Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs

Joana Carlevaro-Fita,[1,2,3,5] Taisia Polidori,[1,2,3,5] Monalisa Das,[1,2,5] Carmen Navarro,[4] Tatjana I. Zoller,[1,2] and Rory Johnson[1,2]

[1]*Department for BioMedical Research (DBMR), University of Bern, 3008 Bern, Switzerland;* [2]*Department of Medical Oncology, Inselspital, University Hospital and University of Bern, 3010 Bern, Switzerland;* [3]*Graduate School of Cellular and Biomedical Sciences, University of Bern, 3012 Bern, Switzerland;* [4]*Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain*

The sequence domains underlying long noncoding RNA (lncRNA) activities, including their characteristic nuclear enrichment, remain largely unknown. It has been proposed that these domains can originate from neofunctionalized fragments of transposable elements (TEs), otherwise known as RIDLs (repeat insertion domains of lncRNA), although just a handful have been identified. It is challenging to distinguish functional RIDL instances against a numerous genomic background of neutrally evolving TEs. We here show evidence that a subset of TE types experience evolutionary selection in the context of lncRNA exons. Together these comprise an enrichment group of 5374 TE fragments in 3566 loci. Their host lncRNAs tend to be functionally validated and associated with disease. This RIDL group was used to explore the relationship between TEs and lncRNA subcellular localization. By using global localization data from 10 human cell lines, we uncover a dose-dependent relationship between nuclear/cytoplasmic distribution and evolutionarily conserved L2b, MIRb, and MIRc elements. This is observed in multiple cell types and is unaffected by confounders of transcript length or expression. Experimental validation with engineered transgenes shows that these TEs drive nuclear enrichment in a natural sequence context. Together these data reveal a role for TEs in regulating the subcellular localization of lncRNAs.

[Supplemental material is available for this article.]

The human genome contains many thousands of long noncoding RNAs (lncRNAs), of which at least a fraction is likely to have evolutionarily selected biological functions (Ulitsky and Bartel 2013). Our current working hypothesis is that, similar to proteins, lncRNA functions are encoded in primary sequence through "domains," or discrete elements that mediate specific aspects of lncRNA activity. Such activities range from molecular interactions to subcellular localization (Guttman and Rinn 2012; Mercer and Mattick 2013; Johnson and Guigó 2014). Experimental support for this domain model is beginning to emerge (Marín-Béjar et al. 2017). Mapping domains in a comprehensive manner is thus a key step toward the understanding and prediction of lncRNA functions.

One possible source of lncRNA domains are transposable elements (TEs) (Johnson and Guigó 2014). TEs are known to have been major contributors to genomic evolution through the insertion and neofunctionalization of sequence fragments, a process known as exaptation (Feschotte 2008; Bourque 2009). This process has contributed to the evolution of diverse features in genomic DNA, including transcriptional regulatory motifs (Johnson et al. 2006; Bourque et al. 2008), microRNAs (Roberts et al. 2014), gene promoters (Faulkner et al. 2009; Huda et al. 2011), and splice sites (Lev-Maor et al. 2003; Sela et al. 2007).

We recently proposed that exaptation also takes place in the context of lncRNAs, with TEs contributing pre-formed functional domains. We termed these repeat insertion domains of lncRNAs (RIDLs) (Johnson and Guigó 2014). As RNAs, TEs are known to interact with a rich variety of proteins, meaning that in the context of lncRNA they could plausibly act as protein-docking sites (Blackwell et al. 2012). Diverse evidence also points to repetitive sequences forming intermolecular Watson–Crick RNA:RNA and RNA:DNA hybrids (Gong and Maquat 2011; Holdt et al. 2013; Johnson and Guigó 2014). However, it is likely that bona fide RIDLs represent a small minority of the many exonic TEs, with the remainder being phenotypically neutral "passengers."

A small but growing number of RIDLs have been described (for review, see Johnson and Guigó 2014). These are found in lncRNAs with clearly demonstrated functions, including the X Chromosome silencing transcript *XIST* (Elisaphenko et al. 2008), the oncogene *ANRIL* (Holdt et al. 2013), and the regulatory antisense *Uchl1os* (also known as *Uchl1as*) (Carrieri et al. 2012). In each case, domains of repetitive origin are necessary for a defined function: The structured A-repeat of *XIST*, of retroviral origin, recruits the PRC2 silencing complex (Elisaphenko et al. 2008); Watson–Crick hybridization between RNA and DNA *Alu* elements recruits *ANRIL* to target genes (Holdt et al. 2013); and a SINEB2 repeat in *Uchl1os* increases translational rate of its sense mRNA (Carrieri et al. 2012). In parallel, transcriptome-wide maps of lncRNA-linked TEs have shown how TEs have contributed extensively to lncRNA gene evolution (Kelley and Rinn 2012; Kapusta et al. 2013; Hezroni et al. 2015; Schmitt et al. 2016). However, there has been no attempt to enrich these maps for RIDLs with evidence of selected functions in the context of mature lncRNA molecules.

[5]These authors contributed equally to this work.
Corresponding author: rory.johnson@dbmr.unibe.ch
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.229922.117. Freely available online through the *Genome Research* Open Access option.

29:208–222 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/19; www.genome.org

Subcellular localization and the domains controlling it are crucial determinants of lncRNA functions (for review, see Chen 2016). For example, transcriptional regulatory lncRNAs must be located in the nucleus and chromatin, whereas those regulating microRNAs or translation should be present in the cytoplasm (Zhang et al. 2014b). Although higher nuclear/cytoplasmic ratios are a hallmark of lncRNAs, a large population of cytoplasmic transcripts also exists (Derrien et al. 2012; Cabili et al. 2015; Carlevaro-Fita et al. 2016; Mas-Ponte et al. 2017; Mukherjee et al. 2017; Benoit Bouvrette et al. 2018). If lessons learned from mRNA are also valid for lncRNAs, then short sequence motifs recognized by RNA binding proteins (RBPs) will be an important localization-regulatory mechanism (Martin and Ephrussi 2009). This was recently demonstrated for the *BORG* lncRNA, in which a pentameric motif was shown to mediate nuclear retention (Zhang et al. 2014a). Similarly, multiple copies of the 156-bp RRD repeat motif mediate nuclear enrichment of the *FIRRE* lncRNA, through binding to HNRNPU (Hacisuleyman et al. 2014, 2016). Another study implicated an inverted pair of *Alu* elements in nuclear retention of *lincRNA-P21* (Chillón and Pyle 2016). This raises the possibility that by "copying and pasting" generic RNA motifs, RIDLs could fine-tune lncRNA localization at a global scale.

The aim of the present study is to create a human transcriptome-wide catalog of putative RIDLs. Supporting its relevance, lncRNAs carrying these RIDLs are enriched for functional genes. Finally, we provide in silico and experimental evidence that certain RIDL types, derived from ancient TEs, promote the nuclear enrichment of their host transcripts.

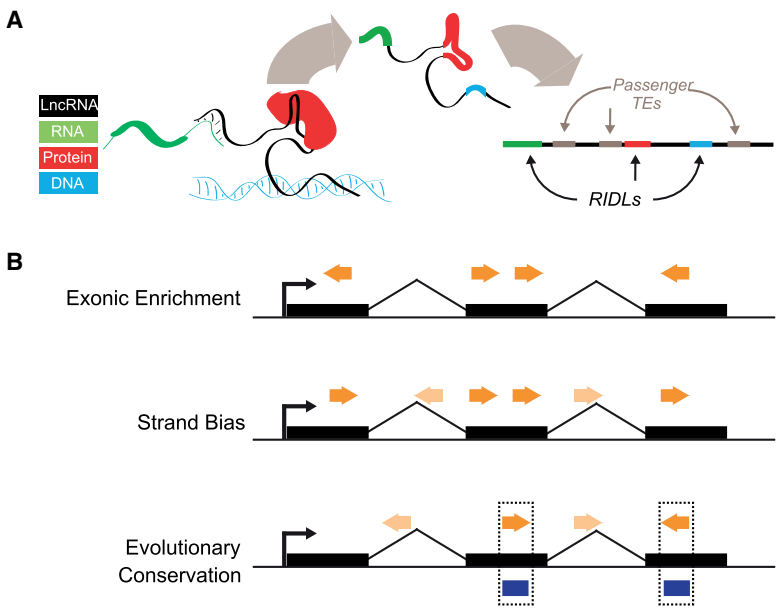

Figure 1. Repeat insertion domains of lncRNAs (RIDLs). (*A*) In the RIDL model, exonically inserted fragments of transposable elements (TEs) contain pre-formed protein-binding (red), RNA-binding (green), or DNA-binding (blue) activities that contribute to the functionality of the host lncRNA (black). RIDLs are likely to be a small minority of exonic TEs, coexisting with large numbers of nonfunctional "passengers" (gray). (*B*) RIDLs (dark orange arrows) will be distinguished from passenger TEs by signals of selection, including (1) simple enrichment in exons, (2) a preference for residing on a particular strand relative to the host transcript, and (3) elevated evolutionary conservation in exons compared with introns. Selection might be identified by comparing exonic TEs to a neutral population, for example, those residing in lncRNA introns (light-colored arrows).

Results

The objective of this study is to create a map of RIDLs and link them to lncRNA functions. We hypothesize that RIDLs could confer such functions through interactions with DNA, RNA, or protein molecules (Fig. 1A; Johnson and Guigó 2014).

Any attempt to map RIDLs must deal with two challenges. First, that they will likely represent a small minority among many phenotypically neutral "passenger" TEs in lncRNA exons (Fig. 1B). Second, many TE instances may be under evolutionary selection but for functions executed at the DNA level (e.g., transcription factor binding sites, enhancer elements) rather than the RNA level (Bassett et al. 2014)

Therefore, it is necessary to identify RIDLs by some signature of selection that is specific for a mature RNA product using an appropriate background model. In this study we use three types of such signatures: exonic enrichment, strand bias (with respect to host gene), and exon-specific evolutionary conservation (Fig. 1B). To estimate background, we use intronic TEs because they should mirror any biases of TE distribution across the genome but are not incorporated into mature lncRNA transcripts.

Resulting RIDL predictions should be considered as "enrichment groups" because of high rates of false-positive predictions, and all downstream analyses should be interpreted accordingly.

A map of exonic TEs in GENCODE version 21 lncRNAs

Our first aim was to create a comprehensive map of TEs within the exons of GENCODE v21 human lncRNAs (Fig. 2A). Altogether 5,520,018 distinct TE insertions were intersected with 48,684 exons from 26,414 transcripts of 15,877 GENCODE version 21 (v21) lncRNA genes, resulting in 46,474 exonic TE insertions in lncRNAs (Fig. 1B). We found 13,121 lncRNA genes (82.6%) carry at least one exonic TE fragment in one or more of their mature transcripts.

We also created a reference data set with 31,004 GENCODE lncRNA introns, resulting in 562,640 intron-overlapping TE fragments (Fig. 2A). By comparing intronic and exonic TE data, we see that lncRNA exons are depleted for TE insertions: 29.2% of exonic nucleotides are of TE origin compared with 43.4% of intronic nucleotides (Fig. 2B), similar to previous studies (Kapusta et al. 2013). This may reflect generalized selection against disruption of functional lncRNA transcripts by TEs. The exonic depletion of TEs in lncRNAs is less pronounced than for protein-coding loci, whereas the intronic TE density of both is similar to the whole-genome average.

