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

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

Genome Res. published online December 26, 2018
Access the most recent version at doi:10.1101/gr.229922.117

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]

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, Bern, Switzerland

4. Department of Computer Science and Artificial Intelligence, University of Granada, Spain

5. Equal contribution

*Correspondence: rory.johnson@dbmr.unibe.ch

Keywords: Transposable element; subcellular localization; long noncoding RNA; lncRNA; evolution; exaptation.

Abstract

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 neofunctionalised fragments of transposable elements (TEs), otherwise known as RIDLs (Repeat Insertion Domains of Long Noncoding RNA), 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 localisation. Using global localisation data from ten 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 localisation of lncRNAs.

Introduction

The human genome contains many thousands of long noncoding RNAs (lncRNAs), of which at least a fraction are 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 localisation (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 towards 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 neofunctionalisation of sequence fragments – a process known as *exaptation* (Bourque 2009)(Feschotte 2008). This process has contributed to the evolution of diverse features in genomic DNA, including transcriptional regulatory motifs (Bourque et al. 2008; Johnson et al. 2006), microRNAs (Roberts et al. 2014), gene promoters (Faulkner et al.; 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 "RIDLs" – *Repeat Insertion Domains of Long noncoding RNAs* (Johnson and Guigó 2014). As RNA, 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 (Johnson and Guigó 2014; Gong and Maquat 2011; Holdt et al. 2013). 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, reviewed in (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 *UchlAS* (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 hybridisation between RNA and DNA *Alu* elements recruits *ANRIL* to target genes (Holdt et al. 2013); a SINEB2 repeat in *Uchl1AS* 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)(Schmitt et al. 2016)(Hezroni et al. 2015). However, there has been no attempt to enrich these maps for RIDLs with evidence of selected functions in the context of mature lncRNA molecules.

Subcellular localisation, and the domains controlling it, are crucial determinants of lncRNA functions (reviewed in (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 (Mukherjee et al. 2017) (Carlevaro-Fita et al. 2016; Derrien et al. 2012)(Cabili et al. 2015; Mas-Ponte et al. 2017)(Benoit Bouvrette et al. 2018). If lessons learned from mRNA are also valid for lncRNAs, then short sequence motifs recognised by RNA binding proteins (RBPs) will be an important localisation-regulatory mechanism (Martin and Ephrussi 2009). This was recently demonstrated for the *BORG* lncRNA, where 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. 2016a) (Hacisuleyman et al. 2014). 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 localisation at a global scale.

The aim of the present study is to create a human transcriptome-wide catalogue 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 transposable elements, promote the nuclear enrichment of their host transcripts.

Results

The objective of this study is to create a map of repeat insertion domains of long noncoding RNAs (RIDLs) and link them to lncRNA functions. We hypothesise that RIDLs could confer such functions through interactions with DNA, RNA or protein molecules (Johnson and Guigó 2014) (Figure 1A).

Any attempt to map RIDLs must deal with two challenges. First, that they will likely represent a small minority amongst many phenotypically-neutral "passenger" transposable elements (TEs) in lncRNA exons (Figure 1B). Second, many TE instances may be under evolutionary selection, but for functions executed at the *DNA level* (eg 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 (Figure 1B). To estimate background, we utilise intronic TEs, since 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", due to high rates of false positive predictions, and all downstream analyses should be interpreted accordingly.

A map of exonic transposable elements in GENCODE v21 lncRNAs

Our first aim was to create a comprehensive map of transposable elements (TEs) within the exons of GENCODE v21 human lncRNAs (Figure 2A). Altogether 5,520,018 distinct TE insertions were intersected with 48684 exons from 26414 transcripts of 15877 GENCODE v21 lncRNA genes, resulting in 46474 exonic TE insertions in lncRNA (Figure 1B). 13121 lncRNA genes (82.6%) carry at least one exonic TE fragment in one or more of their mature transcripts.

We also created a reference dataset with 31,004 GENCODE lncRNA introns, resulting in 562,640 intron-overlapping TE fragments (Figure 2A). 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 to 43.4% of intronic nucleotides (Figure 2B), similar to previous studies (Kapusta et al. 2013). This may reflect generalised 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 transposable elements 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 classified inserted TEs by their contribution to lncRNA gene structure (Figure 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 *Alu*, L1 and ERVL-MaLR classes (Figure 2E). 7497 (28.4%) lncRNA transcripts are terminated by a TE, most commonly by 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%) (Figure 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 transposable element 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, AFA, Hubley, R & Green). We utilise three distinct sources of evidence for selection on TEs: exonic enrichment, strand bias and evolutionary conservation (Figure 1B).

We first asked whether particular TE types are enriched in lncRNA exons, compared to intronic sequence (Kelley and Rinn 2012). Thus, we calculated the ratio of exonic / intronic sequence coverage by TEs (Figure 3A). We found enrichment >2-fold 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 using whole genome, rather than introns, as background (Kelley and Rinn 2012). 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 LINE1 elements are relatively depleted in lncRNA exons (Figure 3A). It is possible that this reflects selection against

disruption to normal gene expression, where numerous weak polyadenylation signals lead to premature transcription termination when the LINE1 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 LINE1 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 artefactual 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 randomisation, with significance defined at $P<0.001$, corresponding to a false discovery rate (FDR) below 5% (similar cutoffs apply to subsequent analyses, more details may be found in Materials and Methods) (Figure 3B). In lncRNA exons, a number of TE types are enriched in either sense or antisense, dominated by LINE1 family members, possibly for the reasons mentioned above. Other significantly enriched TE types include LTR78, MLT1B, and MIRc (Figure 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 Figure S1). This is especially true for LINE1 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. First, 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; second, the more recent “Evolutionarily Conserved Structures” (ECS) set (Smith et al. 2013). Importantly, the phastCons regions are defined based on sequence conservation alone, while 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 randomisation (see inset in Figure 3C). This pipeline was applied independently to the phastCons (placental mammal shown in Figure 3C, primate and vertebrate in Supplemental Figure S1B,C) and ECS (Supplemental Figure S1D) data. The majority of TE types do not exhibit signatures of conservation (grey

points). However, for each conservation type, the method detects significant conservation for a minority of TE types (Figure 3C). This enrichment disappeared when phastCons elements were positionally randomised (Supplemental Figure S2A). It is unlikely that overlap with protein-coding loci biases the results, since equivalent analyses using intergenic lncRNAs yielded similar candidate RIDLs (Supplemental Figure 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 Figure 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) (coloured orange in Figure 3C and Supplemental Figure S1). The cause of this depletion is unclear, although one explanation is enrichment of conservation in intronic TEs due to RNA-independent regulatory roles as observed previously (Su et al. 2014).

All the selection evidence is summarised 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 MIRs 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 (Figure 4A).

After removing redundancy, the final RIDL annotation consists of 5374 elements, located within 3566 distinct lncRNA genes (Figure 3D). These represent 12% (5374/46474) of all exonic TE fragments. The most predominant TE families are MIR and L2 repeats, representing 2329 and 1143 RIDLs (Figure 4B). The majority of both are defined based on evolutionary evidence (Figure 4B, Supplemental Figure

S3). In contrast, RIDLs composed by ERV1, low complexity, satellite and simple repeats families are more frequently identified due to exonic enrichment (Figure 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. 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 Figure S4).

We also examined the evolutionary history of RIDLs. Using 6-mammal alignments, their depth of evolutionary conservation could be inferred (Supplemental Figure S5). 12% of instances appear to be Great Ape-specific, with no orthologous sequence beyond chimpanzee. 47% are primate-specific, while the remaining 40% are identified in at least one non-primate mammal. The wide timeframe for appearance of RIDLs is consistent with the wide diversity of TE types, from ancient MIR elements to presently-active LINE1 (Jurka et al. 1995; Smith et al. 2013; Konkel et al. 2010).

