selected MCF-7 for the subsequent experiments as this cell line has been extensively used for drug screening and studying the effect of xenobiotics on different CYP family genes (Coumoul et al. 2001; Ptak et al. 2010). The copy number of *CYP20A1* is not altered in this line (2n=2) (human *CYP20A1* lacks paralogs (Pan et al. 2016) and the long 3'UTR isoform is expressed at a low level although

CYP20A1 protein is relatively abundant (**Supplementary Figure S4**). Further, 3'RACE followed by nested PCR and amplicon sequencing confirmed the presence of this long 3'UTR (**Figure 2c, supplementary information**).

The sixth exon is skipped in the transcript of our interest but retained in the other isoforms containing shorter 3'UTRs. Using primers encompassing the sixth exon, we could distinguish between the transcripts as this results in 196bp amplicon from the transcript of our interest compared to the 277bp from the shorter 3'UTR isoforms whose expression was also observed to be much higher (**Figure 2d**). Notably, the 277bp amplicon could be obtained using cDNA synthesized using regular RTs, unlike the 196bp amplicon which could only be obtained using cDNA prepared with a very efficient enzyme (such as SuperScript III). 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**). To probe further, we performed six-frame *in silico* translation of all *CYP20A1* isoforms 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 long 3'UTR isoform goes out of frame in the sixth and seventh exon and the resulting truncated 24 amino acid peptide has no hits in the human proteome as evidenced by BLAST analysis. Taken together, these data suggest that the long-3'UTR isoform is unlikely to be protein coding and may represent a novel lncRNA transcribed from the same locus. This presents a classic case of evolutionary sub-functionalization of a gene into two different classes of transcripts that might have evolved for different functions (Force et al. 1999) (**Figure 2e**).

In order to assess the relative contribution of different isoforms to the overall expression of *CYP20A1*, we used RNA-seq data from 15928 single nuclei derived from the different layers of human cerebral cortex (https://doi.org/10.1101/384826). NM_177538 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 S2).** We did not find any reads mapping to the extended 3'UTR in chimpanzee and macaque (prefrontal cortex, CD4+ T cells), suggesting its expression may be human specific. Interestingly, it is expressed in rosehip neurons - a highly specialized cell type in humans(Boldog et al. 2018), suggesting co-evolution (**Supplementary Table S3**).

9kb 3'UTR: an evolving regulatory hub

The 3'UTRs of transcripts can harbor many regulatory motifs that govern their stability (half life), cellular localization and translational fate (Matoulkova et al. 2012). Alu repeats have contributed to many such functional regulatory elements in the primate genomes(Pandey and Mukerji 2011; Elbarbary et al. 2013: 1; Chen and Yang 2017). Hence, we hypothesized that this recently diverged, Alu-rich 3'UTR could be a potential hotspot of many such motifs and act as an evolving regulatory hub. Our earlier study had revealed that 3'UTR-resident Alus contain functional miRNA sites, making exonized isoforms amenable to miRNA-mediated regulation post-transcriptionally(Pandey et al. 2016) . Hence, we wanted to check if this long UTR is also targeted by miRNAs. Using miRTarBase (release 6.0) (Hsu et al. 2011), we identified a set of 169 miRNAs (supported by microarray/ sequencing data) with predicted target sites on *CYP20A1* 3'UTR, of which 46 were also listed as "functional miRNAs" in FuncMir (miRDB) (**Supplementary Table S4**). Surprisingly, approx. 50% of these are either primate-specific or human-specific miRNAs (microRNAviewer) (Kiezun et al. 2012). The occurrence of target sites for human-specific miRNAs in this recently evolved UTR prompted us to carry out further in-depth analysis of miRNA recognition elements (MREs) in this 3'UTR as databases only report a few top target sites for each miRNA and provide limited information from a few experimental datasets.

It is generally recommended to obtain a consensus of MREs predicted using two or more algorithms for reducing the false positives. Since our 3'UTR has diverged across the vertebrate phylogeny, we decided against using algorithms that employ evolutionary conservation as a prediction criteria. Many of them also predict target sites based on seed sequence matches, which is increasingly being considered as an error-prone approach as the length and position of the seed sequence is variable amongst miRNAs (Ellwanger et al. 2011; Mullany et al. 2016). To circumvent this issue of inaccurate MRE prediction in non-conserved regions, we used miRanda which employs a two-step strategy: sequence complementarity, followed by thermodynamic stability of the predicted miRNA-mRNA duplex (Enright et al. 2003). 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 S5**). These MREs are spread across the length of the 3'UTR, although there are several high density pockets (**Figure 3a**). The top 3 miRNAs with maximum number of MREs were miR-6089 (43), miR-762 (36) and miR-4763-3p (33), respectively.

Since Alus harbor regulatory sites, it is possible that the accumulation of so many Alu-MREs in this 3'UTR was due to the retrotransposition or recombination of Alus with pre-existing target sites. However, analysis of a 1000 sets of 23 Alus with similar composition with respect to length and subfamily, taken randomly from the genome, did not show a similar distribution of Alu-MREs. Just 0.5 and 4.2% of these random sets had MREs ≥4742 and 4500, respectively (**Figure 3b**), suggesting that the chance of Alus having retrotransposed into this 3'UTR with pre-existing MREs is extremely low. Given these observations, it is clear that the presence of such large number of MREs is not artifactual and these have been created within Alus post exaptation into *CYP20A1* 3'UTR. This implies a directional evolution of this UTR for a novel function. One possibility is that accumulation of MREs for so many miRNAs could potentiate its function as an endogenous "miRNA sponge" for a regulatory network (**Figure 3c**).

Next, we inquired how many of these 994 miRNAs were expressed in a given cell type. We observed low concordance across replicates and high variation across experiments in publically available miRNA expression datasets. These experiments were mostly microarray based (**supplementary information**). We selected MCF-7 and primary neurons for further experiments. Since *CYP20A1* has also been implicated in neurophysiological roles (Stark et al. 2008: 1) it is convincing that besides its function as protein, lncRNA could have assumed novel role. Also, primary neurons preferentially express longer 3'UTRs which can increase the chances of miRNA-mediated regulation (Miura et al. 2013; Tushev et al. 2018). Using a cut-off of at least 10 MREs on *CYP20A1* 3'UTR and TPM value of 50, we obtained a set of 21 and 9 miRNAs in MCF-7 and neurons, respectively, of which 7 were common to both the cell types (**Table 2**).

CYP20A1 long 3'UTR isoform functions as an endogenous miRNA sponge

To determine if *CYP20A1* long 3'UTR isoform can be a potential miRNA sponge, we compared its features with previously characterized sponge RNAs. An effective sponge is predominantly cytosolic and typically contains 4 to 10 low binding energy MREs for a particular miRNA, each separated by a few nucleotides. It is devoid of destabilizing RNA elements, resistant to cellular nucleases and stably produced at high levels. It contains mispaired bases or bulge within the MREs that allows it to sequester miRNAs without triggering its own degradation and exhibits RNA induced silencing complex (RISC) loading. Most importantly, its expression is positively correlated with the expression of target RNAs that contain cognate MREs for the miRNA it sponges.

Transcripts containing oppositely oriented Alus in close vicinity tend to form double-stranded RNA by intramolecular pairing and are recognized by ADAR1, undergo A-to-I hyperediting (Bazak et al. 2014) and usually retained within nuclear paraspeckles (Chen and Carmichael 2009; Goldstone et al. 2010). Since our transcript of interest fulfils all these criteria, we checked its level using RT-qPCR in both nuclear and cytosolic fractions and found that it is predominantly localized to the cytosol (**Figure 4a**). Even 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 spread across the length of the UTR. We checked for the presence of bulge within the MREs for the 23 prioritized miRNAs (listed in **Table 2**) using miRanda on default parameters. To screen for MREs with efficient miRNA docking without leading to cleavage of *CYP20A1* RNA, two major criteria were followed – complete match in (2-7)6mer seed site and presence of mismatch or insertion at 9-12 position. 6mer sites with wobble base pairing were also retained as 2 wobble-pairs were maximally present in some of the MREs. We found five such sites for miR-6724-5p, 2 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**). In all these sites, insertion was present in *CYP20A1* UTR suggesting bulge formation in the sponge RNA itself.

In order to probe if the alteration in *CYP20A1* long-3'UTR isoform level would be consequential to the expression of transcripts containing cognate MREs, we looked for conditions where it is likely to be altered. 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, killing ~50% more neurons compared to vehicle (**Figure 4b**). Upon treating primary human neurons with HIV1 full length Tat protein, followed by 6hrs recovery, isoform 1 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 3'UTR also carries 17 high confidence binding sites for HSF1, 14 of them within Alus, which show positional conservation in agreement with previous studies (Pandey et al. 2011). This suggests that *CYP20A1* may be amenable to antisense-mediated downregulation during heat shock response as demonstrated by an earlier work from our lab(Pandey et al. 2011). We found isoform 1 to be significantly downregulated in primary neurons upon heat shock, followed by 1hr recovery (**Figure 4d**).

In order to query the expression of other cognate targets of the 9 prioritized miRNAs in these two conditions, we performed stranded RNA sequencing in primary neurons. In accordance with our RT-qPCR results, we found isoform 1 to be significantly downregulated 2.68 folds ($\log_2 FC = -1.42$) upon heat shock recovery (HS) and 1.21 folds upregulated ($\log_2 FC = 0.28$) during Tat response. The latter, however, did not cross our stringent statistical significance threshold. Out of the 3876 genes with significant differential expression in HS or Tat, 380 genes exhibit positively correlated expression pattern as *CYP20A1* (**Figure 5, Supplementary Table S6**). All of these 380 genes contain at least one MRE for one or more of the 9 prioritized miRNAs and majority of their MREs are canonical and not Alu-derived. Whereas most genes contain a few MREs for one or two of these miRNAs, *CYP20A1* is at the 2nd position with a total of 116 total MREs (**Supplementary Table S6**).

Next, we performed enrichment analysis for this gene set using Toppfun and found enrichment of blood coagulation pathways and neuronal development as major hubs in biological processes. The top five processes were hemostasis (28 genes), axon guidance (25 genes), neutrophil degranulation (23 genes), platelet signalling, activation and aggregation (18 genes) and ECM organisation (18 genes). mRNA processing and mitochondria translation, metabolism, amino acid and nucleotide synthesis and antigen presentation were amongst others (**Supplementary Table S7**). Blood coagulation factors have been reported to affect pathophysiology of CNS via coagulation protein mediated signal transduction (De Luca et al. 2017). Although the exact biological role of *CYP20A1* remains be explored, yet enrichment of coagulation pathways in gene set showing correlated expression with *CYP20A1* suggests that it may have some role in neuro-coagulopathy, which remains to be tested.

Discussion

Long non-coding RNAs, as competing endogenous RNAs (ceRNAs), are being implicated in rewiring miRNA gene regulatory networks. These could contribute to lineage specific novelties. For example, lncND has been shown to drive brain expansion in the primate lineage by sequestering miR-143-3p that regulates the expression of Notch-1 and Notch-2 (Rani et al. 2016). ceRNA-mediated regulation has been demonstrated in neurodevelopment, myoblast differentiation, pluripotency maintenance and in various cancers (Wang et al. 2010; Cesana et al. 2011; Fan et al. 2013; Wang et al. 2013) and their expression is spatio-temporally regulated. Circular RNAs, produced by back-splicing, have been previously reported as potent miRNA sponges (Hansen et al. 2013; Barrett and Salzman 2016). However, to the best of our knowledge, TE-derived endogenous miRNA sponge has not been

characterized so far. We present an interesting case of Alu exonization that lengthens the 3'UTR of an orphan protein coding gene *CYP20A1* and creates a novel lncRNA isoform. The initial characterization of this isoform (Isoform 1) reveals exaptation of 23 Alus from different subfamilies in its 3'UTR. These have provided the substrate for evolution of multiple miRNA binding sites and resulted in its neo-functionalization to a potential multi-miRNA sponge. The event seems to be specific to the human lineage.

CYP20A1 is an orphan member of the cytochrome P450 superfamily of oxidoreductases that metabolize and detoxify drugs (Guengerich et al. 2005). These genes are also involved in steroidogenesis. It contains unusual heme-binding motifs and is highly expressed in the liver. In copepods it has been shown to be responsive to stressors like bisphenol A(28) and UV-B (Puthumana et al. 2017). Elevated expression of *CYP20A1* RNA has been reported in the gastric mucosa of splenasthenic syndrome (chronic superficial gastritis) patients as well as in the leukocytes of blunt trauma patients (Brandfellner et al. 2013). Besides, high levels of *CYP20A1* mRNA in both unfertilized zebrafish eggs (Goldstone et al. 2010) and developing mouse embryos (Choudhary et al. 2003) suggest its involvement in early developmental stages. However, these observations are correlational and it is still contentious whether the gene could function as an RNA or a protein.

We report both coding as well as non-coding transcripts from *CYP20A1* locus. While the protein is highly conserved, we observe a novel transcript with a skipped exon that results in an out-of-frame CDS. Further, the novel RNA isoform has a substantially longer UTR with an increased scope of involvement in regulatory events post-transcription. This we confirm and validate experimentally in multiple cell lines. This isoform is also human specific. It is not only expressed in 75% of single nuclei derived from different layers of the human cerebral cortex but also in human specific rosehip neurons. Its specific expression to human lineage is also strengthened by the fact that the 3'UTR genomic region is neither conserved at the DNA level in high order primates nor is present in transcripts derived from RNA seq of closely related primates. The protein coding isoform exhibits a significantly greater conservation across the evolutionary lineages.

miRNA sponges so far have been reported to contain MREs that are targets for a single or a few related miRNA species (Franco-Zorrilla et al. 2007; Poliseno et al. 2010; Hu et al. 2018). Artificially created sponges with MREs >=4 have been effectively used in a few studies(Ebert et al. 2007; Ebert and Sharp 2010) but the *CYP20A1* isoform 1 having >=10 MREs for as many as 140 miRNAs clearly stands out.