Contribution of TEs to lncRNA gene structures

TEs have contributed widely to both coding and noncoding gene structures by the insertion of elements such as promoters, splice sites, and termination sites (Sela et al. 2007). We next

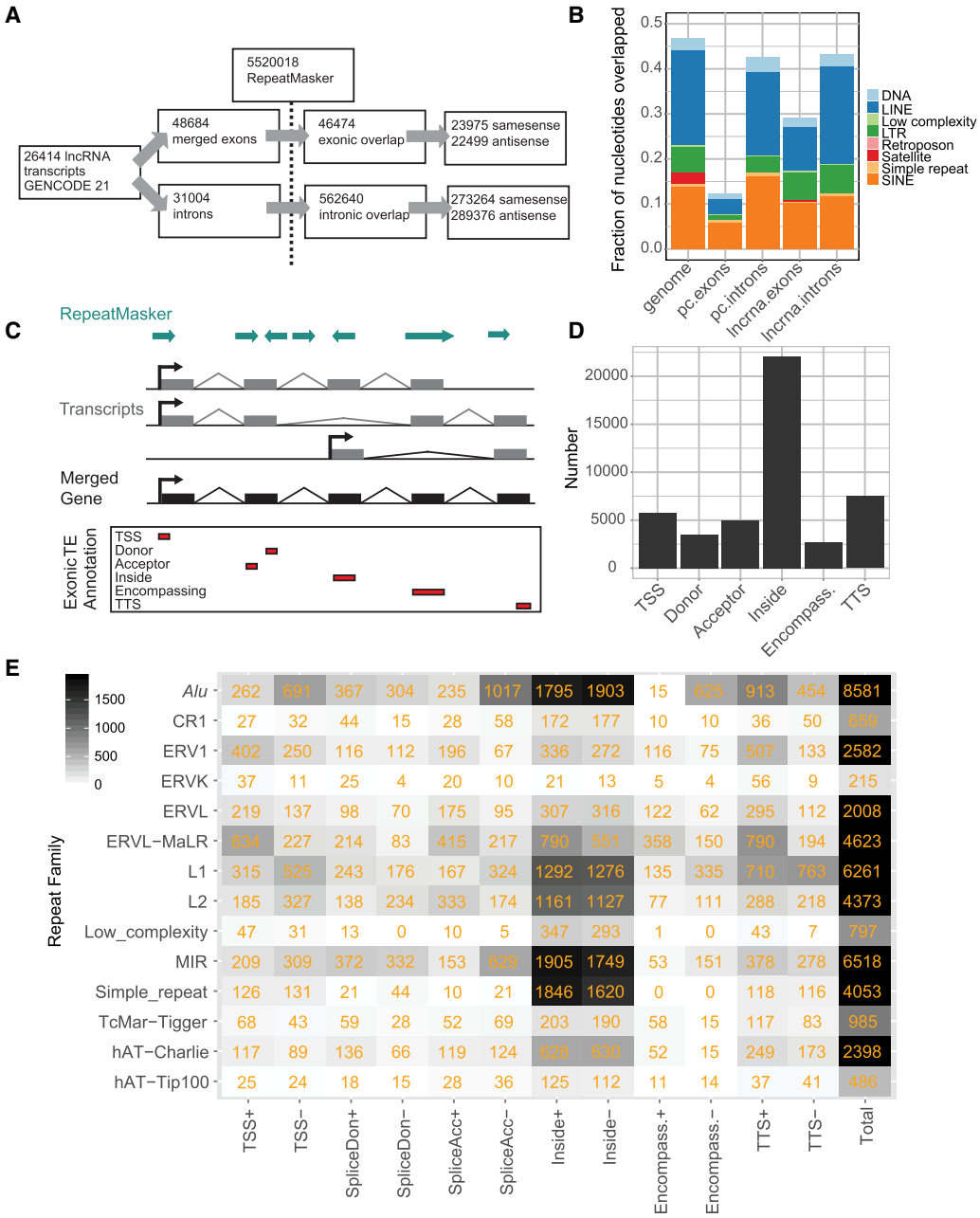

Figure 2. An exonic TE annotation with the GENCODE v21 lncRNA catalog. (*A*) Statistics for the exonic TE annotation process using GENCODE v21 lncRNAs. (*B*) The fraction of nucleotides overlapped by TEs for lncRNA exons and introns, protein-coding introns and exons (pc), and the whole genome. (*C*) Overview of the annotation process. The exons of all transcripts within a lncRNA gene annotation are merged. Merged exons are intersected with the RepeatMasker TE annotation. Intersecting TEs are classified into one of six categories (*bottom*) according to the gene structure with which they intersect and to the relative strand of the TE with respect to the gene: (TSS) overlapping the transcription start site; (donor) splice donor site; (acceptor) splice acceptor site; (inside) the TE boundaries both lie within the exon; (encompassing) the exon boundaries both lie within the TE; and (TTS) the transcription termination site. (*D*) Summary of classification breakdown for exonic TE annotation. (*E*) Classification of TE classes in exonic TE annotation. Numbers indicate instances of each type. (+ or −) Relative strand of the TE with respect to lncRNA transcript.

classified inserted TEs by their contribution to lncRNA gene structure (Fig. 2C,D). It should be borne in mind that this analysis is dependent on the accuracy of underlying GENCODE annotations, which are often incomplete at 5′ and 3′ ends (Lagarde et al. 2017). Altogether 4993 (18.9%) transcripts' promoters lie within a TE, most often those of the *Alu*, L1, and ERVL-MaLR classes (Fig. 2E); 7497 (28.4%) lncRNA transcripts are terminated by a TE, most commonly by the L1, *Alu*, ERVL-MaLR classes; 8494 lncRNA splice sites (32.2%) are of TE origin, and 2681 entire exons are fully contributed by TEs (10.1%) (Fig. 2E). These observations support known contributions of TEs to gene structural features (Sela et al. 2007). Nevertheless, the most frequent case is represented by 22,031 TEs that lie completely within an exon and do not overlap any splice junction (inside).

Evidence for selection on certain exonic TE types

This exonic TE map represents the starting point for the identification of RIDLs, defined as the subset of TEs with evidence for functionality in the context of mature lncRNAs. In this and subsequent analyses, TEs are grouped by type as defined by RepeatMasker (Smit et al. 2013–2015). We use three distinct sources of evidence for selection on TEs: exonic enrichment, strand bias, and evolutionary conservation (Fig. 1B).

We first asked whether particular TE types are enriched in lncRNA exons compared with intronic sequence (Kelley and Rinn 2012). Thus, we calculated the ratio of exonic/intronic sequence coverage by TEs (Fig. 3A). We found enrichment greater than twofold for numerous repeat types, including endogenous retrovirus classes (HERVE-int, HERVK9-int, HERV3-int, LTR12D) in addition to others such as ALR/Alpha, BSR/Beta, and REP522. A number of simple repeats are also enriched in lncRNA, including GC-rich repeats. A weaker but more generalized trend of enrichment is also observed for various MLT repeat classes. These findings are consistent with previous analyses by Kelley and Rinn (2012) using the whole genome, rather than introns, as background. Similarly, both studies agree in finding no difference in *Alu* density between lncRNA exons and intergenic/intronic DNA.

Despite their overall abundance throughout the genome, presently active LINE-1 elements are relatively depleted in lncRNA exons (Fig. 3A). It is possible that this reflects selection against disruption to normal gene expression, in which numerous weak polyadenylation signals lead to premature transcription termination when the LINE-1 element lies on the same strand as the overlapping gene (Perepelitsa-Belancio and Deininger 2003). Other explanations may be low transcriptional processivity exhibited by the LINE-1 ORF2 in the sense strand (Perepelitsa-Belancio and Deininger 2003) or else epigenetic silencing effects (Hollister and Gaut 2009).

As a second source of evidence for selection, we searched for TE types displaying a strand preference relative to host lncRNA (Johnson and Guigó 2014). We were conscious of a major source of bias: As shown above, many TSS and splice sites of lncRNA are contributed by TEs, and such cases would lead to artifactual strand bias. To avoid this, we ignored any TEs that overlap an exon–intron boundary. We calculated the relative strand overlap of all remaining TEs in lncRNA exons. Statistical significance was assessed by randomization, with significance defined at $P < 0.001$, corresponding to a false-discovery rate (FDR) <5% (similar cutoffs apply to subsequent analyses; more details may be found in Methods) (Fig. 3B). In lncRNA exons, a number of TE types are enriched in either sense or antisense, dominated by LINE-1 family members, possibly for the reasons mentioned above. Other significantly enriched TE types include LTR78, MLT1B, and MIRc (Fig. 3B).

To test the specificity of this exonic strand bias, we performed equivalent analysis using introns. Although intronic strand bias is weaker, we did detect a modest yet statistically significant depletion of same-strand TE insertions (Supplemental Fig. S1). This is especially true for LINE-1 elements, possibly for aforementioned reasons. In contrast to exons, almost no TE types were significantly enriched on the same-strand in introns.

To test for TE type–specific conservation, we turned to two sets of predictions of evolutionarily conserved elements: (1) the widely used phastCons conserved elements, based on phylogenetic hidden Markov model (Siepel et al. 2005) calculated separately on primate, placental mammal, and vertebrate alignments; (2) the more recent "evolutionarily conserved structures" (ECS) set (Smith et al. 2013). Importantly, the phastCons regions are defined based on sequence conservation alone, whereas the ECS are defined by phylogenetic analysis of RNA structure evolution.

To look for evidence of evolutionary conservation on exonic TEs, we calculated the fraction of nucleotides overlapped by evolutionarily conserved genomic elements and compared to the equivalent fraction for intronic TEs of the same type. To assess statistical significance, we again used positional randomization (see Fig. 3C, inset). This pipeline was applied independently to the phastCons (placental mammal shown in Fig. 3C; primate and vertebrate in Supplemental Fig. S1B,C) and ECS (Supplemental Fig. S1D) data. The majority of TE types do not exhibit signatures of conservation (gray points). However, for each conservation type, the method detects significant conservation for a minority of TE types (Fig. 3C). This enrichment disappeared when phastCons elements were positionally randomized (Supplemental Fig. S2A). It is unlikely that overlap with protein-coding loci biases the results, because equivalent analyses using intergenic lncRNAs yielded similar candidate RIDLs (Supplemental Fig. S2B). A similar analysis was performed using protein-coding exons, and although a number of significantly conserved TEs were identified, they display limited overlap with those from lncRNAs (Supplemental Fig. S2C). We also found a small number of TEs depleted for signatures of conservation in lncRNA exons, namely, the young *Alu*Sz, *Alu*Sx, and *Alu*Jb (phastCons) and L1M4c and *Alu*Sx1 (ECS) (colored orange in Fig. 3C; Supplemental Fig. S1). The cause of this depletion is unclear, although one explanation is enrichment of conservation in intronic TEs because of RNA-independent regulatory roles as observed previously (Su et al. 2014).

All the selection evidence is summarized in Figure 3D. As might be expected, one observes a high degree of concordance in candidate TEs identified by the three phastCons methods, in addition to a smaller number with both phastCons and ECS evidence, including L2b and MIRb. This is not surprising given the distinct methodologies used to infer conservation. Less concordance is observed between conservation, enrichment, and strand bias candidates, although some TEs are identified by multiple methods, such as MIRc (strand bias and ECS).

An annotation of RIDLs

We next combined all TE classes with evidence of functionality into a draft annotation of RIDLs. This annotation combined altogether 99 TE types with at least one type of selection evidence. For each TE/evidence pair, only those TE instances satisfying that evidence were included. In other words, if MIRb elements were found to be associated with vertebrate phastCons elements, then *only* those instances of exonic MIRb elements overlapping such an element would be included in the RIDL annotation, and all other exonic MIRbs would be excluded. This operation was performed for all three phastCons element types, ECS elements, and strand-bias. An example is *CCAT1* lncRNA oncogene: It carries three exonic MIR elements, of which one is defined as a RIDL based on its overlapping a phastCons element (Fig. 4A).

