Kelley et al. Genome Biology 2014, 15:537
http://genomebiology.com/2014/15/12/537

Genome Biology

RESEARCH Open Access

Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions

David R Kelley[1,2], David G Hendrickson[1,2], Danielle Tenen[1,2] and John L Rinn[1,2,3*]

Abstract

Background: Transposable elements (TEs) have significantly influenced the evolution of transcriptional regulatory networks in the human genome. Post-transcriptional regulation of human genes by TE-derived sequences has been observed in specific contexts, but has yet to be systematically and comprehensively investigated. Here, we study a collection of 75 CLIP-Seq experiments mapping the RNA binding sites for a diverse set of 51 human proteins to explore the role of TEs in post-transcriptional regulation of human mRNAs and lncRNAs via RNA-protein interactions.

Results: We detect widespread interactions between RNA binding proteins (RBPs) and many families of TE-derived sequence in the CLIP-Seq data. Further, alignment coverage peaks on specific positions of the TE consensus sequences, illuminating a diversity of TE-specific RBP binding motifs. Evidence of binding and conservation of these motifs in the nonrepetitive transcriptome suggests that TEs have generally appropriated existing sequence preferences of the RBPs. Depletion assays for numerous RBPs show that TE-derived binding sites affect transcript abundance and splicing similarly to nonrepetitive sites. However, in a few cases the effect of RBP binding depends on the specific TE family bound; for example, the ubiquitously expressed RBP HuR confers transcript stability unless bound to an Alu element.

Conclusions: Our meta-analysis suggests a widespread role for TEs in shaping RNA-protein regulatory networks in the human genome.

Background

The staggering 45 to 60% of nucleotides in the human genome derived from transposable elements (TEs) remain an enigma in our understanding of the function and evolution of the human genome [1,2]. TEs are sequences capable of propagating by self-replication to new positions in the genome [3,4]. This ability comes in many forms, allowing for classification into a multitude of families [5]. The genomic role of TEs has followed an interesting arc — they were initially described as controlling elements in maize, due to the impact of insertions on local gene expression [6]. As their significance was recognized, it was noted that TEs' ability to self-replicate meant that a beneficial functional role was unnecessary to explain their conquest of the genome [3,4]. This led to their well-known categorization as junk DNA.

Recent research has revisited the topic of TE impact on gene expression, noting that the dissemination of highly similar sequence accomplished by TEs is a powerful way to link many diverse genomic regions into a regulatory network [7]. In a number of cases, extant TE sequences have integrated with established genomic functions and been co-opted by the genome for critical roles [7,8]. In the most studied paradigm, some TEs contain DNA binding site motifs for transcription factors and have rewired the transcriptional regulatory networks in which these transcription factors function by introducing many new binding sites via their insertions throughout the genome [9-14].

In the substantial portion of the genome transcribed into RNA [15], TE-derived sequences also appear in RNA transcripts where they can interact with RNA binding proteins (RBPs), which also often have preferred binding site motifs [16]. In perhaps the most understood and interesting

* Correspondence: john_rinn@harvard.edu
[1]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA
[2]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA 02142, USA
Full list of author information is available at the end of the article

example, the antisense strand of Alu elements contains motifs that recruit the cell's splicing machinery and have thus introduced hundreds of novel exons into various protein coding genes [17-19]. Sequence derived from TEs has also been implicated in both degradation [20] and increasing the translation [21] of RNA transcripts. However, the extent to which these examples generalize is unknown, as a comprehensive search for interactions between TEs and RBPs has not yet been performed. Such a search is further justified by the recent appreciation that long noncoding RNAs (lncRNAs), a class of more than 10,000 genes with a rapidly growing list of critical functional roles [22,23], contain TEs at a rate near the high genomic average but in biased proportions of the various individual families, suggesting a possible functional role [24,25].

Crosslinked immunoprecipitation (CLIP)-Seq is the state of the art technique for mapping the direct binding sites of an RBP. It involves crosslinking cells to lock RNA-protein interactions, immunoprecipitating the complexes using an antibody specifically targeted to the RBP, sequencing cDNA reverse transcribed from the captured RNA, and statistically analyzing the aligned sequencing reads [26]. CLIP-Seq has been applied to dozens of RBPs to study splicing regulation [27-29], translation efficiency [30-32], and explore RBPs mutated in neurological disorders [33]. These studies largely focused on uniquely mapping reads and ignored repetitive sequences, leaving the extent of RBP binding to TEs unexplored.

Here, we surveyed evidence for RBP binding to TE-derived RNA sequence in a collection of 75 CLIP-Seq experiments on 51 RBPs performed in human cells. We processed all datasets using a standardized CLIP-Seq analysis pipeline. In these data, RBP interactions with TE-derived sequences were widespread, and we detected hundreds of specific pairwise interactions. Alignment coverage clustered on specific regions of the TE consensus sequences. From these high coverage regions, we extracted a diversity of TE-specific motifs that extensively characterize the *in vivo* binding preferences of the RBPs. The presence of CLIP-Seq coverage and conservation at nonrepetitive instances of these motifs suggest that the TEs appropriated existing binding preferences of the RBPs. RBP binding to TE-derived sites influenced RNA abundance and splicing to a comparable extent as binding to nonrepetitive sites in RBP knockdown experiments. Altogether, our comprehensive meta-analysis suggests a widespread role for TEs in shaping post-transcriptional RNA-protein regulatory networks in the human genome.

Results

CLIP-Seq alignments are enriched in specific transposable elements

To comprehensively survey RBP interactions with TEs in the human genome, we collected and systematically processed all accessible CLIP-Seq datasets, settling on 75 experiments mapping 51 RBPs (Materials and methods). First, we compared the number of aligned reads, relative to total library size, overlapping all instances of each TE family in both orientations in the CLIP-Seq to a null model expectation (Figure 1A). We can split each TE family into its sense and antisense orientations because CLIP-Seq uses a strand-specific library construction method. To control for differing mappability of the TE families and varying expression levels of transcripts containing specific TE families, we compared alignment coverage to simulated coverage from a null model (see Materials and methods).

We found many TE families were enriched for CLIP-Seq aligned reads from specific RBPs (Figure 1B). Enrichments and depletions of RBP-TE associations were broadly similar in mRNA exons, lncRNA exons, and introns (Figure S1 in Additional file 1). We benchmarked our approach by comparing it to previous CLIP studies discussing RBP-TE interactions. STAU1 binding to Alu sequence is a well-known phenomenon [20,34], and we confirmed that interaction here, observing a 3.2- to 4.1-fold enrichment of STAU1 alignments in Alu-derived sequence. Zarnack *et al.* [35] found that hnRNP C preferentially binds antisense Alu elements in RNA, where it prevents U2AF65 binding and aberrant splicing. Our analysis pipeline reproduced that interaction, detecting a 2.3-fold enrichment of hnRNP C alignments in antisense Alu elements throughout the transcriptome. TDP-43 binding to TE-derived sequence had also been previously noted [36]. We observed this phenomenon in the form of significant enrichments to antisense L1 (1.3- to 1.7-fold) and antisense Alu (1.5- to 2.3-fold) elements [37]. Together, these results demonstrate that our mapping and normalization pipeline confirms previously reported TE-driven RNA-protein interactions.

We further detected hundreds of significant novel enrichments between RBPs and TEs. The signal strength from even this low-resolution analysis — considering enrichment over the entirety of TE sequences - suggested that RBP-TE interactions are widespread. Thus, we proceeded to dissect these enrichments and their biological implications.

CLIP-Seq alignments cluster on specific transposable element motifs

Having observed overall enrichment of alignments in families of TE-derived sequence, we next asked whether specific subregions of the repeat consensus sequences drive these enrichments. Many RBPs have specific sequence and/or structure preferences [16,38]. These preferred motifs may appear in TE sequence and manifest as peaks in alignment coverage on the TE consensus. To control for uneven mappability and genomic coverage