Instances of genomic TE insertions typically represent a fragment of the full consensus sequence. We hypothesised 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 (Figure 4C and 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 where RIDL copies in 41 lncRNAs show a strong preference inclusion of the 3' end, in contrast to the general 5' preference observed in introns (Figure 4C). This suggests a possible functional relevance for the 1000-1500 nt region of the SST1 consensus.

To assess whether RIDLs experience purifying evolutionary selection in modern humans, we analysed the derived allele frequency (DAF) spectrum of their overlapping SNPs (Supplemental Figure 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 to other TEs in exons (green bars) or introns (turquoise bars) of the same lncRNAs, and indeed compared to non-RIDL exonic nucleotides (black bars). These differences fail to reach statistical significance, possibly due to small sample sizes. Overall these data are consistent with RIDLs experiencing an elevated rate of purifying evolutionary selection in modern humans compared to nearby neutral sequence, although larger datasets 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 amongst lncRNAs, or else non-randomly 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 Figure 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 (Figure 4E, Supplemental Figure S8). We observed that RIDL-lncRNAs are (1) over-represented in the reference database for functional lncRNAs, lncRNAdb (Quek et al. 2015), (2) are enriched in associations with cancer and other diseases, and (3) enriched in their exons for trait/disease-associated SNPs. In order 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 (Figure 4F). However we did not observed any difference in mean or maximum expression of RIDL-lncRNAs to length matched controls across ten tissues of the Human Body Map RNA-seq dataset (Supplemental Figure S9).

In addition to *CCAT1* (Figure 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 localisation of host transcript

The location of a lncRNA within the cell is of key importance to its molecular function (Cabili et al. 2015; Derrien et al. 2012)(Mas-Ponte et al. 2017), therefore we next investigated whether RIDLs might regulate lncRNA localisation (Zhang et al. 2014a; Hacisuleyman et al. 2016b)(Chillón and Pyle

2016) (Figure 5A). Using subcellular RNA-seq data based on 10 ENCODE cell lines (Djebali et al. 2012), we calculated the relative nuclear/cytoplasmic localisation in $\log_2$ units, or "Relative Concentration Index" (RCI) (Mas-Ponte et al. 2017). Using this dataset, we tested each of the 99 RIDL types for association with localisation 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 (Figure 5B). For example, 44 lncRNAs carrying L2b RIDLs have 6.9-fold higher relative nuclear/cytoplasmic ratio in IMR-90 cells, and this tendency is observed in six different cell types (Figure 5B,C).

The degree of nuclear localisation increases in lncRNAs as a function of the number of RIDLs (L1PA16, L2b, MIRb and MIRc) they carry (Figure 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 2 and 2.3-fold higher relative expression in the cytoplasm of these cells (Supplemental Figure S10).

We were curious whether this relationship with localisation 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 Figure S11). However, the strength of the effect was consistently lower than for RIDLs (Supplemental Figure 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 localisation.

We were concerned that two un-modelled 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 localisation with explanatory variables of RIDL number, transcript length and whole-cell expression (Figure 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 localisation independent of RIDLs (Supplemental Figure S13A,B). We drew similar conclusions from equivalent analyses using partial correlation analysis (Supplemental Figure S13C). In summary, observed RIDLs correlate with lncRNA localisation 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 datasets 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 localisation

LncRNA gene loci are often composed of multiple, differentially-spliced transcript isoforms that partially differ in their mature sequence. We reasoned that differential inclusion of RIDL-containing exons should give rise to differences in localisation amongst 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 (Figure 6, left panel).

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 localisation between their RIDL and non-RIDL isoforms (Figure 6, right panel). 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 1-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 localisation

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

We formulated an assay to compare the localisation of transfected lncRNAs carrying wild-type RIDLs, and mutated versions where the RIDL sequence was randomised without altering sequence composition (“Mutant”) (Figure 7D, full sequences available in Supplemental File S4). Wild-type and Mutant lncRNAs were transfected into cultured cells and their localisation 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 to their endogenous transcripts (Supplemental Figure S15). Fractionation purity was verified by Western blotting (Figure 7E) and qRT-PCR (Figure 7F), and stringent DNase-treatment ensured that plasmid DNA made negligible contributions to our results (Supplemental Figure S16).

With this setup, we compared the nuclear/cytoplasmic localisation of lncRNAs with and without exonic RIDL sequences (Figure 7F). We observed a potent and consistent impact of RIDLs on nuclear/cytoplasmic localisation in HeLa cells: for all three candidates, loss of RIDL sequence resulted in relocalisation of the host transcript from nucleus to cytoplasm (Figure 7F, upper panel). We repeated these experiments in another cell line, A549, and observed similar, albeit less pronounced, effects (Figure 7F, lower panel). This difference may be due to the less nuclear localisation of the endogenous transcripts in A549 (Supplemental Figure S17). To summarise, 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 how such functions are encoded in primary RNA sequences, lag far behind. Two recent conceptual developments offer hope for resolving the sequence-function code of lncRNAs: First, the idea that the subcellular localisation of lncRNAs is a readily quantifiable characteristic that holds important clues to function; Second, that the abundant transposable element (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 localisation data revealed a handful of TEs, most notably LINE2b, MIRb and MIRc, which positively and significantly correlate with the degree of nuclear/cytoplasmic localisation 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 four-fold increase in nuclear/cytoplasmic localisation. There are two principal explanations for this phenomenon. First, an "active" process whereby TEs are recognised by a cellular transport pathway, as demonstrated for *Alus* by Lubelsky and Ulitsky (Lubelsky and Ulitsky 2018). Second, a "passive" process where TEs destabilise 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", transposable elements are co-opted by natural selection to form "Repeat Insertion Domains of LncRNA", 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 (reviewed in (Johnson and Guigó 2014)). In the context of localisation, RIDLs could mediate nuclear retention through hybridisation 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, HuR, KHDRBS1, TARDBP), however we could not find evidence that they contribute to RIDL-lncRNA localisation. Identification of any proteins that mediate RIDLs' localisation activity may be achieved in future through pulldown approaches (Marín-Béjar and Huarte 2015).

The localisation RIDLs discovered – MIR and LINE2 - are both ancient and contemporaneous, being active prior to 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 (Jjingo et al. 2014)(Johnson et al. 2006). 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 (Hezroni et al. 2015)(Hezroni et al. 2017)(Necsulea et al. 2014)(Washietl et al. 2014). This would mean that MIR and LINE2 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 localisation activity is a more modern feature of L2b/MIR RIDLs, which is unrelated to their original roles.

Our approach for identifying localisation-regulating RIDLs has advantages over previous studies (Lubelsky and Ulitsky 2018; Hacisuleyman et al. 2016c) 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 localisation, but cannot be detected using the signals of selection that we have employed. 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 (i) in a focussed re-analysis (Supplemental Figure S11) the effect was observed in multiple cells, and (ii) experimental validation confirmed it in two independent cell types (Figure 7F).

This is further supported by the recent study of Lubelsky and Ulitsky, who performed an experimental screen for localisation motifs in 37 nuclear-enriched lncRNAs, and identified *Alu*Sx as a nuclear-localisation element (Lubelsky and Ulitsky 2018). These 37 lncRNAs are enriched for RIDLs (62% of Lubelsky lncRNAs contain at least one RIDL, compared to 22% for other Gencode v21 lncRNAs, $P=4 \times 10^{-6}$, Fisher's exact test), as well as for the three localisation RIDLs identified here (L2b, MIRb, MIRc: 32% vs 9%, $P=3 \times 10^{-4}$) (Supplemental Figure S18A). Although our bioinformatic screen did not identify *Alu*Sx, a naive unfiltered re-analysis of our data supports Lubelsky's experimental finding that *Alu*Sx-carrying lncRNAs tend to be more nuclear across multiple cell types (Supplemental Figure S18B). Together, these considerations open the possibility that other localisation-controlling TE types may await discovery.

More generally, the RIDL predictions showed rather low concordance between the various selection evidence used (Figure 3D). This likely reflects a number of factors: young evolutionary age of some of the most common TEs, generally low statistical power due to large background of neutral TEs and multiple hypothesis testing, and false positives due to 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 (Figure 3D).