After removing redundancy, the final RIDL annotation consists of 5374 elements, located within 3566 distinct lncRNA genes (Fig. 3D). These represent 12% (5374/46,474) of all exonic TE fragments. The most predominant TE families are MIR and L2 repeats, representing 2329 and 1143 RIDLs (Fig. 4B). The majority of both are defined based on evolutionary evidence (Fig. 4B; Supplemental Fig. S3). In contrast, RIDLs composed by ERV1, low complexity, satellites, and simple repeat families are more frequently identified

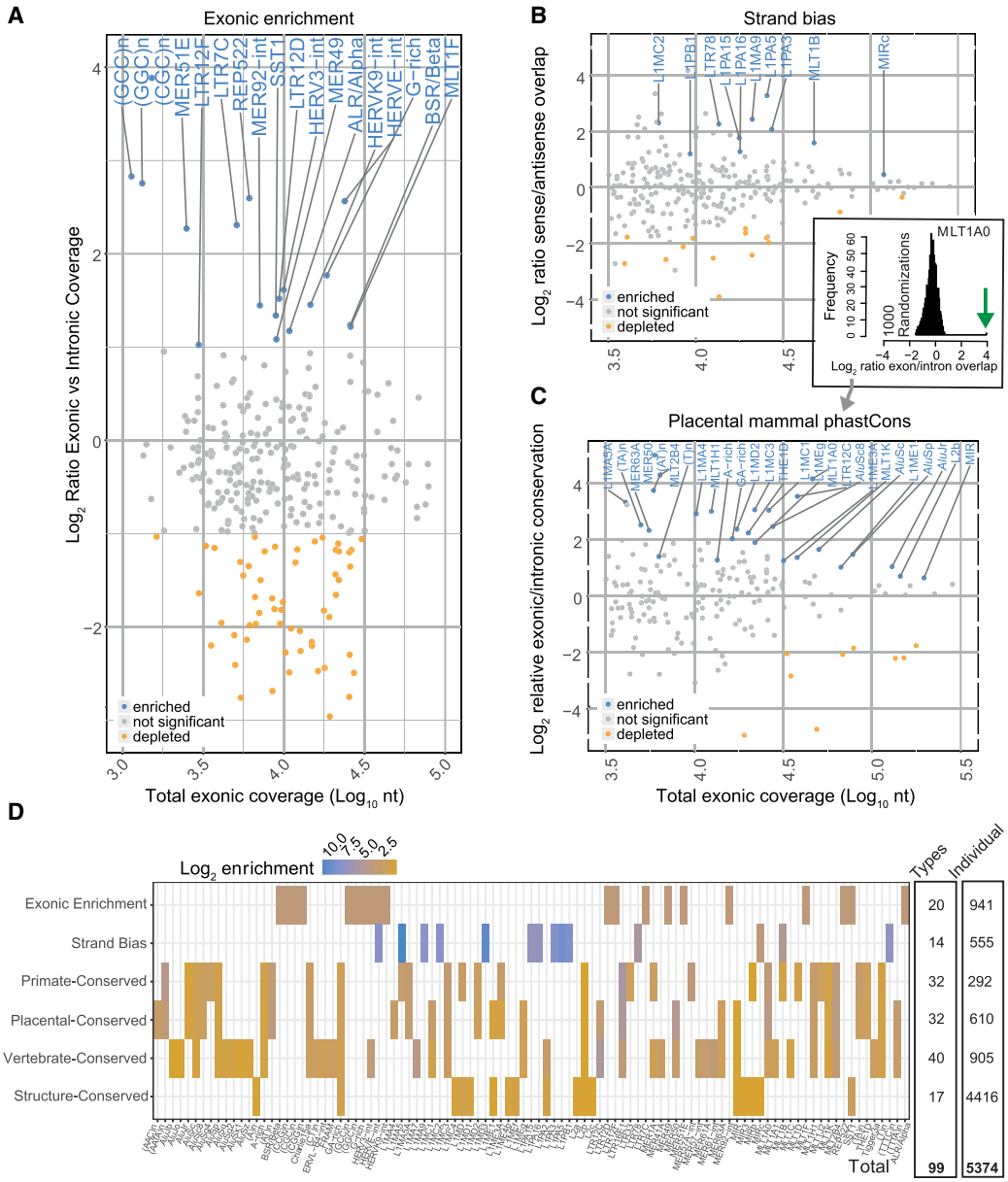

Figure 3. Evidence for selection on TEs in lncRNA exons. (*A*) Figure shows, for every TE type, the enrichment of per nucleotide coverage in exons compared with introns (*y*-axis) and overall exonic nucleotide coverage (*x*-axis). Enriched TE types (at a twofold cutoff) are shown in blue. (*B*) As for *A*, but this time the *y*-axis records the ratio of nucleotide coverage in sense versus antisense configuration. "Sense" here is defined as sense of TE annotation relative to the overlapping exon. Similar results for lncRNA introns may be found in Supplemental Figure S1. Significantly enriched TE types are shown in blue. Statistical significance was estimated by a randomization procedure, and significance is defined at an uncorrected empirical *P*-value <0.001 (see Methods). (*C*) As for *A*, but here the *y*-axis records the ratio of per-nucleotide overlap by phastCons mammalian-conserved elements for exons versus introns. Similar results for three other measures of evolutionary conservation may be found in Supplemental Figure S1. Significantly enriched TE types are shown in blue. Statistical significance was estimated by a randomization procedure, and significance is defined at an uncorrected empirical *P*-value <0.001 (see Methods). An example of significance estimation is shown in the *inset*: The distribution shows the exonic/intronic conservation ratio for 1000 simulations. The green arrow shows the true value, in this case for MLT1A0 type. (*D*) Summary of TE types with evidence of exonic selection. Six distinct evidence types are shown in rows, and TE types in columns. On the *right* are summary statistics for (1) the number of unique TE types identified by each method and (2) the number of instances of exonic TEs from each type with appropriate selection evidence. The latter are henceforth defined as RIDLs.

because of exonic enrichment (Fig. 4B). The entire RIDL annotation is available in Supplemental File S1.

It is important to consider this RIDL annotation as an "enrichment group," with a greater proportion of functional TEs than when using the entire exonic TE set. By using introns as a reference, we conservatively estimate the fraction of true-positive predictions to range from 12% (strand bias) to 40% (phastCons primate) and 78% (exonic enrichment) (Supplemental Fig. S4).

We also examined the evolutionary history of RIDLs. By using six-mammal alignments, their depth of evolutionary conservation could be inferred (Supplemental Fig. S5): 12% of instances appear to be great ape-specific, with no orthologous sequence beyond

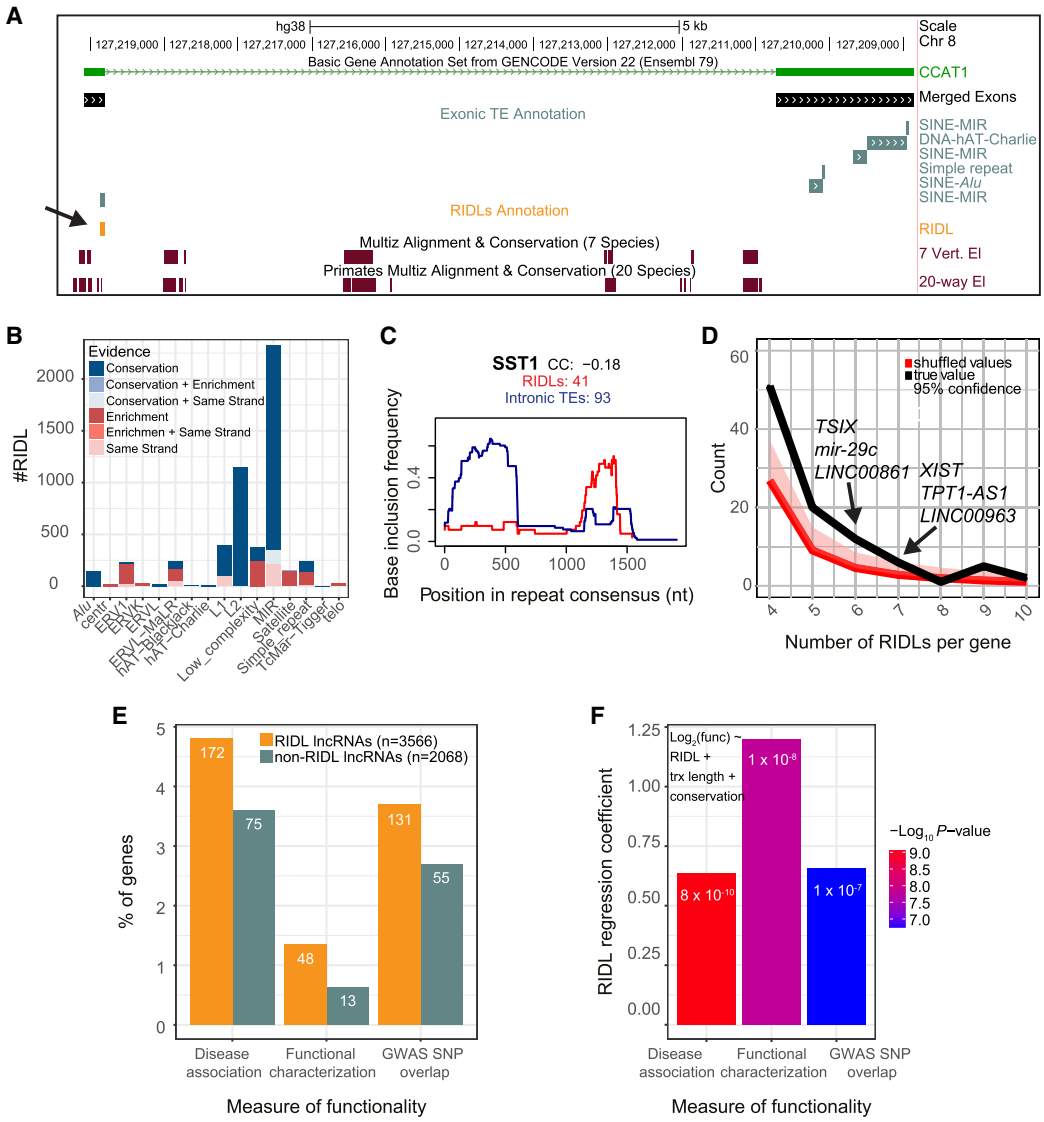

Figure 4. Annotated RIDLs and RIDL-lncRNAs. (*A*) Example of a RIDL-lncRNA gene, *CCAT1*. Of note is that although several exonic TE instances are identified (gray), including three separate MIR elements, only one is defined a RIDL (orange) because of overlap of a conserved element. (*B*) Breakdown of RIDL instances by TE family and evidence sources. (*C*) Insertion profile of SST1 RIDLs (blue) and intronic insertions (red): *x*-axis shows the entire consensus sequence of SST1; *y*-axis indicates the frequency with which each nucleotide position is present in the aggregate of all insertions. (CC) Spearman's correlation coefficient of the two profiles; (RIDLs or intronic TEs) the numbers of individual insertions considered for RIDLs/intronic insertions, respectively. (*D*) Number of lncRNAs (*y*-axis) carrying the indicated number of RIDL (*x*-axis) given the true distribution (black) and randomized distribution (red). The 95% confidence interval was computed empirically by randomly shuffling RIDLs across the entire lncRNA annotation. (*E*) Percentage of RIDL-lncRNAs, and a length-matched set of non-RIDL lncRNAs, which are present in disease- and cancer-associated lncRNA databases (see Methods) or in the lncRNAdb database of functional lncRNAs (functional characterization) or contain at least one trait/disease-associated SNP in an exonic region (GWAS SNP overlap). Numbers denote gene counts. (*F*) Plot shows regression coefficients for the "RIDL" term in the indicated multiple logistic regression model using the same measures of functionality as in *E*. Colors indicate the associated *P*-value. These values assess the correlation between RIDL number and measures of functionality of their host transcript, while accounting for transcript length (trx length) and conservation.

chimpanzee; 47% are primate-specific, whereas the remaining 40% are identified in at least one nonprimate mammal. The wide timeframe for appearance of RIDLs is consistent with the wide diversity of TE types, from ancient MIR elements to presently active LINE-1 (Jurka et al. 1995; Konkel et al. 2010; Smith et al. 2013).

Instances of genomic TE insertions typically represent a fragment of the full consensus sequence. We hypothesized that particular regions of the TE consensus will be important for RIDL activity, introducing selection for these regions that would distinguish them from unselected, intronic copies. To test this, we compared insertion profiles of RIDLs to intronic instances for each TE type and used the correlation coefficient (CC) as a quantitative measure of similarity (Fig. 4C; Supplemental File S2). For 17 cases, a CC < 0.9 points to possible selective forces acting on RIDL insertions. An example is the macrosatellite SST1 repeat in which RIDL copies in 41 lncRNAs show a strong preference inclusion of the 3′ end, in contrast to the general 5′ preference observed in introns (Fig. 4C). This suggests a possible functional relevance for the 1000- to 1500-nt region of the SST1 consensus.

