Ancient exapted transposable elements promote nuclear enrichment of long noncoding RNAs

Joana Carlevaro-Fita[1,2,4]

Taisia Polidori[1,2,4]

Monalisa Das[1,2,4]

Carmen Navarro[3]

Rory Johnson[1,2]*

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

4. Equal contribution

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

Keywords: Transposable element; 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). However, this concept remains largely untested, and just a handful of RIDLs have been identified. We present a transcriptome-wide map of putative RIDLs in human, using evidence from insertion frequency, strand bias and evolutionary conservation of sequence and structure. In the exons of GENCODE v21 lncRNAs, we identify 5374 RIDLs in 3566 loci. These are enriched in functionally-validated and disease-associated genes. This RIDL map 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 exonic LINE2 and MIR 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 L2b, MIRb and MIRc elements drive nuclear enrichment of their host lncRNA. Together these data suggest a global role for exonic 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 (1). Our current thinking 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 as molecular interactions or subcellular localisation (2–4). Mapping these domains is thus a key step towards the understanding and prediction of lncRNA function in healthy and diseased cells.

One possible source of lncRNA domains are transposable elements (TEs) (4). TEs are known to have been major contributors to genomic evolution through the insertion and neofunctionalisation of sequence fragments – a process known as *exaptation* (5)(6). This process has contributed to the evolution of diverse features in genomic DNA, including transcriptional regulatory motifs (7, 8), microRNAs (9), gene promoters (10, 11), and splice sites (12, 13).

We recently proposed that TEs may contribute pre-formed functional domains to lncRNAs. We termed these "RIDLs" – *Repeat Insertion Domains of Long noncoding RNAs* (4). 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 (14). Diverse evidence also points to repetitive sequences forming intermolecular Watson-Crick RNA:RNA and RNA:DNA hybrids (4, 15, 16). 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 (4). These are found in lncRNAs with clearly demonstrated functions, including the X-chromosome silencing transcript *XIST* (17), the oncogene *ANRIL* (16) and the regulatory antisense *UchlAS* (18). 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 (17); Watson-Crick hybridisation between RNA and DNA Alu elements recruits *ANRIL* to target genes (16); a SINEB2 repeat in *Uchl1AS* increases translational rate of its sense mRNA (18). In parallel, transcriptome-wide maps of lncRNA-linked TEs have shown how TEs have contributed extensively to lncRNA gene evolution (19, 20)(21)(22). However, there has been no attempt so far to filter passenger TEs out of these maps, to create annotations of RIDLs with evidence for functionality in the context of mature lncRNA molecules.

Subcellular localisation, and the domains controlling it, are crucial determinants of lncRNA functions (reviewed in (23)). For example, transcriptional regulatory lncRNAs must be located in the nucleus and chromatin, while those regulating microRNAs or translation must be present in the cytoplasm (24). Consistent with this, lncRNAs exhibit diverse patterns of localization (25, 26). If lessons learned from mRNA are also valid for lncRNAs, then short sequence motifs binding to RNA binding proteins (RBPs) will be an important localisation-regulatory mechanism (27). This was recently demonstrated for the *BORG* lncRNA, where a pentameric motif was shown to mediate nuclear retention (28). Most interestingly, another study implicated an inverted pair of Alu elements in nuclear retention of *lincRNA-P21* (29). This raises the possibility that RIDLs may function as regulators of lncRNA localisation.

In the present study, we create a human transcriptome-wide catalogue of putative RIDLs. We show that 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, confer nuclear enrichment on their host transcripts.

Results

The objective of this study was to create a map of repeat insertion domains of long noncoding RNAs (RIDLs) and link them to lncRNA functions. We previously hypothesised that RIDLs could function by mediating interactions between host lncRNA and DNA, RNA or protein molecules (4) (Figure 1A). However, any attempt to map RIDLs is confronted with the problem that they will likely represent a small minority amongst many non-functional "passenger" transposable elements (TEs) residing in lncRNA exons (Figure 1B). Therefore, it is necessary to identify RIDLs by some signature of selection in lncRNA exons. In this study we use three such signatures: exonic enrichment, strand bias, and evolutionary conservation (Figure 1B).

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 against 48684 exons from 26414 transcripts of 15877 GENCODE v21 lncRNA genes, resulting in 46474 exonic TE insertions in lncRNA (Figure 1B). Altogether 13121 lncRNA genes (82.6%) carry at least one exonic TE fragment in one or more of their mature transcripts.

We defined TEs residing in lncRNA introns to represent neutral background, since they do not contribute to functionality of the mature processed transcript, but nevertheless should reflect any large-scale genomic biases in TE abundance. Thus we also created a control TE intersection dataset with 31,004 GENCODE lncRNA introns, resulting in 562,640 intron-overlapping TE fragments (Figure 2A). This analysis is likely to be conservative, since some intronic sequences could contribute to unidentified lncRNA exons.

Comparing intronic and exonic TE data, we see that lncRNA exons are slightly 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 (20). This may reflect generalised selection against disruption of functional lncRNA transcripts by TEs.

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 (13). 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 (J. Lagarde, B. Uszczyńska et al, Manuscript In Press). Altogether 4993 (18.9%) transcripts are promoted from within a TE, most often those of Alu and ERVL-MaLR classes (Figure 2E). 7497 (28.4%) lncRNA transcripts are terminated by a TE, most commonly by Alu, ERVL-MaLR and LINE1 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 (13). Nevertheless, the most frequent case is represented by 22031 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*. 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 more frequently found within lncRNA exons, compared to intronic sequence (19). Thus we calculated the ratio of exonic / intronic sequence coverage by TEs (Figure 3A). We found signals of 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. Surprisingly, quite 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 broadly consistent with previous analyses by Kelley and Rinn using whole genome, rather than introns, as background (19).

Interestingly, presently-active LINE1 elements are 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 (30). Another explanation may be low transcriptional processivity exhibited by the LINE1 ORF2 in the sense strand, leading to reduced gene expression and consequent negative selection (30).