The presence of MREs for multiple miRNAs raises the exciting possibility that it can exert a systemic effect through titrating the cellular levels of many miRNAs simultaneously. A modulated expression of this transcript could alter spatio-temporal dynamics and stoichiometry of the miRNA pools. In a previous study from our lab, we have reported the functional effects of MREs within 3'UTR-resident exonized Alus in fine tuning the p53 regulatory network in heat-shock response (Pandey et al. 2016). In *CYP20A1*, out of 4742 total MREs more that 80% are contained within Alu which could potentially participate in miRNA regulatory network and are not just repetitive sequence artifacts. Further, all the 23 Alus in this 3'UTR have been shown harbour antisense transcripts and are substrate for A-to-I RNA editing (Mandal et al. 2013). These events though dynamic could further disrupt or create new miRNA binding sites from existing MREs thereby increasing the regulatory possibilities. For example, miR-513 and miR-769 targets 3'UTR of DNA fragmentation factor alpha gene in an adenosine deamination dependent manner (Borchert et al. 2009) . Since A-to- I editing events are preponderant in brain (Paz-Yaacov et al. 2010; Rosenthal and Seeburg 2012), these editing events could further contribute to phylogenetic novelties.

We observe that expression of isoform 1 in human primary neurons is inducible in response to HIV1-Tat and decreased during heat shock response. Expression of ceRNAs is also tightly regulated and often, specific to tissue, development stage or some stress conditions (Cesana et al. 2011; Tay et al. 2014). High levels of sponge RNA could yield high expression of the cognate targets as miRISC complexes are sequestered by the sponge (Ebert et al. 2007; Ebert and Sharp 2010), whereas reduced expression of sponge could relieve these miRNA for the latter to bind, inhibit and/or degrade their cognate targets (Figure 3c). We make similar observations in a set of 380 genes which correlate with the expression pattern of *CYP20A1* i.e., downregulated during heat-shock and upregulated upon Tat treatment. These genes mapped to processes that are majorly involved in blood coagulation and neuronal pathways. Neurocoagulopathy has been implicated in the pathophysiology of CNS via coagulation protein mediated signal transduction. Besides coagulation, these proteins interfere with synaptic homeostasis. They also affect neurite outgrowth and morphological changes in neurons, blood brain barrier integrity, ECM stability, ROS generation from astrocytes, secretion of inflammatory cytokines, as well as impair excitability, increase inflammation, mitosis of astrocytes and/or microglia and ultimately, affect neuronal viability. This process is being linked to several neurodegenerative diseases including multiple sclerosis, cancer of the CNS, addiction and mental health (De Luca et al. 2017). 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. The set of 380 genes were also found to be 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. Upon activation by Tat, the sponge could titrate out the miRNA that target the 380 genes and hence modulates all the pathways simultaneously. Validation of this lncRNA 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 NPCs during neurogenesis, neuronal migration during differentiation, etc.

Multiple exonized Alus in *CYP20A1* 3'UTR can potentially make it highly structured and lead to altered bioavailability for MREs. Future studies on secondary structure simulations would allow us to assess the availability and accessibility of these MREs. Closely spaced sites of same or different miRNAs can cooperatively sequester multiple miRNAs thus making the process fast and leading to robust outcomes (Denzler et al. 2016). This would be an exciting possibility to explore. With multiple miRNA that target members of a co-regulated network it could exhibit a systemic response. Also, this could work as an effective regulatory switch for a faster response or return to baseline. We have not yet looked at its turnover rates, however multiple possibilities exist such that it could be transiently over-expressed induced in response to specific cues, regulated through a negative feedback, cleared via transcriptional shutdown by certain miRNAs or have an inherently shorter half-life due to a rapid turnover. Other possibilities such as independent transcription events from this UTR or additional polyadenylation sites, cannot however be ruled out. During the course of this study, we noticed reads mapping beyond the longest 3'UTR annotated in UCSC (Supplementary Figure 5). Beyond its role as miRNA sponge, this UTR can also be involved in sequestering RBPs and deplete the cellular reserves, thereby indirectly affecting other genes (Kelley et al. 2014).

Conclusion

In conclusion, the present study further substantiates the fact that non-random distribution of Alu in human genome can alter conserved regulatory networks and reshape them in lineage specific manner. The contribution of repeats in evolving regulatory networks has been undermined and overlooked until

recently. Extensive Alu mediated 3'UTR lengthening in case of *CYP20A1* provides a platform for creation of multiple MREs which makes it capable of functioning as multi-miRNA sponge. *CYP20A1* provides an interesting model for studying lncRNA derived ceRNAs that could co-regulate multiple genes and participate in complex networks that govern intermediate phenotypes such as neuro-coagulation. These could play important roles as regulatory switches in response to biological cues. This could result in differential cellular outcomes due to rapid release or sequestering of miRNAs.

Materials and Methods

Genome-wide 3'UTR length distribution

The coordinates for all human NCBI RefSeq transcripts were downloaded from RefSeq version 74 (hg38). Both NM and XM ID containing transcripts were included in the analysis. The 3'UTR length was computed from its annotated translation end site through the transcription end site and for every gene, only the longest 3'UTR was considered. The summary statistics for the distribution were calculated using R scripts.

Conservation Analyses

DNA

DNA sequence conservation across different species was checked using UCSC genome browser. We have used multiple alignment across 20 species generated by multiz (Blanchette 2004) which uses a combination of two algorithms (phastCons and phyloP) to estimate evolutionary conservation, with the null hypothesis of neutral evolution. Both gaps as well as unaligned sequences were treated as 'missing' data.

For calculating Jukes-Cantor (JC) divergence, DNA sequences with 1kb upstream and downstream were downloaded from UCSC for human *CYP20A1* 5'UTR, 3'UTR and 10kb promoter proximal region (chr2:203,228,449-203,238,448; hg38). The corresponding regions were fetched for pair-wise alignment between human-chimp (panTro4), human-gorilla (gorGor4), human-bonobo (panPan1) and human-macaque (rheMac8), which were subsequently used to calculate the JC-score.

Protein

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). It uses seeded guide trees and hidden Markov model (HMM) to estimate the similarity among three or more sequences.

Ka/Ks ratio was calculated using the method detailed in a previous publication by our group (Gautam et al. 2015). Protein sequence IDs of CYP20A1 ortholog for each species was identified as the first hit of NCBI pBLAST using human NP-ID as the query term. The NM-IDs of corresponding mRNAs were also obtained from NCBI and sequences were fetched for these NP and NM-IDs. For a pair of orthologous genes, the protein-coding DNA alignment between human and each of the other species was constructed using ParaAT (a Parallel Alignment and back-Translation tool) (Zhang et al. 2012). The protein sequence information was used to match each triplet codon with its respective amino acid. Aligned coding DNA between a pair (human vs. another species) was provided as an input to KaKs_Calculator (Zhang et al. 2006). Model averaging (MA), which takes an average of substitution rates of seven models, was used as a maximum likelihood method to calculate the Ka/Ks value.

Cell culture

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, 10569044) supplemented with 10% heat inactivated FBS (Gibco, US origin, 10082147), HEPES (Gibco, 11560496) and 1X antibiotic-antimycotic (Gibco, 15240096). The culture was maintained at 70-80% confluency at 37°C, 5% CO_2. We had authenticated the cell line identity by STR profiling (**supplementary information**). The cells were routinely screened for mycoplasma using MycoAlert Mycoplasma detection kit (Lonza, LT07-218.) as per manufacturer's protocol and confirmed to be free from contamination. Additionally, 16S rRNA PCR was performed using both genomic DNA from cells and cell-free DNA from spent media to confirm the absence of bacterial contamination.

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, P7886) coated flasks. The cells were maintained in neurobasal media (Gibco, 21103049) containing N2 supplement (Gibco, 17502048), Neural Survival Factor 1 (Lonza, CC-4323), 25ng/ml bovine fibroblast growth factor (bFGF) (Sigma, F0291) and 20ng/ml human epidermal growth factor (hEGF) (Sigma, 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; 10ng/ml each of PDGF(Sigma, P3326) and BDNF(Sigma, B3795) were instead added to the media cocktail. Differentiation of NPCs to astrocytes required Minimum Essential Medium (MEM) (Sigma, 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 S6**).

RNA isolation, cDNA synthesis and RT-qPCR

Total RNA was isolated using TRIzol (Ambion, 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 [3 units of TURBO DNase (Invitrogen, AM1907) was used for 5µg RNA; incubated at 37°C for 20'] using 300 units SuperScript III RT (Invitrogen, 18080-044) (50°C for 90', 70°C for 15'). RNA template was digested from the cDNA using 2 units of *E. coli* RNaseH (37°C for 20') (Invitrogen, 18021071). Primers were designed using Primer3 (version 4.0.0) and were synthesized by Sigma (**supplementary information**). To insure against spurious amplification, we had designed two pairs of overlapping primers both on the 5' as well as 3'ends of our transcript of interest and run 'minus-RT' controls in every reaction. Additionally, we had sequenced three amplicons (1, 5 and 10) with both forward and reverse primers to check the specificity of amplification. 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 in 10µl volume using 5µl 2X SYBR Green I master mix (Kappa; KK3605), 0.2µl each of 10µM forward and reverse primers and cDNA (stock: ~80ng/µl RNA equivalent, diluted 1:40 in the final reaction). The reaction was carried out in Roche Light Cycler 480 (USA) and the cycling conditions were: 95°C for 3', 45 cycles of (95°C for 30", 58°C for 30", 72°C for 40"), 72°C for 3'; followed by melting: 95°C for 5", 65°C for 1'. Melting curves were confirmed to contain a single peak and the fold change was calculated by ΔΔCt method. MIQE guidelines were followed for data analysis.

3'RACE

cDNA for 3'RACE was prepared using RLM-RACE kit (Ambion, AM1700) with 1µg MCF-7 total RNA as per the manufacturer's recommendation. Nested PCR was performed with FP10 and inner primer using the amplicon produced by FP9 and outer primer. The product of this nested PCR was run on 2% agarose gel and four major bands were observed, which were gel-eluted (As per manufacturer's protocol, Qiaquick gel extraction kit, Qiagen, 28704) and subsequently sequenced. BLASTN (NCBI; 2.6.0+) against the 861bp *in silico* amplicon revealed 98% and 95% identity for the ~800bp and 700 bp amplicons, respectively; BLAT on hg38 showed that these sequences (including a part of AluSc) map to *CYP20A1* 3'UTR with 97% identity.

Sanger sequencing

The sequencing reaction was carried out using Big dye Terminator v3.1 cycle sequencing kit (ABI, 4337454) in 10µl volume (containing 2.5µl purified DNA, 0.8µl sequencing reaction mix, 2µl 5X dilution buffer and 0.6µl forward/ reverse primer) with the following cycling conditions - 3 mins at 95°C, 40 cycles of (10 sec at 95°C, 10 sec at 55°C, 4 mins at 60°C) and 10 mins at 4°C. Subsequently, the PCR product was purified by mixing with 12µl of 125mM EDTA (pH 8.0) and incubating at RT for 5 mins. 50µl of absolute ethanol and 2µl of 3M NaOAc (pH 4.8) were then added, incubated at RT for 10 mins and centrifuged at 3800rpm for 30 mins, followed by invert spin at <300rpm to discard the supernatant. The pellet was washed twice with 100µl of 70% ethanol at 4000rpm for 15 mins and supernatant was discarded by invert spin. The pellet was air dried, dissolved in 12µl of Hi-Di formamide (Thermo fisher, 4311320), denatured at 95°C for 5 mins followed by snapchill, and linked to ABI 3130xl sequencer. Base calling was carried out using sequencing analysis software (v5.3.1) (ABI, US) and sequence was analyzed using Chromas v2.6.5 (Technelysium, Australia).

CYP20A1 expression in non-human primates

In order to query if the long-3'UTR isoform is expressed in other primate species, we checked for reads mapping to the 3'UTR in publicly available chimp and macaque RNA-seq datasets [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 skipped exon 6 in our isoform of interest. In *CYP20A1* long-3'UTR isoform, exon 5 and exon 7 are contiguous and if we have reads mapping to the region spanning the junction of these two exons, we can reliably say that this isoform is expressed. Towards this, we performed BLAST against human datasets in SRA using a 289bp sequence reconstructed by joining exon 5 and 7 and further confirmed the hits by checking for read mapping to the 3'UTR.

miRNA target prediction

miRNA target sites (MREs) on *CYP20A1* 3'UTR were predicted using miRanda (version 3.3a) (Enright et al. 2003), 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).