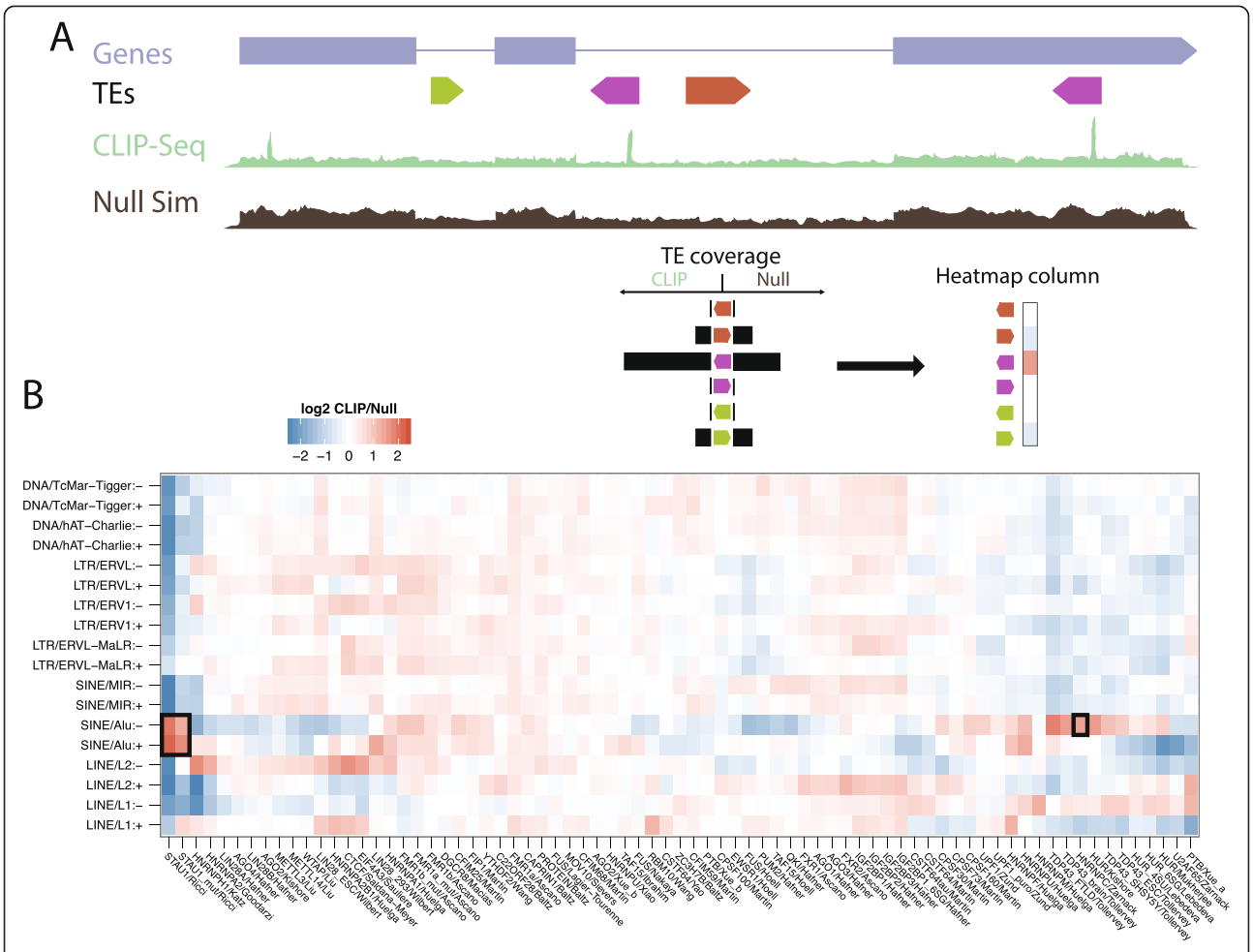

Figure 1 CLIP-Seq alignments are enriched and depleted in specific TEs. (A) Across the entire transcriptome (including introns), we counted overlapping aligned reads in the CLIP-Seq and null model simulation for every TE family in both orientations. The top track draws an example gene with blue boxes representing exons. The track below draws a set of example TEs, colored by family and with arrowhead describing orientation. Alignment coverage from CLIP-Seq and null simulation experiments are drawn below, with CLIP-Seq coverage spiking primarily on the purple TE in antisense orientation to the gene. **(B)** Heatmap in which we plotted the log2 ratio of the CLIP-Seq and null model counts, normalized by library size, for every pair of TE family in both orientations on the y-axis and RBP (with publication first author) on the x-axis. The heatmap was column mean normalized. Black rectangles highlight well-characterized interactions between STAU1 and Alu elements and hnRNP C and antisense Alu elements. TcMar-Tigger and hAT-Charlie are DNA transposons, which transpose via a cut-and-paste mechanism. ERVL, ERV1, and ERVL-MaLR are endogenous retrovirus families that have long terminal repeats (LTRs) on both ends. Both long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) are non-LTR retrotransposons, which mobilize via an RNA intermediate using a copy-and-paste mechanism.

of TEs, we compared CLIP-Seq alignment coverage to coverage from the uniform null model across TE consensus sequences (see Materials and methods).

Strikingly, both enriched RBP-TE pairs and many others without enrichment showed strong evidence of RBP binding to specific subregions within the TE consensus. For example, hnRNP H1 CLIP-Seq alignments clustered on two particular subregions of the antisense consensus of L2 elements (Figure 2A,B). Even sharper peaks appeared in both the sense and antisense orientations of the DNA transposon Tigger1 (Figure 2C,D), where the large size of the approximately 2,500-nucleotide element hid the interaction in Figure 1. In each case, the underlying sequence was AG-rich, in line with prior studies on hnRNP H1 binding preferences [39,40].

Having witnessed peaked coverage within TE subregions, we next compared and contrasted the preferences for these subregions among the many RBPs analyzed. In some cases, RBPs shared a preference - for example, a number of RBPs shared hnRNP C's affinity for two poly-U tracts of antisense Alu elements, which are well-studied in their sense poly-A form (Figure S3 in Additional file 1) [41]. However, even for antisense Alu, we observed a diversity of binding profiles among enriched RBPs, with STAU1, FMR1, and hnRNP U preferring different subregions (Figure S3 in Additional file 1). RBPs sharing a

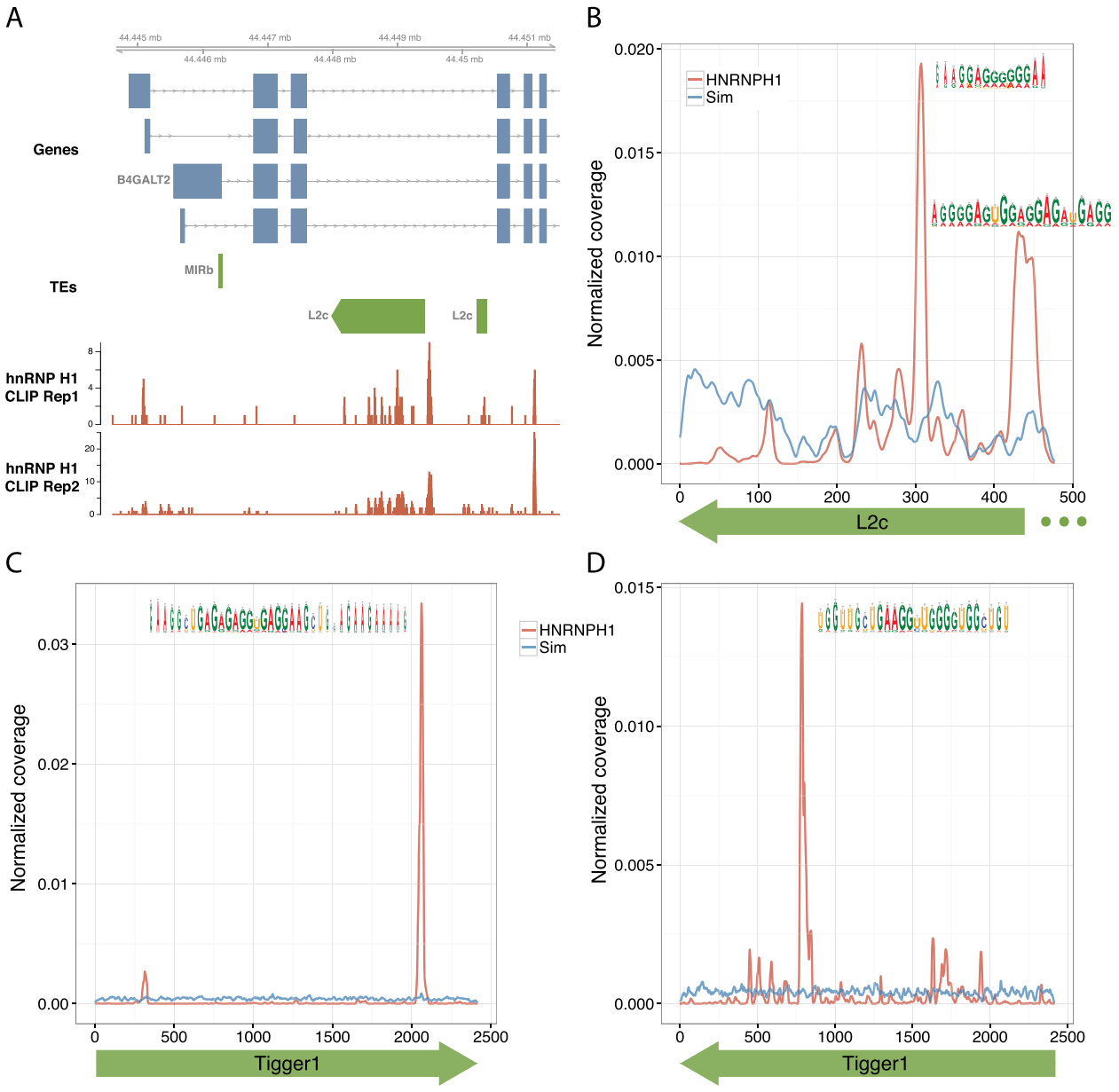

Figure 2 hnRNP H1 CLIP-Seq coverage clusters at specific TE motifs. (A) Two replicates of hnRNP H1 CLIP-Seq alignments show an alignment coverage peak at the same position of an antisense L2c element in *B4GALT2*. **(B)** hnRNP H1 alignments are enriched for antisense L2c overall, and coverage on the consensus 3′ end of the element clusters on two AG-rich positions. Coverage was normalized to sum to one across the TE span. **(C,D)** In contrast, Tigger1 is depleted for hnRNP H1 overall, but nearly all alignments cluster on particular positions in both the sense **(C)** and antisense orientation **(D)**.

similar binding profile in one TE did not generalize to common profiles in all TEs. For example, the splicing factor U2AF65 bound near hnRNP C in antisense Alu elements, where aberrant splicing is repressed [35], but the two RBPs bind apart in antisense L1 elements (Figure S4 in Additional file 1).

RBPs mapped using the same experimental protocol within a single study offer a valuable opportunity to ask whether the binding profiles are broadly similar between RBPs, which might occur if sequence composition biases of the protocol overwhelm the true signal. The differences between hnRNP C and U2AF65, both mapped by Zarnack *et al.* [35], in antisense L1 elements described above generalized to other major elements (Figure S4 in Additional file 1), lending credence to the authenticity of the detected interactions. As an additional example, CLIP-Seq HuR and Ago2 coverage profiles in Kishore *et al.* [42] diverge drastically (Figure S5 in Additional file 1).

Based on our observation that RBP alignment coverage clusters on specific subregions of TEs, we sought to systematically identify the underlying sequence motifs. To this end, we segmented coverage peaks where CLIP-Seq coverage was greater than three-fold more than the null model and refer to them hereafter as TE-specific motifs (TESMs). To standardize the TESMs for further analysis, we focused on nine nucleotide motifs centered at the maximum coverage nucleotide of each peak region. Although many known RBP motifs are shorter [16], we chose nine to include additional surrounding context and add specificity in studying the motif occurrences.

The TESM sequences broadly matched the known binding preferences of the RBPs. In accordance with prior work, TE enrichments of the ubiquitously expressed stabilizing RBP HuR were driven by binding to U-rich regions (Figure 3) [43,44]. HuR appears to bind these U-rich motifs in prevalent antisense Alu, antisense L1, and sense L2 elements. Its affinity for antisense Alu focuses on the two poly-U tracts (Figure 3B).