In a similar way, we investigated whether exonic TEs display a strand preference relative to host lncRNA (4). 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 strand bias in the resulting exonic TE sequences. To avoid this, we removed any TE insertions 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 (see Materials and Methods) (Figure 3B). In lncRNA exons, a number of TE types are enriched in either sense or antisense. They are 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. Here we observed far weaker strand bias. Nevertheless, we did detect a widespread, modest yet statistically-significant depletion of same-strand TE insertions (Supplementary Figure S1). This is especially true for LINE1 elements, possibly for reasons mentioned above. In contrast to exons, almost no TE types were significantly enriched on the sense 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 (31) calculated on primate, placental mammal and vertebrate alignments; second, the more recently published “Evolutionarily Conserved Structures” (ECS) set (32). 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 frequency of their overlap with evolutionarily conserved genomic elements compared to 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 Supplementary Figure S1B,C) and ECS (Supplementary Figure S1D) data. The majority of TEs do not exhibit significantly different conservation when found in exons or introns (Grey points). However, depending on conservation data used, the method detects significant conservation for 17 to 40 TE types (Figure 3C). This enrichment disappeared when PhastCons elements were positionally randomised, as expected (Supplementary Figure S1E). We also found a small number of TEs depleted for signatures of conservation in lncRNA exons, namely the young AluSz, AluSx and AluJb (PhastCons) and L1M4c and AluSx1 (ECS) (coloured red in Figure 3C and Supplementary 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 (33).

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. 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 a PhastCons element would be included in the RIDL annotation, and all other exonic MIRs would be excluded. 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). It is important to understand that this RIDL annotation should be considered an "enrichment group", with a greater proportion of functional TEs than when using the entire exonic TE set.

After removing redundancy, the final RIDL annotation consists of 5374 elements, located within 3566 distinct lncRNA genes (Figure 3D). 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). 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 Supplementary File S1.

We also examined the evolutionary history of RIDLs. Using 6-mammal alignments, their depth of evolutionary conservation could be inferred (Supplementary Figure S2). 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 (32, 34, 35).

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 Supplementary 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.

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) (Supplementary Figure S3).

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 non-RIDL lncRNAs (Supplementary Figure S4). RIDL-lncRNAs are (1) over-represented in the reference database for functional lncRNAs, lncRNAdb (36) (Figure 4E) and (2) are significantly enriched for lncRNAs associated with cancer and other diseases (Figure 4F). Finally, we found that RIDL-lncRNAs are enriched for trait/disease-associated SNPs in both their promoter and exonic regions (Figure 4G).

In addition to *CCAT1* (Figure 4A) (37) 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 (4), these include an intriguing 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 (4). 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 of RIDL-carrying lncRNAs can be found in Supplementary 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 (25, 38)(26). We next asked whether RIDLs might regulate lncRNA localisation (28, 39)(29) (Figure 5A). Using subcellular RNAseq data based on 10 ENCODE cell lines (40), we calculated the relative nuclear/cytoplasmic localisation in log2 units, or "Relative Concentration Index" (RCI) (26). Next, we comprehensively tested each of the 99 RIDL types for association with localisation of their host transcript.

After correcting for multiple hypothesis testing, this approach identified four distinct TEs that significantly correlate with nuclear enrichment: L1PA16, L2b, MIRb and MIRc (Figure 5B). For example, 44 lncRNAs carrying L2b RIDLs have 6.9-fold higher relative nuclear/cytoplasmic ratio in IMR90 cells, and this tendency is observed in six different cell types (Figure 5B,C). Furthermore, 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-RIDL containing lncRNAs have between 2 and 2.3-fold higher relative expression in cytoplasm in these cells (Supplementary Figure S5).

We hypothesised that both transcript length and whole-cell gene expression may be the true factors influencing localisation. If either correlated with TE number, then the apparent correlation between TEs and localisation could be an artefact. To address this, we constructed a linear model 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 remove confounding effects. Training such models independently 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 (Supplementary Figure S6). We drew similar conclusions from equivalent analyses using partial correlation analysis (data not shown).

Given that L2b and MIR elements predate human-mouse divergence, we attempted perform similar analyses in mouse cells. However given that just two equivalent datasets are available at present (41)(42), as well as the relatively low number of annotated lncRNAs in mouse, we were unable to draw statistically-robust conclusions regarding the conservation of this phenomenon.

LncRNA gene loci are often composed of multiple, differentially-spliced transcript isoforms that partially differ in their mature sequence. We reasoned that if RIDLs are indeed causative in the observed nuclear enrichment of their host transcripts, then we would expect this effect to also be observed *within* gene loci. In other words, for RIDL-lncRNA gene loci having multiple transcript isoforms, then those isoforms *with* a RIDL should display greater nuclear enrichment than those isoforms *without* a RIDL (Figure 6A, 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 localisation between their RIDL and non-RIDL isoforms (Figure 6A). In every case, 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. An example of this effect is shown for *LINC00173* in Figure 6. From its three expressed transcripts (Figure 6B), the Trx1 isoform carrying two RIDLs (MIRb and MIRc) has the highest nuclear enrichment median across cells, followed by the Trx4 isoform carrying one (L2b). In contrast, the non-RIDL Trx3 isoform has the lowest overall nuclear/cytoplasmic ratio (Figure 6C).

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 in human cells.

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 host lncRNA. We selected three lncRNAs, based on: (i) presence of L2b, MIRb and MIRc RIDLs; (ii) moderate expression; (iii) nuclear localisation, as inferred from RNAseq (Figure 7A,B and Supplementary Figure S7). Nuclear localisation of these candidates could be validated in HeLa cells using qRTPCR (Figure 7C).

We formulated an assay to compare the localisation of transfected lncRNAs carrying wild-type RIDLs, and mutated versions where RIDL sequence has been randomised without altering sequence composition ("Mutant") (Figure 7D, full sequences available in Supplementary File S4). Wild-type and Mutant lncRNAs were transfected into cultured cells and their localisation evaluated by fractionation. qRT-PCR primers were designed in such a way 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 (Supplementary Figure S8). Fractionation quality was verified by Western blotting (Figure 7E) and qRT-PCR (Figure 7F), and our stringent DNase-treatments meant that plasmid DNA contamination in cDNA samples made negligible contributions to our results (Supplementary Figure S9).

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 (Supplementary Figure S10). To summarise, these results demonstrate a role for exonic L2b, MIRb and MIRc elements in 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 have a two- to four-fold effect on nuclear/cytoplasmic localisation.

These data support our published hypothesis that exonic TE elements may function as 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 (4). In the context of localisation, RIDLs could mediate nuclear retention due to their described interactions with nuclear proteins (43), although it is also conceivable that the effect occurs through hybridisation to complementary repeats in genomic DNA.