To assess whether RIDLs experience purifying evolutionary selection in modern humans, we analyzed the derived allele frequency (DAF) spectrum of their overlapping SNPs (Supplemental Fig. S6; Haerty and Ponting 2013; Tan et al. 2017). This showed that RIDLs (orange bars) have a greater proportion of rare (DAF < 0.1) alleles compared with other TEs in exons (green bars) or introns (turquoise bars) of the same lncRNAs and, indeed, compared with non-RIDL exonic nucleotides (black bars). These differences fail to reach statistical significance, possibly because of small sample sizes. Overall these data are consistent with RIDLs experiencing an elevated rate of purifying evolutionary selection in modern humans compared with nearby neutral sequence, although larger data sets will be required before this can be stated conclusively.

RIDL-carrying lncRNAs are enriched for functions and disease roles

We next looked for evidence to support the RIDL annotation by investigating the properties of their host lncRNAs. We first asked whether RIDLs are randomly distributed among lncRNAs or are nonrandomly clustered in a smaller number of genes. Figure 4D shows that the latter is the case, with a significant deviation of RIDLs from a random distribution. These lncRNAs carry a mean of 1.15 RIDLs/kb of exonic sequence (median, 0.84 RIDLs/kb) (Supplemental Fig. S7).

Are RIDL-lncRNAs more likely to be functional? To address this, we compared lncRNA genes carrying one or more RIDLs to a length-matched set of control lncRNAs (Fig. 4E; Supplemental Fig. S8). We observed that RIDL-lncRNAs are (1) overrepresented in the reference database for functional lncRNAs, lncRNAdb (Quek et al. 2015); (2) enriched in associations with cancer and other diseases; and (3) enriched in their exons for trait/disease-associated SNPs. To estimate the impact of carrying RIDLs on the functional-associated outcomes mentioned above while controlling for potential biases from conservation and length, we performed multiple logistic regression analysis. In each case, the overlap with RIDL-lncRNAs was positive and statistically significant (Fig. 4F). However, we did not observe any difference in mean or maximum expression of RIDL-lncRNAs to length-matched controls across 10 tissues of the human body map RNA-seq data set (Supplemental Fig. S9).

In addition to *CCAT1* (Fig. 4A; Nissan et al. 2012), there are a number of deeply studied RIDL-containing genes. *XIST*, the X Chromosome silencing RNA, contains seven internal RIDL elements. As we pointed out previously (Johnson and Guigó 2014), these include an array of four similar pairs of MIRc/L2b repeats. The prostate cancer–associated *UCA1* gene has a transcript isoform promoted from an LTR7c, as well as an additional internal RIDL, thereby making a potential link between cancer gene regulation and RIDLs. The *TUG1* gene, involved in neuronal differentiation, contains highly evolutionarily conserved RIDLs, including Charlie15k and MLT1K elements (Johnson and Guigó 2014). Other RIDL-containing lncRNAs include *MEG3*, *MEG9*, *SNHG5*, *ANRIL*, *NEAT1*, *CARMEN1*, and *SOX2OT*. *LINC01206*, located adjacent to *SOX2OT*, also contains numerous RIDLs. A full list can be found in Supplemental File S3.

Correlation between RIDLs and subcellular localization of the host transcript

The location of a lncRNA within the cell is of key importance to its molecular function (Derrien et al. 2012; Cabili et al. 2015; Mas-Ponte et al. 2017); therefore, we next investigated whether RIDLs might regulate lncRNA localization (Fig. 5A; Zhang et al. 2014a; Chillón and Pyle 2016; Hacisuleyman et al. 2016). By using subcellular RNA-seq data based on 10 ENCODE cell lines (Djebali et al. 2012), we calculated the relative nuclear/cytoplasmic localization in $\log_2$ units, or relative concentration index (RCI) (Mas-Ponte et al. 2017). By using this data set, we tested each of the 99 RIDL types for association with localization of their host transcript.

After correcting for multiple hypothesis testing using the Benjamini–Hochberg method (Benjamini and Hochberg 1995), this approach identified four distinct RIDL types: L1PA16, L2b, MIRb, and MIRc (Fig. 5B). For example, 44 lncRNAs carrying L2b RIDLs have a 6.9-fold higher relative nuclear/cytoplasmic ratio in IMR-90 cells, and this tendency is observed in six different cell types (Fig. 5B,C).

The degree of nuclear localization increases in lncRNAs as a function of the number of RIDLs (L1PA16, L2b, MIRb, and MIRc) they carry (Fig. 5D). We also found a significant relationship between GC-rich elements and cytoplasmic enrichment across three independent cell samples. The GC-rich-containing lncRNAs have between two- and 2.3-fold higher relative expression in the cytoplasm of these cells (Supplemental Fig. S10).

We were curious whether this relationship with localization is only a property of RIDLs or, conversely, holds true when considering any instances of L1PA16, L2b, MIRb, and MIRc. Indeed, when the preceding analysis was repeated with unfiltered TE instances, the latter was observed (Supplemental Fig. S11). However, the strength of the effect was consistently lower than for RIDLs (Supplemental Fig. S12). This difference between RIDLs and unfiltered TEs supports both the usefulness of the RIDL identification method and the idea that RIDLs are under selection as a result of their effect on localization.

We were concerned that two unmodeled confounding factors that positively correlated with TE number could explain the observed data: transcript length and whole-cell gene expression. To address this, we performed multiple linear regression for localization with explanatory variables of RIDL number, transcript length, and whole-cell expression (Fig. 5E). Such a model accounts independently for each variable, enabling one to eliminate confounding effects. Training such models for each cell type/RIDL pair, we observed positive and statistically significant contributions for RIDL number in most cases. We also observed weaker but significant contributions from transcript length and whole-cell expression terms, indicating that our intuition was correct that these factors influence localization independently of RIDLs (Supplemental Fig. S13A,B). We drew similar conclusions from equivalent analyses using partial correlation (Supplemental Fig. S13C). In summary, observed RIDLs correlate with lncRNA localization even when controlling for other factors.

Given that L2b and MIR elements predate human–mouse divergence, we attempted to perform similar analyses in mouse cells. However, given that just two equivalent data sets are available at present (Bahar Halpern et al. 2015; Tan et al. 2015), as well as the relatively low number of annotated lncRNAs in mouse, we were unable to draw statistically robust conclusions regarding the evolutionary conservation of this phenomenon.

Intra-gene correlation between RIDLs and subcellular localization

LncRNA gene loci are often composed of multiple, differentially spliced transcript isoforms that partially differ in their mature

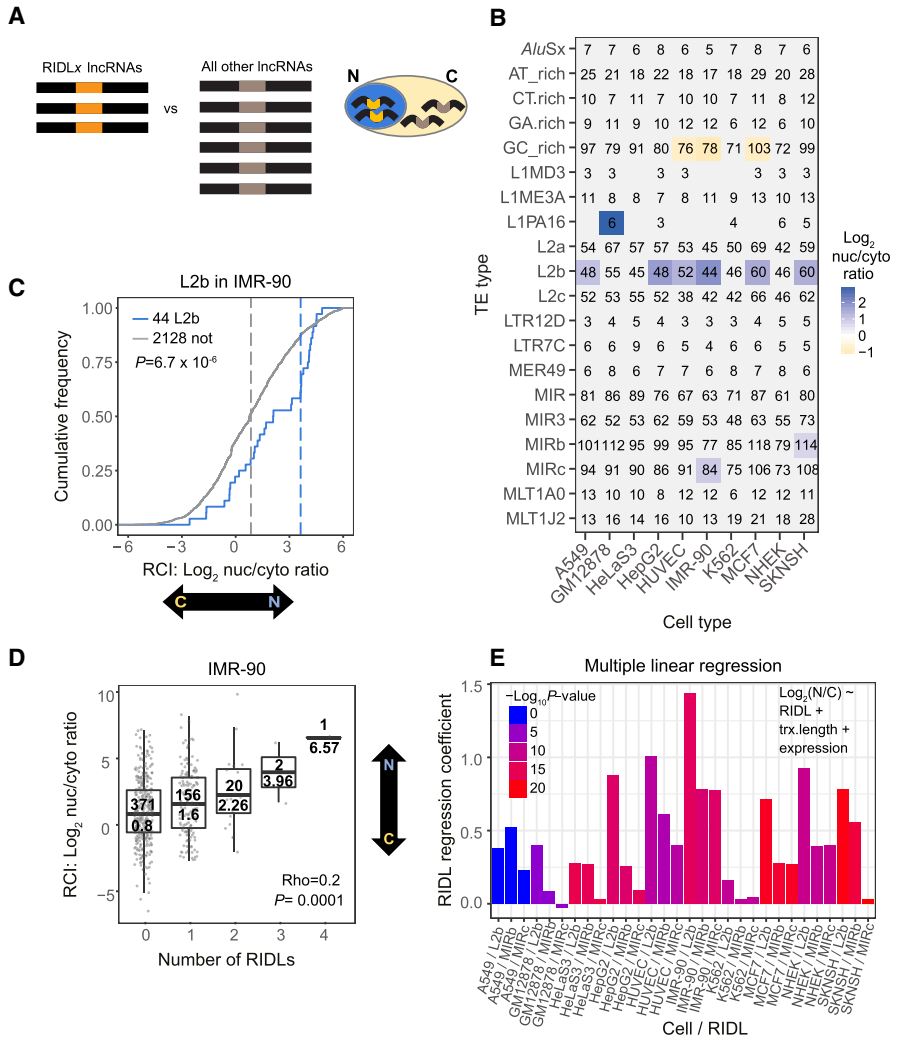

Figure 5. Correlation between RIDLs and host lncRNA nuclear/cytoplasmic localization. (*A*) Outline of in silico screen for localization-regulating RIDLs. For each RIDL-type/cell-type combination, the nuclear/cytoplasmic localization of RIDL-lncRNAs is compared to all other detected lncRNAs. (*B*) Results of screen. (Rows) RIDL types; (columns) cell types. Significant RIDL–cell type combinations are colored (Benjamini–Hochberg corrected *P*-value <0.01; Wilcoxon test). Color scale indicates the nuclear/cytoplasmic ratio mean of RIDL-lncRNAs. Numbers in cells indicate the number of considered RIDL-lncRNAs. Analyses were performed using a single representative transcript isoform from each gene locus, being that with the greatest number of exons. (*C*) The nuclear/cytoplasmic localization of lncRNAs carrying L2b RIDLs in IMR-90 cells. Blue indicates lncRNAs carrying one or more RIDLs; gray indicates all other detected lncRNAs (not). Dashed lines represent medians. Significance was calculated using Wilcoxon test (*P*). (*D*) The nuclear/cytoplasmic ratio of lncRNAs as a function of the number of RIDLs that they carry (L1PA16, L2b, MIRb, MIRc). CC (Rho) and the corresponding *P*-value (*P*) were calculated using Spearman correlation, two-sided test. In each box, the *upper* value indicates the number of lncRNAs; *lower* value, the median. (*E*) Plot shows regression coefficients for the RIDL term in the indicated linear model using L2b, MIRb, and MIRc RIDLs (see Methods). Colors indicate the associated *P*-value. These values assess the correlation between RIDL number and nuclear/cytoplasmic localization (Log_2(N/C)) of their host transcript while accounting for possible confounding factors of transcript length (trx.length) or whole-cell expression levels (expression).

sequence. We reasoned that differential inclusion of RIDL-containing exons should give rise to differences in localization among transcripts from the same gene locus. In other words, for RIDL-lncRNA gene loci having multiple transcript isoforms, those isoforms with a RIDL should display greater nuclear enrichment than those isoforms without a RIDL (Fig. 6, left).

We tested this individually for each cell type. For every appropriate RIDL-lncRNA locus (numbers shown inside boxplot), we calculated the difference in the mean of the localization between their RIDL and non-RIDL isoforms (Fig. 6, right). For every cell line, the median difference was positive, indicating that RIDL-carrying transcript isoforms are more nuclear enriched than their non-RIDL cousins from the same gene locus. Given our a priori hypothesis that RIDLs promote nuclear enrichment, statistical significance was tested by comparison to zero using a one-sided *t*-test. Altogether, these data point to a consistent correlation between the presence of certain exonic TE elements—L1PA16, L2b, MIRb, and MIRc—and the nuclear enrichment of their host lncRNA.