This work marks a step in the ongoing efforts to map the domains of lncRNAs. Previous studies have utilised a variety of approaches, from integrating experimental protein-binding data (Van Nostrand et al. 2016)(Hu et al. 2017)(Li et al. 2014), 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 (Kapusta et al. 2013)(Kelley and Rinn 2012)(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 a substantial false positive (and likely also false negative) rates that will have to be improved in 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 to mRNAs (Clark et al. 2012)(Ulitsky and Bartel 2013)(Derrien et al. 2012)(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 localisation (Chillón and Pyle 2016)(Zhang et al. 2014a). 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 localisation in lncRNAs.

Materials and Methods

All operations were carried out 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 31st 2014 (Smit, AFA, Hubley, R & Green), and GENCODE version 21 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, 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 6 categories: TSS (transcription start site), 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, where an entire exon lies within the TE; TTS (transcription termination site), overlapping the last nucleotide of the last exon. In every case, the TEs are separated by strand relative to the host gene: + where both gene and TE are annotated on the same strand, otherwise -. The result is the "Exonic TE Annotation" (Supplemental File S5).

RIDL identification

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 employed, 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 >2-fold 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 non-splice junction crossing TE instances ("noSJ"). This additional filter was employed to guard against false positive enrichments for TEs known to provide splice sites (Sela et al. 2007; Lev-Maor et al. 2003). 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 re-positioned in the genome using *BEDTools shuffle*, and the intersection with the entire *RepeatMasker* annotation was re-calculated. 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 LINE1 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 artefact 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), while the ECS conserved regions from obtained from Supplemental Data of Smith et al (Smith et al. 2013) (see Supplemental File S6 for summary). 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 to introns:

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

Where C is conserved TE nucleotides, N is non-conserved 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 normalises for different lengths of exons and introns (Supplemental Figure 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 where the true REIC was greater or less than every one of 1000 randomised 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 Figure S2D). We also tested an alternative approach for estimating significance whereby conserved elements were held constant, and TEs were positionally randomised. While there was a significant overlap in identified candidate RIDLs, this method displayed a bias towards longer TEs (Supplemental Figure S2D), and therefore was not employed further.

We chose to randomise conserved elements, rather than TEs because the former are enriched in lncRNA exons (Pegueroles and Gabaldón 2016). Thus, using randomised TEs to estimate background REIC would lead to overestimation of exonic TE conservation, and hence underestimate 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.

False discovery rates (FDR) 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, where 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 “fdr” setting (Rackham et al. 2011; R Core Team). Accordingly, 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; ECS: 0.034. This analysis is conservative, since 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. This data is shown in Supplemental Figure S4.

RIDL orthology analysis

In order to assess evolutionary history of RIDLs, we used chained alignments of human to chimp (hg19ToPanTro4), macaque (hg19ToRheMac3), mouse (hg19ToMm10), rat (hg19ToRn5), and cow

(hg19ToBosTau7). Due to 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).

Derived allele frequency (DAF) analysis

We used allele frequencies from African population provided by the 1000 Genomes Project (The 1000 Genomes Project Consortium, 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 analysed. Ancestral repeats (AR) were defined as human repeats (excluding simple repeats) intersecting at least 1 nucleotide of mouse repeats defined by LiftOver, and falling within 5kb of but not overlapping RIDL-containing genes.

Comparing RIDL-carrying lncRNAs versus other lncRNAs

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

A) Functionally-characterised lncRNAs: lncRNAs from GENCODE v21 that are present in lncRNAdb (Quek et al. 2015).
B) Disease-associated genes: 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. bioRxiv doi: 10.1101/152769)
C) GWAS SNPs: We collected SNPs from the NHGRI-EBI Catalog of published genome-wide association studies (Welter et al. 2014; Hindorff et al. 2009) (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 Figure S8). We performed sampling using *matchDistribution* script: https://github.com/julienlag/matchDistribution. In order to simultaneously control for both conservation and length, we performed multiple logistic regression analysis using glm R function (R Core Team), with the following structure:

Functional-association outcome ~ RIDLs + transcript length + exonic conservation

Where functional-association outcome indicates A, B and C traits defined above; RIDLs indicates the number of RIDL instances in the host gene; transcript length indicates the projected exonic length;

conservation indicates the percent 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 (VIF) <1.1). We used the 'VIF' command from the R package 'fmsb' (Nakazawa 2018).

Subcellular localisation analysis

Processed RNA-seq data from human cell fractions were obtained from ENCODE in the form of RPKM (reads per kilobase per million mapped reads) quantified against the GENCODE 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 non-zero values (at Irreproducible Discovery Rate (IDR) between samples < 1) in both were considered. These ratios were $\log_2$-transformed, to yield the Relative Concentration Index (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 False Discovery Rate 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) , at the level of lncRNA transcripts with the form:

localisation ~ RIDL + transcript length + expression

Localisation refers to nuclear/cytoplasmic RCI; RIDL denotes the number of instances of a given RIDL in a transcript; 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 variance inflation factors (VIF) using the 'VIF' command from the R package 'fmsb' (Nakazawa 2018). In no case did VIF exceed 1.1, below values raising concern of multicollinearity (>4).

611 Cell lines and reagents

612 Human cervical cancer cell line HeLa and human lung cancer cell line A549 were cultured in
613 Dulbecco's Modified Eagle's medium (Sigma-Aldrich, # D5671) supplemented with 10% FBS and 1%
614 penicillin streptomycin at 37ºC and 5 % CO_2. Anti-GAPDH antibody (Sigma-Aldrich, # G9545) and anti-
615 histone H3 antibody (Abcam, # ab24834) were used for Western blot analysis.

616

617 Gene synthesis and cloning of lncRNAs

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

625

626 LncRNA transfection and sub-cellular fractionation

627 Wild-type and mutant lncRNA clones for each tested gene were transfected independently in
628 separate wells of a 6-well plate. Transfections and subsequent analysis were repeated as biological
629 replicates (four for HeLa, four for A549), defined as transfections performed on different days with
630 different cell passages. Transfections were carried out with 2 μg of total plasmid DNA in each well using
631 Lipofectamine 2000. 48 h post-transfection, cells from each well were harvested, pooled and re-seeded
632 into a 10 cm dish and allowed to grow till 100% confluence. Expression of transgenes was check by qRT-
633 PCR using specific primers, and found to typically be several-fold greater than endogenous copies (HeLa)
634 or from 0.2- to 1-fold (A549) (Supplemental Figure S15).

635 The nuclear and cytoplasmic fractionation was carried out as described previously (Suzuki et al.
636 2010) with minor modifications. In brief, cells from 10 cm dishes were harvested by scraping and washed
637 with 1x ice-cold PBS. For fractionation, cell pellet was re-suspended in 900 μl of ice-cold 0.1% NP-40 in
638 PBS and triturated 7 times using a p1000 micropipette. 300 μl of the cell lysate was saved as the whole
639 cell lysate. The remaining 600 μl of the cell lysate was centrifuged for 30 sec on a table top centrifuge and
640 the supernatant was collected as "cytoplasmic fraction". 300 μl from the cytoplasmic supernatant was
641 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 of 0.1% NP-40 in PBS. The nuclear pellet was re-suspended in 200 µl 1X PBS and subjected to a quick sonication of 3 pulses with 2 sec ON-2 sec OFF to lyse the nuclei and prepare the "nuclear fraction". 100 µl 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 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 employed, due to 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 indicated primers (Supplemental File S7) and GoTaq qPCR master mix (Promega, # A6001). In order 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 (see Supplemental Figure S16 for a representative example).

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 4ºC with anti-GAPDH antibody as a cytoplasmic marker and anti p-histone H3 antibody as nuclear marker. Membranes were washed with PBS-T (1X 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

672 "*transposon.profiler*", is available on Github at https://github.com/gold-lab/shared_scripts and in
673 Supplemental Code.