Small RNA sequencing and analyses

cDNA libraries were prepared as per manufacturer's protocol using TruSeq small RNA library prep kit (Illumina, RS-122-2202) from 1µg total RNA (RIN ≥8.5). Briefly, 3' and 5' adapters were ligated, followed by enrichment of RNA fragments with adapter molecules on both ends. Reverse transcription followed by amplification using primers that anneal to the adapter ends created the cDNA libraries which was then purified and size selected (~147nt) to enrich for small RNAs by running on 6% native PAGE. Subsequently, the libraries were concentrated by ethanol precipitation and quality checked on

Agilent technologies 2100 Bioanalyzer (USA) using DNA high sensitivity chips (Agilent, 5067-4626). Using 10mM Tris-HCl (pH 8.5) libraries were normalized to 2nM, denatured and applied to cBot for cluster generation (Illumina, PE-401-3001), which were then sequenced on HiSeq2000 (Illumina, California, USA) using TruSeq SBS kit (Illumina, FC-401-2202) for 50 cycles.

The data were quality checked using FastQC (version 0.11.2), followed by adapter trimming by cutadapt (version 1.18); reads were not discarded. As expected, around 95% of the of adapter trimming events happened on the 3'end of the reads. Length filtering and quality filtering were carried out by cutadapt; Q30 reads with sequence length >15 but < 35nt. were retained for mapping. ~80% of the reads were retained after these filtering steps. Size distribution of the reads and kmer position were (21-25, 28-32) and 26-28 respectively. Subsequently, these reads were mapped onto hg38 using Bowtie2. On an average, 61% of the reads were uniquely mapped. miRDeep2 was run to obtain the read counts as TPM.

Nuclear cytosolic RNA isolation

Nuclear and cytosolic RNA were isolated using PARIS kit (Ambion, AM1921) as per manufacturer's protocol. Briefly, ~10 million cells were resuspended in fractionation buffer, incubated on ice and centrifuged at 4°C to separate the nuclear and cytosolic fractions. The nuclear pellet was additionally treated with 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 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.

Cell lysate preparation and Western Blotting

Cells were harvested with 0.25% trypsin-EDTA (Gibco, 25200-072) and pelleted. 30% of this pellet (2-3 million cells) was lysed in 100µl RIPA (Sigma, R0278) (reconstituted with 0.1M DTT and 1µl protease inhibitor cocktail (Sigma, P8340) and incubated on ice for 30' with intermittent vortexing. The total cell lysate was purified by centrifugation at 14000rpm for 45' (at 4°C) and the supernatant was stored at -20°C.

For nuclear cytosolic fractionation, ~2 million cells were gently lysed with 200µl cell lysis buffer (10mM Tris-HCl pH 7.5, 0.15% NP-40 and 150mM NaCl) and incubated on ice for 20'. The lysate was then layered on top of 2.5 volumes of chilled 24% sucrose cushion (prepared in lysis buffer) and centrifuged at 14000rpm for 10' at 4°C. The supernatant (cytosolic fraction) was collected and stored at -20°C. The nuclear fraction was obtained by re-suspending the pellet in 200µl pre-chilled glycerol buffer (20mM Tris-HCl pH 7.5, 75mM NaCl, 0.5mM EDTA, 50% glycerol and 0.85mM DTT) and an equal volume of nuclear lysis buffer (10mM Tris-HCl pH 7.5, 1mM DTT, 75mM $MgCl_2$, 0.2mM EDTA, 0.3M NaCl, 1M urea and 1% NP-40) with brief vortexing.

The protein concentration was determined by Pierce BCA Protein Assay Kit (Thermo Scientific, PI23225). 25µg of denatured protein (90°C for 10') was loaded on 15% SDS PAGE and was resolved at 60V, followed by overnight transfer (30V at 4°C) onto 0.2µm PVDF membrane (MDI, SVFX8301XXXX101). The transfer was confirmed by Ponceau staining (USB, 32819). Blocking was done in 5% skimmed milk (Himedia, GRM1254) for 1hr, followed by incubation with primary and secondary Ab. The blot was developed by ECL (Abcam, AB133406) staining and analyzed using ImageJ (NIH, US) (antibody details provided in **supplementary information**).

Immunostaining

Cells were seeded on coverslips and cultured in 4-well plates for this experiment. Spent media was discarded and cells were briefly washed with 1X PBS (to remove residual media completely). The cells were then fixed with 4% PFA (freshly prepared in 1X PBS) for 20 minutes, followed by 3 washes in 1X PBS (5 minutes each) to completely remove traces of PFA. Blocking was done with 4% BSA containing 0.5% Triton-X (to permeabilize the cells) for 1hr with gentle rocking. Incubation in primary antibody (MAP2, Tuj1, NeuN, Nestin, GFAP; all diluted 1:200 in 0.1% BSA) was done overnight at 4°C on a rocker, except GFAP (1hr at RT), followed by 3 washes in 1X PBS (5 minutes each). Incubation with fluorescently labeled secondary antibody (diluted 1:1000 in 0.1% BSA) was performed for 1hr in the dark with shaking. This was followed by five washes in 1X PBS (5 minutes per wash) on a rocker. The coverslips were mounted in DAPI hardset (Vectashield, USA) on clean glass slides. Images were captured from 7-10 random fields using AxioImager.Z1 microscope (Carl Zeiss, Germany).

Stress treatment

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' 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 S7**). Finally, 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/5% CO_2 for another 6hrs prior to harvesting.

TUNEL Assay

The assay was performed with *in situ* Cell Death Detection kit, TMR red (Roche, Germany) which uses a proprietary terminal deoxynucleotidyl transferase (TdT) to incorporate a fluorescently tagged nucleotide at the free 3'-OH ends of damaged DNA molecules, a hallmark of late apoptotic cells.

~20,000 cells were seeded per well (on coverslips) in 12-well plates. Post Tat treatment, cells were washed once with 1X PBS and instantly fixed with 4% PFA, followed by three brief washes in 1X PBS, permeabilization and blocking with 4% BSA containing 0.5% Triton-X 100, incubation with TdT for 1hr in the dark and three thorough washes with 1X PBS. Coverslips were then mounted on clean glass slides using hardset mounting media containing DAPI (Vectashield, USA). 6-8 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). At least 1000 cells were scored for each replicate.

Library preparation, RNA sequencing and data analyses

Libraries were prepared following Illumina's TruSeq stranded total RNA protocol as per manufacturer's recommendations, using three biological replicates per experimental condition. Integrity of the RNA samples was checked on Agilent 2100 Bioanalyzer (Agilent RNA 6000 Nano

Kit); RIN values ranged from 8.6-9.6 for our samples. 500ng of total RNA was taken in a volume of 10µl and ribosomal RNA was removed using Ribozero depletion kit. Briefly, total RNA was denatured at 68°C for 5', followed by removal of rRNA using magnetic beads, subsequent clean up with RNAClean XP beads (Beckman Coulter, A63987), two consecutive washes in freshly prepared 70% ethanol and then elution (by heating at 94°C for 8'; to fragment and prime the RNA for cDNA synthesis). First strand synthesis was carried out with random hexamers using the following program (25°C for 10', 42°C and 70°C for 15' in each). This was followed by second strand cDNA synthesis at 16°C for 1hr for elimination of the template RNA, replacement of dUTP by dTTP and to finally produce blunt-ended cDNA, which was then cleaned up using AMPure XP beads (Beckman Coulter, A63881) and two washes in 80% ethanol. Next, the 3'ends of these blunt cDNA molecules were adenylated to create a single 'A' overhang (incubation at 37°C for 30', followed by 70°C for 5') such that these can form complementary base pairs with the 'T's on the adapter ends and also do not concatenate with each other to form chimera during adapter ligation. Then, adapters were ligated (30°C for 10') at both the ends of these double-stranded cDNA molecules so that these can hybridize to the probes on the sequencing flow-cell. The fragments with adapter ligated on both ends (libraries) were enriched and amplified using a 15-cycle PCR (the PCR primer cocktail anneal to both ends of the adapters; PCR conditions: 98°C for 30", 15 cycles of (98°C for 10", 60°C for 30", 72°C for 30"), final extension 72°C for 5'). This was followed by their clean up using AMPure XP beads and two washes in 80% ethanol. The libraries were quantified using Qubit fluorometer (Invitrogen) and subsequently validated on Bioanalyzer (Agilent DNA 1000 kit) to confirm the insert size. The size of the libraries was in the range of 260-280bp. This was followed by normalization, dilution of the libraries to 10nM and pooling 3 samples together. The 10nM libraries were denatured using NaOH, diluted to 8pM and clusters were generated on a paired end flowcell (1 pool/lane) using TruSeq PE cluster kit v3-cBOT-HS on cBot system. Paired end sequencing was carried out on HiSeq 2000 using TruSeq SBS kit v3-HS (200 cycles).

Sample-wise data obtained in the RNAseq has been summarized in the supplementary. Briefly, .fastq files were checked using FASTQC and overall Q score was found to be >20. There was 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).

MTT assay

~10,000 cells were seeded per well in 96 well plate. Post treatment, freshly prepared MTT (HiMedia, TC191) solution (in 1X PBS) was added to the cells at a final concentration of 0.5mg/ml. Cells were incubated for 3hrs at 37°C in dark and then checked for formazan crystals under the microscope. After removing the solution, the formazan crystals were dissolved in 100µl DMSO (Sigma, D2650) at RT for 5-10 minutes and then absorbance was acquired at 560nm on Tecan.

Figure Legends

Figure 1: *CYP20A1* contains a unique 3'UTR with Alu-driven divergence

a. UCSC track representing the four transcript isoforms of *CYP20A1* with varying 3'UTR length. Only isoform 1 (NM_177538) contains the full length 8932bp 3'UTR. The RepeatMasker track shows this 3'UTR to be populated with 23 Alu repeats belonging to different subfamilies.

b. Genome wide 3'UTR length distribution shows *CYP20A1* to be an outlier. Mean and median 3'UTR lengths were 1553bp and 1007bp, respectively.

c. Cladogram representing the *CYP20A1* protein sequence divergence among different classes of vertebrates suggest that it has diverged minimally.

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 (also see Supplementary Figure S3).

Figure 2: *CYP20A1* full length transcript with 9kb 3'UTR is expressed and may be a long non coding RNA

a. A schematic representation of the primers designed on the full length *CYP20A1* long-3'UTR transcript. Representative gel images showing this isoform is expressed in multiple cell lines of different tissue origin using primer pairs 1 and 10 from 5'UTR and 3' UTR, respectively. Amplicons (1, 5 and 10) were also confirmed by Sanger sequencing.

* genomic DNA was used as positive control for PCR

b. RT-qPCR for *CYP20A1* expression in cancerous versus non cancerous cell types of neuronal origin. Fold change was calculated with respect to SKNSH, 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 (** → $p<0.01$, *** → $p<0.005$; 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* 3'UTR, suggesting that the full length transcript is expressed in untreated MCF-7 cells (N=3).

d. Differentiating the long-3'UTR transcript from other isoforms: the schematic highlights the skipped exon 6 and the position of flanking primers on shared exons.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* isoform 1 and isoforms 2 and 3.

Figure 3: *CYP20A1*-long 3'UTR putative lncRNA may act as a miRNA sponge

a. Circos plot representing the MREs for the 994 miRNAs on *CYP20A1* 3'UTR. miRNAs are grouped on the basis of number of MREs. 23 Alus in this 3'UTR contribute to 65% of its length and are distributed throughout the UTR. Only 14% of miRNAs have MREs >10 (117 and 23 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. Schematic representation illustrates *CYP20A1* acting as an miRNA sponge. In the condition where *CYP20A1* is highly expressed, it will recruit multiple miRISC complexes thereby relieving the repression of cognate targets and subsequently leading to their translation; whereas in case of its low expression, those miRISC complexes remain free to load on the cognate targets and affect translational

repression or promote mRNA degradation. *CYP20A1* has the potential to sponge multiple miRNAs at the same time thereby regulating a large repertoire of transcripts.

Figure 4: Features of *CYP20A1* long-3'UTR isoform for being a potential sponge RNA

a. Cytosolic localisation of *CYP20A1* isoform 1 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 were used for each experiment. Quality controls for assessing the purity of nuclear (*MALAT1*) and cytosolic (*GAPDH*) fractions are shown on the right.

The RT-qPCR data were analyzed in accordance with the MIQE guidelines (Bustin et al. 2009) (see supplementary information for details).

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 ~10% 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 represent the mean and SD of three independent experiments and >1000 nuclei were scored per condition for each experiment.

c and d. Expression of *CYP20A1* isoform 1 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) but lost statistical significance as the cells continued to recover (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 geomean of *GAPDH*, *ACTB* and 18S rRNA in [c] and with respect to non heat shock treated cells, after internal normalization with the geomean of *GAPDH* and *ACTB* in [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: Log2FC values of 380 genes upregulated in response to Tat treatment (red) and downregulated during heat shock recovery (green) in primary neurons, mirroring the exact trend exhibited by *CYP20A1* isoform 1. All of them contain one or more MREs for the 9 miRNAs that can potentially be sponged by isoform 1 in neurons and represent potential cognate targets whose expression can be regulated by *CYP20A1* isoform 1 perturbation.
Genes are plotted in order as supplementary table S6.

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

Acknowledgements

Financial support from CSIR in the form of fellowships to AB and KS and through MLP-901 project to MM 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 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.

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, Dr. Amit Chaurasia for Jukes Cantor divergence analysis and Drs. Debojyoti Chakraborty and Rakesh Dey for providing reagents and many fruitful discussions.

Conflict of interest: None to disclose

Authors contribution

MM and AB designed the study and co-wrote the manuscript along with KS. 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 seminally 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 sn-RNA seq data. BP contributed reagents, helped in troubleshooting experiments and provided critical inputs. MM carried out data analysis and supervised the overall study.

All authors have read and approved the final version of the manuscript.