The localisation RIDLs discovered – MIR and LINE2 - are both ancient and contemporaneous, being active prior to the mammalian radiation (44). Both have previously been associated with acquired roles in the context of genomic DNA (45)(8). 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 (22)(46)(47)(48). This would mean that MIR and LINE2 RIDLs were pre-existing sequences that were exapted by newly-born lncRNAs. This corresponds to the "latent" exaptation model proposed by Feschotte and colleagues(49). However it is also possible that for other cases the reverse could be true – a pre-existing lncRNA exapts a newly-inserted TE. Our use of evolutionary conservation to identify RIDLs likely biases our analysis against the discovery of modern TE families, such as Alus.

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.

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 (50)(51)(52), to evolutionarily conserved segments (32)(53). Previous maps of TEs have highlighted their profound roles in lncRNA gene evolution (20)(54)(22). 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 most likely contains substantial false positive and false negative rates that will be improved in future.

This study may help to explain a longstanding and unexplained property of lncRNAs: their nuclear enrichment (38)(1). Although they are readily detected in the cytoplasm, it is nevertheless true that lncRNAs tend to have higher nuclear/cytoplasmic ratios compared to mRNAs (55)(1)(38)(26). This is true across various human and mouse cell types. Although this may partially be explained by decreased stability (56), it is likely that RNA sequence motifs also contribute to nuclear localisation (29)(28). 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 mRNAs may help explain their greater cytoplasmic abundance (57). 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 (58). Together, this evidence may point to 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, and GENCODE version 21 lncRNA annotations in GTF format were downloaded from www.GENCODEgenes.org (59). We developed a pipeline, called *Transposon.Profile*, to annotate exonic and intronic TEs of a given gene annotation. Transposon.Profile is largely based on BEDTools' *intersect* and *merge* functionalities(60). *Transposon.Profile* merges the exons of all transcripts belonging to the given gene annotation, 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" (Supplementary File S5).

RIDL identification

Using this Exonic TE Annnotation, we identified the subset of individual TEs with evidence for functionality. For certain analysis, an equivalent Intronic TE Annotation was also employed, being the *Transposon.Profile* 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 (12, 13). 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 randonly 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 or negatively) all of 1000 simulations. All exonic instances of these TE types that also have the appropriate sense or antisense strand orientation are defined as RIDLs.

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(31), while the ECS conserved regions from obtained from Supplementary Data of Smith et al (32) (see Supplementary 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. For each TE type we calculated the exonic/intronic conservation ratio. To do this we used *IntersectBED* (60) to overlap exonic locations with TEs, and calculate the total number of nucleotides overlapping. We performed a similar operation for intronic regions. Finally, we calculated the following ratio:

r_{ei}=(nt of overlap between conserved elements and exons)/(nt of overlap between conserved elements and introns)

To estimate the background, the exonic and intronic BED files were positionally randomized 1000 times using *bedtools shuffle*, each time recalculating r_{ei}. We considered to be significantly conserved those TE types where the true r_{ei} was greater or less than every one of 1000 randomised r_{ei} values. All exonic instances of these TE types that also intersect the appropriate evolutionarily conserved element are defined as RIDLs.

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 Supplementary File S1.

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), 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.

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 (36).
B) Disease-associated genes: lncRNAs from GENCODE v21 that are present in at least in one of the following databases or public sets: LncRNADisease (61), Lnc2Cancer (62), 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 (63, 64) (https://www.ebi.ac.uk/gwas/home). We intersected its coordinates with lncRNA exons and promoters (promoters defined as 1000nt upstream and 500nt downstream of GENCODE-annotated TSS).

For defining a comparable set of "other lncRNAs" we sampled from the rest of GENCODE v21 lncRNAs a set of lncRNAs matching RIDL-carrying lncRNAs exonic length distribution (Supplementary Figure S4). We performed the subsampling using *matchDistribution* script: https://github.com/julienlag/matchDistribution.

Subcellular localisation analysis

Processed RNAseq 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 (26, 40). 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 polyA+ RPKM)/(cytoplasmic polyA+ RPKM), and only transcripts having non-zero values in both were considered. These ratios were log2-transformed, to yield the Relative Concentration Index (RCI) (26). 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 Wilcox 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 (65).

Linear model analysis

Linear models were created in R using the "lm" package, 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 RNAseq in units of RPKM. Equivalent partial correlation analyses were performed, using the R "pcor.test" function (Spearman correlation), correlating RCI with RIDL number, while controlling for transcript length and expression.

Cell lines and reagents

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 37ºC 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 corresponding mutated versions where RIDL sequence has been randomised ("Mutant"), were

synthesized commercially (BioCat GmbH). The sequences were cloned into pcDNA 3.1 (+) vector within the *Nhe*I and *Xho*I 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 Supplementary File S4.

LncRNA transfection and sub-cellular fractionation

Wild-type and mutant lncRNA clones were transfected independently in separate wells of a 6-well plate. Transfections were carried out with 2 µg of total plasmid DNA in each well using Lipofectamine 2000. 48 h post-transfection, cells from each well were harvested, pooled and re-seeded into a 10 cm dish and allowed to grow till 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- to 1-fold (A549) (Supplementary Figure S8).

The nuclear and cytoplasmic fractionation was carried out as described previously (66) with minor modifications. In brief, cells from 10 cm dishes were harvested by scraping and washed with 1X ice cold PBS. For fractionation, cell pellet was re-suspended in 900 µl of ice-cold 0.1% NP40 in PBS and triturated 7 times using a p1000 micropipette. 300 µl 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". 300 µl 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 of 0.1% NP40 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 (Supplementary 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 Supplementary Figure S9 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 Scientific, # 34077).

Acknowledgements

We acknowledge Deborah Re (DBMR), Silvia Roesselet (DBMR) and Marianne Zahn (Inselspital) for administrative support. We wish to thank Roderic Guigo (CRG), Marc Friedlaender (SciLife Lab) and Marta Mele (Harvard) for many helpful discussions and Roberta Esposito (DBMR) for help and advice regarding experiments procedure and analysis. Samir Ounzain (CHUV) also contributed suggestions regarding experimental design. Julien Lagarde (CRG) kindly provided help in gene sampling analysis. Carlos Pulido (DBMR) assisted with RNAseq analysis. CN is supported by grant TIN-2013-41990-R from the Spanish Ministry of Economy, Industry and Competitiveness (MINECO). This research was supported by the NCCR RNA & Disease funded by the Swiss National Science Foundation.