RIDLs play a causative role in lncRNA nuclear localization

To more directly test whether RIDLs play a causative role in nuclear localization, we designed an experimental approach to quantify the effect of exonic TEs on localization of a transfected lncRNA. We selected three lncRNAs, based on (1) presence of L2b, MIRb, and MIRc RIDLs; (2) moderate expression; and (3) nuclear localization as inferred from RNA-seq (Fig. 7A,B; Supplemental Fig. S14). Nuclear localization of these candidates could be validated in HeLa cells using qRT-PCR (Fig. 7C).

We formulated an assay to compare the localization of transfected lncRNAs carrying wild-type RIDLs and of mutated versions in which the RIDL sequence was randomized without altering sequence composition (mutant) (Fig. 7D; full sequences available in Supplemental File S4). Wild-type and mutant lncRNAs were transfected into cultured cells and their localization evaluated by fractionation. qRT-PCR primers were designed to distinguish transfected wild-type and mutant transcripts from endogenously expressed copies. Transgenes were typically expressed in a range of 0.2- to 10-fold compared with their endogenous transcripts (Supplemental Fig. S15). Fractionation purity was verified by western blotting (Fig. 7E) and qRT-PCR (Fig. 7F), and stringent DNase-treatment ensured that plasmid DNA made negligible contributions to our results (Supplemental Fig. S16).

With this setup, we compared the nuclear/cytoplasmic localization of lncRNAs with and without exonic RIDL sequences (Fig. 7F). We observed a potent and consistent impact of RIDLs on nuclear/cytoplasmic localization in HeLa cells: For all three candidates, the loss of RIDL sequence resulted in relocalization of the host transcript from nucleus to cytoplasm (Fig. 7F, *top*). We

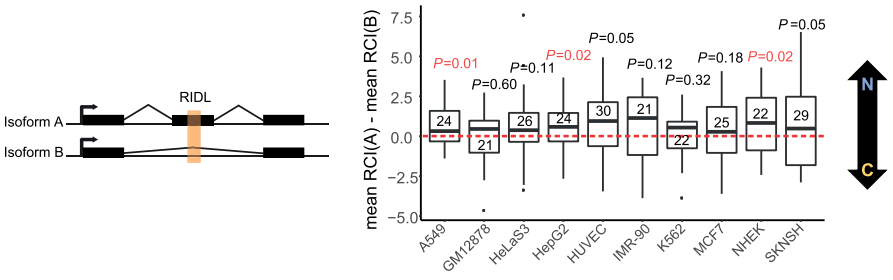

Figure 6. RIDLs correlate with differential localization of lncRNA transcripts from the same locus. Distribution of differences between RCI mean of transcripts with nuclear RIDL (mean RCI(A)) and RCI mean of transcripts without nuclear RIDL (mean RCI(B)). A positive value indicates that RIDL-carrying transcripts are more nuclear-enriched than non-RIDL transcripts. Data were calculated individually for every gene that has one or more RIDL-transcript and one or more non-RIDL transcript expressed in a given cell line. Numbers inside the boxplots indicate the number of gene loci analyzed for each cell line. Horizontal bar indicates the median. Here nuclear RIDL refers to L1AP16, MIRb, MIRc, and L2b. *P*-values obtained from one-sided *t*-test are shown (in red when $P < 0.05$).

repeated these experiments in another cell line, A549, and observed similar, albeit less pronounced, effects (Fig. 7F, *bottom*). This difference may be because of the less nuclear localization of the endogenous transcripts in A549 (Supplemental Fig. S17). To summarize, exonic L2b, MIRb, and MIRc elements promote the nuclear enrichment of host lncRNAs.

Discussion

Recent years have seen a rapid increase in the number of annotated lncRNAs. However, our understanding of their molecular functions and of how such functions are encoded in primary RNA sequences lags far behind. Two recent conceptual developments offer hope for resolving the sequence-function code of lncRNAs: (1) The subcellular localization of lncRNAs is a readily quantifiable characteristic that holds important clues to function and (2) the abundant TE content of lncRNAs may contribute to functionality.

In this study, we have linked these two ideas by showing evidence that certain TEs can drive the nuclear enrichment of lncRNAs. A global correlation analysis of TEs and RNA localization data revealed a handful of TEs, most notably LINE2b, MIRb, and MIRc, that positively and significantly correlate with the degree of nuclear/cytoplasmic localization of their host transcripts. This correlation is observed in multiple cell types and scales with the number of TEs present. A causative link was established experimentally, confirming that the indicated TEs are sufficient for a two- to fourfold increase in nuclear/cytoplasmic localization. There are two principal explanations for this phenomenon: (1) an "active" process whereby TEs are recognized by a cellular transport pathway, as demonstrated for *Alus* by Lubelsky and Ulitsky (2018); and (2) a "passive" process in which TEs destabilize transcripts, leading to a concentration gradient from nucleus to cytoplasm. Although future studies will examine this question in detail, the fact that we do not observe a constant difference in steady-state levels in TE/mutated transgenes would be more consistent with the active model.

These data support the hypothesis that exonic TE elements can act as functional lncRNA domains. In this RIDL hypothesis, TEs are co-opted by natural selection to form RIDLs, that is, fragments of sequence that confer adaptive advantage through some change in the activity of their host lncRNA. We proposed that RIDLs may serve as binding sites for proteins or other nucleic acids, and indeed, a growing body of evidence supports this (for review, see Johnson and Guigó 2014). In the context of localization, RIDLs could mediate nuclear retention through hybridization to complementary repeats in genomic DNA or through their described interactions with nuclear proteins (Kelley et al. 2014). In the course of this study, we bioinformatically identified five candidate proteins (HNRNPU, HNRNPH2, ELAVL1, KHDRBS1, TARDBP); however, we could not find evidence that they contribute to RIDL-lncRNA localization. Identification of any proteins that mediate RIDLs' localization activity may be achieved in the future through pulldown approaches (Marín-Béjar and Huarte 2015).

The localization RIDLs discovered—MIR and LINE-2—are both ancient and contemporaneous, being active before the mammalian radiation (Cordaux and Batzer 2009). Both have previously been associated with acquired roles in the context of genomic DNA but not, to our knowledge, in RNA (Johnson et al. 2006; Jjingo et al. 2014). Although the evolutionary history of lncRNAs remains an active area of research and accurate dating of lncRNA gene birth is challenging, it appears that the majority of human lncRNAs were born after the mammalian radiation (Necsulea et al. 2014; Washietl et al. 2014; Hezroni et al. 2015, 2017). This would mean that MIR and LINE-2 RIDLs were pre-existing sequences that were exapted by newly born lncRNAs, corresponding to the "latent" exaptation model proposed by Feschotte and colleagues (Chuong et al. 2017). However, it is also possible that for other cases the reverse could be true: A pre-existing lncRNA exapts a newly inserted TE. Given that nuclear retention is at odds with the primary needs of natural TE transcripts to be exported to the cytoplasm, we propose that the observed nuclear localization activity is a more modern feature of L2b/MIR RIDLs, which is unrelated to their original roles.

Our approach for identifying localization-regulating RIDLs has advantages over previous studies (Hacisuleyman et al. 2016; Lubelsky and Ulitsky 2018) in terms of its genome-wide scale. However, an unavoidable consequence of our use of evolutionary conservation as a filter is that it likely biases our analysis against recently evolved TEs such as *Alus*. It remains entirely possible that modern TEs also influence lncRNA localization but cannot be detected using the signals of selection that we have used. On the other hand, MIRb and MIRc were only identified in one cell type each. We expect this reflects low sensitivity of the statistical screen rather than cell-type specificity alone because (1) in a focused reanalysis (Supplemental Fig. S11) the effect was observed in multiple cells, and (2) experimental validation confirmed it in two independent cell types (Fig. 7F).

This is further supported by the recent study of Lubelsky and Ulitsky (2018), who performed an experimental screen for localization motifs in 37 nuclear-enriched lncRNAs and identified *Alu*Sx as a nuclear-localization element. These 37 lncRNAs are enriched for RIDLs (62% of Lubelsky lncRNAs contain at least one RIDL compared with 22% for other GENCODE v21 lncRNAs, $P = 4 \times 10^{-6}$, Fisher's exact test), as well as for the three localization RIDLs identified here (L2b, MIRb, MIRc: 32% vs. 9%, $P = 3 \times 10^{-4}$) (Supplemental Fig. S18A). Although our bioinformatic

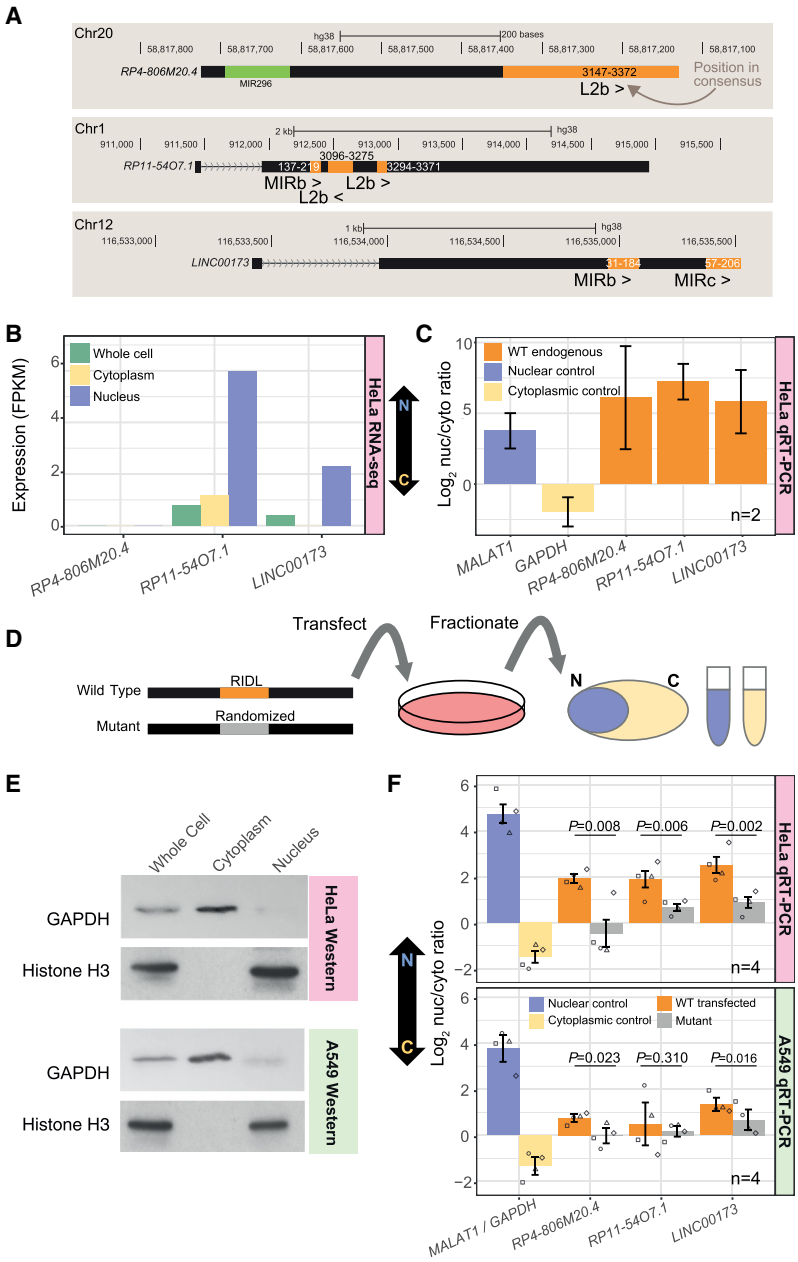

Figure 7. Disruption of RIDLs results in lncRNA relocalization from nucleus to cytoplasm. (*A*) Structures of candidate RIDL-lncRNAs. Orange indicates RIDL positions. For each RIDL, numbers indicate the position within the TE consensus, and its orientation with respect to the lncRNA is indicated by arrows. (>) same strand; (<) opposite strand. (*B*) Expression of the three lncRNA candidates as inferred from HeLa RNA-seq (Djebali et al. 2012). (*C*) Nuclear/cytoplasmic localization of endogenous candidate lncRNA copies in wild-type HeLa cells as measured by qRT-PCR. (*D*) Experimental design. (*E*) The purity of HeLa and A549 subcellular fractions was assessed by western blotting against specific markers. GAPDH/Histone H3 proteins are used as cytoplasmic/nuclear markers, respectively. (*F*) Nuclear/cytoplasmic localization of transfected candidate lncRNAs in HeLa (*top*) and A549 (*bottom*). *GAPDH/MALAT1* are used as cytoplasmic/nuclear controls, respectively. (N) Number of biological replicates (values from all replicates are plotted; each replicate is represented by a different dot shape). Error bars, SEM. *P*-values for paired *t*-test (one tail) are shown.

screen did not identify *Alu*Sx, a naive unfiltered reanalysis of our data supports Lubelsky's experimental finding that *Alu*Sx-carrying lncRNAs tend to be more nuclear across multiple cell types (Supplemental Fig. S18B). Together, these considerations open the possibility that other localization-controlling TE types may await discovery.