Uridine is known to crosslink more efficiently in PAR-CLIP experiments [42,45], which was used by all three groups to map HuR [42-44]. To address the possible concern that this may produce false positives, especially in the

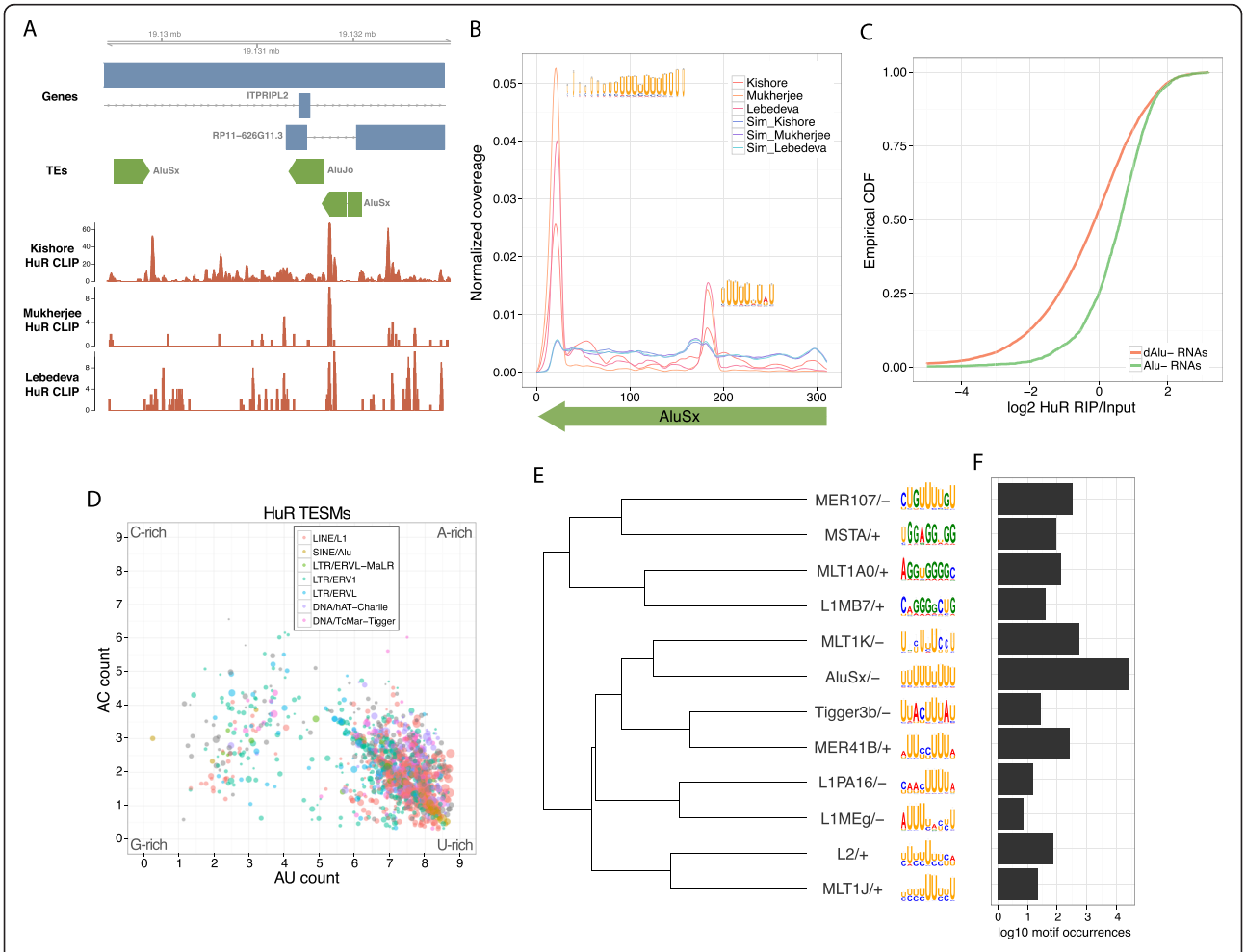

Figure 3 HuR binds U-rich motifs in antisense Alu elements. (A) CLIP-Seq alignments from three different studies demonstrated *HuR* binding within an antisense AluSx in the 3′ UTR of *ITPRIPL2*, **(B)** which generalized to genome-wide binding to two poly-U tracts in the antisense Alu consensus, shown here as alignment coverage normalized by dataset to sum to one across the element. **(C)** In a formaldehyde RNA immunoprecipitation (fRIP)-Seq of HuR, genes containing exonic antisense Alu elements (Alu- RNAs) were strongly enriched in the RIP over input compared with genes devoid of antisense Alu (dAlu- RNAs). Genes targeted via intronic antisense Alu elements also showed strong evidence of binding (Figure S6 in Additional file 1). **(D)** Plotting all HuR TE-specific motifs by their nucleotide composition revealed a diversity of motif compositions but a strong tendency towards Us. The x-axis specifies the expected number of As and Us in the motif model, and the y-axis specifies the expected number of As and Cs. Point size is proportional to the log2 of the number of CLIP-Seq alignments overlapped by the motif. **(E,F)** To better reveal the relationships between the strongest motifs, we collapsed highly redundant motifs into 12 representatives and hierarchically clustered them using information coverage Euclidean distance (see Materials and methods) [75].

prominent binding to U-rich antisense Alu-derived sequence, we performed an HuR formaldehyde crosslinked RNA immunoprecipitation and sequencing (fRIP-Seq), which does not suffer from a uridine bias (see Materials and methods). Consistent with the enriched CLIP-Seq alignment coverage, genes containing antisense Alu elements had far greater fold changes in our fRIP over input RNA than devoid genes (Figure 3C), which was true for both exonic and intronic occurrences of antisense Alu (Figure S6 in Additional file 1). In fact, antisense Alu elements were the strongest predictor of *HuR* binding among all TE families and orientations.

By leveraging the repetitive nature of TEs and combining CLIP-Seq signal over many copies of similar sequences, we extensively characterized the *in vivo* binding preferences of these RBPs. Plotting the nucleotide composition of HuR TESMs clearly shows their uridine richness, with a slight bend towards adenosines, matching prior expectations of HuR as a binder of AU-rich elements (Figure 3D) [46]. Surprisingly, we discovered multiple HuR TESMs with a different, somewhat G-rich composition (Figure 3E). These motifs only account for a small proportion of the alignments, but there is a strong enrichment for CLIP-Seq alignments at both their TE and nonrepetitive occurrences (Figure S7 in Additional file 1). Though the other HuR CLIP experiments offer a mixed view of the relevance of these sites (Figure S7 in Additional file 1), our HuR fRIP-Seq suggests their validity (Figure S8 in Additional file 1).

Altogether, we delineated 15,424 TESMs from the 75 datasets. The distribution of motif number varied widely between RBPs because the datasets differ in sequencing depth and enrichment of bound RNA over input (Figure S9 in Additional file 1). Clustering the datasets by their TESM coverage profiles revealed a diversity of RBP binding preferences, with a substantial group of AU-rich binders (Figure S9 in Additional file 1).

Overall, we found CLIP-Seq alignment coverage on TEs is highly nonuniform, clustering on thousands of TESMs, which generally matched the known RBP binding preferences but also uncover possible alternative binding modes. Enumerating and comparing the TESMs produces an extensive characterization of the *in vivo* binding preferences of the RBPs.

Transposable element-specific RBP motifs are bound and conserved in the nonrepetitive transcriptome

We next compared the binding preferences of the RBPs within and outside of TEs. If a TESM is prevalent outside of the TE and attracts CLIP-Seq alignment coverage, it would serve to validate the RBP affinity for that motif. It might also suggest that the TEs, which are typically newer entrants into the genome, appropriated existing RBP sequence preferences.

For this analysis, we considered only the top 300 TESMs per dataset, which were chosen by collapsing highly redundant motifs (for example, from homologous positions of Alu subfamilies) and ranking by coverage enrichment over the null model (see Materials and methods). Across all datasets, we elucidated 5,546 TESMs with evidence of RBP binding. We mapped these motifs in the nonrepetitive portion of the transcriptome to study their properties.

To assess CLIP-Seq coverage of the TESMs outside of repeats, we compared coverage directly at the motif with that in a surrounding 200 nucleotide region. We observed strong evidence that these motifs are bound outside of TE-derived sequences; 87% had increased coverage at the motif (Figure 4A), exemplified here by PTB CLIP-Seq coverage on nonrepetitive occurrences of a motif found in antisense L1MC4a (Figure 4B). Thus, RBP sequence preferences in TEs resemble those outside of repeats.

To further explore the potential for function in these nonrepetitive TESM occurrences, we considered their conservation using PhyloP (Figure 4C) [47]. Due to the severely different PhyloP backgrounds in the various annotation classes, we separated the analysis into introns, lncRNAs, and 3′ UTRs, ignoring coding sequence due to its much higher conservation signal. Seventy percent of motifs had a mean PhyloP above the intron baseline mean, exemplified again by an L1MC4a motif found for PTB (Figure 4D). The discovery of most motifs was driven by intronic sequencing coverage; accordingly, fewer motifs show constraint in the exonic sequence of 3′ UTRs (47%) and lncRNAs (45%). Nevertheless, for all annotation classes, TESM PhyloP distributions were significantly greater than expected by random sampling of 9-mers (Figure S10 in Additional file 1), despite the fact that 9-mers that appear in TE consensus sequences (approximately the set from which these TESMs were discovered) have a severely decreased PhyloP distribution overall (Figure S11 in Additional file 1). Motif conservation and CLIP-Seq coverage were not strongly related (Figure S12 in Additional file 1).

We next asked whether the PhyloP distributions differed by nucleotide position within a TESM. Indeed, plotting these distributions revealed nonuniform conservation of the motifs. Though in some cases constraint was present across the entire motif (Figure S13a in Additional file 1), in other cases only a subset of the nucleotides showed evidence of constraint in interesting and often symmetrical patterns (Figure S13b,c in Additional file 1). For a final set, we detected a high mutation rate across the motif (Figure S13d in Additional file 1), suggesting that many sequences throughout the nonrepetitive transcriptome have mutated towards these motifs.