Acknowledgements

We wish to thank Roderic Guigó (CRG), Marc Friedlaender (SciLife Lab) 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. CN 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.

685 **References**

686

687 Bahar Halpern K, Caspi I, Lemze D, Levy M, Landen S, Elinav E, Ulitsky I, Itzkovitz S. 2015. Nuclear
688 Retention of mRNA in Mammalian Tissues. *Cell Rep* **13**: 2653–2662.
689 Bassett AR, Akhtar A, Barlow DP, Bird AP, Brockdorff N, Duboule D, Ephrussi A, Ferguson-Smith AC,
690 Gingeras TR, Haerty W, et al. 2014. Considerations when investigating lncRNA function in vivo.
691 *Elife* **3**: e03058.
692 Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful
693 Approach to Multiple Testing. *J R Stat Soc Ser B* **57**: 289–300.
694 Benoit Bouvrette LP, Cody NAL, Bergalet J, Lefebvre FA, Diot C, Wang X, Blanchette M, Lécuyer E.
695 2018. CeFra-seq reveals broad asymmetric mRNA and noncoding RNA distribution profiles in
696 *Drosophila* and human cells. *RNA* **24**: 98–113.
697 Blackwell BJ, Lopez MF, Wang J, Krastins B, Sarracino D, Tollervey JR, Dobke M, Jordan IK, Lunyak
698 V V. 2012. Protein interactions with piALU RNA indicates putative participation of retroRNA in
699 the cell cycle, DNA repair and chromatin assembly. *Mob Genet Elements* **2**: 26–35.
700 Bourque G. 2009. Transposable elements in gene regulation and in the evolution of vertebrate genomes.
701 *Curr Opin Genet Dev* **19**: 607–612.
702 Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew J-L, Ruan Y, Wei C-L, Ng HH,
703 et al. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable
704 elements. *Genome Res* **18**: 1752–62.
705 Cabili MN, Dunagin MC, McClanahan PD, Biaesch A, Padovan-Merhar O, Regev A, Rinn JL, Raj A.
706 2015. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule
707 resolution. *Genome Biol* **16**: 20.
708 Carlevaro-Fita J, Rahim A, Guigó R, Vardy LA, Johnson R. 2016. Cytoplasmic long noncoding RNAs
709 are frequently bound to and degraded at ribosomes in human cells. *RNA* **22**: 867–82.
710 Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, Pesce E, Ferrer I, Collavin L,
711 Santoro C, et al. 2012. Long non-coding antisense RNA controls Uchl1 translation through an
712 embedded SINEB2 repeat. *Nature* **491**: 454–7.
713 Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, Zhang Q, Yan G, Cui Q. 2013. LncRNADisease: a
714 database for long-non-coding RNA-associated diseases. *Nucleic Acids Res* **41**: D983–D986.
715 Chen L-L. 2016. Linking Long Noncoding RNA Localization and Function. *Trends Biochem Sci*.
716 Chillón I, Pyle AM. 2016. Inverted repeappcor elements in the human lincRNA-p21 adopt a conserved
717 secondary structure that regulates RNA function. *Nucleic Acids Res* gkw599.
718 Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts

to benefits.

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**: 885–98.

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

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**: 1775–89.

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.

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.

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, et al. The regulated retrotransposon transcriptome of mammalian cells.

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

Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, Lee S, Muzny D, Reid JG, Zhu Y, et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74.

Gong C, Maquat LE. 2011. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature* **470**: 284–8.

Guttman M, Rinn JL. 2012. Modular regulatory principles of large non-coding RNAs. *Nature* **482**: 339–46.

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.

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

Hacisuleyman E, Shukla CJ, Weiner CL, Rinn JL, Koning AP de, Gu W, Castoe TA, Batzer MA, Pollock DD, Wicker T, et al. 2016b. Function and evolution of local repeats in the Firre locus. *Nat Commun* **7**: 11021.

Hacisuleyman E, Shukla CJ, Weiner CL, Rinn JL, Koning AP de, Gu W, Castoe TA, Batzer MA, Pollock DD, Wicker T, et al. 2016c. Function and evolution of local repeats in the Firre locus. *Nat Commun* **7**: 11021.

753 Haerty W, Ponting CP. 2013. Mutations within lncRNAs are effectively selected against in fruitfly but
754 not in human. *Genome Biol* **14**: R49.
755 Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D,
756 Zadissa A, Searle S, et al. 2012. GENCODE: The reference human genome annotation for The
757 ENCODE Project. *Genome Res* **22**: 1760–1774.
758 Hezroni H, Ben-Tov Perry R, Meir Z, Housman G, Lubelsky Y, Ulitsky I. 2017. A subset of conserved
759 mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. *Genome Biol* **18**:
760 162.
761 Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. 2015. Principles of Long
762 Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell*
763 *Rep* **11**: 1110–1122.
764 Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential
765 etiologic and functional implications of genome-wide association loci for human diseases and traits.
766 *Proc Natl Acad Sci* **106**: 9362–9367.
767 Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS,
768 Harte RA, Hsu F, et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids*
769 *Res* **34**: D590–D598.
770 Holdt LM, Hoffmann S, Sass K, Langenberger D, Scholz M, Krohn K, Finstermeier K, Stahringer A,
771 Wilfert W, Beutner F, et al. 2013. Alu Elements in ANRIL Non-Coding RNA at Chromosome 9p21
772 Modulate Atherogenic Cell Functions through Trans-Regulation of Gene Networks. *PLoS Genet* **9**.
773 Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced
774 transposition and deleterious effects on neighboring gene expression. *Genome Res* **19**: 1419–28.
775 Hu B, Yang Y-CT, Huang Y, Zhu Y, Lu ZJ. 2017. POSTAR: a platform for exploring post-transcriptional
776 regulation coordinated by RNA-binding proteins. *Nucleic Acids Res* **45**: D104–D114.
777 Huda A, Bowen NJ, Conley AB, Jordan IK. 2011. Epigenetic regulation of transposable element derived
778 human gene promoters. *Gene* **475**: 39–48.
779 Jjingo D, Conley AB, Wang J, Mariño-Ramírez L, Lunyak V V, Jordan IK. 2014. Mammalian-wide
780 interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. *Mob*
781 *DNA* **5**: 14.
782 Johnson R, Guigó R. 2014. The RIDL hypothesis: transposable elements as functional domains of long
783 noncoding RNAs. *RNA* **20**: 959–76.
784 Johnson R, W.B. D, B.T. L, J.R. A, S.L. G, M. M, G.J. W, C.M. M, B.T. L, R. R. 2006. Identification of
785 the REST regulon reveals extensive transposable element-mediated binding site duplication. *Nucleic*
786 *Acids Res* **34**: 3862–3877.

Jurka J, Zietkiewicz E, Labuda D. 1995. Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. *Nucleic Acids Res* **23**: 170–5.

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

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. ed. H.E. Hoekstra. *PLoS Genet* **9**: e1003470.

Kelley D, Rinn J. 2012. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol* **13**: 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.

Kim S. 2015. ppcor: Partial and Semi-Partial (Part) Correlation.

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

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.

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 (80-)* **300**.

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**: D92–D97.

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

Marín-Béjar O, Huarte M. 2015. RNA Pulldown Protocol for In Vitro Detection and Identification of RNA-Associated Proteins. In *Methods in molecular biology (Clifton, N.J.)*, Vol. 1206 of, pp. 87–95.

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.

Martin KC, Ephrussi A. 2009. mRNA localization: gene expression in the spatial dimension. *Cell* **136**: 719–30.

Mas-Ponte D, Carlevaro-Fita J, Palumbo E, Hermoso Pulido T, Guigo R, Johnson R. 2017. LncATLAS database for subcellular localisation of long noncoding RNAs. *RNA* rna.060814.117.

Mercer TR, Mattick JS. 2013. Structure and function of long noncoding RNAs in epigenetic regulation.

Nat Publ Gr **20**.