References

1. Ulitsky,I. and Bartel,D.P. (2013) lincRNAs: Genomics, Evolution, and Mechanisms. *Cell*, **154**, 26–46.
2. Guttman,M. and Rinn,J.L. (2012) Modular regulatory principles of large non-coding RNAs. *Nature*, **482**, 339–346.
3. Mercer,T.R. and Mattick,J.S. (2013) Structure and function of long noncoding RNAs in epigenetic regulation. *Nat. Publ. Gr.*, **20**.
4. Johnson,R. and Guigó,R. (2014) The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA*, **20**, 959–76.
5. Bourque,G. (2009) Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Curr. Opin. Genet. Dev.*, **19**, 607–612.
6. Feschotte,C. (2008) Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.*, **9**, 397–405.
7. Bourque,G., Leong,B., Vega,V.B., Chen,X., Lee,Y.L., Srinivasan,K.G., Chew,J.-L., Ruan,Y., Wei,C.-L., Ng,H.H., *et al.* (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.*, **18**, 1752–62.
8. 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. and R.,R. (2006) Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. *Nucleic Acids Res.*, **34**, 3862–3877.
9. Roberts,J.T., Cardin,S.E. and Borchert,G.M. (2014) Burgeoning evidence indicates that microRNAs were initially formed from transposable element sequences. *Mob. Genet. Elements*, **4**, e29255.
10. Faulkner,G.J., Kimura,Y., Daub,C.O., Wani,S., Plessy,C., Irvine,K.M., Schroder,K., Cloonan,N., Steptoe,A.L., Lassmann,T., *et al.* The regulated retrotransposon transcriptome of mammalian cells. 10.1038/ng.368.
11. Huda,A., Bowen,N.J., Conley,A.B. and Jordan,I.K. (2011) Epigenetic regulation of transposable element derived human gene promoters. *Gene*, **475**, 39–48.
12. Lev-Maor,G., Sorek,R., Shomron,N. and Ast,G. (2003) The Birth of an Alternatively Spliced Exon: 3' Splice-Site Selection in Alu Exons. *Science (80-.).*, **300**.
13. Sela,N., Mersch,B., Gal-Mark,N., Lev-Maor,G., Hotz-Wagenblatt,A. and 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**.
14. Blackwell,B.J., Lopez,M.F., Wang,J., Krastins,B., Sarracino,D., Tollervey,J.R., Dobke,M., Jordan,I.K. and Lunyak,V. V Protein interactions with piALU RNA indicates putative participation of retroRNA in the cell cycle, DNA repair and chromatin assembly. 10.4161/mge.19032.
15. Gong,C. and Maquat,L.E. (2011) lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature*, **470**, 284–8.
16. Holdt,L.M., 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**.
17. Elisaphenko,E.A., Kolesnikov,N.N., Shevchenko,A.I., Rogozin,I.B., Nesterova,T.B., Brockdorff,N. and Zakian,S.M. (2008) A Dual Origin of the Xist Gene from a Protein-Coding Gene and a Set of Transposable Elements. 10.1371/journal.pone.0002521.
18. 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**, 454–7.
19. Kelley,D. and Rinn,J. (2012) Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.*, **13**, R107.
20. Kapusta,A., Kronenberg,Z., Lynch,V.J., Zhuo,X., Ramsay,L., Bourque,G., Yandell,M. and Feschotte,C. (2013) Transposable elements are major contributors to the origin, diversification, and

regulation of vertebrate long noncoding RNAs. *PLoS Genet.*, **9**, e1003470.
21. Schmitt,A.M., Chang,H.Y., Abdelmohsen,K., Panda,A., Kang,M.J., Xu,J., Selimyan,R., Yoon,J.H., Martindale,J.L., De,S., *et al.* (2016) Long Noncoding RNAs in Cancer Pathways. *Cancer Cell*, **29**, 452–463.
22. Hezroni,H., Koppstein,D., Schwartz,M.G., Avrutin,A., Bartel,D.P. and Ulitsky,I. (2015) Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell Rep.*, **11**, 1110–1122.
23. Chen,L.-L. (2016) Linking Long Noncoding RNA Localization and Function. *Trends Biochem. Sci.*, 10.1016/j.tibs.2016.07.003.
24. Zhang,K., Shi,Z.-M., Chang,Y.-N., Hu,Z.-M., Qi,H.-X. and Hong,W. (2014) The ways of action of long non-coding RNAs in cytoplasm and nucleus. *Gene*, **547**, 1–9.
25. Cabili,M.N., Dunagin,M.C., McClanahan,P.D., Biaesch,A., Padovan-Merhar,O., Regev,A., Rinn,J.L. and Raj,A. (2015) Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol.*, **16**, 20.
26. Mas-Ponte,D., Carlevaro-Fita,J., Palumbo,E., Hermoso Pulido,T., Guigo,R. and Johnson,R. (2017) LncATLAS database for subcellular localisation of long noncoding RNAs. *RNA*, 10.1261/rna.060814.117.
27. Martin,K.C. and Ephrussi,A. (2009) mRNA localization: gene expression in the spatial dimension. *Cell*, **136**, 719–30.
28. Zhang,B., Gunawardane,L., Niazi,F., Jahanbani,F., Chen,X. and Valadkhan,S. (2014) A Novel RNA Motif Mediates the Strict Nuclear Localization of a Long Noncoding RNA. *Mol. Cell. Biol.*, **34**, 2318–2329.
29. Chillón,I. and Pyle,A.M. (2016) Inverted repeat *Alu* elements in the human lincRNA-p21 adopt a conserved secondary structure that regulates RNA function. *Nucleic Acids Res.*, 10.1093/nar/gkw599.
30. Perepelitsa-Belancio,V. and Deininger,P. (2003) RNA truncation by premature polyadenylation attenuates human mobile element activity. *Nat. Genet.*, **35**, 363–366.
31. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S., *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–50.
32. Smith,M.A., Gesell,T., Stadler,P.F. and Mattick,J.S. (2013) Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res.*, **41**, 8220–8236.
33. Su,M., Han,D., Boyd-Kirkup,J., Yu,X. and Han,J.-D.J. (2014) Evolution of Alu elements toward enhancers. *Cell Rep.*, **7**, 376–85.
34. Jurka,J., Zietkiewicz,E. and Labuda,D. (1995) Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. *Nucleic Acids Res.*, **23**, 170–5.
35. Konkel,M.K., Walker,J.A. and Batzer,M.A. (2010) LINEs and SINEs of primate evolution. *Evol. Anthropol.*, **19**, 236–249.
36. Quek,X.C., Thomson,D.W., Maag,J.L. V, Bartonicek,N., Signal,B., Clark,M.B., Gloss,B.S. and Dinger,M.E. (2015) lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.*, **43**, D168-73.
37. Nissan,A., Stojadinovic,A., Mitrani-Rosenbaum,S., Halle,D., Grinbaum,R., Roistacher,M., Bochem,A., Dayanc,B.E., 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.
38. Derrien,T., Johnson,R., Bussotti,G., Tanzer,A., Djebali,S., Tilgner,H., Guernec,G., Martin,D., Merkel,A., Knowles,D.G., *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.
39. Hacisuleyman,E., Shukla,C.J., Weiner,C.L., Rinn,J.L., Koning,A.P. de, Gu,W., Castoe,T.A., Batzer,M.A., Pollock,D.D., Wicker,T., *et al.* (2016) Function and evolution of local repeats in the Firre locus. *Nat. Commun.*, **7**, 11021.