More generally, the RIDL predictions showed rather low concordance between the various selection evidence used (Fig. 3D). This likely reflects a number of factors: young evolutionary age of some of the most common TEs, generally low statistical power because of large background of neutral TEs and multiple hypothesis testing, and false positives because of TEs that promote transcription or splicing of lncRNAs. However, it is worthy of note that validated candidates L2b, MIRb, and MIRc are all implicated by multiple, independent evidence sources (Fig. 3D).

This work marks a step in the ongoing efforts to map the domains of lncRNAs. Previous studies have used a variety of approaches, from integrating experimental protein-binding data (Li et al. 2014; Van Nostrand et al. 2016; Hu et al. 2017) to evolutionarily conserved segments (Smith et al. 2013; Seemann et al. 2017). Previous maps of TEs have highlighted their profound roles in lncRNA gene evolution (Kelley and Rinn 2012; Kapusta et al. 2013; Hezroni et al. 2015). However, the present RIDL annotation stands apart in attempting to identify the subset of TEs with evidence for selection. We hope that this RIDL map will prove a resource for future studies to better understand functional domains of lncRNAs. Although various evidence suggests that the RIDL annotation is a useful enrichment group of functional TE elements, it contains substantial false-positive (and likely also false-negative) rates that will have to be improved in the future.

This study may help to explain a longstanding and unexplained property of lncRNAs: their nuclear enrichment (Derrien et al. 2012; Ulitsky and Bartel 2013). Although they are readily detected in the cytoplasm, lncRNAs general tendency is to have higher nuclear/cytoplasmic ratios compared with those of mRNAs (Clark et al. 2012; Derrien et al. 2012; Ulitsky and Bartel 2013; Mas-Ponte et al. 2017). This is true across various human and mouse cell types. Although this may partially be explained by decreased stability (Mukherjee et al. 2017), it is likely that RNA sequence motifs also contribute to nuclear localization (Zhang et al. 2014a; Chillón and Pyle 2016). Here we show that this is the case and that the enrichment of certain RIDL types in lncRNA mature sequences is likely to be a major contributor to lncRNA nuclear

retention. In contrast, the far lower exonic content of TEs in protein-coding mRNAs may help explain their greater cytoplasmic abundance (Kapusta and Feschotte 2014). Indeed, even within the cytoplasm, there is evidence that TE content may also influence the efficiency with which lncRNAs are trafficked to the translation machinery (Carlevaro-Fita et al. 2016). Together, this evidence may reflect unknown cellular quality control mechanisms that vet RNAs based on their TE content, tending to retain TE-rich sequences (including lncRNAs or incorrectly processed mRNAs) in the nucleus, and promote the cytoplasmic export and ribosomal loading of canonical TE-poor mRNAs.

In summary, therefore, we have made available a first annotation of selected RIDLs in lncRNAs and described a new paradigm for TE-derived fragments as drivers of nuclear localization in lncRNAs.

Methods

All operations were performed on human genome version GRCh38/hg38, unless stated otherwise.

Exonic TE curation

RepeatMasker annotations were downloaded from the UCSC Genome Browser (version hg38) on December 31, 2014 (Smit et al. 2013–2015), and GENCODE v21 lncRNA annotations in GTF format were downloaded from www.gencodegenes.org (Harrow et al. 2012). Annotations were not filtered further. The transposon.profiler script, largely based on BEDTools' intersect and merge functionalities (Quinlan and Hall 2010), was used to annotate exonic and intronic TEs of the given gene annotation (Supplemental Code). Exons of all transcripts belonging to the given gene annotation were merged and are henceforth referred to as exons. The set of introns was curated by subtracting the merged exonic sequences from the full gene spans and only retaining those introns that belonged to a single gene. Intronic regions were assigned the strand of the host gene.

The RepeatMasker annotation file was intersected with exons and classified into one of six categories: transcription start site (TSS), overlapping the first exonic nucleotide of the first exon; splice acceptor, overlapping exon 5′ end; splice donor, overlapping exon 3′ end; internal, residing within an exon and not overlapping any intronic sequence; encompassing, in which an entire exon lies within the TE; and transcription termination site (TTS), overlapping the last nucleotide of the last exon. In every case, the TEs are separated by strand relative to the host gene: positive, in which both gene and TE are annotated on the same strand, otherwise negative. The result is the Exonic TE Annotation (Supplemental File S5).

RIDL identification

By using this Exonic TE Annotation, we identified the subset of individual TEs with evidence for functionality. For certain analysis, an Intronic TE Annotation was also used, being the output for the equivalent intron annotation described above. Three different types of evidence were used: enrichment, strand bias, and evolutionary conservation.

In enrichment analysis, the exon/intron ratio of the fraction of nucleotide coverage by each repeat type was calculated. Any repeat type with greater than twofold exon/intron ratio was considered as a candidate. All exonic TE instances belonging to such TE types are defined as RIDLs.

In strand bias analysis, a subset of Exonic TE Annotation was used, being the set of nonsplice junction crossing TE instances (noSJ). This additional filter was used to guard against false-positive enrichments for TEs known to provide splice sites (Lev-Maor et al. 2003; Sela et al. 2007). For all TE instances, the relative strand was calculated: positive, if the annotated TE strand matches that of the host transcript; negative, if not. Then for every TE type, the ratio of relative strand sense/antisense was calculated. Statistical significance was calculated empirically: Entire gene structures were randomly repositioned in the genome using BEDTools shuffle, and the intersection with the entire RepeatMasker annotation was recalculated. For each iteration, sense/antisense ratios were calculated for all TE types. A TE type was considered to have significant strand bias if its true ratio exceeded (positively) all of 1000 simulations. All exonic instances of these TE types that also have the same strand orientation to the host transcript are defined as RIDLs. On the other hand, after inspection of the data, we decided to exclude TEs with significant antisense enrichment. This is because most instances were from the LINE-1 class, which are known to interfere with gene expression when falling on the same strand (Perepelitsa-Belancio and Deininger 2003). Therefore, we considered it likely that observed antisense enrichment is simply an artifact of selection against insertion on the same strand and, in the interests of controlling the false-positive prediction rate, decided to exclude these cases.

In evolutionary analysis, four different annotations of evolutionarily conserved regions were treated similarly, using unfiltered Exonic TE Annotations. Primate, placental mammal, and vertebrate phastCons elements based on 46-way alignments were downloaded as BED files from UCSC Genome Browser (Siepel et al. 2005), whereas the ECS conserved regions were obtained from the Supplemental Data of Smith et al. (2013) (for summary, see Supplemental File S6). Because at the time of analysis phastCons elements were only available for hg19 genome build, we mapped them to hg38 using liftOver utility (Hinrichs et al. 2006). For each TE type, we calculated the exonic/intronic conservation ratio. To do this, we used IntersectBED (Quinlan and Hall 2010) to overlap exonic locations with TEs and calculate the total number of nucleotides overlapping. We performed a similar operation for intronic regions. Then for each TE type, we calculated the ratio of conserved TE nucleotides for exons compared with introns:

$$\text{Relative exonic-intronic conservation (REIC)} = (C_e/(C_e + N_e))/(C_i/(C_i + N_i)),$$

in which C is conserved TE nucleotides, N is nonconserved TE nucleotides, and subscripts e and i denote exonic and intronic, respectively. Note that because it calculates fractional overlap of TEs by conserved elements, REIC normalizes for different lengths of exons and introns (Supplemental Fig. S19).

To estimate the background, the conserved element BED files were positionally randomized 1000 times using BEDTools shuffle, each time recalculating REIC. We considered to be significantly conserved those TE types in which the true REIC was greater or less than every one of 1000 randomized REIC values. All exonic instances of these TE types that also intersect the appropriate evolutionarily conserved element are defined as RIDLs. This approach of shuffling conserved elements displayed no apparent bias in the length of TEs it identifies (Supplemental Fig. S2D). We also tested an alternative approach for estimating significance, whereby conserved elements were held constant and TEs were positionally randomized. Although there was a significant overlap in identified candidate RIDLs, this method displayed a bias toward longer TEs (Supplemental Fig. S2D) and therefore was not used further.

We chose to randomize conserved elements rather than TEs because the former are enriched in lncRNA exons (Pegueroles and Gabaldón 2016). Thus, using randomized TEs to estimate

background REIC would lead to overestimation of exonic TE conservation and, hence, underestimation of the rate of conservation of TEs in real data.

All RIDL predictions were then merged using mergeBED, and any instances with length <10 nt were discarded. The outcome, a BED format file with coordinates for hg38, is found in Supplemental File S1.

FDRs were estimated for RIDL predictions. TE-type FDR estimates were based on shuffling simulations described above. Empirical *P*-values for true data were estimated according to *P* = (rank in distribution)/(1 + number of simulations). For significant cases, in which the true value exceeded all $n = 1000$ simulations, this value was conservatively defined to be $P = 0.001$. These empirical *P*-values were then converted to FDR using the R command *p.adjust* with the *fdr* setting (Rackham et al. 2011; R Core Team 2015). Accordingly, the empirical significance cutoff ($P < 0.001$) mentioned in the main text corresponds to the following FDR values: strand bias, 0.027; vertebrate phastCons, 0.013; placental phastCons, 0.014; primate phastCons, 0.009; and ECS, 0.034. This analysis is conservative because empirical *P*-values of candidates are rounded up in every case to 0.001.

FDR rates were also estimated at the element level. Here, the set of significant TEs were grouped for each evidence type. Then the frequency of overlap of these TEs with the evidence type was compared for lncRNA exons and introns. These data are shown in Supplemental Figure S4.

RIDL orthology analysis

To assess evolutionary history of RIDLs, we used chained alignments of human to chimp (hg19ToPanTro4), macaque (hg19ToRheMac3), mouse (hg19ToMm10), rat (hg19ToRn5), and cow (hg19ToBosTau7). Because of the availability of chain files, RIDL coordinates were first converted from hg38 to hg19. Orthology was defined by liftOver utility used at default settings (Hinrichs et al. 2006).

DAF analysis

We used allele frequencies from African population provided by the 1000 Genomes Project (The 1000 Genomes Project Consortium et al. 2015), as performed previously by (Haerty and Ponting 2013). DAF was determined for human common SNPs from dbSNP (build 150) (Sherry et al. 2001) for every group analyzed. Ancestral repeats (ARs) were defined as human repeats (excluding simple repeats) intersecting at least one nucleotide of mouse repeats defined by liftOver and falling within 5 kb of but not overlapping RIDL-containing genes.

Comparing RIDL-carrying lncRNAs versus other lncRNAs

To test for functional enrichment among lncRNAs hosting RIDLs, we tested for statistical enrichment of the following traits in RIDL-carrying lncRNAs compared with other lncRNAs (see below) by Fisher's exact test:

- Functionally characterized lncRNAs are lncRNAs from GENCODE v21 that are present in lncRNAdb (Quek et al. 2015).
- Disease-associated genes are lncRNAs from GENCODE v21 that are present in at least in one of the following databases or public sets: LncRNADisease (Chen et al. 2013), Lnc2Cancer (Ning et al. 2016), Cancer LncRNA Census (CLC) (Carlevaro-Fita et al. 2017).
- For GWAS SNPs, we collected SNPs from the NHGRI-EBI Catalog of published genome-wide association studies (Hindorff et al. 2009; Welter et al. 2014; https://www.ebi.ac.uk/gwas/home). We intersected its coordinates with lncRNA exons coordinates.