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.

Nakazawa M. 2018. fmsb: Functions for Medical Statistics Book with some Demographic Data.

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**: 635–40.

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.

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.

Pegueroles C, Gabaldón T. 2016. Secondary structure impacts patterns of selection in human lncRNAs. *BMC Biol* **14**: 60.

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

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**: D168-73.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–2.

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**: 2085–93.

R Core Team. R: A Language and Environment for Statistical Computing.

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

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.

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.

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. **8**.

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.

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**: 1034–50.

Smit, AFA, Hubley, R & Green P. 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.

Su M, Han D, Boyd-Kirkup J, Yu X, Han J-DJ. 2014. Evolution of Alu elements toward enhancers. *Cell Rep* **7**: 376–85.

Suzuki K, Bose P, Leong-Quong RYY, Fujita DJ, Riabowol K. 2010. REAP: A two minute cell fractionation method. *BMC Res Notes* **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.

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.

Ulitsky I, Bartel DP. 2013. lincRNAs: Genomics, Evolution, and Mechanisms. *Cell* **154**: 26–46.

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.

Washietl S, Kellis M, Garber M. 2014. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res* **24**: 616–28.

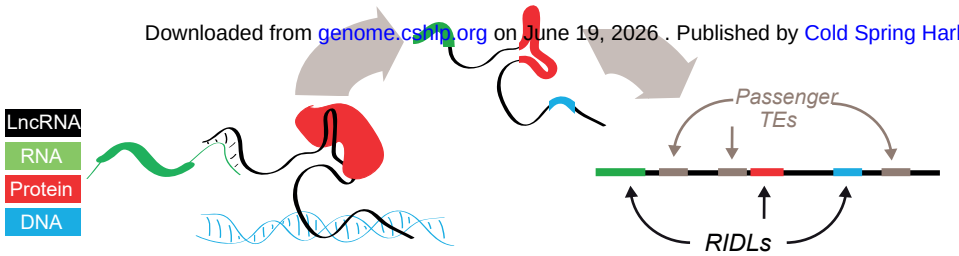
Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**: D1001–D1006.

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.

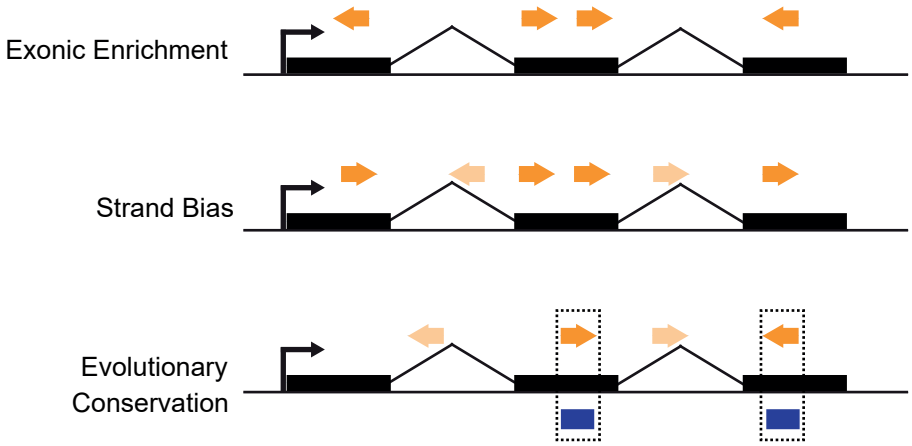
Zhang K, Shi Z-M, Chang Y-N, Hu Z-M, Qi H-X, Hong W. 2014b. The ways of action of long non-coding RNAs in cytoplasm and nucleus. *Gene* **547**: 1–9.