40. Djebali,S., Davis,C.A., 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.
41. Bahar Halpern,K., Caspi,I., Lemze,D., Levy,M., Landen,S., Elinav,E., Ulitsky,I. and Itzkovitz,S. (2015) Nuclear Retention of mRNA in Mammalian Tissues. *Cell Rep.*, **13**, 2653–2662.
42. Tan,J.Y., Sirey,T., Honti,F., Graham,B., Piovesan,A., Merkenschlager,M., Webber,C., Ponting,C.P. and Marques,A.C. (2015) Extensive microRNA-mediated crosstalk between lncRNAs and mRNAs in mouse embryonic stem cells. *Genome Res.*, **25**, 655–666.
43. Kelley,D.R., Hendrickson,D.G., Tenen,D. and Rinn,J.L. (2014) Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. *Genome Biol.*, **15**, 537.
44. Cordaux,R. and Batzer,M.A. The impact of retrotransposons on human genome evolution. 10.1038/nrg2640.
45. Jjingo,D., Conley,A.B., Wang,J., Mariño-Ramírez,L., Lunyak,V. V and Jordan,I.K. (2014) Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. *Mob. DNA*, **5**, 14.
46. Hezroni,H., Ben-Tov Perry,R., Meir,Z., Housman,G., Lubelsky,Y. and Ulitsky,I. (2017) A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. *Genome Biol.*, **18**, 162.
47. Necsulea,A., Soumillon,M., Warnefors,M., Liechti,A., Daish,T., Zeller,U., Baker,J.C., Grützner,F. and Kaessmann,H. (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, **505**, 635–40.
48. Washietl,S., Kellis,M. and Garber,M. (2014) Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res.*, **24**, 616–28.
49. Chuong,E.B., Elde,N.C. and Feschotte,C. (2017) Regulatory activities of transposable elements: from conflicts to benefits. 10.1038/nrg.2016.139.
50. Van Nostrand,E.L., Pratt,G.A., Shishkin,A.A., Gelboin-Burkhart,C., Fang,M.Y., Sundararaman,B., Blue,S.M., Nguyen,T.B., 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.
51. Hu,B., Yang,Y.-C.T., Huang,Y., Zhu,Y. and Lu,Z.J. (2017) POSTAR: a platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res.*, **45**, D104–D114.
52. Li,J.-H., Liu,S., Zhou,H., Qu,L.-H. and 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.
53. Seemann,S.E., Mirza,A.H., Hansen,C., Bang-Berthelsen,C.H., Garde,C., Christensen-Dalsgaard,M., Torarinsson,E., Yao,Z., Workman,C.T., Pociot,F., *et al.* (2017) The identification and functional annotation of RNA structures conserved in vertebrates. *Genome Res.*, **27**, 1371–1383.
54. Kelley,D. and Rinn,J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. 10.1186/gb-2012-13-11-r107.
55. Clark,M.B., Johnston,R.L., Inostroza-Ponta,M., Fox,A.H., Fortini,E., Moscato,P., Dinger,M.E. and Mattick,J.S. (2012) Genome-wide analysis of long noncoding RNA stability. *Genome Res.*, **22**, 885–98.
56. Mukherjee,N., Calviello,L., Hirsekorn,A., de Pretis,S., Pelizzola,M. and Ohler,U. (2017) Integrative classification of human coding and noncoding genes through RNA metabolism profiles. *Nat. Struct. Mol. Biol.*, **24**, 86–96.
57. Kapusta,A. and Feschotte,C. (2014) Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends Genet.*, **30**, 439–452.
58. Carlevaro-Fita,J., Rahim,A., Guigó,R., Vardy,L.A. and Johnson,R. (2016) Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells. *RNA*, **22**, 867–82.
59. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L.,

Barrell,D., Zadissa,A., Searle,S., *et al.* (2012) GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.

60. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–2.
61. Chen,G., Wang,Z., Wang,D., Qiu,C., Liu,M., Chen,X., Zhang,Q., Yan,G. and Cui,Q. (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.*, **41**, D983–D986.
62. 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.
63. 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.
64. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 9362–7.
65. Benjamini,Y. and 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.
66. Suzuki,K., Bose,P., Leong-Quong,R.Y.Y., Fujita,D.J. and Riabowol,K. (2010) REAP: A two minute cell fractionation method. *BMC Res. Notes*, **3**, 294.