For defining a comparable set of "other lncRNAs," we sampled from the rest of GENCODE v21 a set of lncRNAs matching RIDL-lncRNAs' exonic length distribution (Supplemental Fig. S8). We performed sampling using the matchDistribution script (https://github.com/julienlag/matchDistribution). To simultaneously control for both conservation and length, we performed multiple logistic regression analysis using the *glm* R function (R Core Team 2015), with the following structure:

$$\text{Functional-association outcome} \sim \text{RIDLs} + \text{transcript length} + \text{exonic conservation},$$

in which functional-association outcome indicates the traits defined above; RIDLs indicates the number of RIDL instances in the host gene; transcript length indicates the projected exonic length; and conservation indicates the percentage of exonic lncRNA nucleotides overlapping the union of primate, placental mammal, and vertebrate phastCons elements. We did not find evidence for multicollinearity in any case (variance inflation factors [VIFs] <1.1). We used the "VIF" command from the R package fmsb (Nakazawa 2018).

Subcellular localization analysis

Processed RNA-seq data from human cell fractions were obtained from ENCODE in the form of reads per kilobase per million mapped reads (RPKM) quantified against the GENCODE version 19 (v19) annotation (Djebali et al. 2012; Mas-Ponte et al. 2017). Only transcripts common to both the v21 and v19 annotations were considered. For the following analysis, only one transcript per gene was considered, defined as the one with largest number of exons. Nuclear/cytoplasmic ratio expression for each transcript was defined as (nuclear poly(A)+ RPKM)/(cytoplasmic poly(A)+ RPKM), and only transcripts having nonzero values (at irreproducible discovery rate [IDR] between samples <1) in both were considered. These ratios were $\log_2$-transformed to yield the RCI (Mas-Ponte et al. 2017). For each RIDL type and cell type in turn, the nuclear/cytoplasmic ratio distribution of RIDL-containing to non-RIDL-containing lncRNAs was compared using Wilcoxon test. Only RIDLs having at least three expressed transcripts in at least one cell type were tested. Resulting *P*-values were globally adjusted to FDR using the Benjamini–Hochberg method (Benjamini and Hochberg 1995).

Multiple linear regression and partial correlation analysis

Linear models were created in R using the "lm" function (R Core Team 2015), at the level of lncRNA transcripts with the form:

$$\text{localization} \sim \text{RIDL} + \text{transcript length} + \text{expression}.$$

Localization refers to nuclear/cytoplasmic RCI; RIDL denotes the number of instances of a given RIDL in a transcript; and expression denotes the whole-cell expression level as inferred from RNA-seq in units of RPKM. Equivalent partial correlation analyses were performed using the R pcor.test function from the ppcor package (Spearman correlation) (Kim 2015), correlating RCI with RIDL number while controlling for transcript length and expression. We checked all regression models for multicollinearity by searching for VIFs using the VIF command from the R package fmsb (Nakazawa 2018). In no case did VIF exceed 1.1, thus not raising concern of multicollinearity (>4).

Cell lines and reagents

The human cervical cancer cell line HeLa and human lung cancer cell line A549 were cultured in Dulbecco's Modified Eagle's Medium (Sigma-Aldrich D5671) supplemented with 10% FBS and 1% penicillin/streptomycin at 37°C and 5% CO_2. Anti-GAPDH antibody (Sigma-Aldrich G9545) and anti-histone H3 antibody (Abcam ab24834) were used for western blot analysis.

Gene synthesis and cloning of lncRNAs

The three lncRNA sequences (*RP11-5407*, *LINC00173*, *RP4-806M20.4*) containing wild-type RIDLs and the corresponding mutated versions in which RIDL sequence has been randomized ("mutant") were synthesized commercially (BioCat GmbH). For each gene locus, only one transcript contained the RIDL(s) and was chosen for experimental study. The sequences were cloned into pcDNA 3.1 (+) vector within the NheI and XhoI restriction enzyme sites. The clones were checked by restriction digestion and Sanger sequencing. The sequence of the wild-type and mutant clones are provided in Supplemental File S4.

LncRNA transfection and subcellular fractionation

Wild-type and mutant lncRNA clones for each tested gene were transfected independently in separate wells of a six-well plate. Transfections and subsequent analysis were repeated as biological replicates (four for HeLa, four for A549), defined as transfections performed on different days with different cell passages. Transfections were performed with 2 µg total plasmid DNA in each well using Lipofectamine 2000. Forty-eight hours post-transfection, cells from each well were harvested, pooled, and reseeded into a 10-cm dish and allowed to grow until 100% confluence. Expression of transgenes was check by qRT-PCR using specific primers and found to typically be several-fold greater than endogenous copies (HeLa) or from 0.2-fold to onefold (A549) (Supplemental Fig. S15).

The nuclear and cytoplasmic fractionation was performed as described previously (Suzuki et al. 2010) with minor modifications. In brief, cells from 10-cm dishes were harvested by scraping and washed with 1× ice-cold PBS. For fractionation, a cell pellet was resuspended in 900 µL ice-cold 0.1% NP-40 in PBS and triturated seven times using a p1000 micropipette. Three hundred microliters of the cell lysate was saved as the whole-cell lysate. The remaining 600 µL of the cell lysate was centrifuged for 30 sec on a table top centrifuge, and the supernatant was collected as cytoplasmic fraction. Three hundred microliters from the cytoplasmic supernatant was kept for RNA isolation, and the remaining 300 µL was saved for protein analysis by western blot. The pellet containing the intact nuclei was washed with 1 mL 0.1% NP-40 in PBS. The nuclear pellet was resuspended in 200 µL 1× PBS and subjected to a quick sonication of three pulses with 2-sec on/2-sec off to lyse the nuclei and prepare the "nuclear fraction." One hundred microliters of nuclear fraction was saved for RNA isolation, and the remaining 100 µL was kept for western blot.

RNA isolation and real-time PCR

The RNA from each nuclear and cytoplasmic fraction was isolated using a Quick-RNA MiniPrep Kit (ZYMO Research R1055). The RNAs were subjected to on-column DNase I treatment and clean-up using the manufacturer's protocol. For A549 samples, additional units of DNase were used because of residual signal in –RT samples. The RNA from each fraction was converted to cDNA using GoScript reverse transcriptase (Promega A5001) and random hexamer primers. The expression of each of the individual transcripts was quantified by qRT-PCR (Applied Biosystems 7500 Real-Time) using the indicated primers (Supplemental File S7) and GoTaq qPCR master mix (Promega A6001). To distinguish expression of transfected wild-type genes from endogenous copies, we designed forward primers against a transcribed region of the expression vector backbone. Human *GAPDH* mRNA and *MALAT1* lncRNA were used as cytoplasmic and nuclear markers, respectively. The absence of contaminating plasmid DNA in cDNA was checked for all samples using qPCR (for a representative example, see Supplemental Fig. S16).

Western blotting

The protein concentration of each of the fractions was determined, and equal amounts of protein (50 µg) from whole-cell lysate, cytoplasmic fraction, and nuclear fraction were resolved on 12% Tris-glycine SDS-polyacrylamide gels and transferred onto polyvinylidene fluoride (PVDF) membranes (VWR 1060029). Membranes were blocked with 5% skimmed milk and incubated overnight at 4°C with anti-GAPDH antibody as a cytoplasmic marker and anti-p-histone H3 antibody as a nuclear marker. Membranes were washed with PBS-T (1× PBS with 0.1 % Tween 20) followed by incubation with HRP-conjugated anti-rabbit or anti-mouse secondary antibodies, respectively. The bands were detected using SuperSignal West Pico chemiluminescent substrate (Thermo Fisher Scientific 34077).

Software availability

transposon.profiler is available on GitHub at https://github.com/gold-lab/shared_scripts and in the Supplemental Code.

Acknowledgments

We thank Roderic Guigó (CRG), Marc Friedlaender (SciLifeLab), and Marta Melé (Harvard) for many helpful discussions. Roberta Esposito (DBMR) and Samir Ounzain (CHUV) contributed valuable suggestions regarding experimental design and analysis. Julien Lagarde (CRG) kindly provided help in gene sampling analysis. Carlos Pulido (DBMR) assisted with RNA-seq analysis, and Reza Sodaie (CRG) helped with combinatorial analysis of TEs. We acknowledge Deborah Re (DBMR), Silvia Roesselet (DBMR), and Marianne Zahn (Inselspital) for administrative support. C.N. is supported by grants TIN-2013-41990-R and DPI-2017-84439-R from the Spanish Ministry of Economy, Industry and Competitiveness (MINECO). This research was funded by the NCCR "RNA & Disease" funded by the Swiss National Science Foundation and by the Medical Faculty of the University and University Hospital of Bern.

References

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526:** 68–74. doi:10.1038/nature15393

Bahar Halpern K, Caspi I, Lemze D, Levy M, Landen S, Elinav E, Ulitsky I, Itzkovitz S. 2015. Nuclear retention of mRNA in mammalian tissues. *Cell Rep* **13:** 2653–2662. doi:10.1016/j.celrep.2015.11.036

Bassett AR, Akhtar A, Barlow DP, Bird AP, Brockdorff N, Duboule D, Ephrussi A, Ferguson-Smith AC, Gingeras TR, Haerty W, et al. 2014. Considerations when investigating lncRNA function in vivo. *eLife* **3:** e03058. doi:10.7554/eLife.03058

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* **57:** 289–300. doi:10.2307/2346101

Benoit Bouvrette LP, Cody NAL, Bergalet J, Lefebvre FA, Diot C, Wang X, Blanchette M, Lécuyer E. 2018. CeFra-seq reveals broad asymmetric mRNA and noncoding RNA distribution profiles in *Drosophila* and human cells. *RNA* **24:** 98–113. doi:10.1261/rna.063172.117

Blackwell BJ, Lopez MF, Wang J, Krastins B, Sarracino D, Tollervey JR, Dobke M, Jordan IK, Lunyak VV. 2012. Protein interactions with *piALU* RNA

indicates putative participation of retroRNA in the cell cycle, DNA repair and chromatin assembly. *Mob Genet Elements* **2:** 26–35. doi:10.4161/mge.19032

Bourque G. 2009. Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Curr Opin Genet Dev* **19:** 607–612. doi:10.1016/j.gde.2009.10.013

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew J-L, Ruan Y, Wei C-L, Ng HH, et al. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* **18:** 1752–1762. doi:10.1101/gr.080663.108

Cabili MN, Dunagin MC, McClanahan PD, Biaesch A, Padovan-Merhar O, Regev A, Rinn JL, Raj A. 2015. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol* **16:** 20. doi:10.1186/s13059-015-0586-4

Carlevaro-Fita J, Rahim A, Guigó R, Vardy LA, Johnson R. 2016. Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells. *RNA* **22:** 867–882. doi:10.1261/rna.053561.115

Carlevaro-Fita J, Camaioni AAL, Feuerbach L, Hong C, Mas-Ponte D, Guigo R, Pedersen JS, Johnson R. 2017. Unique genomic features and deeply-conserved functions of long non-coding RNAs in the Cancer LncRNA Census (CLC). bioRxiv doi:10.1101/152769

Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, Pesce E, Ferrer I, Collavin L, Santoro C, et al. 2012. Long non-coding antisense RNA controls *Uchl1* translation through an embedded SINEB2 repeat. *Nature* **491:** 454–457. doi:10.1038/nature11508

Chen LL. 2016. Linking long noncoding RNA localization and function. *Trends Biochem Sci* **41:** 761–772. doi:10.1016/j.tibs.2016.07.003

Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, Zhang Q, Yan G, Cui Q. 2013. LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res* **41:** D983–D986. doi:10.1093/nar/gks1099

Chillón I, Pyle AM. 2016. Inverted repeat *Alu* elements in the human lincRNA-p21 adopt a conserved secondary structure that regulates RNA function. *Nucleic Acids Res* **44:** 9462–9471. doi:10.1093/nar/gkw599

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* **18:** 71–86. doi:10.1038/nrg.2016.139

Clark MB, Johnston RL, Inostroza-Ponta M, Fox AH, Fortini E, Moscato P, Dinger ME, Mattick JS. 2012. Genome-wide analysis of long noncoding RNA stability. *Genome Res* **22:** 885–898. doi:10.1101/gr.131037.111

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10:** 691–703. doi:10.1038/nrg2640

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22:** 1775–1789. doi:10.1101/gr.132159.111

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* **489:** 101–108. doi:10.1038/nature11233