References

1. An HJ, Lee D, Lee KH, Bhak J. 2004. The association of Alu repeats with the generation of potential AU-rich elements (ARE) at 3' untranslated regions. BMC Genomics 5:97.
2. Bakshi A, Herke SW, Batzer MA, Kim J. 2016. DNA methylation variation of human-specific Alu repeats. Epigenetics 11:163–173.
3. Barrett SP, Salzman J. 2016. Circular RNAs: analysis, expression and potential functions. Development 143:1838–1847.
4. Bazak L, Levanon EY, Eisenberg E. 2014. Genome-wide analysis of Alu editability. Nucleic Acids Res. 42:6876–6884.
5. Blanchette M. 2004. Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. Genome Res. 14:708–715.
6. Boldog E, Bakken TE, Hodge RD, Novotny M, Aevermann BD, Baka J, Bordé S, Close JL, Diez-Fuertes F, Ding S-L, et al. 2018. Transcriptomic and morphophysiological evidence for a specialized human cortical GABAergic cell type. Nat. Neurosci. 21:1185–1195.
7. Borchert GM, Gilmore BL, Spengler RM, Xing Y, Lanier W, Bhattacharya D, Davidson BL. 2009. Adenosine deamination in human transcripts generates novel microRNA binding sites. Hum. Mol. Genet. 18:4801–4807.
8. Brandfellner HM, Ruparel SB, Gelfond JA, Hargreaves KM. 2013. Major blunt trauma evokes selective upregulation of oxidative enzymes in circulating leukocytes. Shock Augusta Ga 40:182–187.
9. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL, et al. 2009. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin. Chem. 55:611–622.
10. Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, Chinappi M, Tramontano A, Bozzoni I. 2011. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell 147:358–369.
11. Chen H, Chen L, Wu Y, Shen H, Yang G, Deng C. 2017. 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. 34:3216–3231.
12. Chen L-L, Carmichael GG. 2009. Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA. Mol. Cell 35:467–478.
13. Chen L-L, Yang L. 2017. ALUternative Regulation for Gene Expression. Trends Cell Biol. 27:480–490.
14. Choudhary D, Jansson I, Schenkman JB, Sarfarazi M, Stoilov I. 2003. Comparative expression profiling of 40 mouse cytochrome P450 genes in embryonic and adult tissues. Arch. Biochem. Biophys. 414:91–100.

15. Coumoul X, Diry M, Robillot C, Barouki R. 2001. 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. 61:3942–3948.
16. De Luca C, Virtuoso A, Maggio N, Papa M. 2017. Neuro-Coagulopathy: Blood Coagulation Factors in Central Nervous System Diseases. Int. J. Mol. Sci. [Internet] 18. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5666810/
17. Denzler R, McGeary SE, Title AC, Agarwal V, Bartel DP, Stoffel M. 2016. Impact of MicroRNA Levels, Target-Site Complementarity, and Cooperativity on Competing Endogenous RNA-Regulated Gene Expression. Mol. Cell 64:565–579.
18. Di Ruocco F, Basso V, Rivoire M, Mehlen P, Ambati J, De Falco S, Tarallo V. 2018. Alu RNA accumulation induces epithelial-to-mesenchymal transition by modulating miR-566 and is associated with cancer progression. Oncogene 37:627–637.
19. Ebert MS, Neilson JR, Sharp PA. 2007. MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. Nat. Methods 4:721–726.
20. Ebert MS, Sharp PA. 2010. MicroRNA sponges: Progress and possibilities. RNA 16:2043–2050.
21. Elbarbary RA, Li W, Tian B, Maquat LE. 2013. STAU1 binding 3' UTR IRAlus complements nuclear retention to protect cells from PKR-mediated translational shutdown. Genes Dev. 27:1495–1510.
22. Ellwanger DC, Büttner FA, Mewes H-W, Stümpflen V. 2011. The sufficient minimal set of miRNA seed types. Bioinformatics 27:1346–1350.
23. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. 2003. MicroRNA targets in Drosophila. Genome Biol. 5:R1.
24. Fan M, Li X, Jiang W, Huang Y, Li J, Wang Z. 2013. A long non-coding RNA, PTCSC3, as a tumor suppressor and a target of miRNAs in thyroid cancer cells. Exp. Ther. Med. 5:1143–1146.
25. Fatima M, Prajapati B, Saleem K, Kumari R, Mohindar Singh Singal C, Seth P. 2017. Novel insights into role of miR-320a-VDAC1 axis in astrocyte-mediated neuronal damage in neuroAIDS. Glia 65:250–263.
26. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545.
27. Franco-Zorrilla JM, Valli A, Todesco M, Mateos I, Puga MI, Rubio-Somoza I, Leyva A, Weigel D, García JA, Paz-Ares J. 2007. Target mimicry provides a new mechanism for regulation of microRNA activity. Nat. Genet. 39:1033–1037.
28. Gautam P, Chaurasia A, Bhattacharya A, Grover R, Indian Genome Variation Consortium, Mukerji M, Natarajan VT. 2015. Population diversity and adaptive evolution in keratinization genes: impact of environment in shaping skin phenotypes. Mol. Biol. Evol. 32:555–573.
29. Goldstone JV, McArthur AG, Kubota A, Zanette J, Parente T, Jönsson ME, Nelson DR, Stegeman JJ. 2010. Identification and developmental expression of the full complement of Cytochrome P450 genes in Zebrafish. BMC Genomics 11:643.
30. Grover D, Kannan K, Brahmachari SK, Mukerji M. 2005. ALU-ring elements in the primate genomes. Genetica 124:273–289.
31. Grover D, Majumder PP, B Rao C, Brahmachari SK, Mukerji M. 2003. Nonrandom distribution of alu elements in genes of various functional categories: insight from analysis of human chromosomes 21 and 22. Mol. Biol. Evol. 20:1420–1424.
32. Grover D, Mukerji M, Bhatnagar P, Kannan K, Brahmachari SK. 2004. Alu repeat analysis in the complete human genome: trends and variations with respect to genomic composition. Bioinforma. Oxf. Engl. 20:813–817.
33. Guengerich FP, Wu Z-L, Bartleson CJ. 2005. Function of human cytochrome P450s: characterization of the orphans. Biochem. Biophys. Res. Commun. 338:465–469.
34. Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, Kjems J. 2013. Natural RNA circles function as efficient microRNA sponges. Nature 495:384–388.
35. Häsler J, Samuelsson T, Strub K. 2007. Useful "junk": Alu RNAs in the human transcriptome. Cell. Mol. Life

Sci. CMLS 64:1793–1800.
36. Häsler J, Strub K. 2006. Alu elements as regulators of gene expression. Nucleic Acids Res. 34:5491–5497.
37. Hoffman Y, Pilpel Y, Oren M. 2014. microRNAs and Alu elements in the p53-Mdm2-Mdm4 regulatory network. J. Mol. Cell Biol. 6:192–197.
38. Hsu S-D, Lin F-M, Wu W-Y, Liang C, Huang W-C, Chan W-L, Tsai W-T, Chen G-Z, Lee C-J, Chiu C-M, et al. 2011. miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res. 39:D163-169.
39. Hu H, Liu J-M, Hu Z, Jiang X, Yang X, Li J, Zhang Y, Yu H, Khaitovich P. 2018. Recently Evolved Tumor Suppressor Transcript TP73-AS1 Functions as Sponge of Human-Specific miR-941. Mol. Biol. Evol. 35:1063–1077.
40. Hu S-B, Xiang J-F, Li X, Xu Y, Xue W, Huang M, Wong CC, Sagum CA, Bedford MT, Yang L, et al. 2015. Protein arginine methyltransferase CARM1 attenuates the paraspeckle-mediated nuclear retention of mRNAs containing IRAlus. Genes Dev. 29:630–645.
41. Kelley DR, Hendrickson DG, Tenen D, Rinn JL. 2014. Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. Genome Biol. 15:537.
42. Kiezun A, Artzi S, Modai S, Volk N, Isakov O, Shomron N. 2012. miRviewer: a multispecies microRNA homologous viewer. BMC Res. Notes 5:92.
43. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.
44. Lee JY, Ji Z, Tian B. 2008. Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3'-end of genes. Nucleic Acids Res. 36:5581–5590.
45. Lin L, Shen S, Tye A, Cai JJ, Jiang P, Davidson BL, Xing Y. 2008. Diverse splicing patterns of exonized Alu elements in human tissues. PLoS Genet. 4:e1000225.
46. Lynch VJ, Leclerc RD, May G, Wagner GP. 2011. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat. Genet. 43:1154–1159.
47. Mandal AK, Pandey R, Jha V, Mukerji M. 2013. Transcriptome-wide expansion of non-coding regulatory switches: evidence from co-occurrence of Alu exonization, antisense and editing. Nucleic Acids Res. 41:2121–2137.
48. Matoulkova E, Michalova E, Vojtesek B, Hrstka R. 2012. The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells. RNA Biol. 9:563–576.
49. Mullany LE, Herrick JS, Wolff RK, Slattery ML. 2016. MicroRNA Seed Region Length Impact on Target Messenger RNA Expression and Survival in Colorectal Cancer. PLoS ONE [Internet] 11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4849741/
50. Pan S-T, Xue D, Li Z-L, Zhou Z-W, He Z-X, Yang Y, Yang T, Qiu J-X, Zhou S-F. 2016. Computational Identification of the Paralogs and Orthologs of Human Cytochrome P450 Superfamily and the Implication in Drug Discovery. Int. J. Mol. Sci. 17:1020.
51. Pandey R, Bhattacharya A, Bhardwaj V, Jha V, Mandal AK, Mukerji M. 2016. Alu-miRNA interactions modulate transcript isoform diversity in stress response and reveal signatures of positive selection. Sci. Rep. 6:32348.
52. Pandey R, Mandal AK, Jha V, Mukerji M. 2011. Heat shock factor binding in Alu repeats expands its involvement in stress through an antisense mechanism. Genome Biol. 12:R117.
53. Pandey R, Mukerji M. 2011. From 'JUNK' to Just Unexplored Noncoding Knowledge: the case of transcribed Alus. Brief. Funct. Genomics 10:294–311.
54. Payer LM, Steranka JP, Ardeljan D, Walker J, Fitzgerald KC, Calabresi PA, Cooper TA, Burns KH. 2019. Alu insertion variants alter mRNA splicing. Nucleic Acids Res. 47:421–431.
55. Payer LM, Steranka JP, Yang WR, Kryatova M, Medabalimi S, Ardeljan D, Liu C, Boeke JD, Avramopoulos D, Burns KH. 2017. Structural variants caused by Alu insertions are associated with risks for many human diseases. Proc. Natl. Acad. Sci. U. S. A. 114:E3984–E3992.

56. Paz-Yaacov N, Levanon EY, Nevo E, Kinar Y, Harmelin A, Jacob-Hirsch J, Amariglio N, Eisenberg E, Rechavi G. 2010. Adenosine-to-inosine RNA editing shapes transcriptome diversity in primates. Proc. Natl. Acad. Sci. 107:12174–12179.
57. Polak P, Domany E. 2006. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. BMC Genomics 7:133.
58. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. 2010. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 465:1033–1038.
59. Ptak A, Ludewig G, Rak A, Nadolna W, Bochenek M, Gregoraszczuk EL. 2010. 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. 36:935–941.
60. Puthumana J, Lee M-C, Park JC, Kim H-S, Hwang D-S, Han J, Lee J-S. 2017. Ultraviolet B radiation induces impaired lifecycle traits and modulates expression of cytochrome P450 (CYP) genes in the copepod Tigriopus japonicus. Aquat. Toxicol. Amst. Neth. 184:116–122.
61. Rani N, Nowakowski TJ, Zhou H, Godshalk SE, Lisi V, Kriegstein AR, Kosik KS. 2016. A Primate lncRNA Mediates Notch Signaling during Neuronal Development by Sequestering miRNA. Neuron 90:1174–1188.
62. Rebollo R, Romanish MT, Mager DL. 2012. Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. Annu. Rev. Genet. 46:21–42.
63. Rosenthal JJC, Seeburg PH. 2012. A-to-I RNA editing: effects on proteins key to neural excitability. Neuron 74:432–439.
64. Roy-Engel AM, El-Sawy M, Farooq L, Odom GL, Perepelitsa-Belancio V, Bruch H, Oyeniran OO, Deininger PL. 2005. Human retroelements may introduce intragenic polyadenylation signals. Cytogenet. Genome Res. 110:365–371.
65. Shen S, Lin L, Cai JJ, Jiang P, Kenkel EJ, Stroik MR, Sato S, Davidson BL, Xing Y. 2011. Widespread establishment and regulatory impact of Alu exons in human genes. Proc. Natl. Acad. Sci. 108:2837–2842.
66. Sobczak K, Krzyzosiak WJ. 2002. Structural determinants of BRCA1 translational regulation. J. Biol. Chem. 277:17349–17358.
67. Sorek R. 2009. When new exons are born. Heredity 103:279–280.
68. Sorek R, Ast G, Graur D. 2002. Alu-containing exons are alternatively spliced. Genome Res. 12:1060–1067.
69. Spengler RM, Oakley CK, Davidson BL. 2014. Functional microRNAs and target sites are created by lineage-specific transposition. Hum. Mol. Genet. 23:1783–1793.
70. Stark K, Wu Z-L, Bartleson CJ, Guengerich FP. 2008. mRNA distribution and heterologous expression of orphan cytochrome P450 20A1. Drug Metab. Dispos. Biol. Fate Chem. 36:1930–1937.
71. Suzuki A, Makinoshima H, Wakaguri H, Esumi H, Sugano S, Kohno T, Tsuchihara K, Suzuki Y. 2014. Aberrant transcriptional regulations in cancers: genome, transcriptome and epigenome analysis of lung adenocarcinoma cell lines. Nucleic Acids Res. 42:13557–13572.
72. Tay Y, Rinn J, Pandolfi PP. 2014. The multilayered complexity of ceRNA crosstalk and competition. Nature 505:344–352.
73. Toll-Riera M, Castelo R, Bellora N, Albà MM. 2009. Evolution of primate orphan proteins. Biochem. Soc. Trans. 37:778–782.
74. Tristán-Flores FE, Guzmán P, Ortega-Kermedy MS, Cruz-Torres G, de la Rocha C, Silva-Martínez GA, Rodríguez-Ríos D, Alvarado-Caudillo Y, Barbosa-Sabanero G, Sayols S, et al. 2018. Liver X Receptor–Binding DNA Motif Associated With Atherosclerosis-Specific DNA Methylation Profiles of Alu Elements and Neighboring CpG Islands. J. Am. Heart Assoc. Cardiovasc. Cerebrovasc. Dis. [Internet] 7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5850253/
75. Trizzino M, Park Y, Holsbach-Beltrame M, Aracena K, Mika K, Caliskan M, Perry GH, Lynch VJ, Brown CD. 2017. Transposable elements are the primary source of novelty in primate gene regulation. Genome Res. 27:1623–1633.
76. Versteeg R, van Schaik BDC, van Batenburg MF, Roos M, Monajemi R, Caron H, Bussemaker HJ, van Kampen