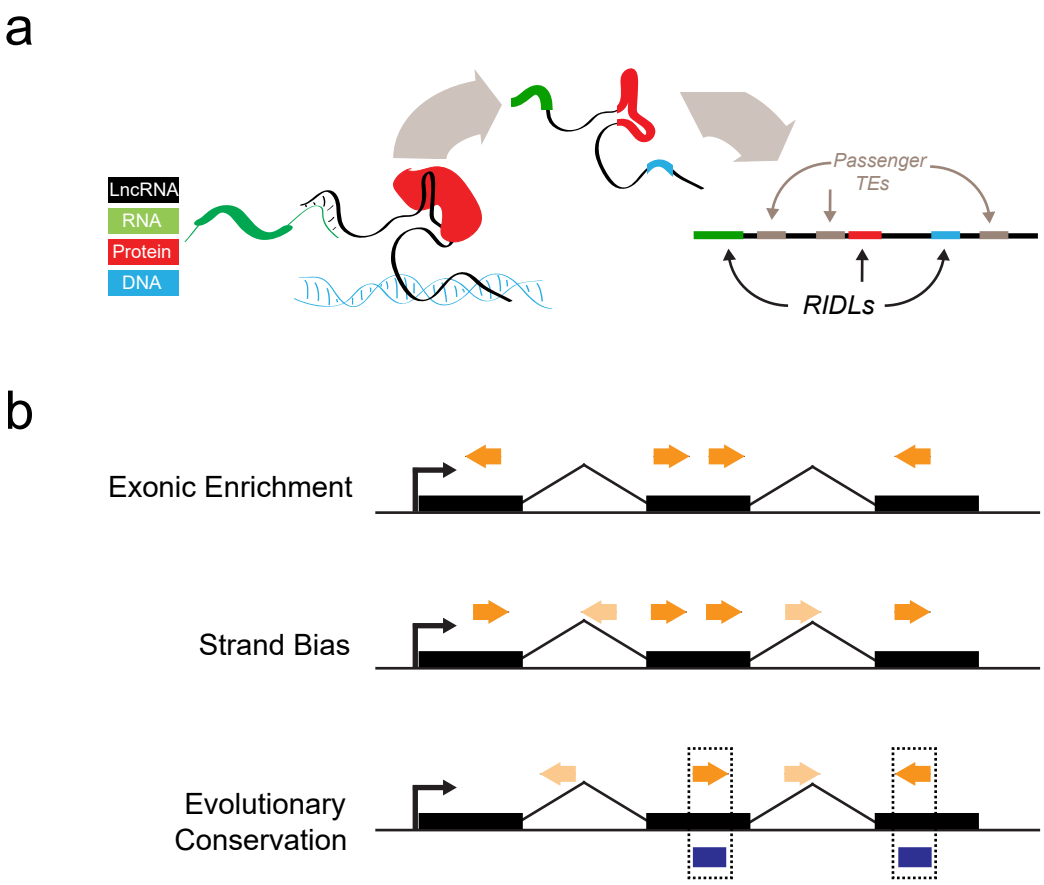

Figure 1: Repeat insertion domains of lncRNAs (RIDLs).
(a) In the "Repeat Insertion Domain of LncRNAs" (RIDL) model, exonically-inserted fragments of transposable elements contain pre-formed protein-binding (red), RNA-binding (green) or DNA-binding (blue) activities, that contribute to functionality 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) evolutionary conservation. Selection might be identified by comparing exonic TEs to a neutral population, for example those residing in lncRNA introns (light coloured arrows).

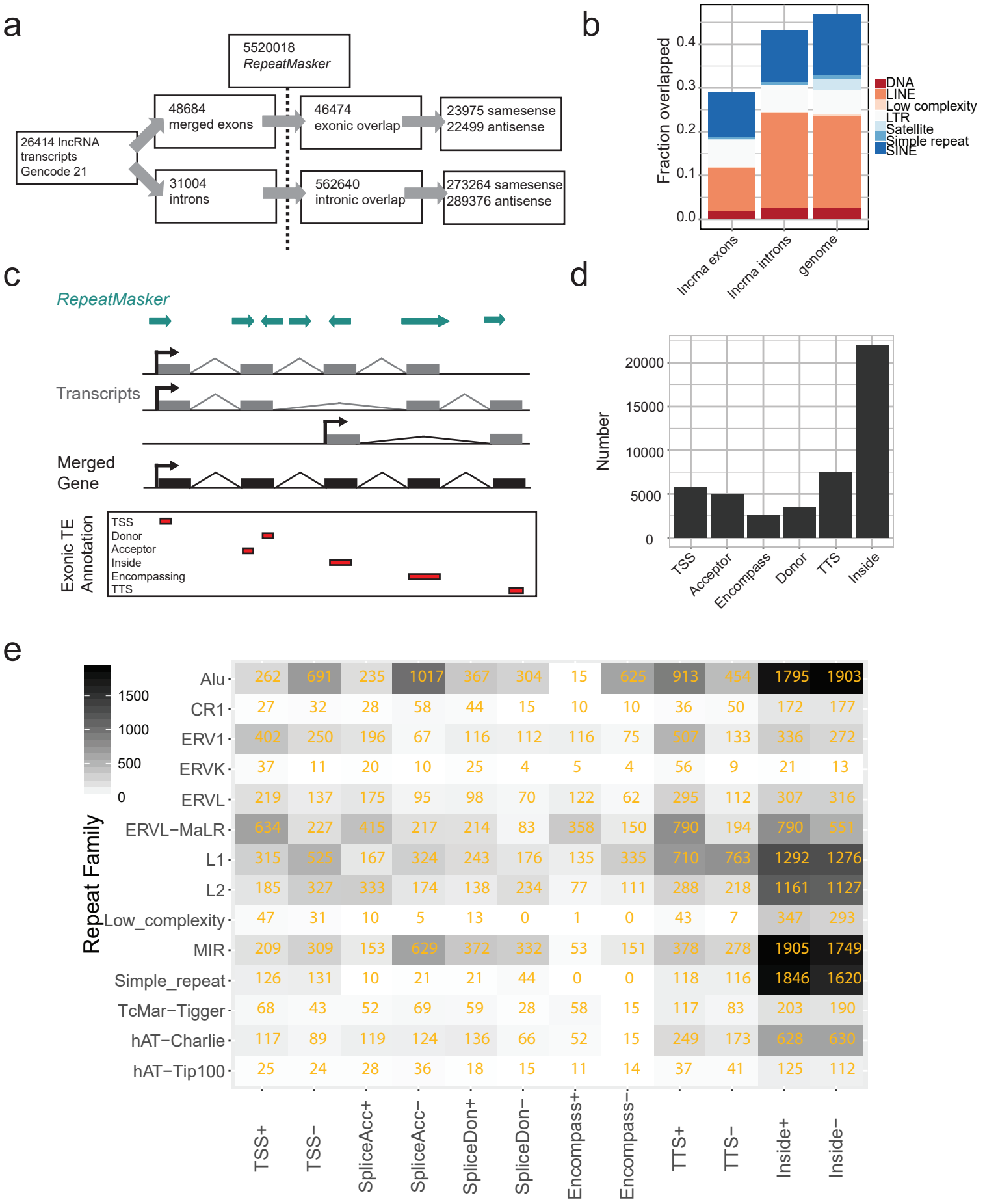

Figure 2: An exonic transposable element annotation with the GENCODE v21 lncRNA catalogue.
(a) Statistics for the exonic TE annotation process using GENCODE v21 lncRNAs. (b) The fraction of nucleotides overlapped by TEs for lncRNA exons and introns, and the whole genome. (c) Overview of the annotation 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. (d) Summary of classification breakdown for exonic TE annotation. (e) Classification of TE classes in exonic TE annotation. Numbers indicate instances of each type.