Altogether, TESMs show evidence of binding to constrained sequences in the nonrepetitive transcriptome,

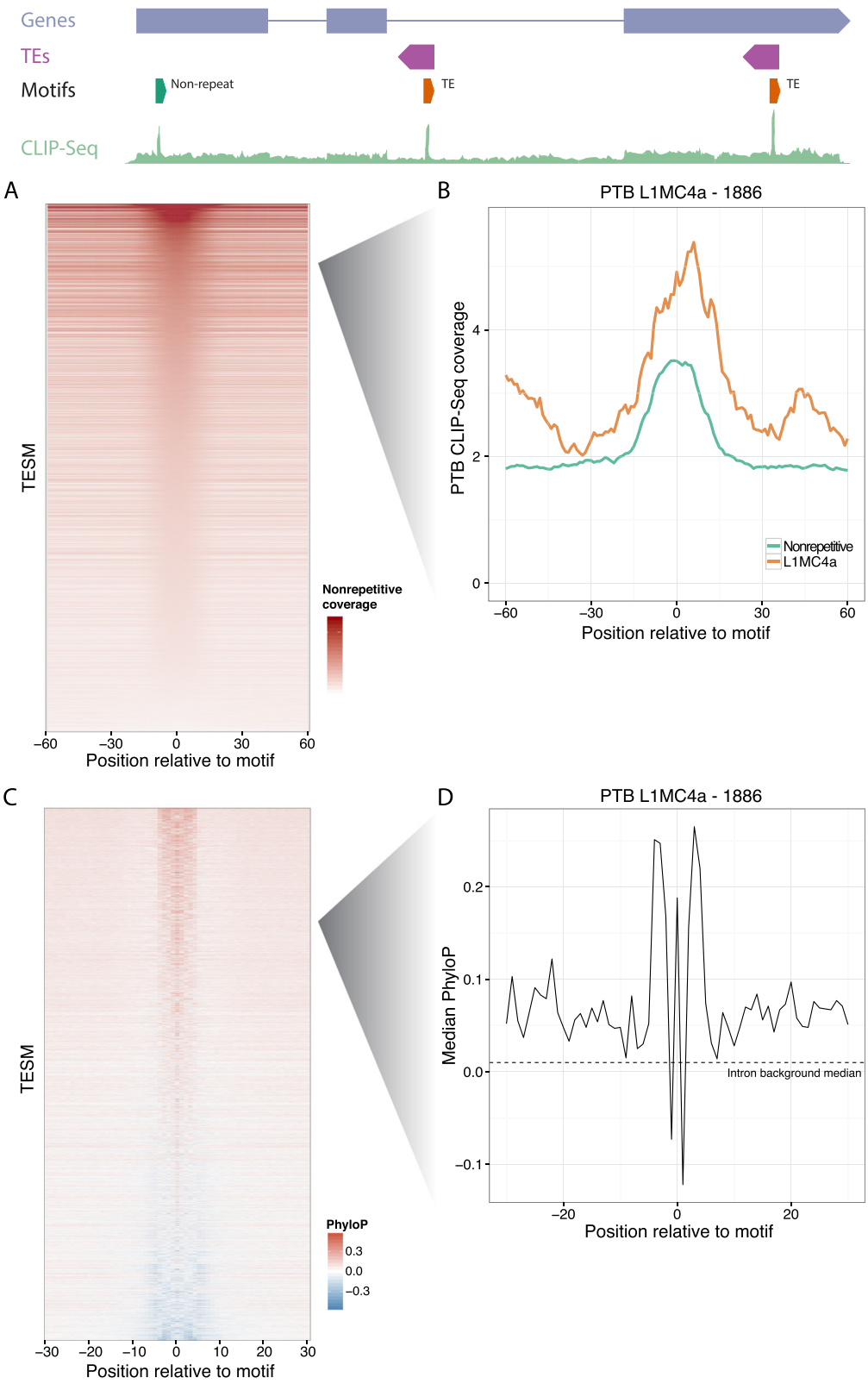

Figure 4 (See legend on next page.)

(See figure on previous page.)
Figure 4 TE-specific motifs are bound and conserved outside of TEs. We mapped 5,546 TE-specific motifs around the nonrepetitive transcriptome. **(A)** CLIP-Seq alignment coverage indicated that most motifs were bound; 87% of motif instances showed increased coverage at the motif versus the surrounding 200 nucleotides. **(B)** A motif discovered in PTB CLIP-Seq on antisense L1MB2 exemplified this, with coverage both in and out of L1MC4a. **(C)** PhyloP conservation scores indicated that many motifs were also conserved. The heatmap plots the median PhyloP score across all intronic motif instances. **(D)** At finer resolution, mutation rates differed by position within the motif, exemplified here by a PTB antisense L1MC4a motif.

suggesting that TEs have primarily appropriated the existing conserved binding preferences of these RBPs rather than binding through alternative mechanisms.

Transposable element binding affects RNA abundance and splicing

The evidence that TEs have intercepted existing RBP binding preferences, coupled with the widespread binding of RBPs to TEs, begs the question of whether TE binding sites are functionally similar to nonrepetitive sites. To investigate this question, we collected RBP knockdown RNA-Seq experiments matching 12 of the analyzed CLIP-Seq datasets. These experiments can detect changes in the abundances and splicing of genes determined to be targeted by the RBP in the CLIP-Seq and were used in the original studies to understand the RBPs' functions. We identified target genes with the CLIP-Seq using an enhanced version of established statistical procedures to call binding sites, considering both exons and introns, and computed differential expression between the paired RNA-Seq samples using Cuffdiff (see Materials and methods).

We first asked whether hnRNP C knockdown impacted TE binding sites similarly to nonrepetitive binding sites. Target genes tended to be upregulated after RNA interference-mediated depletion of hnRNP C (Figure 5A), suggesting a destabilizing effect on bound transcripts that increases with the number of binding sites (Figure 5B). To separately measure the effect of TE and nonrepetitive binding sites, we plotted the cumulative distributions of Cuffdiff's differential expression test statistic for genes bound only in nonrepetitive sequence or only in TEs (Figure 5C). As hypothesized, genes bound only in nonrepetitive sequence and only in TEs were similarly upregulated. This result held up separately for mRNAs and lncRNAs (Figure S14 in Additional file 1).

To better understand the effect of the various types of binding sites in genes targeted at multiple sites, we computed a linear regression to predict the differential expression test statistic using the logarithms of the number of binding sites in each class, further divided by exon and intron. The positive model coefficients augment the case that TE-derived hnRNP C binding sites repress target genes to a similar magnitude as nonrepetitive sites (Figure 5D).

To determine if TE-derived binding sites affect alternative splicing, we examined Cuffdiff *P*-values for the Materials and methodsstatistical significance of an isoform switch (see Materials and methods). Misregulation of splicing in antisense Alu elements was the primary phenotype described for hnRNP C in these data [35]; accordingly, we found that genes bound by hnRNP C had lesser splicing difference *P*-values, indicating more alternative splicing, after hnRNP C knockdown (Figure S15a in Additional file 1). Further, genes with more sites had more evidence for splicing differences (Figure S15b in Additional file 1). Binding sites in TEs affected splicing of their genes similarly to nonrepetitive sites (Figure S15c, d in Additional file 1). Alu and non-Alu TE sites were indistinguishable, suggesting the novel insight that hnRNP C's function as a splicing repressor generalizes beyond Alu elements.

We next examined the effect of HuR depletion on genes with TE-derived binding sites in RNA-Seq experiments from two studies [42,44]. Both found that HuR stabilized target genes, as genes targeted in the CLIP-Seq were significantly downregulated upon HuR knockdown. We reproduced these results but focused further analysis on the Kishore *et al.* dataset because bound genes were more affected by the knockdown (Figure 6A; Figure S16 in Additional file 1).

HuR target sites in TEs generally function similarly to nonrepetitive sites, but depend on the family bound. Genes targeted via non-Alu TEs were similarly downregulated after HuR knockdown (Figure 6C). Surprisingly, downregulation was absent for genes targeted only in Alu elements, which tended to change less than unbound genes in both directions (Figure 6C). These effects were apparent in both mRNAs and lncRNAs separately (Figure S14 in Additional file 1). As above, we computed a linear regression to quantify the effect of binding sites in these various annotation classes. The opposing model coefficients furthered the case that non-Alu TE-derived HuR binding sites stabilize the gene, but Alu binding sites do not (Figure 6D,E).

The remaining knockdown experiments further corroborated the significant effect of TE binding sites. Binding to TE-derived sites by both hnRNP H1 (Figure S17 in Additional file 1) and hnRNP U (Figure S18 in Additional file 1) stabilizes transcripts, while TE binding by PTB (Figure S19 in Additional file 1), WTAP (Figure S20 in Additional file 1), and METTL3 (Figure S21 in Additional file 1) represses transcripts. METTL3 serves as an additional example where activity depends on the TE family bound; L1 sites buck the general trend and appear to stabilize the transcript (Figure S21 in Additional file 1).