AHC. 2003. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. Genome Res. 13:1998–2004.
77. Wang J, Liu X, Wu H, Ni P, Gu Z, Qiao Y, Chen N, Sun F, Fan Q. 2010. CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. Nucleic Acids Res. 38:5366–5383.
78. Wang L, Rishishwar L, Mariño-Ramírez L, Jordan IK. 2017. Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. Nucleic Acids Res. 45:2318–2328.
79. Wang Y, Xu Z, Jiang J, Xu C, Kang J, Xiao L, Wu M, Xiong J, Guo X, Liu H. 2013. Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. Dev. Cell 25:69–80.
80. Xie H, Wang M, Bonaldo M de F, Smith C, Rajaram V, Goldman S, Tomita T, Soares MB. 2009. High-throughput sequence-based epigenomic analysis of Alu repeats in human cerebellum. Nucleic Acids Res. 37:4331–4340.
81. Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer MA, et al. 2009. Mobile elements create structural variation: analysis of a complete human genome. Genome Res. 19:1516–1526.
82. Zhang Z, Li J, Zhao X-Q, Wang J, Wong GK-S, Yu J. 2006. KaKs_Calculator: Calculating Ka and Ks Through Model Selection and Model Averaging. Genomics Proteomics Bioinformatics 4:259–263.
83. Zhang Z, Xiao J, Wu J, Zhang H, Liu G, Wang X, Dai L. 2012. ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments. Biochem. Biophys. Res. Commun. 419:779–781.

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.

S.no.	miRNA	MIM ID	Expression(≥50 TPM)		No. of MREs with binding energy≤ -25Kcal/mol
			MCF-7	Pr. Neurons	
1	hsa-miR-941	MIMAT0004984	119639	950.33	10
2	hsa-miR-3677-3p	MIMAT0018101	2892.33	51.67	12
3	hsa-miR-1304-3p	MIMAT0022720	1922.5	80.33	10
4	hsa-miR-4446-3p	MIMAT0018965	1839.33	-	13
5	hsa-miR-296-3p	MIMAT0004679	1406.5	70.33	10
6	hsa-miR-1254	MIMAT0005905	1235	-	10
7	hsa-miR-6724-5p	MIMAT0025856	330.17	-	20
8	hsa-miR-619-5p	MIMAT0026622	193.17	89	26
9	hsa-miR-1908-3p	MIMAT0026916	191.67	-	16
10	hsa-miR-3944-3p	MIMAT0018360	158.67	-	14
11	hsa-miR-6842-3p	MIMAT0027587	158.17	175.67	10
12	hsa-miR-4767	MIMAT0019919	129.83		18
13	hsa-miR-5096	MIMAT0020603	98.5	72.33	14
14	hsa-miR-7703	MIMAT0030018	97.17	-	13
15	hsa-miR-939-5p	MIMAT0004982	81.83	-	19
16	hsa-miR-3620-5p	MIMAT0022967	81.33	-	21
17	hsa-miR-1226-5p	MIMAT0005576	81.17	-	18
18	hsa-miR-1915-5p	MIMAT0007891	78.5	-	14
19	hsa-miR-6732-3p	MIMAT0027366	57.5	-	13
20	hsa-miR-1273g-3p	MIMAT0022742	56.67	-	11
21	hsa-miR-4707-3p	MIMAT0019808	53	-	10
22	hsa-miR-668-3p	MIMAT0003881	-	175.33	10
23	hsa-miR-370-3p	MIMAT0000722	-	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.*

S.no.	miRNA name	Total no. of MREs on CYP20A1 3'UTR	Average of Overall complemntarity of miRNA with CYP20A1 3'UTR	MRES with seed match(2-7 nt) including wobble	Number of MREs with mismatch or insertion at bulge position	Features of MREs		
						Forward score	Binding energy	Alignment length
1	hsa-miR-941	29	59.37	2	1	146	-22.05	21
2	hsa-miR-3677-3p	57	54.14	3	-	-	-	-
3	hsa-miR-1304-3p	28	59.57	12	-	-	-	-
4	hsa-miR-4446-3p	34	60.56	3	1	120	-27.05	22
5	hsa-miR-296-3p	42	54.76	5	1	146	-25.23	25
6	hsa-miR-1254	45	59.72	18	2	126,128	-22.05, -23.75	25,23
7	hsa-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
8	hsa-miR-619-5p	45	67.17	27	1	115	-21.76	20
9	hsa-miR-1908-3p	43	52.60	1	-	-	-	-
10	hsa-miR-3944-3p	64	54.82	23	-	-	-	-
11	hsa-miR-6842-3p	39	59.20	11	1	104	-24.14	29
12	hsa-miR-4767	72	54.71	9	2	101,124	-25.3, -23.93	27, 16
13	hsa-miR-5096	22	65.80	8	-	-	-	-
14	hsa-miR-7703	51	57.20	4	-	-	-	-
15	hsa-miR-939-5p	63	51.32	9	-	-	-	-
16	hsa-miR-3620-5p	77	53.71	32	2	108,130	-21.24, -22.89	22,20
17	hsa-miR-1226-5p	73	56.95	6	1	124	-26.76	29
18	hsa-miR-1915-5p	40	54.88	1	-	-	-	-
19	hsa-miR-6732-3p	35	60.37	0	-	-	-	-
20	hsa-miR-1273g-3p	30	67.30	12	-	-	-	-
21	hsa-miR-4707-3p	30	58.03	6	-	-	-	-
22	hsa-miR-668-3p	38	52.86	12	-	-	-	-
23	hsa-miR-370-3p	56	56.57	2	-	-	-	-

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

Supplementary Information

Mapping 319 genes onto the apoptosis pathway

A transcriptome-wide scan for the co-occurrence of Alu exonization, editing and antisense at the single-exon level had revealed a set of 319 genes, enriched for apoptosis (Mandal et al. 2013). Because enrichment analyses are heavily confounded by background data used for normalization, we found it prudent to ask how many of these 319 genes are actually involved in apoptotic cell death and where in the apoptosis pathway they mapped. To begin with, we listed all the important steps in this pathway: arrest of division is usually the first cellular response to any apoptotic trigger, which is followed by cessation of DNA damage repair (DDR), initiation of p53 signaling and polymerase halt. At the later stages of apoptosis, endonucleases are activated and nuclear DNA fragmentation occurs, mitochondrial ROS may be generated and many proteins are ubiquitinylated and marked for proteasomal degradation. The final stages are marked by blebbing and membrane flipping to attract phagocytes to engulf the dying cell. A flow diagram was constructed incorporating all the nodes and pathways involved in apoptosis. We performed an extensive, systematic manual curation of these 319 genes using published literature and other web resources with the intent of binning them in the major events of apoptosis listed above. We asked i.) which of these genes are involved in apoptosis, and ii.) for those that are involved, where in the pathway they occur. We mapped all those genes which interact with known apoptotic genes (like *BCL2* family members) or contain known motifs (e.g., WD repeats), or are a part of apoptotic signaling pathway, or have apoptosis (or any of its sub-readouts) as the major phenotype upon their perturbation. Many of the genes that we mapped are involved in cell division, DNA replication and ubiquitinylation; perturbations in these genes often lead to cell cycle arrest, impaired DDR and protein misfolding/ degradation, respectively. Many genes are involved in multiple, interlinked functions but we had mapped them uniquely. Mapping was based on the major function of the gene, its mutant phenotype or the function affected by perturbation (knockdown, knock-out or over-expression), disease association and evolutionary conservation (wherever orthologues could be reliably ascertained). To check the accuracy of our curation and mapping, we used GeneMANIA (attributes: physical and genetic interactions, co-localization, same biological pathway) to fetch the top 100 interacting partners of our 91 apoptosis genes. There was no overlap between these interacting partners and the 228 genes that we have not mapped to apoptosis, reassuring us that our mapping stringency has been optimal.

Disparities in *CYP20A1* annotation across different portals

There exists a huge disparity in the annotation of *CYP20A1* transcript isoforms in different databases. There are 10, 9 and 4 annotated isoforms in Ensembl, NCBI and UCSC, respectively. All the four UCSC isoforms (GENCODE v24) are mapped onto a single RefSeq transcript NM_177538. Among the 10 isoforms listed in Ensembl, three are targeted for NMD and another three are processed transcripts. In NCBI, out of the 9 transcripts, two are miscellaneous RNAs, six are predicted protein-coding and only one is experimentally validated (NM_177538.2), which is the largest isoform and also annotated as the principal isoform (by APPRIS). It matches with the longest isoform on UCSC in terms of length (10.94kb) as well. Ensembl previously mapped it to ENST00000611416.4 (which has been removed from the current build - release 94) but currently maps this onto its transcript which is just 1949bp long. Notably, the size of the CYP20A1 protein remains the same (462aa) on every portal. These annotations are based on microarray or RNA-seq data, both of which have their own limitations. Array imputes the expression based on one or few probes and cross hybridization is a potential confound. RNA-seq gives an FPKM value for a transcript regardless of the reads aligning over the entire length of the transcript. The scenario is further complicated because 65% of the long-3'UTR of *CYP20A1* is made up of Alu repeats which is typically not captured by array probes and where sequencing reads do not map uniquely. Also, from these data, one cannot discriminate if the full length 3'UTR containing isoform is expressed as a single transcript or as multiple short RNAs.

Primers used in this study

Oligo Name	Sequence (5'→3')	Length (nt)	Tm (°C)	GC%	Amplicon size (bp)	
CYP20A1_5'UTR_FP1	TGTCTGAAATCGTGTGCAC	19	56.8	47.4	222	used as the "5'UTR primer"
CYP20A1_5'UTR_RP1	CCAATTTAAGGCATAGCGTG	20	55.7	45.0		
CYP20A1_5'UTR_FP2	ACAGAGCACTACTAACTCCT	20	55.5	45.0	214	
CYP20A1_5'UTR_RP2	CAGAGGAAAGAGCAATGGAT	20	55.7	45.0		
CYP20A1_CDS_FP3a	TCACTGTTGATGAGAATGGG	20	55.6	45.0	655	
CYP20A1_CDS_FP3b	GTATTGGTGAAGAGACTGCA	20	55.7	45.0	470	
CYP20A1_3'UTR_RP3	ACAAACATGATTCCCCAACT	20	55.7	40.0		
CYP20A1_3'UTR_FP4	TCCCTTCCCACTTATTTTCT	20	54.1	40.0	166	
CYP20A1_3'UTR_RP4	TTGCTTTAAACTCCACCTCT	20	55.1	40.0		
CYP20A1_3'UTR_FP5	ACCAAGTGTCAGATCAGATG	20	55.4	45.0	177	
CYP20A1_3'UTR_RP5	GCTCTCTTTTCCTTGATTCC	20	54.6	45.0		
CYP20A1_3'UTR_FP6	GAGAGCATAGAGGAATCAGC	20	55.7	50.0	216	
CYP20A1_3'UTR_RP6	GTGAAGTGGCCAAAATTTGA	20	55.5	40.0		
CYP20A1_3'UTR_FP7	TTGTCACTCTGAATGTAGGC	20	55.7	45.0	207	
CYP20A1_3'UTR_RP7	AAAATGAGCTGGGTGTGATG	20	56.9	45.0		
CYP20A1_3'UTR_FP8	GCCACTACACTCAGCTAATT	20	55.7	45.0	238	
CYP20A1_3'UTR_RP8	GTTGATTGAGCCCTCCTTAA	20	55.6	45.0		
CYP20A1_3'UTR_FP9	TTTGTAAGAGCTTCAGGGAA	20	54.8	40.0	209	used as the "3'UTR primer"
CYP20A1_3'UTR_RP9	CACCGAACTGTAAACCAATT	20	54.7	40.0		
CYP20A1_3'UTR_FP10	GCTTCAGGGAATAGGCTTTT	20	56.3	45.0	177	
CYP20A1_3'UTR_RP10	TTGGGAGTAGAACTGGAAGA	20	55.7	45.0		
CYP20A1_Exon5_FP	GGAAGTCATTCGCTTCCAGA	20	64.3	50.0	277 (including exon6) & 196 (excluding exon 6)	for discrimating the transcript isoforms
CYP20A1_Exon8_RP	TATGCAACTGGCCAGAGAAA	20	63.3	45.0		
MALAT1_FP	CTTCCTGTGGCAGGAGAGAC	20	64.1	60.0	217	
MALAT1_RP	CGCTTGAGATTTGGGCTTTA	20	64	45.0		
GAPDH_FP	CGACCACTTTGTCAAGCTCA	20	58.99	50.0	138	
GAPDH_RP	CTTCCTCTTGTGCTCTTGCTG	21	59.77	52.4		
Spike-in FP	CCTCTTGTATCTCACAGCTCAA	22	54.1	45.5	141	zebrafish lncRNA durga IVT product
Spike-in RP	CTGGAGACAATAGAAGGCAAT	21	52.2	42.9		
ACTB_FP	GTCTTCCCCTCCATCGTG	18	63.3	61.1	126	
ACTB_RP	GATGGGGTACTTCAGGGTGA	20	63.8	55.0		
18S rRNA_FP	GGCCCTGTAATTGGAATGAGTC	22	59.63	50	146	
18S rRNA_RP	CCAAGATCCAACTACGAGCTT	21	58.63	47.6		