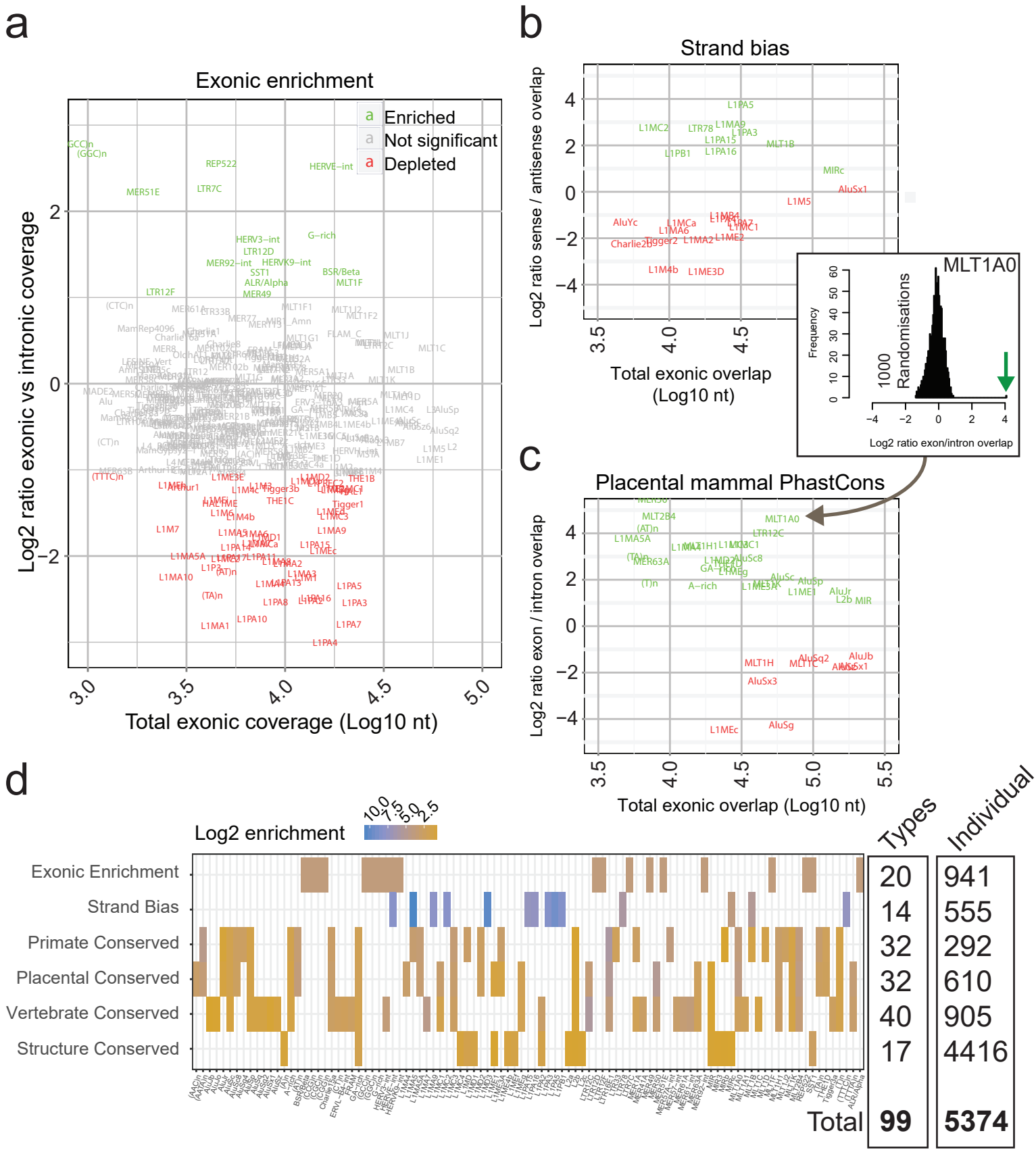

Figure 3: Evidence for selection on transposable elements in lncRNA exons.
(a) Figure shows, for every TE type, the enrichment of per nucleotide coverage in exons compared to introns (*y* axis) and overall exonic nucleotide coverage (*x* axis). Enriched TE types (at a 2-fold cutoff) are shown in green. (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 Supplementary Figure S1. Significantly-enriched TE types are shown in green. 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 Supplementary Figure S1. Significantly-enriched TE types are shown in green. 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".