Elisaphenko EA, Kolesnikov NN, Shevchenko AI, Rogozin IB, Nesterova TB, Brockdorff N, Zakian SM. 2008. A dual origin of the *Xist* gene from a protein-coding gene and a set of transposable elements. *PLoS One* **3:** e2521. doi:10.1371/journal.pone.0002521

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **41:** 563–571. doi:10.1038/ng.368

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9:** 397–405. doi:10.1038/nrg2337

Gong C, Maquat LE. 2011. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements. *Nature* **470:** 284–288. doi:10.1038/nature09701

Guttman M, Rinn JL. 2012. Modular regulatory principles of large non-coding RNAs. *Nature* **482:** 339–346. doi:10.1038/nature10887

Hacisuleyman E, Goff LA, Trapnell C, Williams A, Henao-Mejia J, Sun L, McClanahan P, Hendrickson DG, Sauvageau M, Kelley DR, et al. 2014. Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat Struct Mol Biol* **21:** 198–206. doi:10.1038/nsmb.2764

Hacisuleyman E, Shukla CJ, Weiner CL, Rinn JL. 2016. Function and evolution of local repeats in the *Firre* locus. *Nat Commun* **7:** 11021. doi:10.1038/ncomms11021

Haerty W, Ponting CP. 2013. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biol* **14:** R49. doi:10.1186/gb-2013-14-5-r49

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22:** 1760–1774. doi:10.1101/gr.135350.111

Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. 2015. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep* **11:** 1110–1122. doi:10.1016/j.celrep.2015.04.023

Hezroni H, Ben-Tov Perry R, Meir Z, Housman G, Lubelsky Y, Ulitsky I. 2017. A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. *Genome Biol* **18:** 162. doi:10.1186/s13059-017-1293-0

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci* **106:** 9362–9367. doi:10.1073/pnas.0903103106

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al. 2006. The UCSC genome browser database: update 2006. *Nucleic Acids Res* **34:** D590–D598. doi:10.1093/nar/gkj144

Holdt LM, Hoffmann S, Sass K, Langenberger D, Scholz M, Krohn K, Finstermeier K, Stahringer A, Wilfert W, Beutner F, et al. 2013. Alu elements in *ANRIL* non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through *trans*-regulation of gene networks. *PLoS Genet* **9:** e1003588. doi:10.1371/journal.pgen.1003588

Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res* **19:** 1419–1428. doi:10.1101/gr.091678.109

Hu B, Yang Y-CT, Huang Y, Zhu Y, Lu ZJ. 2017. POSTAR: a platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res* **45:** D104–D114. doi:10.1093/nar/gkw888

Huda A, Bowen NJ, Conley AB, Jordan IK. 2011. Epigenetic regulation of transposable element derived human gene promoters. *Gene* **475:** 39–48. doi:10.1016/j.gene.2010.12.010

Jjingo D, Conley AB, Wang J, Mariño-Ramírez L, Lunyak VV, Jordan IK. 2014. Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. *Mob DNA* **5:** 14. doi:10.1186/1759-8753-5-14

Johnson R, Guigó R. 2014. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA* **20:** 959–976. doi:10.1261/rna.044560.114

Johnson R, Gamblin RJ, Ooi L, Bruce AW, Donaldson IJ, Westhead DR, Wood IC, Jackson RM, Buckley NJ. 2006. Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. *Nucleic Acids Res* **34:** 3862–3877. doi:10.1093/nar/gkl525

Jurka J, Zietkiewicz E, Labuda D. 1995. Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. *Nucleic Acids Res* **23:** 170–175. doi:10.1093/nar/23.1.170

Kapusta A, Feschotte C. 2014. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends Genet* **30:** 439–452. doi:10.1016/j.tig.2014.08.004

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C. 2013. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet* **9:** e1003470. doi:10.1371/journal.pgen.1003470

Kelley D, Rinn J. 2012. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol* **13:** R107. doi:10.1186/gb-2012-13-11-r107

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. doi:10.1186/s13059-014-0537-5

Kim S. 2015. ppcor: an R package for a fast calculation to semi-partial correlation coefficients. *Commun Stat Appl Methods* **22:** 665–674. doi:10.5351/CSAM.2015.22.6.665

Konkel MK, Walker JA, Batzer MA. 2010. LINEs and SINEs of primate evolution. *Evol Anthropol* **19:** 236–249. doi:10.1002/evan.20283

Lagarde J, Uszczynska-Ratajczak B, Carbonell S, Pérez-Lluch S, Abad A, Davis C, Gingeras TR, Frankish A, Harrow J, Guigo R, et al. 2017. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat Genet* **49:** 1731–1740. doi:10.1038/ng.3988

Lev-Maor G, Sorek R, Shomron N, Ast G. 2003. The birth of an alternatively spliced exon: 3′ splice-site selection in *Alu* exons. *Science* **300:** 1288–1291. doi:10.1126/science.1082588

Li J-H, Liu S, Zhou H, Qu L-H, Yang J-H. 2014. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* **42**(Database issue)**:** D92–D97. doi:10.1093/nar/gkt1248

Lubelsky Y, Ulitsky I. 2018. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature* **555:** 107–111. doi:10.1038/nature25757

Marín-Béjar O, Huarte M. 2015. RNA pulldown protocol for in vitro detection and identification of RNA-associated proteins. *Methods Mol Biol* **1206:** 87–95. doi:10.1007/978-1-4939-1369-5_8

Marín-Béjar O, Mas AM, González J, Martinez D, Athie A, Morales X, Galduroz M, Raimondi I, Grossi E, Guo S, et al. 2017. The human lncRNA LINC-PINT inhibits tumor cell invasion through a highly conserved sequence element. *Genome Biol* **18:** 202. doi:10.1186/s13059-017-1331-y

Martin KC, Ephrussi A. 2009. mRNA localization: gene expression in the spatial dimension. *Cell* **136:** 719–730. doi:10.1016/j.cell.2009.01.044

Mas-Ponte D, Carlevaro-Fita J, Palumbo E, Hermoso Pulido T, Guigo R, Johnson R. 2017. LncATLAS database for subcellular localization of long noncoding RNAs. *RNA* **23:** 1080–1087. doi:10.1261/rna. 060814. 117

Mercer TR, Mattick JS. 2013. Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol* **20:** 300–307. doi:10.1038/nsmb.2480

Mukherjee N, Calviello L, Hirsekorn A, de Pretis S, Pelizzola M, Ohler U. 2017. Integrative classification of human coding and noncoding genes through RNA metabolism profiles. *Nat Struct Mol Biol* **24:** 86–96. doi:10.1038/nsmb.3325

Nakazawa M. 2018. fmsb: Functions for Medical Statistics Book with some Demographic Data. Available at: https://cran.r-project.org/web/packages/fmsb/fmsb.pdf. Package version 0.6.3.

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505:** 635–640. doi:10.1038/nature12943

Ning S, Zhang J, Wang P, Zhi H, Wang J, Liu Y, Gao Y, Guo M, Yue M, Wang L, et al. 2016. Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res* **44:** D980–D985. doi:10.1093/nar/gkv1094

Nissan A, Stojadinovic A, Mitrani-Rosenbaum S, Halle D, Grinbaum R, Roistacher M, Bochem A, Dayanc BE, Ritter G, Gomceli I, et al. 2012. Colon cancer associated transcript-1: a novel RNA expressed in malignant and pre-malignant human tissues. *Int J Cancer* **130:** 1598–1606. doi:10.1002/ijc.26170

Pegueroles C, Gabaldón T. 2016. Secondary structure impacts patterns of selection in human lncRNAs. *BMC Biol* **14:** 60. doi:10.1186/s12915-016-0283-0

Perepelitsa-Belancio V, Deininger P. 2003. RNA truncation by premature polyadenylation attenuates human mobile element activity. *Nat Genet* **35:** 363–366. doi:10.1038/ng1269

Quek XC, Thomson DW, Maag JL V, Bartonicek N, Signal B, Clark MB, Gloss BS, Dinger ME. 2015. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res* **43:** D168–D173. doi:10.1093/nar/gku988

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

R Core Team. 2015. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Rackham O, Shearwood A-MJ, Mercer TR, Davies SMK, Mattick JS, Filipovska A. 2011. Long noncoding RNAs are generated from the mitochondrial genome and regulated by nuclear-encoded proteins. *RNA* **17:** 2085–2093. doi:10.1261/rna.029405.111

Roberts JT, Cardin SE, Borchert GM. 2014. Burgeoning evidence indicates that microRNAs were initially formed from transposable element sequences. *Mob Genet Elements* **4:** e29255. doi:10.4161/mge.29255

Schmitt AM, Chang HY, Abdelmohsen K, Panda A, Kang MJ, Xu J, Selimyan R, Yoon JH, Martindale JL, De S, et al. 2016. Long noncoding RNAs in cancer pathways. *Cancer Cell* **29:** 452–463. doi:10.1016/j.ccell.2016.03.010

Seemann SE, Mirza AH, Hansen C, Bang-Berthelsen CH, Garde C, Christensen-Dalsgaard M, Torarinsson E, Yao Z, Workman CT, Pociot F, et al. 2017. The identification and functional annotation of RNA structures conserved in vertebrates. *Genome Res* **27:** 1371–1383. doi:10.1101/gr.208652.116

Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, Ast G. 2007. Comparative analysis of transposed element insertion within human and mouse genomes reveals *Alu*'s unique role in shaping the human transcriptome. *Genome Biol* **8:** R127. doi:10.1186/gb-2007-8-6-r127

Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29:** 308–311. doi:10.1093/nar/29.1.308

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034–1050. doi:10.1101/gr.3715005

Smit AFA, Hubley R, Green P. 2013–2015. RepeatMasker Open-4.0. http://www.repeatmasker.org.

Smith MA, Gesell T, Stadler PF, Mattick JS. 2013. Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res* **41:** 8220–8236. doi:10.1093/nar/gkt596

Su M, Han D, Boyd-Kirkup J, Yu X, Han J-DJ. 2014. Evolution of Alu elements toward enhancers. *Cell Rep* **7:** 376–385. doi:10.1016/j.celrep.2014.03.011

Suzuki K, Bose P, Leong-Quong RYY, Fujita DJ, Riabowol K. 2010. REAP: a two-minute cell fractionation method. *BMC Res Notes* **3:** 294. doi:10.1186/1756-0500-3-294

Tan JY, Sirey T, Honti F, Graham B, Piovesan A, Merkenschlager M, Webber C, Ponting CP, Marques AC. 2015. Extensive microRNA-mediated crosstalk between lncRNAs and mRNAs in mouse embryonic stem cells. *Genome Res* **25:** 655–666. doi:10.1101/gr.181974.114

Tan JY, Smith AAT, Ferreira da Silva M, Matthey-Doret C, Rueedi R, Sönmez R, Ding D, Kutalik Z, Bergmann S, Marques AC. 2017. *cis*-Acting complex-trait-associated lincRNA expression correlates with modulation of chromosomal architecture. *Cell Rep* **18:** 2280–2288. doi:10.1016/j.celrep.2017.02.009

Ulitsky I, Bartel DP. 2013. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154:** 26–46. doi:10.1016/j.cell.2013.06.020

Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K, et al. 2016. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13:** 508–514. doi:10.1038/nmeth.3810

Washietl S, Kellis M, Garber M. 2014. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res* **24:** 616–628. doi:10.1101/gr.165035.113

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42:** D1001–D1006. doi:10.1093/nar/gkt1229

Zhang B, Gunawardane L, Niazi F, Jahanbani F, Chen X, Valadkhan S. 2014a. A novel RNA motif mediates the strict nuclear localization of a long noncoding RNA. *Mol Cell Biol* **34:** 2318–2329. doi:10.1128/MCB.01673-13

Zhang K, Shi Z-M, Chang Y-N, Hu Z-M, Qi H-X, Hong W. 2014b. The ways of action of long non-coding RNAs in cytoplasm and nucleus. *Gene* **547:** 1–9. doi:10.1016/j.gene.2014.06.043

Received January 23, 2018; accepted in revised form December 18, 2018.

Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs

Joana Carlevaro-Fita, Taisia Polidori, Monalisa Das, et al.

Genome Res. 2019 29: 208-222 originally published online December 26, 2018
Access the most recent version at doi:10.1101/gr.229922.117