Antibodies used in WB

Sl.no.	Company	Catalogue no.	Antibody name	Diltuion	Incubation
1	abcam	136094	Rabbit polyclonal to *CYP20A1*	1:500	Overnight at 4°, without shaking
2	Santa cruz	sc32233	Mouse monoclonal anti *GAPDH*	1:2000	RT for 2 hrs, with shaking
3	abcam	176842	Rabbit monoclonal anti H3	1:1000	Overnight at 4°, with shaking
4	Santa cruz	sc 2004	anti rabbit IgG-HRP conjugated rasied in goat	1:10,000	RT for 1 hour, with shaking
5	Santa cruz	sc 2005	anti mouse IgG-HRP conjugated rasied in goat	1:10,000	RT for 1 hour, with shaking

Details of 3'RACE amplicons

Sl. No.	Band size (bp)	Primer	Seq. length (nt)	NCBI BLAST (hg38)			UCSC BLAT (hg38)
				Top hit	Identity (%)	Query cover (%)	
1	~1500	FP10	546	*CYP20A1* mRNA (NM_177538.2)	99	34	Mapped to *CYP20A1* 3'UTR with high score
3	>800	FP10	181	*CYP20A1* mRNA (NM_177538.2)	98	100	Mapped to *CYP20A1* 3'UTR with high score
5	>700	FP10	656	*CYP20A1* mRNA (NM_177538.2)	99	28	Mapped to *CYP20A1* 3'UTR with high score
7	>600	FP10	592	*CYP20A1* mRNA (NM_177538.2)	98	32	Mapped to Chr.6(-) with low score
9	~500	FP10	362	*MFSD1*	95	99	Mapped to *MFSD1* exon 16 (3'UTR) with very high score
* no significant similarity was found with inner primer for any of the bands							

Details of nuclear cytosolic RT-qPCR

Experiment	Lowest(#) Ct difference between +RT and -RT					Ct value in NTC					Ct variation in Spike-In +RT(#)
	3'UTR primer	5'UTR primer	*MALAT1*	*GAPDH*	Spike-In	3'UTR primer	5'UTR primer	*MALAT1*	*GAPDH*	Spike-In	
1	ND	7.19	9.34	16.44	ND	ND	35.12	33.35*	39.5	35.39*	0.20
2	7.31	6.45	9.04	11.77	6.66	40.93*	33.99	33.3*	33.55	31.67*	0.06
3	2.87	7.40	8.93	13.03	4.88	32.03	32.56	33.49*	34.57	34.99*	0.58
4	3.81	6.71	9.56	14.13	4.85						1.09

Noise band was set at 1.00 in all experiments
among the total, nuclear and cytosolic fractions
* Ct value detected in one technical replicate only
ND = Ct Not detected in -RT; NTC = no template control
RT-qPCR for both experiment 3 and 4 were put in the same plate, therefore single NTC value

miRNA expression in MCF-7 cells

We wanted to use publicly available expression data to find out which of the 994 miRNAs with predicted MREs on *CYP20A1* 3'UTR, are expressed in our study system. In order to reduce false positives, we decided to take a consensus of miRNAs found to be expressed in MCF-7 in both microarray (doi:10.1371/journal.pone.0049067) as well as NGS (GSE68246) datasets. In the microarray data, 104 out of our 994 miRNAs of interest were expressed (averaged signal range 7.6 to 974.63). miRNome data of MCF-7 monolayer culture (using Illumina TruSeq small RNA on HiSeq2000) revealed the expression of 102 out of our 994 miRNAs.

Although we found around 100 (out of 994) miRNAs to be expressed in each of these data yet there was absolutely no overlap between the two sets.

Details of small RNA sequencing data

Sample	Total # of reads	% of 5'adapter +ve reads	% of 3'adapter +ve reads	# of reads post length filtering	# of reads post quality filtering	# of reads mapped
Pr. neuron Set 1	58507953	3.52	96.48	45873165	44297019	36653535
Pr. neuron Set 2	65588157	4.07	95.93	52787771	51364925	40148127
Pr. neuron Set 3	35409327	3.57	96.43	31974327	31426458	25742827
MCF-7 Set 1	49174023	3.76	96.24	43824299	41480248	36657502
MCF-7 Set 2	94370695	4.27	95.73	83124005	78450081	68786702
MCF-7 Set 3	59179178	8.05	91.95	50024126	46832593	46832593
MCF-7 Set 4	75156681	7.77	92.23	62389991	57937708	49243044
MCF-7 Set 5	62465203	8.56	91.44	54622720	51028477	42745201
MCF-7 Set 6	73825714	3.73	96.27	64923259	60749505	53921524

Sample-wise data generated in primary neurons RNA-seq experiment

Sample Name	Read 1	Read 2	Total data (in GB)
Untreated Set 1	7.8	7.9	15.7
Untreated Set 2	5.2	5.2	10.4
Untreated Set 3	11.8	11.2	23
HS treated + 1hr recovery Set 1	7.5	7.6	15.1
HS treated + 1hr recovery Set 2	8.8	8.8	17.6
HS treated + 1hr recovery Set 3	9.8	9.9	19.7
Tat treated + with recovery Set 1	7.2	7.3	14.5
Tat treated + with recovery Set 2	7.8	7.9	15.7
Tat treated + with recovery Set 3	6.8	6.9	13.7

Supplementary Table Legends

Table S1: 91 genes (out of 319 containing the cooccurrence of Alu exonization, A-to-I editing and cis-antisense at single exon level) were mapped onto the apoptosis network. Out of these, 68 clustered around three central hubs (termed ‘mito.’, ‘p53’ and ‘ubi’).

Table S2: RNA-seq reads from *CYP20A1* different transcript isoforms mapping to 15928 single nuclei from different regions of human cerebral cortex.

Table S3: Expression of *CYP20A1* in human specific Rosehip neurons.

Table S4: 46 out of the 169 miRNAs with MREs on *CYP20A1* 3'UTR were present in miRDB as FuncMir. 19 of them occur exclusively in one or more primate species; some of them are also stress responsive.

Table S5: List of 994 miRNAs with predicted MREs on *CYP20A1* long-3'UTR. Out of these 994, 140 miRNAs have ≥10 MREs.
The miRNAs with ≥20 MREs or that are human specific or contain human specific seed changes were curated and available literature was collated to gain insights into the biological contexts where these might be functionally relevant.

Table S6: 380 genes that show concordant directionality of expression as *CYP20A1* long-3'UTR isoform in heat shock and Tat response.

Table S7: Major biological pathways enriched in toppfun analysis of 380 genes using Benjamini-Hochberg corrected FDR (q- value 0.05).

Supplementary Figure Legends

Figure S1: Out of 38300 genes, 512 have their longest 3'UTR greater than 6kb but only 98 of these contain Alu exonization events.

Figure S2: Snapshot (1-60aa) of multiple sequence alignment of CYP20A1 proteins depicts high degree of sequence conservation across vertebrates. As expected, the primates show the highest similarity to the human protein, which reduces as we move from the mammals to other non-mammalian vertebrates. The similarity is least for zebrafish which is also the most evolutionarily distant vertebrate from humans. (* exact aa conserved, : replaced by aa containing similar functional group)

Figure S3: UCSC track representing DNA conservation analysis of 10kb upstream of 5'UTR and 10 kb downstream of 3'UTR region among 20 mammals.

Figure S4: Expression of *CYP20A1* long-3'UTR isoform in MCF-7 cells was checked by RT-qPCR using two different primer pairs and normalized to the geometric mean of β-actin, *GAPDH* and 18S rRNA. The error bars represent the SD of four biological replicates. Average of three technical replicates was taken for each biological replicate.
In contrast to the low expression of our transcript of interest, CYP20A1 protein is highly expressed in MCF-7 cells (~50% of GAPDH level).

Figure S5: Transcript models built by Cufflinks based on RNA-seq data suggest that 3'UTRs longer than 9kb, occur for *CYP20A1*. TCONS_00101353 denotes the isoform studied by us.

Figure S6: >95% of the cells are immuno-positive for neuronal markers like MAP2, Tuj1 and NeuN, while no cells are stained with GFAP, a glial cell marker, ensuring it is a pure neuron culture.

Figure S7: **A.** 'Kill-curve' of different doses of HIV1-Tat on primary neurons was prepared using viability and cytotoxicity parameters (ApoToxGlo assay; Promega) after 24hrs of treatment.
B. Subsequently, the protocol was standardized for 6hrs treatment. ~40% cell death occurs in primary neurons upon 100ng/ml Tat treatment for 6hrs. The data represent the mean (±S.D.) of two independent experiments.

Figure S8: **A.** MTT assay was performed to quantify the cytotoxicity/ cytostaticity of paclitaxel on MCF-7 cells. The data represent the mean (±S.D.) of six independent experiments.
Student's t-test *→$p<0.05$, **→$p<0.01$, ***→$p<0.001$, ns → not significant
B. At inhibitory doses of paclitaxel, CYP20A1 protein is downregulated.

Figure 1

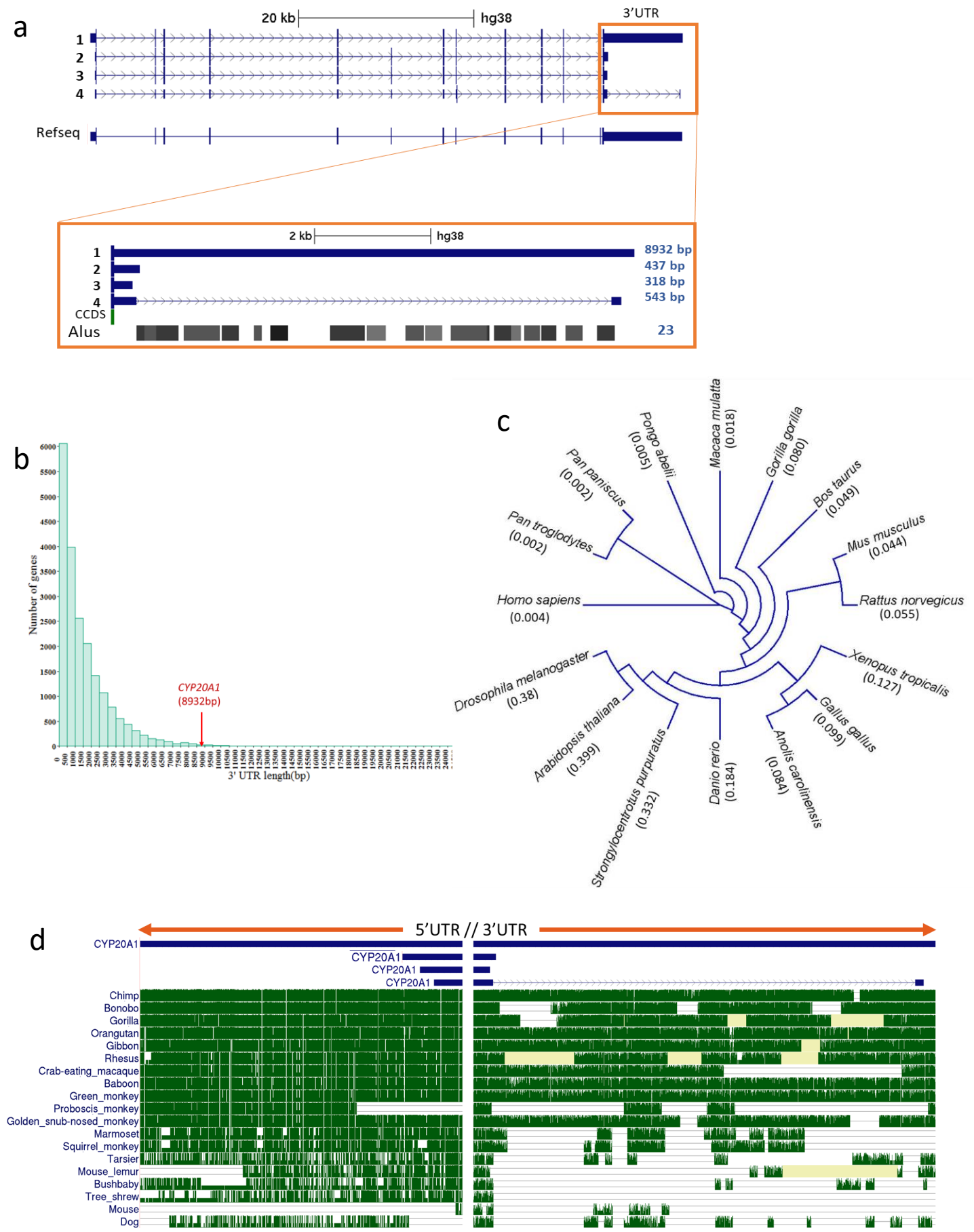
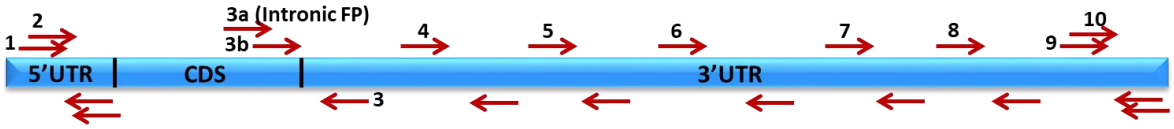
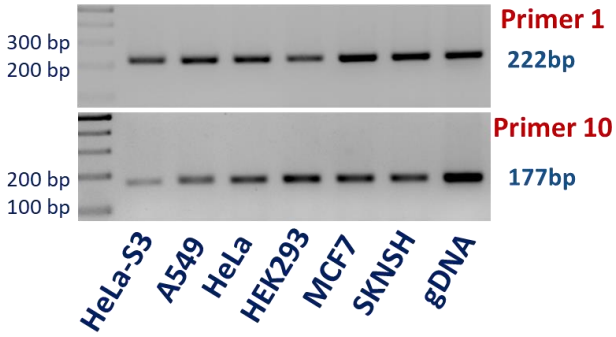