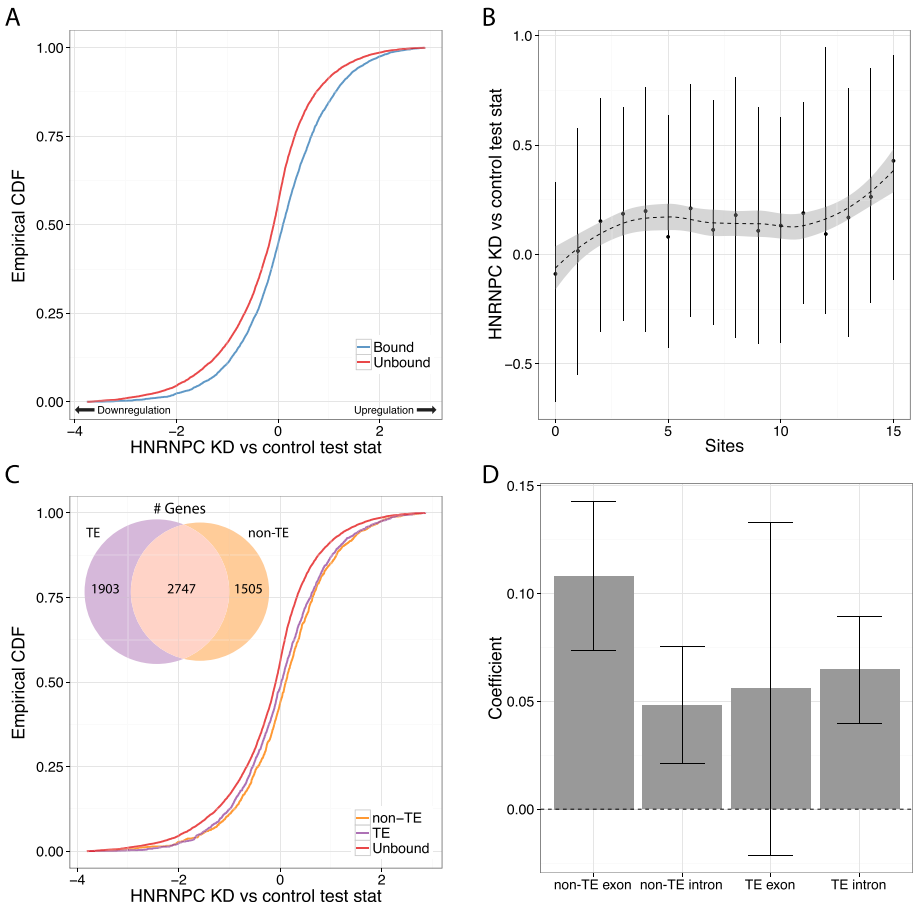

Figure 5 hnRNP C-TE binding sites repress genes similarly to nonrepetitive sites. (A) Genes targeted by hnRNP C were upregulated after hnRNP C knockdown (KD) compared to unbound genes, shown here as the cumulative distributions of the Cuffdiff differential expression test statistic (CDF). Positive values indicated greater abundance in the knockdown. **(B)** The test statistic distribution, plotted as the median and interquartile range, increased with the number of binding sites identified in the gene span. **(C)** Genes targeted only in TE sites were upregulated similarly to genes targeted only in nonrepetitive sites. The Venn diagram indicates the wide scale of TE binding sites, depicting the number of genes bound only in TEs, only in non-TEs, and in both. **(D)** A linear regression on the logarithm of the number of sites in each class showed that upregulation had a positive relationship with site number for all categories. Error bars represent a 95% confidence interval. The TE exon coefficient has large error bars because there were few examples to learn from.

Splicing analysis for most experiments was underpowered by having performed only single replicates, but METTL3 also showed a phenotype, with TE-derived sites increasing the likelihood of an isoform switch after knockdown (Figure S21 in Additional file 1).

In summary, RBP knockdown gene expression analyses establish that TE-derived and nonrepetitive RBP binding sites affect RNA state similarly, with interesting counter-examples, like Alu-HuR interactions, where the TE binding context may alter function. Extrapolating these results, the thousands of RBP-TE binding sites discovered in this analysis are candidates for function via RNA-protein interaction.

Discussion

Recent research has described a substantial role for TEs in the evolution of gene regulation at the transcriptional level; for example, TEs have dispersed transcriptional regulatory signals around the genome, and many sites have been co-opted for essential functions [7]. However, the influence of TEs on post-transcriptional regulation has previously been limited to a few promising examples. Here, we globally and systematically studied binding of RBPs to TE-derived sequence in human RNAs using a diverse set of CLIP-Seq experiments. We discovered widespread enrichment of RBPs on individual TE families, driven by sequence composition preferences of the RBPs for specific regions of those TEs. We described and studied thousands of these TESMs.

Many RBPs preferred U-rich TESMs, which was of notable concern because uridine is known to crosslink more efficiently in some CLIP experiments [42,45]. In most cases, the RBPs preference for U-rich sequence was previously known, such as HuR, hnRNP C, FUS, among others.

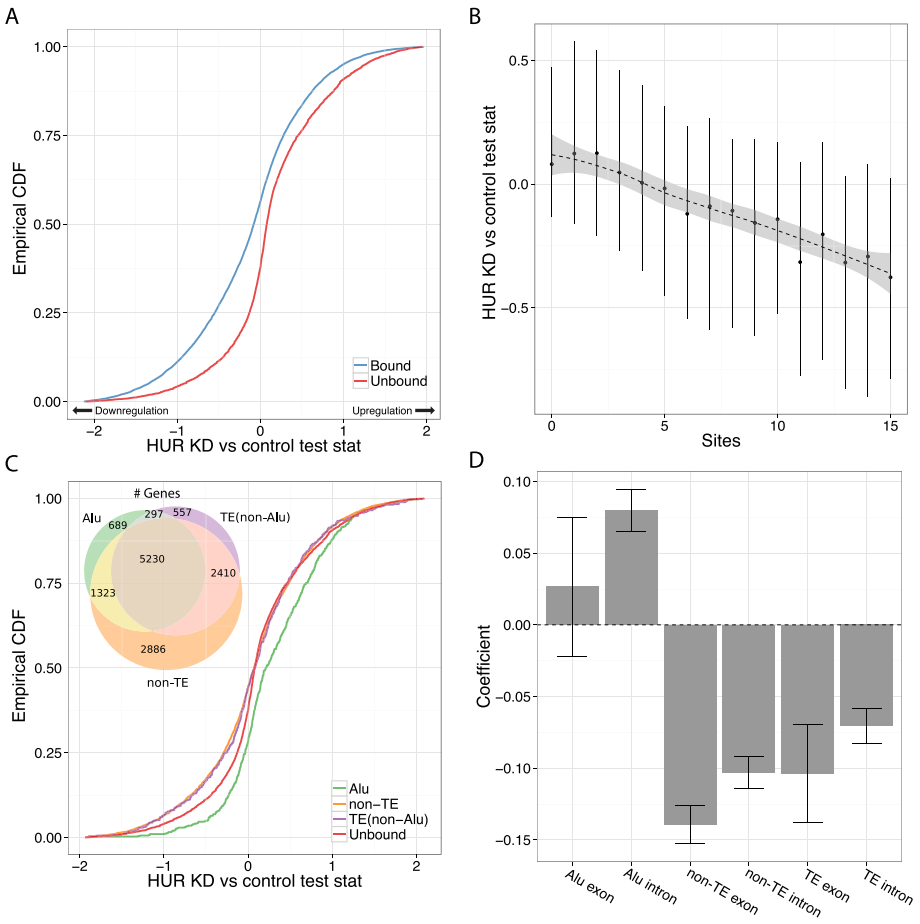

Figure 6 HuR-TE binding sites stabilize genes, unless in Alu elements. (A) As expected for the known transcript stabilizer HuR, targeted genes were downregulated after HuR knockdown (KD) compared with unbound genes, shown here as the cumulative distributions of the Cuffdiff differential expression test statistic (CDF). **(B)** HuR's stabilizing effect depended strongly on the number of binding sites in the gene span, shown through the decreasing medians and interquartile ranges of the test statistic distribution. **(C)** Unexpectedly, binding sites in Alu elements had the opposite effect; genes targeted only in Alu elements were upregulated after HuR knockdown. All other TE binding sites had the expected effect with similar magnitude as nonrepetitive sites. The Venn diagram indicates the wide scale of TE binding sites, depicting the number of genes bound only in Alu elements, only in non-Alu TEs, only in non-TEs, and in various mixtures. **(D)** A linear regression on the logarithm of the number of sites in each class verified that non-Alu TE and nonrepetitive sites predicted downregulation, but Alu sites did not. Error bars represent a 95% confidence interval. The Alu exon and TE exon coefficients have large error bars because there were few examples to learn from.

But to more definitively test the many U-rich interactions of HuR, we performed a fRIP-Seq, which does not suffer from more efficient crosslinking of uridines. The experiment validated HuR interaction with many U-rich sequences like antisense Alu elements as major contributors to HuR binding.

We accumulated compelling evidence that TESMs are relevant not only within TEs but also in the nonrepetitive transcriptome. For most TESMs, CLIP-Seq alignment coverage increased over motif occurrences outside of repeats. These motifs also showed greater than expected conservation in nonrepetitive 3′ UTR, intron, and lncRNA sequences. Together, these observations suggest that the RBPs' sequence preferences for these motifs were already established, and TE-derived instances of the motifs intercepted these preferences upon entry into the genome.

Despite this potentially 'uninvited' entry into post-transcriptional regulatory networks, we found that most RBP-TE binding sites affect RNA state with the same effect and to a similar magnitude as binding sites in nonrepetitive sequence. In addition to reproducing the impact of hnRNP C binding to antisense Alu elements on splicing [35], we discovered that many more hnRNP C binding sites on other TE families also affect splicing and transcript abundance upon hnRNP C depletion. Depletion of additional RBPs (hnRNP H1, hnRNP U, PTB, WTAP, and METTL3) introduced more evidence for functional TE binding as genes targeted via TE-derived sites had similarly altered abundance to genes targeted via nonrepetitive

sites. Thus, we have gathered here considerable evidence that often-ignored TE binding sites should be considered alongside nonrepetitive sites for their potential to modulate RNA abundance and splicing.

We also observed cases of TE-dependent regulation, such as HuR binding to antisense Alu elements. HuR binding sites stabilize the RNA such that HuR depletion causes downregulation of target genes. While we observed this influence for most TE-derived binding sites, the many Alu-derived sites were a conspicuous exception. Alu-bound genes were both less downregulated and less upregulated than unbound genes, suggesting that the abundance level of these genes is resistant to change in either direction. Most likely, we are missing the full picture at these Alu sites and combinatorial binding of multiple RBPs determine the effect on abundance. Follow-up experiments are needed to unravel this complex case.

The evidence here that RBP binding to TEs is widespread and can produce measurable and sometimes complex effects on gene abundance and splicing begs the question of how disrupting these interactions might affect the health of cells and organisms. RBPs have been implicated in numerous genetic diseases [48], including recent associations with neurodegenerative disorders [49] such as amyotrophic lateral sclerosis [50]. TEs, too, are a substantial focus of disease research, primarily with respect to deleterious novel transposition events [51,52], but also via misregulation of TE RNA [53]. Disruption of RBP-TE interaction homeostasis via RBP mutations or TE misregulation is a new and important avenue to consider in the etiology of human disease.

A considerable proportion of TEs in the exonic transcriptome lies in lncRNAs [24,25]. The myriad lncRNAs implicated for critical roles in development [54-56] and disease [57,58] emphasize the need for improved understanding of lncRNA function. A modular domain structure for lncRNAs has been hypothesized [59] but has, thus far, eluded a thorough characterization. Our observation that TE sequences contain functional RBP binding sites represents an important step towards characterizing TEs as one type of modular domain in lncRNAs where RBP-TE interactions may function.