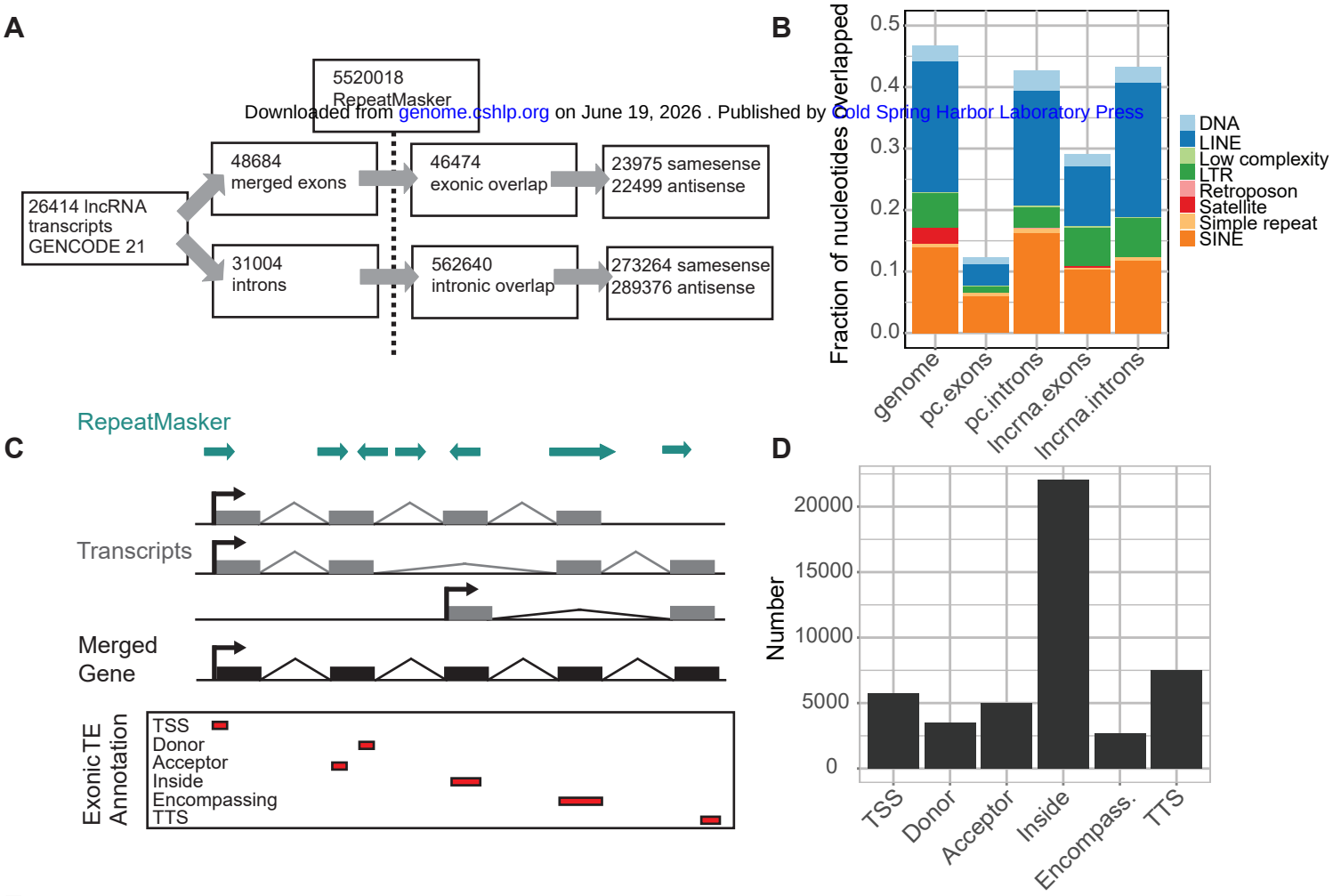
889

A



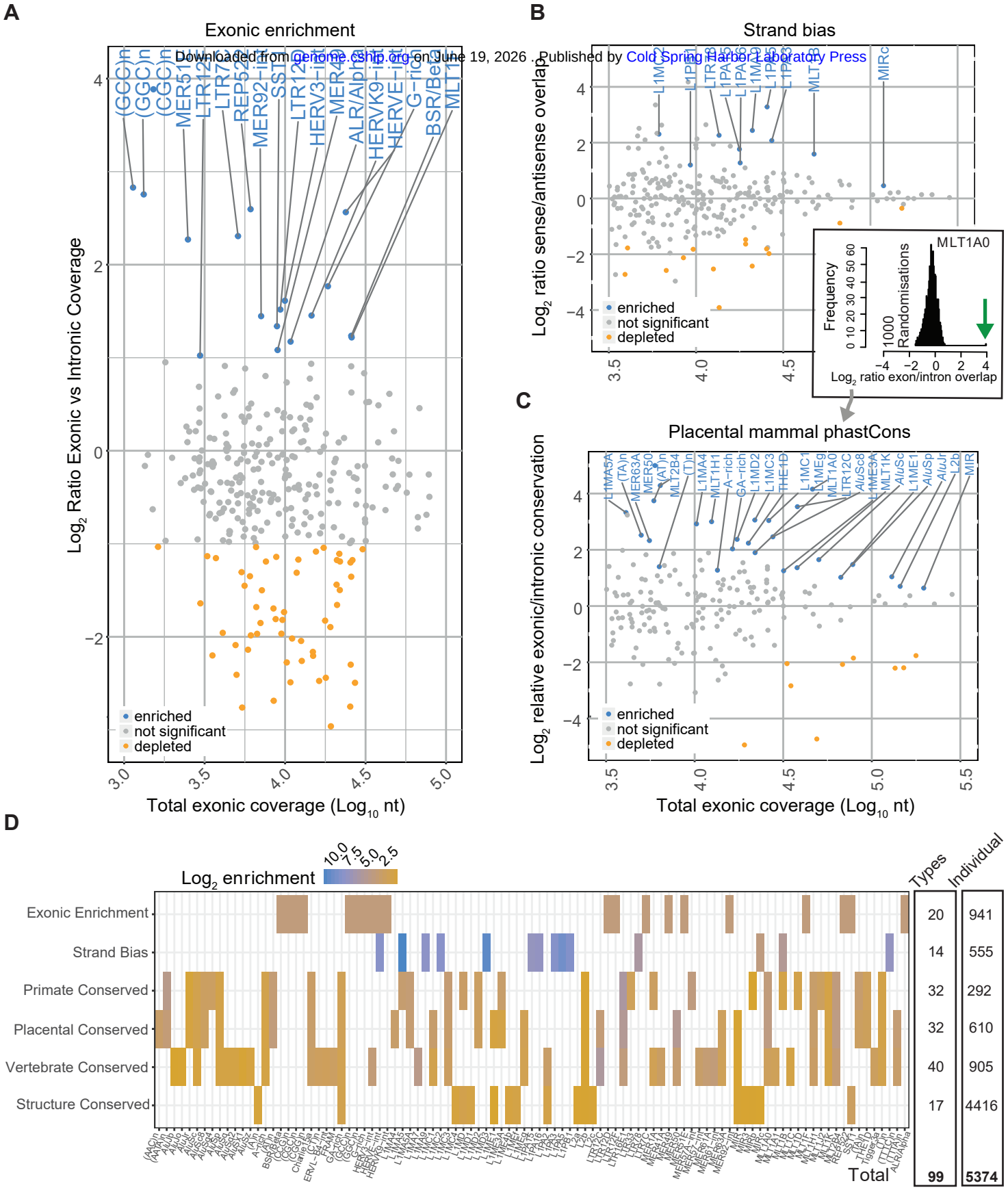
B

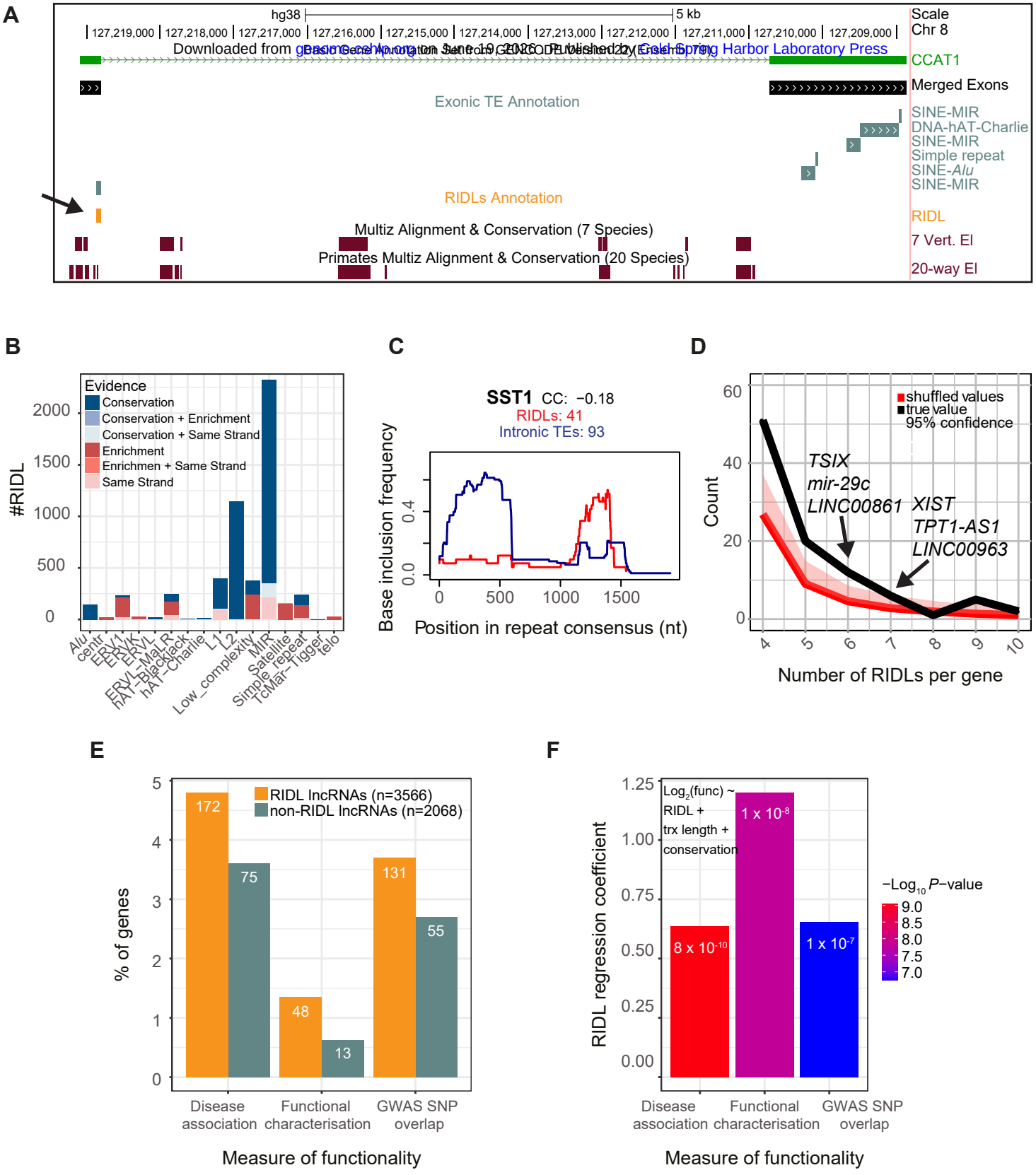


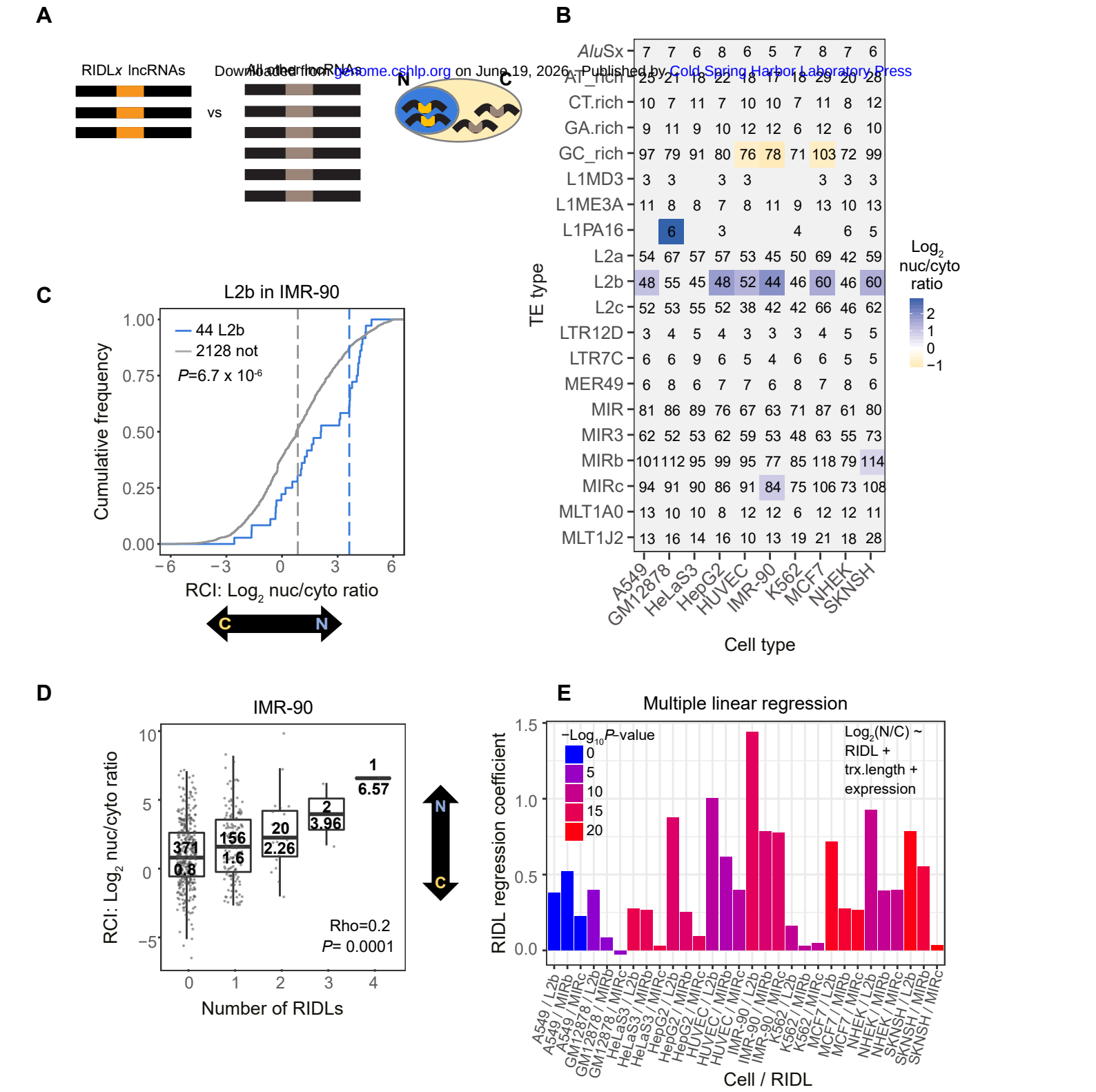


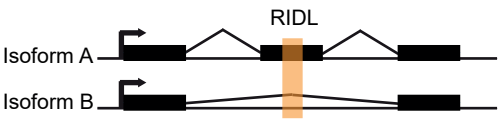
E

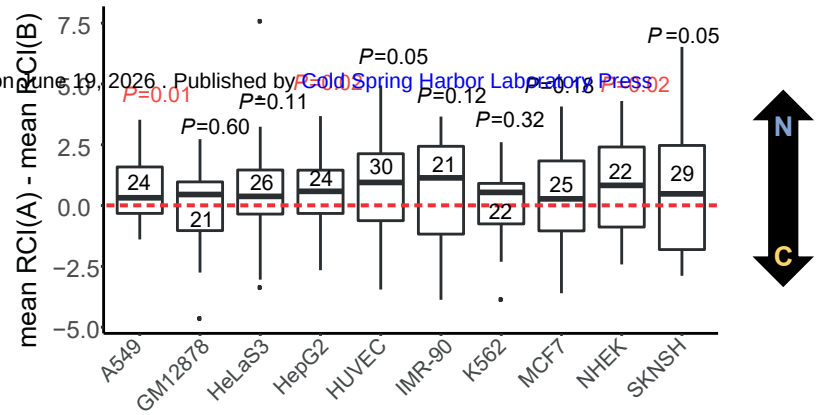
Repeat Family	TSS+	TSS−	SpliceDon+	SpliceDon−	SpliceAcc+	SpliceAcc−	Inside+	Inside−	Encompass.+	Encompass.−	TTS+	TTS−	Total
Alu	262	691	367	304	235	1017	1795	1903	15	625	913	454	8581
CR1	27	32	44	15	28	58	172	177	10	10	36	50	659
ERV1	402	250	116	112	196	67	336	272	116	75	507	133	2582
ERVK	37	11	25	4	20	10	21	13	5	4	56	9	215
ERVL	219	137	98	70	175	95	307	316	122	62	295	112	2008
ERVL−MaLR	634	227	214	83	415	217	790	551	358	150	790	194	4623
L1	315	525	243	176	167	324	1292	1276	135	335	710	763	6261
L2	185	327	138	234	333	174	1161	1127	77	111	288	218	4373
Low_complexity	47	31	13	0	10	5	347	293	1	0	43	7	797
MIR	209	309	372	332	153	629	1905	1749	53	151	378	278	6518
Simple_repeat	126	131	21	44	10	21	1846	1620	0	0	118	116	4053
TcMar−Tigger	68	43	59	28	52	69	203	190	58	15	117	83	985
hAT−Charlie	117	89	136	66	119	124	628	630	52	15	249	173	2398
hAT−Tip100	25	24	18	15	28	36	125	112	11	14	37	41	486




A
hg38
5 kb
Scale
Chr 8
127,219,000 | 127,218,000 | 127,217,000 | 127,216,000 | 127,215,000 | 127,214,000 | 127,213,000 | 127,212,000 | 127,211,000 | 127,210,000 | 127,209,000