a
20 kb
hg38
3'UTR
1
2
3
4
Refseq
2 kb
hg38
8932 bp
437 bp
318 bp
543 bp
CCDS
Alus
23
b
Number of genes
CYP20A1
(8932bp)
3' UTR length(bp)
c
Macaca mulatta (0.018)
Gorilla gorilla (0.080)
Bos taurus (0.049)
Mus musculus (0.044)
Rattus norvegicus (0.055)
Xenopus tropicalis (0.127)
Gallus gallus (0.099)
Anolis carolinensis (0.084)
Danio rerio (0.184)
Strongylocentrotus purpuratus (0.332)
Arabidopsis thaliana (0.399)
Drosophila melanogaster (0.38)
Homo sapiens (0.004)
Pan troglodytes (0.002)
Pan paniscus (0.002)
Pongo abelii (0.005)
d
5'UTR // 3'UTR
CYP20A1
Chimp
Bonobo
Gorilla
Orangutan
Gibbon
Rhesus
Crab-eating_macaque
Baboon
Green_monkey
Proboscis_monkey
Golden_snub-nosed_monkey
Marmoset
Squirrel_monkey
Tarsier
Mouse_lemur
Bushbaby
Tree_shrew
Mouse
Dog

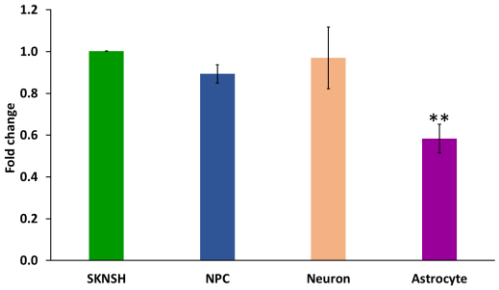
Figure 2

a

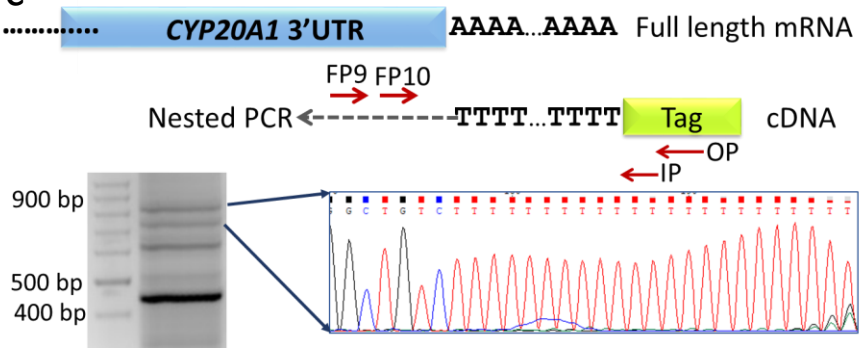




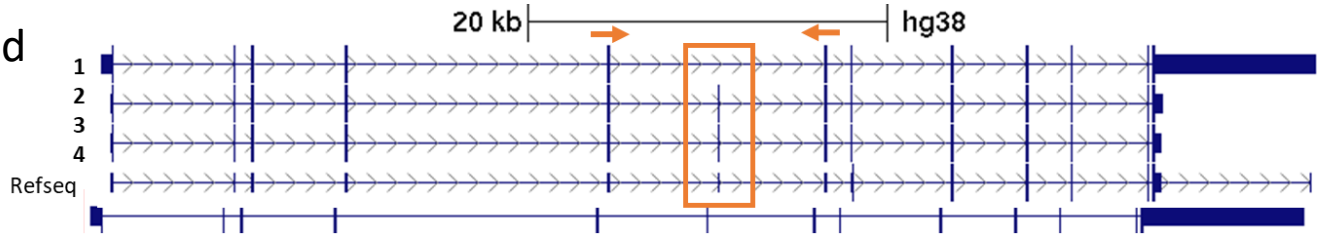
b

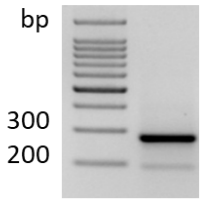


c



d



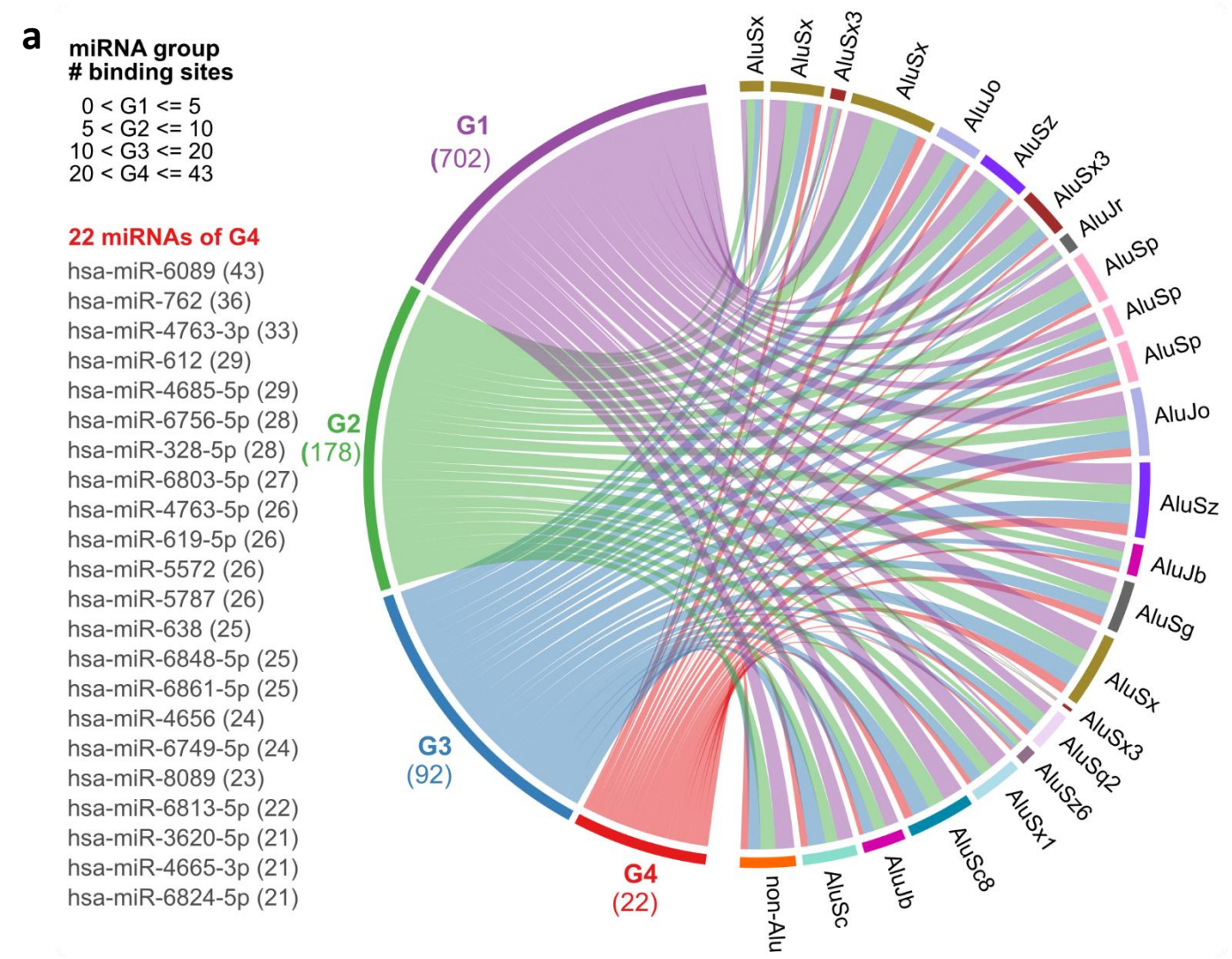


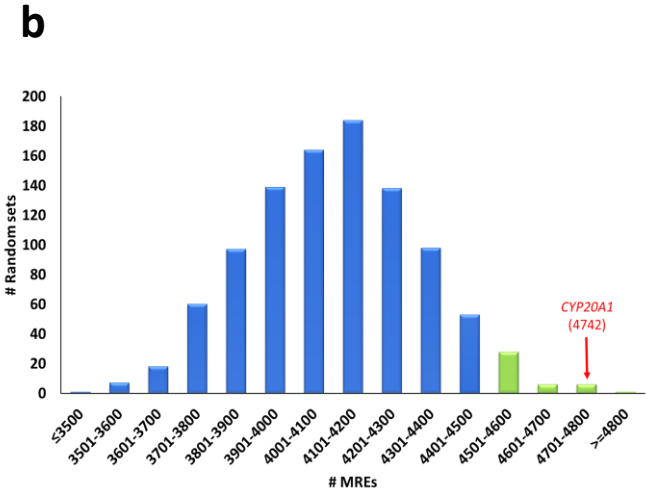
>gi|29171730|ref|NP_803882.1| cytochrome P450 20A1 [*Homo sapiens*]
Exon6: (LCIFTYFQ) VWSEIGKGFLDGSLDKNMTRKKQYEDA
Exon7: ALMQLESVLRNIIKERKGRNFSQHI//FIDSLVQGNLNDQQ

e

Isoforms	Schematic representation	Exon 6	ORF	3'UTR (bp)	Function
1		Skipped	Disrupted	8932	lncRNA
2 & 3		Retained	Maintained	437 & 318	Protein Coding

Figure 3


a
miRNA group
binding sites
0 < G1 <= 5
5 < G2 <= 10
10 < G3 <= 20
20 < G4 <= 43
22 miRNAs of G4
hsa-miR-6089 (43)
hsa-miR-762 (36)
hsa-miR-4763-3p (33)
hsa-miR-612 (29)
hsa-miR-4685-5p (29)
hsa-miR-6756-5p (28)
hsa-miR-328-5p (28)
hsa-miR-6803-5p (27)
hsa-miR-4763-5p (26)
hsa-miR-619-5p (26)
hsa-miR-5572 (26)
hsa-miR-5787 (26)
hsa-miR-638 (25)
hsa-miR-6848-5p (25)
hsa-miR-6861-5p (25)
hsa-miR-4656 (24)
hsa-miR-6749-5p (24)
hsa-miR-8089 (23)
hsa-miR-6813-5p (22)
hsa-miR-3620-5p (21)
hsa-miR-4665-3p (21)
hsa-miR-6824-5p (21)
G1
(702)
G2
(178)
G3
(92)
G4
(22)
AluSx
AluSx
AluSx3
AluSx
AluJo
AluSz
AluSx3
AluJr
AluSp
AluSp
AluSp
AluJo
AluSz
AluJb
AluSg
AluSx
AluSx3
AluSq2
AluSz6
AluSx1
AluSc8
AluJb
AluSc
non-Alu


b
Random sets
200
180
160
140
120
100
80
60
40
20
0
CYP20A1
(4742)
≤3500
3501-3600
3601-3700
3701-3800
3801-3900
3901-4000
4001-4100
4101-4200
4201-4300
4301-4400
4401-4500
4501-4600
4601-4700
4701-4800
>=4800
MREs