Alterations to transcriptional regulatory networks are a major driver of evolutionary change [60,61]. TEs containing transcription factor binding sites play a substantial role in creating the variation that provides the raw material for selection to operate on [9,10,24,62]. Comparisons across species show that lineage-specific TEs can rapidly rewire regulatory networks [11-13,63]. Ultimately, these transcriptional regulatory site changes drive morphological change because they modulate protein abundance. Through a variety of mechanisms acting on RNA, such as splicing, localization, and degradation, post-transcriptional regulation by RBPs also modulates protein abundance. Our observation that RBP binding to TEs is widespread and can produce measurable effects on gene abundance and splicing suggests that TEs may also provide variation for post-transcriptional regulatory evolution. Mapping RBP-TE interactions in more species and placing them in the context of development and the adaptive responses of adult cells will elucidate the degree to which these interactions, too, are a major driver of evolutionary change.

Materials and methods

CLIP-Seq data and processing

We downloaded 75 CLIP-Seq datasets from 31 studies mapping 51 RBPs from the Gene Expression Omnibus (GEO), Sequence Read Archive, and EMBL ArrayExpress.

CLIP-Seq experiments typically build sequencing libraries using a small RNA protocol, which attaches a sequence adapter to the suffix of the short read. Unfortunately, these adapters are rarely reported in the manuscript or public database entry. Rather than take on the impractical, and perhaps impossible, task of acquiring accurate adapter information for 75 datasets, we implemented an adapter-ignorant strategy to align the prefixes of these reads up to the putative adapter sequence.

We aligned with TopHat 2.0.9 [64] to human genome assembly hg19, providing GENCODE v18 as reference annotation [65]. To align read prefixes, we carried out the following steps. First, we attempted to align the first 20 nucleotides of the read. For every read, if a unique alignment was found, we returned that alignment. If multiple alignments were found, we added it back to a set for re-alignment with an additional nucleotide added back to the end of the read prefix. We repeated this procedure, re-aligning ambiguous read prefixes up to the full read length. When a read that aligned in one iteration fails to align in the next, we presumably encountered the adapter and returned the read's alignment(s) from the previous iteration. Open source Python code implementing this strategy is available from [66].

The primary error that this pipeline can make is to distribute a highly repetitive read in a biased manner to genomic positions where the nucleotides downstream of the true read match the prefix of the adapter by chance. This error is not problematic for the analyses here where we merely need reads from TE-derived sequence to be aligned to some instance(s) of the TE family.

Due to low input material, CLIP-Seq experiments tend to have many PCR duplicated reads. We found allowing two alignments per chromosomal position struck a good balance between throwing away misleading PCR duplicates and keeping informative alignments from highly expressed genes and enriched clusters where redundancy is expected. Finally, we merged replicate experiments into one alignment file.

After alignment and filtering duplicates, numerous datasets contained so few reads that their reliability for the downstream analyses was questionable. Thus, we removed any dataset containing <200,000 aligned reads.

Multi-mapping reads

Careful interpretation of multi-mapping reads is critical to studying repetitive TE-derived sequences. We output 20 alignments per read with TopHat, which was found in a ChIP-Seq analysis of multi-mapping reads to be approximately the point where accuracy levels off [67]. That is, for reads with greater than 20 alignments, 20 are randomly chosen to be output. In all counting analyses described, we normalized alignments by the number of alignments of the read to account for the uncertainty of the true source alignment. For example, a read with 20 alignments will count 1/20 at each aligned position. As mentioned above, we note that most analyses performed only require that the read aligned to any instance of a TE family, which may or may not be the true source instance of the read.

Transposable element alignment enrichment/depletion

We computed the number of CLIP-Seq reads in each dataset that overlap each TE family from RepeatMasker [68] in both orientations using BEDTools [69] and compared the counts with a null model that accounts for the differing abundances of transcripts and assumes uniform coverage along those transcripts. CLIP-Seq experiments typically have substantial background read alignments, which allowed us to approximate these abundance estimates by running Cufflinks [70] on the CLIP-Seq alignments themselves, using the –multi-read-correct option to more accurately distribute multi-mapping reads. In order to account for introns, we augmented the GENCODE v18 annotation with unspliced pre-RNA isoforms [65]. Using these abundance estimates, we simulated new reads uniformly along the transcripts and mapped these reads back to the genome. We computed enrichment/depletion as the log ratio of the proportion of reads in the true and null model datasets overlapping each TE family in both orientations.

Transposable element consensus alignment coverage

To plot read coverage along the consensus sequence for each RBP-TE pair, we aligned all reads overlapping each TE family to its DFAM profile hidden Markov model [71] using HMMer [72]. To adjust for the influence of both mappability and the nonuniform presence of the TE consensus (for example, genomic instances of LINE1 often include only the 3′ end [73]), we normalized the actual read coverage by coverage from the null model simulated reads described above.

Transposable element-specific motifs

We characterized sequence motifs underlying the alignment coverage peaks by identifying regions of the TE consensus profile hidden Markov model for which CLIP-Seq alignment coverage was more than three-fold greater than the null simulation alignment coverage. For each of these coverage peaks, we represented the motif as a position weight matrix, with column frequencies defined by the multiple sequence alignment of aligned reads. We primarily studied nine nucleotide motifs, centered at the maximum alignment coverage nucleotide of each peak. We mapped motifs throughout the transcriptome using PoSSuM and *P*-value threshold 1e-5 [74].

Transposable element-specific motif clustering

At multiple stages of the motif analysis, we wanted to better understand the relationship between motifs and collapse highly similar motifs (for example, from homologous positions of Alu subfamilies) to avoid redundant computation. We chose information coverage Euclidean distance, an effective distance computed on position weight matrices, to quantify motif similarity [75]. Information coverage measures how informative a column in the position weight matrix is. For example, a uniform distribution of the four nucleotides would have zero information, and a column with only one valid nucleotide would have maximal information. Given two position weight matrices, we find their ungapped alignment with the minimum sum of Euclidean distances between column nucleotide distributions, weighted by the columns' information coverages. That is, we more strongly consider similar nucleotide distributions at informative over uninformative columns.

In our analysis of the full set of TESMs across datasets, we collapsed motifs within each dataset by computing pairwise distances as above and performing an average linkage hierarchical clustering, flattening the clusters at a threshold of 0.15. To form the final set, we chose the top 300 TESMs per dataset after ranking by coverage enrichment over the null model.

HuR formaldehyde RIP-Seq

Cell culture and cross-linking

K562 cells (ATCC catalog number CCL-243) were grown in RPMI 1640 (Invitrogen; Carlsbad, CA USA; catalog number 22400105) with 10% fetal bovine serum and 1% Antibiotic-Antimycotic 100X (Invitrogen; Carlsbad, CA USA; catalog number 15240062). We collected cells with a gentle 5 minute spin (500 g) and washed them with room temperature phosphate-buffered saline. We re-suspended at 5e6 cells per ml in room temperature RPMI media sans fetal bovine serum or Antibiotic-Antimycotic and added formaldehyde to a final concentration of 0.1%. We crosslinked at room temperature for

10 minutes and then halted it by quenching for 5 minutes at room temperature after adding glycine to a final concentration of 125 mM at a medium pace. We spun cells for 5 minutes at 500 g and washed twice in 4°C phosphate-buffered saline. We flash froze pellets of 10e6 cells and stored them at -80°C.

fRIP

We re-suspended frozen pellets in 1 ml of RIPA lysis buffer (50 mM Tris (pH 8), 150 mM KCl, 0.1% SDS, 1% Triton-X, 5 mM EDTA, 0.5% sodium deoxycholate, 0.5 mM dithiothreitol (add fresh) plus protease inhibitor cocktail (Thermo Scientific; Waltham, MA, USA; PI-87785) plus 100 U/ml RNaseOUT™ (Life Technologies; Woburn, MA, USA; catalog number 10777-019). We incubated cells at 4°C for 10 minutes before lysing on a Branson® digital sonifier (Emerson Industrial Automation; St. Louis, MO, USA) using 10% amplitude for 0.7 s on and 1.3 s off at 30 s intervals for a total of 90 s. We used chilled tube holders and swapped them out between shearing runs to reduce temperature elevation. After lysis, we spun the lysate at 4°C at maximum speed for 10 minutes. We collected supernatant and diluted by adding equal volume of fRIP binding/wash buffer (150 mM KCl, 25 mM Tris (pH 7.5), 5 mM EDTA, 0.5% NP-40, 0.5 mM DTT (add fresh), 1× protease inhibitor cocktail (add fresh), 100 U/mL RNaseOUT (add fresh)). At this point, we removed 50 μl of lysate for input sample and stored it at -20°C for later RNA purification and library construction. After dilution, we clarified the lysate by passage through a 0.45 μM syringe filter. We then 'pre-cleared' filtered lysate by incubating with Dynabeads® Protein G (Life Technologies; Woburn, MA, USA; catalog number 10004D) at a concentration of 25 μl of beads per 5 million cells for 30 minutes at 4°C with slow rotation. We flash froze pre-cleared lysate in 1 ml aliquots of approximately 5 million cells and stored it at -80°C. For fRIP, we thawed lysate on ice and added 6 μg of HuR antibody (Santa Cruz Biotechnology; Dallas, TX, USA; catalog number sc-5483). After addition of antibody, we rotated lysate at 4°C for 2 h before adding 50 μl of Dynabeads® Protein G. We rotated beads and lysate at 4°C for 1 h before washing twice with 1 ml of fRIP binding/washing buffer plus 1× protease inhibitor cocktail and 100 U/mL RNaseOUT. After the final wash, we removed the supernatant and froze and stored the beads at -20°C.