Downloaded from genome.ucsc.edu on June 19, 2026 (UCSC version 22) © Cold Spring Harbor Laboratory Press
CCAT1
Merged Exons
Exonic TE Annotation
SINE-MIR
DNA-hAT-Charlie
SINE-MIR
Simple repeat
SINE-Alu
SINE-MIR
RIDLs Annotation
RIDL
Multiz Alignment & Conservation (7 Species)
7 Vert. El
Primates Multiz Alignment & Conservation (20 Species)
20-way El
B
Evidence
Conservation
Conservation + Enrichment
Conservation + Same Strand
Enrichment
Enrichmen + Same Strand
Same Strand
#RIDL
2000
1500
1000
500
0
Alu
centr
ERV1
ERVK
ERVL
ERVL-MaLR
hAT-Blackjack
hAT-Charlie
L1
L2
Low_complexity
MIR
Satellite
Simple_repeat
TcMar-Tigger
telo
C
SST1 CC: −0.18
RIDLs: 41
Intronic TEs: 93
Base inclusion frequency
0.4
0.0
0
500
1000
1500
Position in repeat consensus (nt)
D
shuffled values
true value
95% confidence
Count
60
40
20
0
TSIX
mir-29c
LINC00861
XIST
TPT1-AS1
LINC00963
4
5
6
7
8
9
10
Number of RIDLs per gene
E
RIDL lncRNAs (n=3566)
non-RIDL lncRNAs (n=2068)
% of genes
5
4
3
2
1
0
172
75
48
13
131
55
Disease association
Functional characterisation
GWAS SNP overlap
Measure of functionality
F
Log2(func) ~ RIDL + trx length + conservation
RIDL regression coefficient
1.25
1.00
0.75
0.50
0.25
0.00
1 x 10-8
8 x 10-10
1 x 10-7
−Log10 P−value
9.0
8.5
8.0
7.5
7.0
Disease association
Functional characterisation
GWAS SNP overlap
Measure of functionality