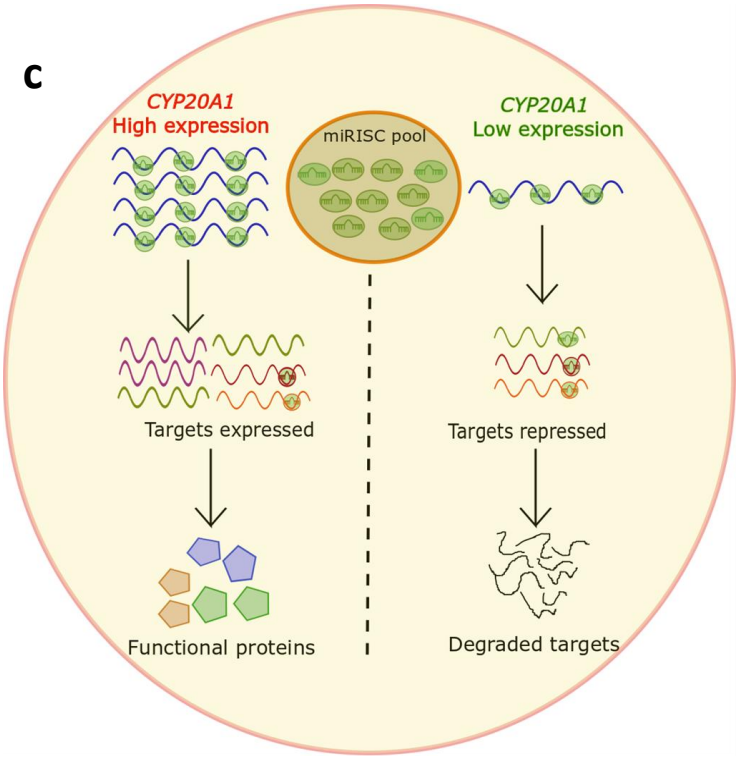
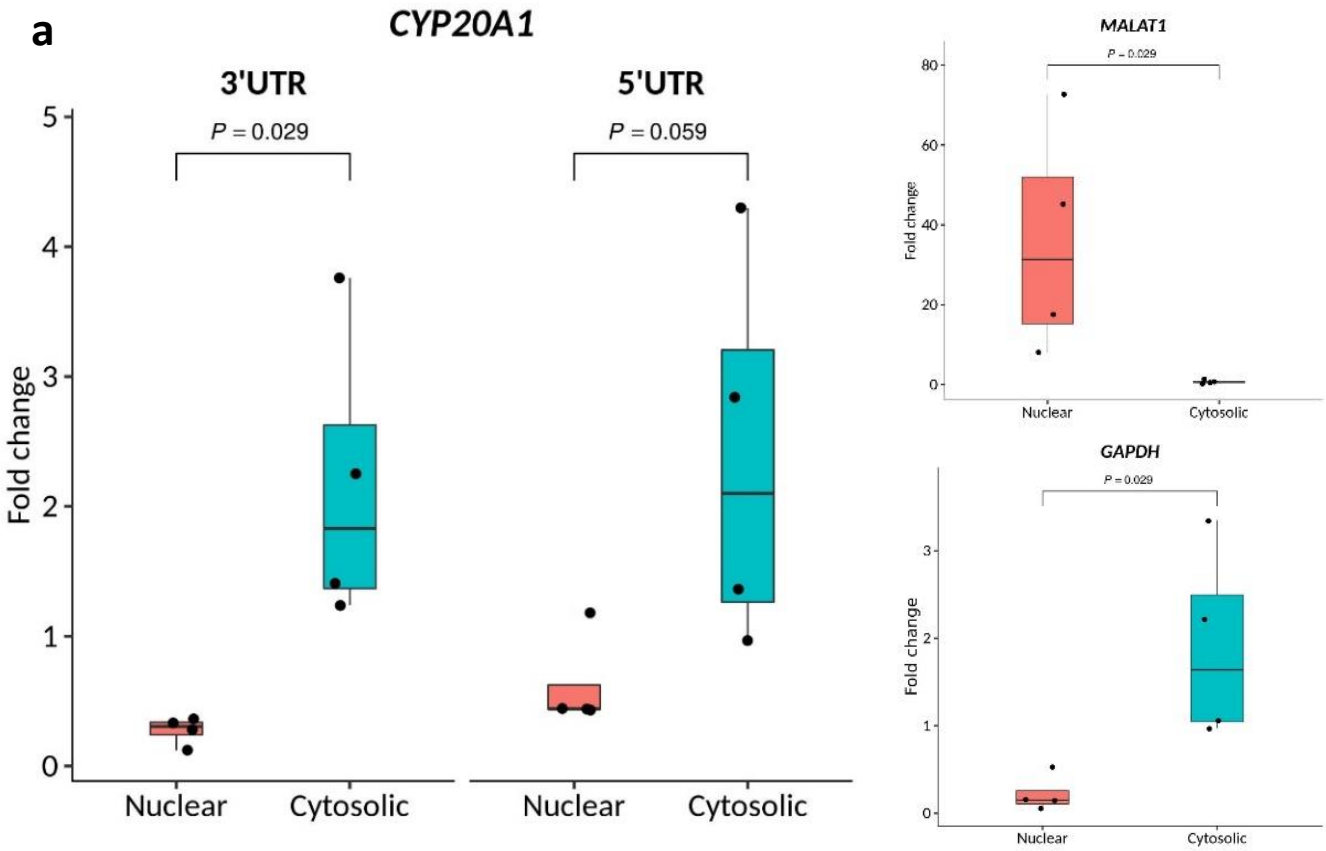
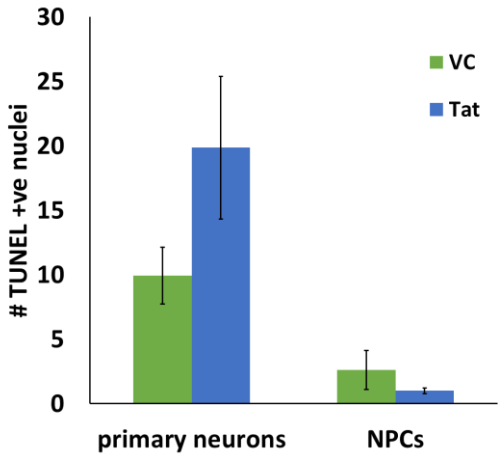
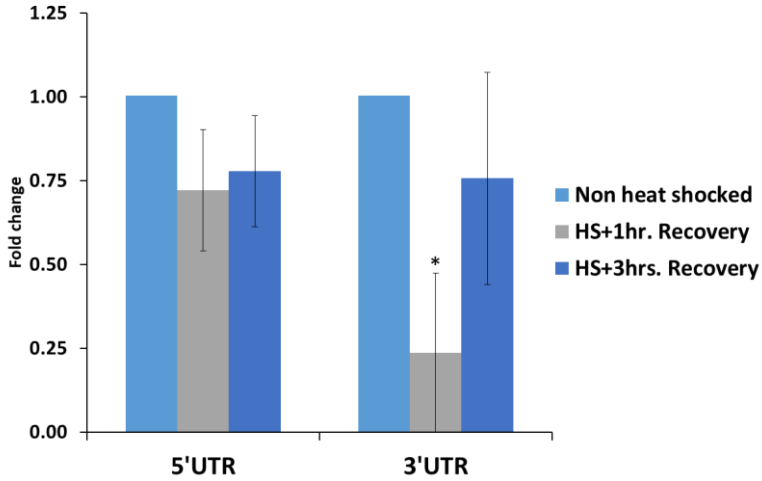
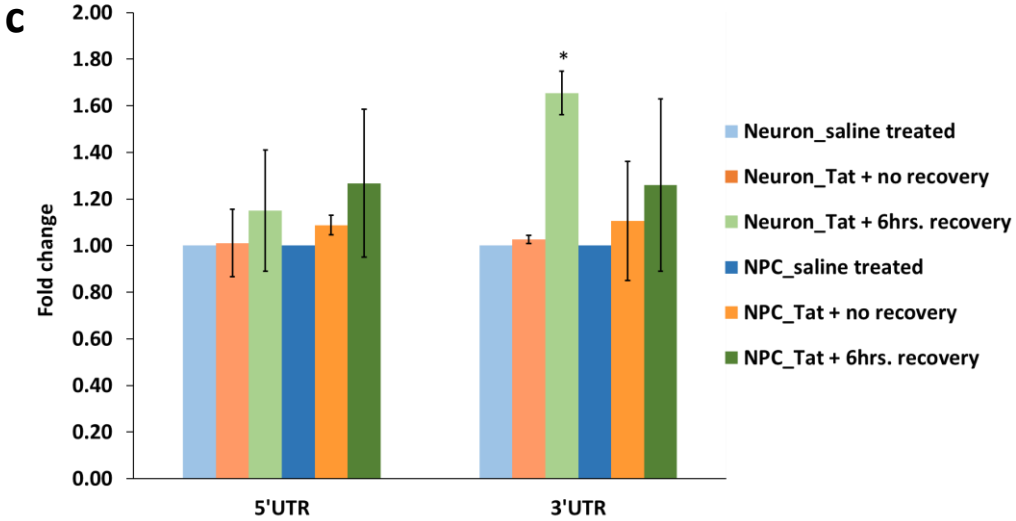

c
CYP20A1
High expression
miRISC pool
CYP20A1
Low expression
Targets expressed
Targets repressed
Functional proteins
Degraded targets

Figure 4


a
CYP20A1
3'UTR
5'UTR
P = 0.029
P = 0.059
Fold change
Nuclear
Cytosolic
MALAT1
P = 0.029
GAPDH
P = 0.029


b
TUNEL +ve nuclei
VC
Tat
primary neurons
NPCs


d
Fold change
Non heat shocked
HS+1hr. Recovery
HS+3hrs. Recovery
5'UTR
3'UTR


c
Fold change
Neuron_saline treated
Neuron_Tat + no recovery
Neuron_Tat + 6hrs. recovery
NPC_saline treated
NPC_Tat + no recovery
NPC_Tat + 6hrs. recovery
5'UTR
3'UTR

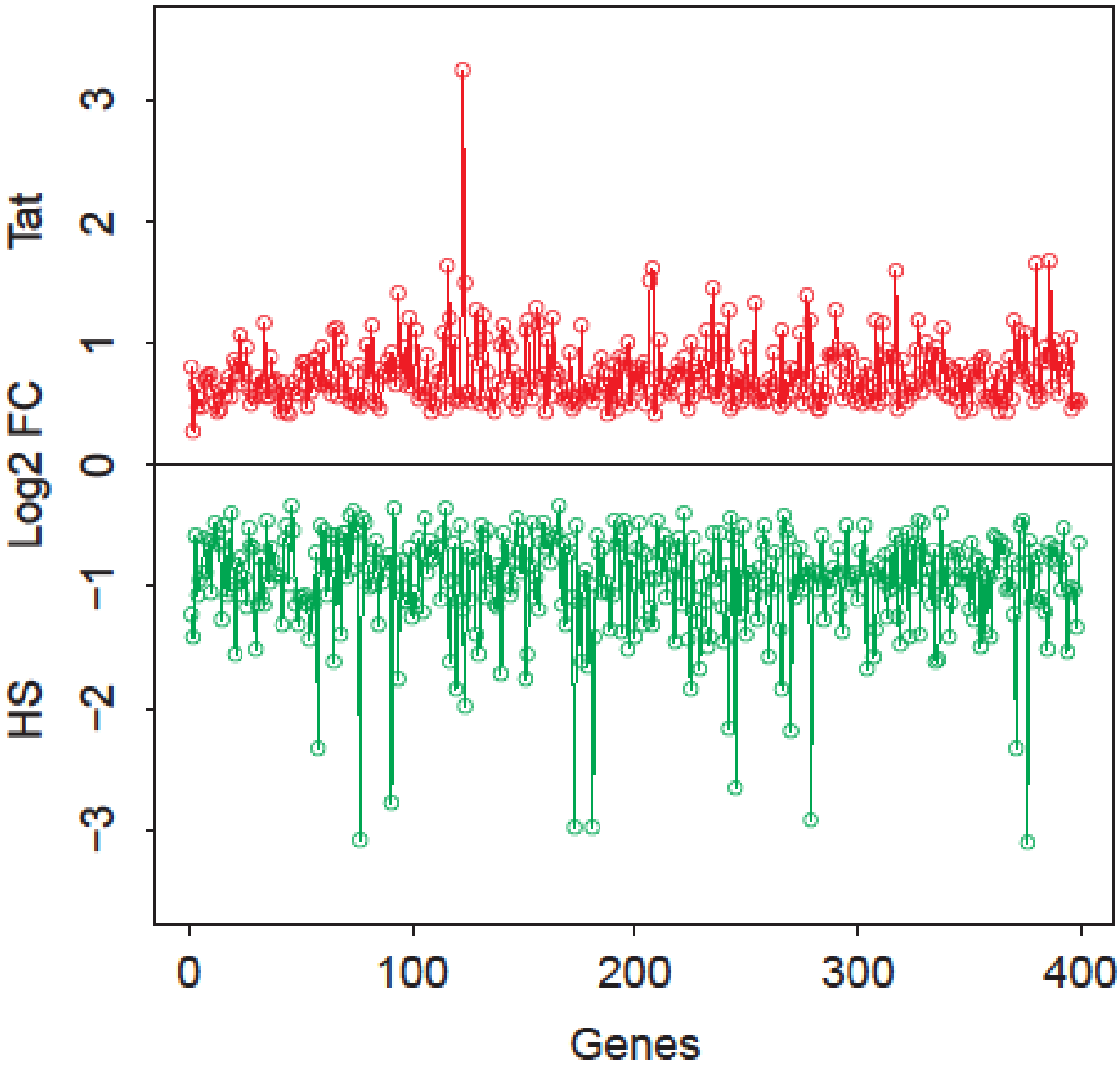



Figure 5