RNA purification and library construction

We resuspended the frozen beads in 56 μl of RNase-free water and added 33 μl of 3× reverse-crosslinking buffer (3× phosphate-buffered saline (without Mg or Ca), 6% N-lauroyl sarcosine, 30 mM EDTA, 15 mM dithiothreitol (add fresh)), 10 μl of Proteinase K (Life Technologies; Woburn, MA, USA; catalog number AM9516), and 1 μl of RNaseOUT to both the re-suspended beads and input sample. We performed protein degradation and reverse-crosslinking for 1 h at 42°C, then another 1 h at 55°C. We added beads and reaction buffer to 1 ml of TriZol (Life Technologies; Woburn, MA, USA; catalog number 15596-026). After agitation, we added 200 μl of chloroform followed by approximately 15 s of vigorous agitation and a 20 minute microcentrifuge spin at 4°C at maximum speed. We collected the aqueous layer, added it to 750 μl of ethanol plus 1 μl GlycoBlue™, and ran it over a Qiagen RNeasy® min-elute column (Qiagen; Valencia, CA, USA; catalog number 74204). We extracted RNA using the buffer RWT/3X isopropanol modification detailed in 'Appendix B: Optional On-Column DNAse Digestion…' of the Qiagen miRNeasy® Mini Handbook. We eluted RNA in 15 μl of RNase-free water. To remove ribosomal RNA, we fed ≥70 ng of input and fRIP RNA into the Ribo-Zero™ Magnetic Gold Kit (Epicentre; Madison, WI, USA; catalog number MRZG12324) followed by a cleanup using Agencourt RNAClean XP beads (Beckman Coulter; Brea, CA, USA; catalog number A63987) and elution with 19.5 μl of Elute, Prime, Fragment mix from the TruSeq RNA Sample Preparation Kit (Illumina; San Diego, CA, USA; catalog number RS-122-2001). We performed library construction per the vendor's instructions, starting with the 'Incubate RFP' step. We pooled the resulting cDNA libraries and subjected them to paired-end sequencing on an Illumina HiSeq 2500 at a depth of 31 base pairs per read.

Computational analysis

We aligned fRIP-Seq reads to hg19 and GENCODE v18 reference annotation using TopHat 2.0.9 [64] and ran Cuffdiff 2.1.1 [76] to estimate gene abundances and perform statistical comparisons between the fRIP versus input alignments. Raw reads and Cuffdiff output have been deposited in GEO as record GSE61238.

CLIP-Seq peak calling

To study the impact of RBP knockdown, we needed to define bound and unbound genes. We did so by annotating binding sites from the CLIP-Seq alignment coverage using a method based on prior CLIP-Seq scan statistic-based peak calling strategies [77], but with enhanced modeling of the multi-isoform structure of most human genes. A software implementation is available at [66].

Our peak calling strategy proceeded as follows. To avoid false positive peak calls from the very frequent PCR duplications without grossly betraying the scan statistic model assumptions (that is, that duplicate reads occur naturally), we first capped the number of reads aligning to the same chromosome and position at two. Next, to parameterize the scan statistic model, we estimated

isoform abundances using Cufflinks and the –multi-read-correct and –compatible-hits-norm options. As described above, we augmented the reference transcriptome with unspliced pre-RNA isoforms in order to capture intron binding sites. We computed enriched 30-nucleotide windows using a Poisson scan statistic approach [78], where each window was parameterized based on the abundances of the overlapping isoforms. We weighted multi-mapping read alignments by the inverse of their read's total number of alignments. Merging enriched windows produced the final peak calls. This procedure can be framed as an isoform-aware version of the Poisson-based methods commonly used for CLIP-Seq peak calling [77]. Finally, to focus this analysis on high confidence RBP targets, we removed peak calls overlapping particularly challenging genomic regions using a precomputed index, similar to the Genome Mappability Score [79], but computed with TopHat.

Knockdown differential expression

We determined differentially expressed genes after RBP knockdown by aligning RNA-Seq reads to hg19 and GENCODE v18 reference annotation using TopHat 2.0.9 [64] and running Cuffdiff 2.1.1 [76] to compare RNA-Seq alignments. Because most experiments were performed as only single replicates, and were thus underpowered to detect significant changes, we primarily studied the differential expression test statistic assigned to every gene, which quantifies the significance of the observed change in the number of fragments per kilobase per million reads (FPKM).

Cuffdiff analyzes differential splicing by computing the Jensen-Shannon metric between the two conditions' distributions of FPKM among the multiple isoforms from a transcription start site. Again, due to underpowered experiments, we primarily studied the *P*-values assigned to each gene transcription start site, which measure the significance of the observed splicing difference.

Visualization

Genome browser figures were constructed with GViz [80] or IGV [81].

Data availability

All CLIP-Seq datasets are publicly available with accession numbers specified in Additional file 2. HuR fRIP-Seq reads and Cuffdiff output are available as record GSE61238 in GEO.

Additional files

Additional file 1: Supplementary figures.

Additional file 2: CLIP-Seq dataset descriptions and public database accessions.

Abbreviations

CLIP: crosslinked immunoprecipitation; fRIP: formaldehyde RNA immunoprecipitation; GEO: Gene Expression Omnibus; lncRNA: long noncoding RNA; PCR: polymerase chain reaction; RBP: RNA binding protein; TE: transposable element; TESM: transposable element-specific motif; UTR: untranslated region.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DK and JR conceived the study and wrote the manuscript. DK carried out the analyses. DH and DT performed the HuR fRIP-Seq. All authors have read and approved the manuscript for publication.

Acknowledgements

The authors acknowledge Cole Trapnell for TopHat and Cufflinks assistance; Cole Trapnell, Chinmay Shukla, and Stephanie Sasse for feedback on the manuscript. DK was supported by NIH K25 award ES022984-02. JLR is Alvin and Esta Star Associate Professor. Work was supported by R01ES020260.

Author details

[1]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA. [2]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA 02142, USA. [3]Beth Israel Deaconess Medical Center, Boston, MA 02215, USA.

Received: 17 September 2014 Accepted: 7 November 2014
Published online: 03 December 2014

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, *et al*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860–921.
2. de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD: **Repetitive elements may comprise over two-thirds of the human genome.** *PLoS Genet* 2011, **7:**e1002384.
3. Doolittle WF, Sapienza C: **Selfish genes, the phenotype paradigm and genome evolution.** *Nature* 1980, **284:**601–603.
4. Orgel LE, Crick FH: **Selfish DNA: the ultimate parasite.** *Nature* 1980, **284:**604–607.
5. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH: **A unified classification system for eukaryotic transposable elements.** *Nat Rev Genet* 2007, **8:**973–982.
6. McClintock B: **Controlling elements and the gene.** *Cold Spring Harb Symp Quant Biol* 1956, **21:**197–216.
7. Feschotte C: **Transposable elements and the evolution of regulatory networks.** *Nat Rev Genet* 2008, **9:**397–405.
8. Cordaux R, Batzer MA: **The impact of retrotransposons on human genome evolution.** *Nat Rev Genet* 2009, **10:**691–703.
9. Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK, Haussler D: **Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53.** *Proc Natl Acad Sci U S A* 2007, **104:**18613–18618.
10. Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew J-L, Ruan Y, Wei C-L, Ng H-H, Liu ET: **Evolution of the mammalian transcription factor binding repertoire via transposable elements.** *Genome Res* 2008, **18:**1752–1762.
11. Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, Chan Y-S, Ng H-H, Bourque G: **Transposable elements have rewired the core regulatory network of human embryonic stem cells.** *Nat Genet* 2010, **42:**631–634.
12. Lynch VJ, Leclerc RD, May G, Wagner GP: **Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals.** *Nat Genet* 2011, **43:**1154–1159.
13. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT: **Waves of retrotransposon**

expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* 2012, **148:**335–348.
14. Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, Cai H, Besser D, Prigione A, Fuchs NV, Schumann GG, Chen W, Lorincz MC, Ivics Z, Hurst LD, Izsvák Z: **Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells.** *Nature*, in press.
15. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, *et al*: **Landscape of transcription in human cells.** *Nature* 2012, **489:**101–108.
16. Lunde BM, Moore C, Varani G: **RNA-binding proteins: modular design for efficient function.** *Nat Rev Mol Cell Biol* 2007, **8:**479–490.
17. Sorek R, Ast G, Graur D: **Alu-containing exons are alternatively spliced.** *Genome Res* 2002, **12:**1060–1067.
18. Lev-Maor G, Wang T, Sorek R, Zeng J, Shomron N, Lowe CB, Ast G, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK, Haussler D: **The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons.** *Science* 2003, **300:**1288–1291.
19. Shen S, Lin L, Cai JJ, Jiang P, Kenkel EJ, Stroik MR, Sato S, Davidson BL, Xing Y: **Widespread establishment and regulatory impact of Alu exons in human genes.** *Proc Natl Acad Sci U S A* 2011, **108:**2837–2842.
20. Gong C, Maquat LE: **lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements.** *Nature* 2011, **470:**284–288.
21. Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, Pesce E, Ferrer I, Collavin L, Santoro C, Forrest ARR, Carninci P, Biffo S, Stupka E, Gustincich S: **Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat.** *Nature* 2012, **491:**454–457.
22. Ulitsky I, Bartel DP: **lincRNAs: genomics, evolution, and mechanisms.** *Cell* 2013, **154:**26–46.
23. Rinn JL, Chang HY: **Genome regulation by long noncoding RNAs.** *Annu Rev Biochem* 2012, **81:**145–166.
24. Kelley DR, Rinn J: **Transposable elements reveal a stem cell-specific class of long noncoding RNAs.** *Genome Biol* 2012, **13:**R107.
25. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C: **Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs.** *PLoS Genet* 2013, **9:**e1003470.
26. König J, Zarnack K, Luscombe NM, Ule J: **Protein-RNA interactions: new genomic technologies and perspectives.** *Nat Rev Genet* 2011, **13:**77–83.
27. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB: **CLIP identifies Nova-regulated RNA networks in the brain.** *Science* 2003, **302:**1212–1215.
28. Ule J, Stefani G, Mele A, Ruggiu M, Wang X, Taneri B, Gaasterland T, Blencowe BJ, Darnell RB: **An RNA map predicting Nova-dependent splicing regulation.** *Nature* 2006, **444:**580–586.
29. Wang Z, Kayikci M, Briese M, Zarnack K, Luscombe NM, Rot G, Zupan B, Curk T, Ule J: **iCLIP predicts the dual splicing effects of TIA-RNA interactions.** *PLoS Biol* 2010, **8:**e1000530.
30. Darnell JC, Van Driesche SJ, Zhang C, Hung KYS, Mele A, Fraser CE, Stone EF, Chen C, Fak JJ, Chi SW, Licatalosi DD, Richter JD, Darnell RB: **FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism.** *Cell* 2011, **146:**247–261.
31. Ascano M, Mukherjee N, Bandaru P, Miller JB, Nusbaum JD, Corcoran DL, Langlois C, Munschauer M, Dewell S, Hafner M, Williams Z, Ohler U, Tuschl T: **FMRP targets distinct mRNA sequence elements to regulate protein expression.** *Nature* 2012, **492:**382–386.
32. Cho J, Chang H, Kwon SC, Kim B, Kim Y, Choe J, Ha M, Kim YK, Kim VN: **LIN28A is a suppressor of ER-associated translation in embryonic stem cells.** *Cell* 2012, **151:**765–777.
33. Lagier-Tourenne C, Polymenidou M, Hutt KR, Vu AQ, Baughn M, Huelga SC, Clutario KM, Ling S-C, Liang TY, Mazur C, Wancewicz E, Kim AS, Watt A, Freier S, Hicks GG, Donohue JP, Shiue L, Bennett CF, Ravits J, Cleveland DW, Yeo GW: **Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs.** *Nat Neurosci* 2012, **15:**1488–1497.
34. Ricci EP, Kucukural A, Cenik C, Mercier BC, Singh G, Heyer EE, Ashar-Patel A, Peng L, Moore MJ: **Staufen1 senses overall transcript secondary structure to regulate translation.** *Nat Struct Mol Biol* 2014, **21:**26–35.
35. Zarnack K, König J, Tajnik M, Martincorena I, Eustermann S, Stévant I, Reyes A, Anders S, Luscombe NM, Ule J: **Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements.** *Cell* 2013, **152:**453–466.
36. Li W, Jin Y, Prazak L, Hammell M, Dubnau J: **Transposable elements in TDP-43-mediated neurodegenerative disorders.** *PLoS One* 2012, **7:**e44099.
37. Tollervey JR, Curk T, Rogelj B, Briese M, Cereda M, Kayikci M, König J, Hortobágyi T, Nishimura AL, Zupunski V, Patani R, Chandran S, Rot G, Zupan B, Shaw CE, Ule J: **Characterizing the RNA targets and position-dependent splicing regulation by TDP-43.** *Nat Neurosci* 2011, **14:**452–458.
38. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, Na H, Irimia M, Matzat LH, Dale RK, Smith SA, Yarosh CA, Kelly SM, Nabet B, Mecenas D, Li W, Laishram RS, Qiao M, Lipshitz HD, Piano F, Corbett AH, Carstens RP, Frey BJ, Anderson RA, Lynch KW, Penalva LOF, *et al*: **A compendium of RNA-binding motifs for decoding gene regulation.** *Nature* 2013, **499:**172–177.
39. Caputi M, Zahler AM: **Determination of the RNA binding specificity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H′/F/2H9 family.** *J Biol Chem* 2001, **276:**43850–43859.
40. Han K, Yeo G, An P, Burge CB, Grabowski PJ: **A combinatorial code for splicing silencing: UAGG and GGGG motifs.** *PLoS Biol* 2005, **3:**e158.
41. Deininger P: **Alu elements: know the SINEs.** *Genome Biol* 2011, **12:**236.
42. Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, Zavolan M: **A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins.** *Nat Methods* 2011, **8:**559–564.
43. Mukherjee N, Corcoran DL, Nusbaum JD, Reid DW, Georgiev S, Hafner M, Ascano M, Tuschl T, Ohler U, Keene JD: **Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability.** *Mol Cell* 2011, **43:**327–339.
44. Lebedeva S, Jens M, Theil K, Schwanhäusser B, Selbach M, Landthaler M, Rajewsky N: **Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR.** *Mol Cell* 2011, **43:**340–352.
45. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jungkamp A-C, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T: **Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP.** *Cell* 2010, **141:**129–141.
46. Peng SS, Chen CY, Xu N, Shyu AB: **RNA stabilization by the AU-rich element binding protein, HuR, an ELAV protein.** *EMBO J* 1998, **17:**3461–3470.
47. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A: **Detection of nonneutral substitution rates on mammalian phylogenies.** *Genome Res* 2010, **20:**110–121.
48. Castello A, Fischer B, Hentze MW, Preiss T: **RNA-binding proteins in Mendelian disease.** *Trends Genet* 2013, **29:**318–327.
49. Ramaswami M, Taylor JP, Parker R: **Altered ribostasis: RNA-protein granules in degenerative disorders.** *Cell* 2013, **154:**727–736.
50. Xu Z, Poidevin M, Li X, Li Y, Shu L, Nelson DL, Li H, Hales CM, Gearing M, Wingo TS, Jin P: **Expanded GGGGCC repeat RNA associated with amyotrophic lateral sclerosis and frontotemporal dementia causes neurodegeneration.** *Proc Natl Acad Sci U S A* 2013, **110:**7778–7783.
51. Belancio VP, Hedges DJ, Deininger P: **Mammalian non-LTR retrotransposons: For better or worse, in sickness and in health.** *Genome Res* 2012, **18:**343–348.
52. Burns KH, Boeke JD: **Human transposon tectonics.** *Cell* 2012, **149:**740–752.
53. Kaneko H, Dridi S, Tarallo V, Gelfand BD, Fowler BJ, Cho WG, Kleinman ME, Ponicsan SL, Hauswirth WW, Chiodo VA, Karikó K, Yoo JW, Lee D-K, Hadziahmetovic M, Song Y, Misra S, Chaudhuri G, Buaas FW, Braun RE, Hinton DR, Zhang Q, Grossniklaus HE, Provis JM, Madigan MC, Milam AH, Justice NL, Albuquerque RJC, Blandford AD, Bogdanovich S, Hirano Y, *et al*: **DICER1 deficit induces Alu RNA toxicity in age-related macular degeneration.** *Nature* 2011, **471:**325–330.
54. Sauvageau M, Goff LA, Lodato S, Bonev B, Groff AF, Gerhardinger C, Sanchez-Gomez DB, Hacisuleyman E, Li E, Spence M, Liapis SC, Mallard W, Morse M, Swerdel MR, D'Ecclessis MF, Moore JC, Lai V, Gong G, Yancopoulos GD, Frendewey D, Kellis M, Hart RP, Valenzuela DM, Arlotta P, Rinn JL: **Multiple knockout mouse models reveal lincRNAs are required for life and brain development.** *Elife* 2013, **2:**e01749.
55. Fatica A, Bozzoni I: **Long non-coding RNAs: new players in cell differentiation and development.** *Nat Rev Genet* 2014, **15:**7–21.
56. Li L, Chang HY: **Physiological roles of long noncoding RNAs: insight from knockout mice.** *Trends Cell Biol* 2014, **24:**594–602.
57. Wapinski O, Chang HY: **Long noncoding RNAs and human disease.** *Trends Cell Biol* 2011, **21:**354–361.

58. Prensner JR, Iyer MK, Sahu A, Asangani IA, Cao Q, Patel L, Vergara IA, Davicioni E, Erho N, Ghadessi M, Jenkins RB, Triche TJ, Malik R, Bedenis R, McGregor N, Ma T, Chen W, Han S, Jing X, Cao X, Wang X, Chandler B, Yan W, Siddiqui J, Kunju LP, Dhanasekaran SM, Pienta KJ, Feng FY, Chinnaiyan AM: **The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex.** *Nat Genet* 2013, **45:**1392–1398.
59. Guttman M, Rinn JL: **Modular regulatory principles of large non-coding RNAs.** *Nature* 2012, **482:**339–346.
60. Carroll SB: **Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution.** *Cell* 2008, **134:**25–36.
61. Shibata Y, Sheffield NC, Fedrigo O, Babbitt CC, Wortham M, Tewari AK, London D, Song L, Lee B-K, Iyer VR, Parker SCJ, Margulies EH, Wray GA, Furey TS, Crawford GE: **Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection.** *PLoS Genet* 2012, **8:**e1002789.
62. Thornburg BG, Gotea V, Makałowski W: **Transposable elements as a significant source of transcription regulating signals.** *Gene* 2006, **365:**104–110.
63. Chuong EB, Rumi MAK, Soares MJ, Baker JC: **Endogenous retroviruses function as species-specific enhancer elements in the placenta.** *Nat Genet* 2013, **45:**325–329.
64. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25:**1105–1111.
65. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, *et al*: **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome Res* 2012, **22:**1760–1774.
66. **CLIP-Seq peak calling.** [https://github.com/davek44/CLIP-Seq]
67. Newkirk D, Biesinger J, Chon A, Yokomori K, Xie X: **AREM: aligning short reads from ChIP-sequencing by expectation maximization.** *J Comput Biol* 2011, **18:**1495–1505.
68. ***Repeatmasker.*** [http://www.repeatmasker.org]
69. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26:**841–842.
70. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol* 2010, **28:**511–515.
71. Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, Smit AFA, Finn RD: **Dfam: a database of repetitive DNA based on profile hidden Markov models.** *Nucleic Acids Res* 2013, **41:**D70–D82.
72. Wheeler TJ, Eddy SR: **nhmmer: DNA homology search with profile HMMs.** *Bioinformatics* 2013, **29:**2487–2489.
73. Smit AF, Tóth G, Riggs AD, Jurka J: **Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences.** *J Mol Biol* 1995, **246:**401–417.
74. Beckstette M, Homann R, Giegerich R, Kurtz S: **Fast index based algorithms and software for matching position specific scoring matrices.** *BMC Bioinformatics* 2006, **7:**389.
75. Stegmaier P, Kel A, Wingender E, Borlak J: **A discriminative approach for unsupervised clustering of DNA sequence motifs.** *PLoS Comput Biol* 2013, **9:**e1002958.
76. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: **Differential analysis of gene regulation at transcript resolution with RNA-seq.** *Nat Biotechnol* 2013, **31:**46–53.
77. Polymenidou M, Kunarso G, Lagier-Tourenne C, Chia N-Y, Hutt KR, Jeyakani J, Huelga SC, Hwang C, Moran J, Lu X, Liang TY, Chan Y-S, Ling S-C, Ng H-H, Sun E, Bourque G, Wancewicz E, Mazur C, Kordasiewicz H, Sedaghat Y, Donohue JP, Shiue L, Bennett CF, Yeo GW, Cleveland DW: **Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43.** *Nat Neurosci* 2011, **14:**459–468.
78. Glaz J, Pozdnyakov V, Wallenstein S: *Scan statistics: Methods and applications.* New York: Springer Press; 2001.
79. Lee H, Schatz MC: **Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score.** *Bioinformatics* 2012, **28:**2097–2105.
80. **Gviz: Plotting data and annotation information along genomic coordinates.** [http://www.bioconductor.org/packages/release/bioc/html/Gviz.html]
81. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer.** *Nat Biotechnol* 2011, **29:**24–26.

doi:10.1186/s13059-014-0537-5
Cite this article as: Kelley *et al.*: Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. *Genome Biology* 2014 **15**:537.