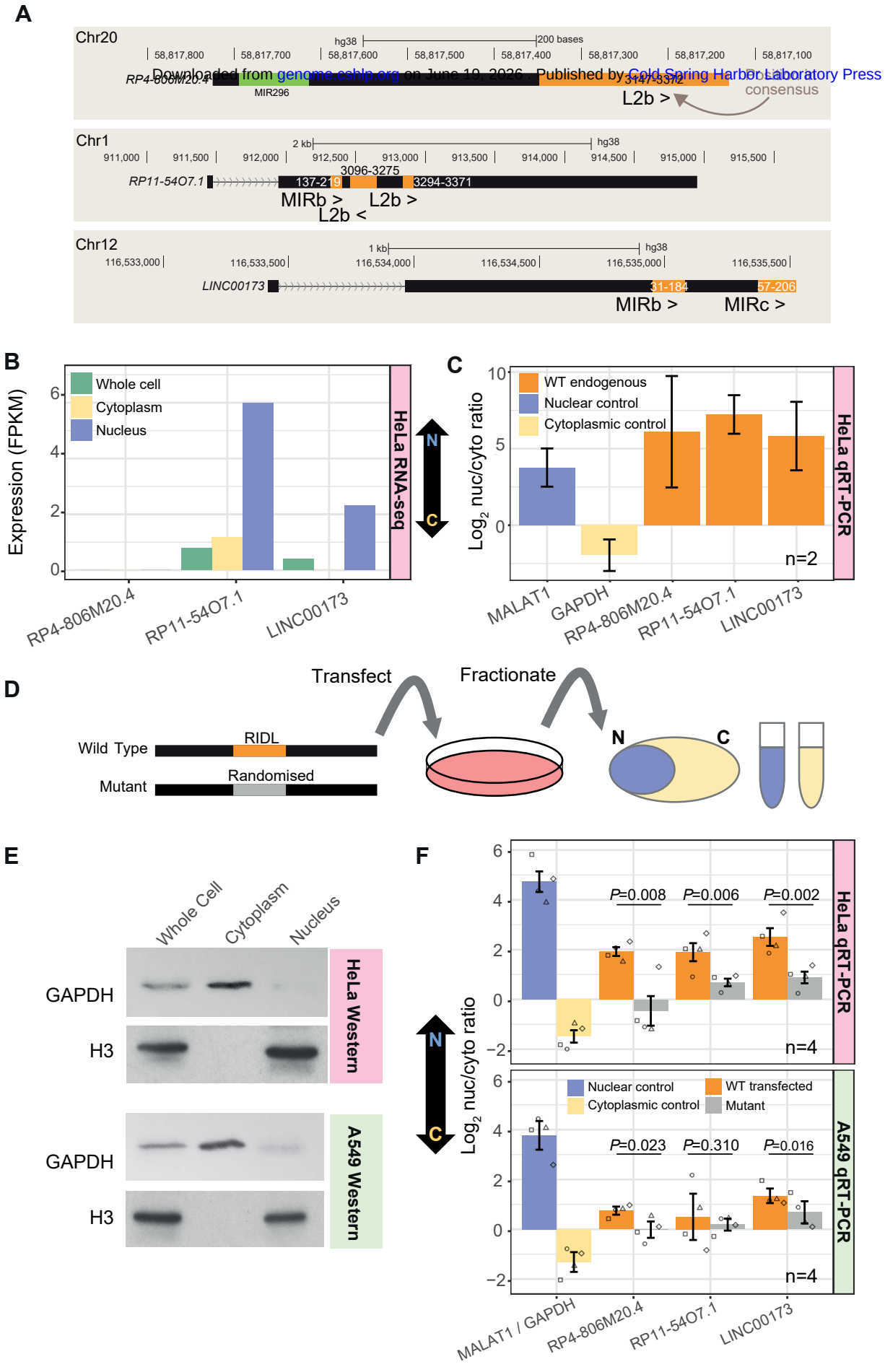


Figure Legends:

Figure 1: Repeat insertion domains of lncRNAs (RIDLs).

(A) In the "Repeat Insertion Domain of LncRNAs" (RIDL) model, exonically-inserted fragments of transposable elements contain pre-formed protein-binding (red), RNA-binding (green) or DNA-binding (blue) activities, that contribute to functionality of the host lncRNA (black). RIDLs are likely to be a small minority of exonic TEs, coexisting with large numbers of non-functional "passengers" (grey). (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; (3) elevated evolutionary conservation in exons compared to introns. Selection might be identified by comparing exonic TEs to a neutral population, for example those residing in lncRNA introns (light coloured arrows).

Figure 2: An exonic transposable element annotation with the GENCODE v21 lncRNA catalogue

(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 panel) according to the gene structure with which they intersect, and 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; "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. +/- indicate the relative strand of the TE with respect to lncRNA transcript.

Figure 3: Evidence for selection on transposable elements in lncRNA exons.

(A) Figure shows, for every TE type, the enrichment of per nucleotide coverage in exons compared to introns (*y* axis) and overall exonic nucleotide coverage (*x* axis). Enriched TE types (at a 2-fold cutoff) are shown in blue. (B) As for (A), but this time the *y* axis records the ratio of nucleotide coverage in sense vs 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 randomisation procedure, and significance is defined at an uncorrected empirical *p*-value < 0.001 (See Material and Methods). (C) As for (A), but here the *y* axis records the ratio of per-nucleotide overlap by phastCons mammalian-conserved elements for exons vs 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 randomisation procedure, and significance is defined at an uncorrected empirical *p*-value < 0.001 (See Material and 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 (i) the number of unique TE types identified by each method, and (ii) the number of instances of exonic

TEs from each type with appropriate selection evidence. The latter are henceforth defined as "RIDLs".

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 (grey), including three separate MIR elements, only one is defined a RIDL (orange) due to 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 correlation coefficient of the two profiles. "RIDLs" / "Intronic TEs" indicate 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 Materials and Methods), in the lncRNAdb database of functional lncRNAs (functional characterisation), 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 than in (E). Colours 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.

Figure 5: Correlation between RIDLs and host lncRNA nuclear/cytoplasmic localisation.

(A) Outline of in silico screen for localisation-regulating RIDLs. For each RIDL-type / cell-type combination, the nuclear/cytoplasmic localisation of RIDL-lncRNAs is compared to all other detected lncRNAs. (B) Results of an in silico screen. Rows: RIDL types; Columns: Cell types. Significant RIDL-cell type combinations are coloured (Benjamini-Hochberg corrected *p*-value < 0.01; Wilcoxon test). Colour 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 ≥1 RIDLs, grey 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). Correlation coefficient (Rho) and corresponding *p*-value (*P*) were calculated using Spearman correlation, two-sided test. In each box, upper value indicates the number of lncRNAs, and 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). Colours indicate the associated *p*-value. These values assess the correlation between RIDL number and nuclear/cytoplasmic localisation (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).

Figure 6: RIDLs correlate with differential localisation 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 ≥1 RIDL-transcript and ≥1 non-RIDL transcript expressed in a given cell line. Numbers inside the boxplots indicate the number of gene loci analysed 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 $P < 0.05$).

Figure 7: Disruption of RIDLs results in lncRNA relocalisation 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 (">" for same strand, "<" for opposite strand). (B) Expression of the three lncRNA candidates as inferred from HeLa RNA-seq (40). (C) Nuclear/cytoplasmic localisation 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 localisation of transfected candidate lncRNAs in HeLa (upper panel) and A549 (lower panel). GAPDH/MALAT1 are used as cytoplasmic/nuclear controls, respectively. N indicates the number of biological replicates (values from all replicates are plotted, each replicate is represented by a different dot shape), and error bars represent standard error of the mean. *P*-values for paired *t*-test (1 tail) are shown.