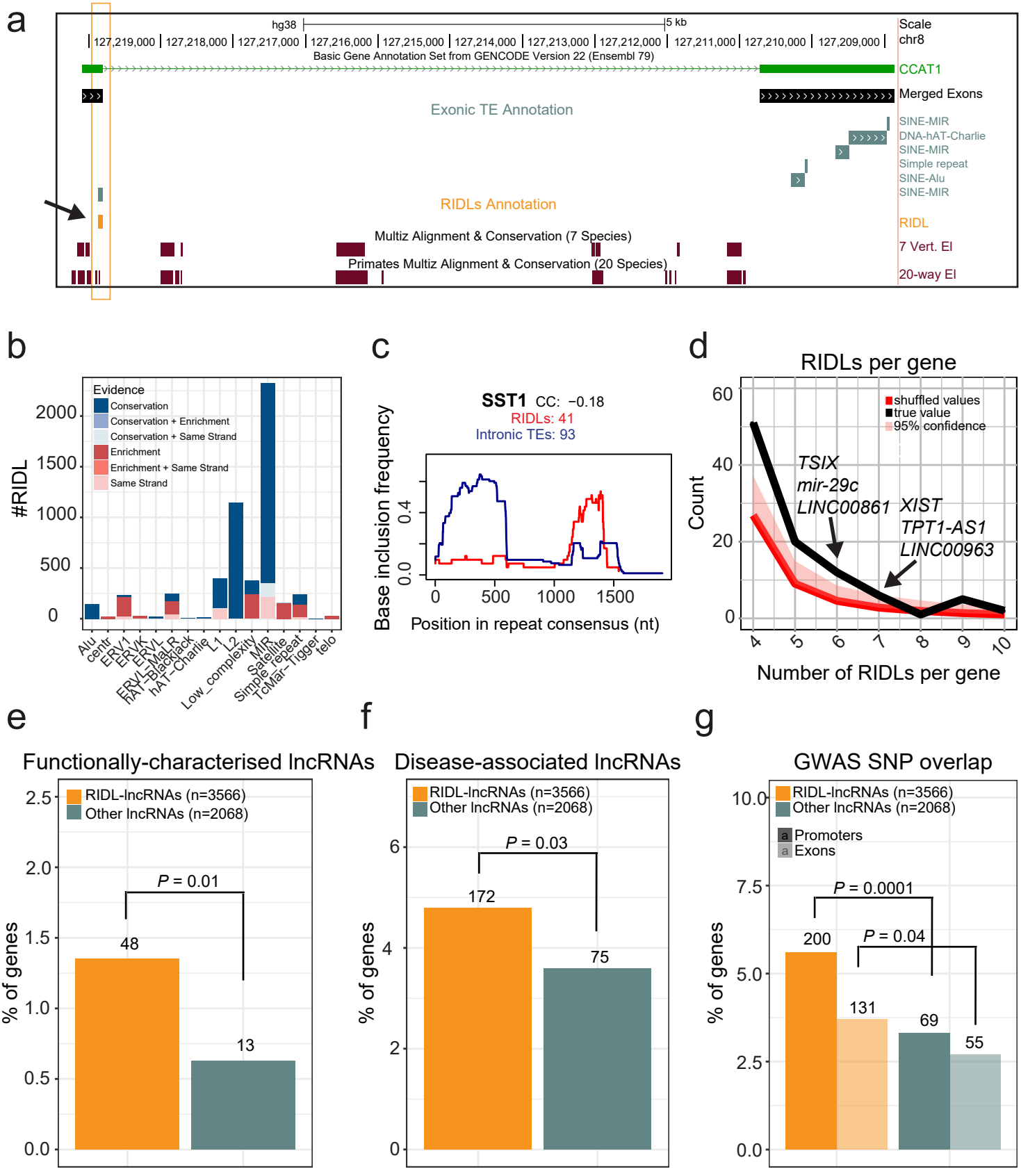

Figure 4: Annotated RIDLs and RIDL-lncRNAs

(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) The number of RIDL-lncRNAs, and a length-matched set of non-RIDL lncRNAs, which are present in the lncRNAdb database of functional lncRNAs (36). Numbers denote gene counts. Significance was calculated using Fisher's exact test. (f) As for (e) for lncRNAs present in disease- and cancer-associated lncRNA databases (see Materials and Methods). (g) As for (e) for lncRNAs containing at least one trait/disease-associated SNP in their promoter regions (darker colors) or exonic regions (lighter colors) (see Materials and Methods).

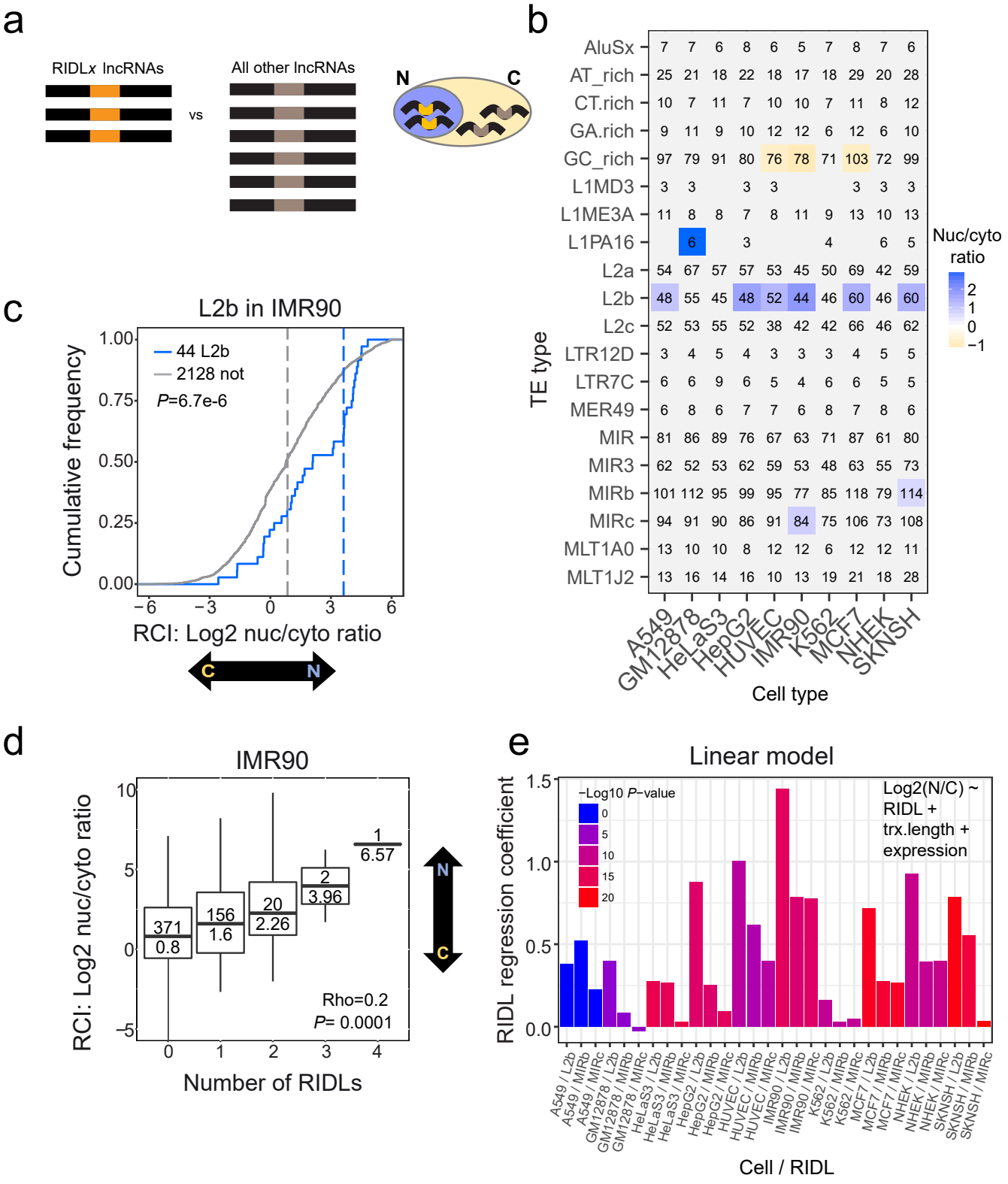

Figure 5: Correlation between RIDLs and host lncRNA nuclear/cytoplasmic localisation.
(a) Outline of in silico screen for localisation-regulating RIDLs. For each RIDL-type / cell-type combination, the nuclear/cytoplasmic localisation of RIDL-lncRNAs is compared to all other detected lncRNAs. (b) Results 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 IMR90 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 (Log2(N/C)) of their host transcript, while accounting for possible confounding factors of transcript length (trx.length) or whole-cell expression levels (expression).

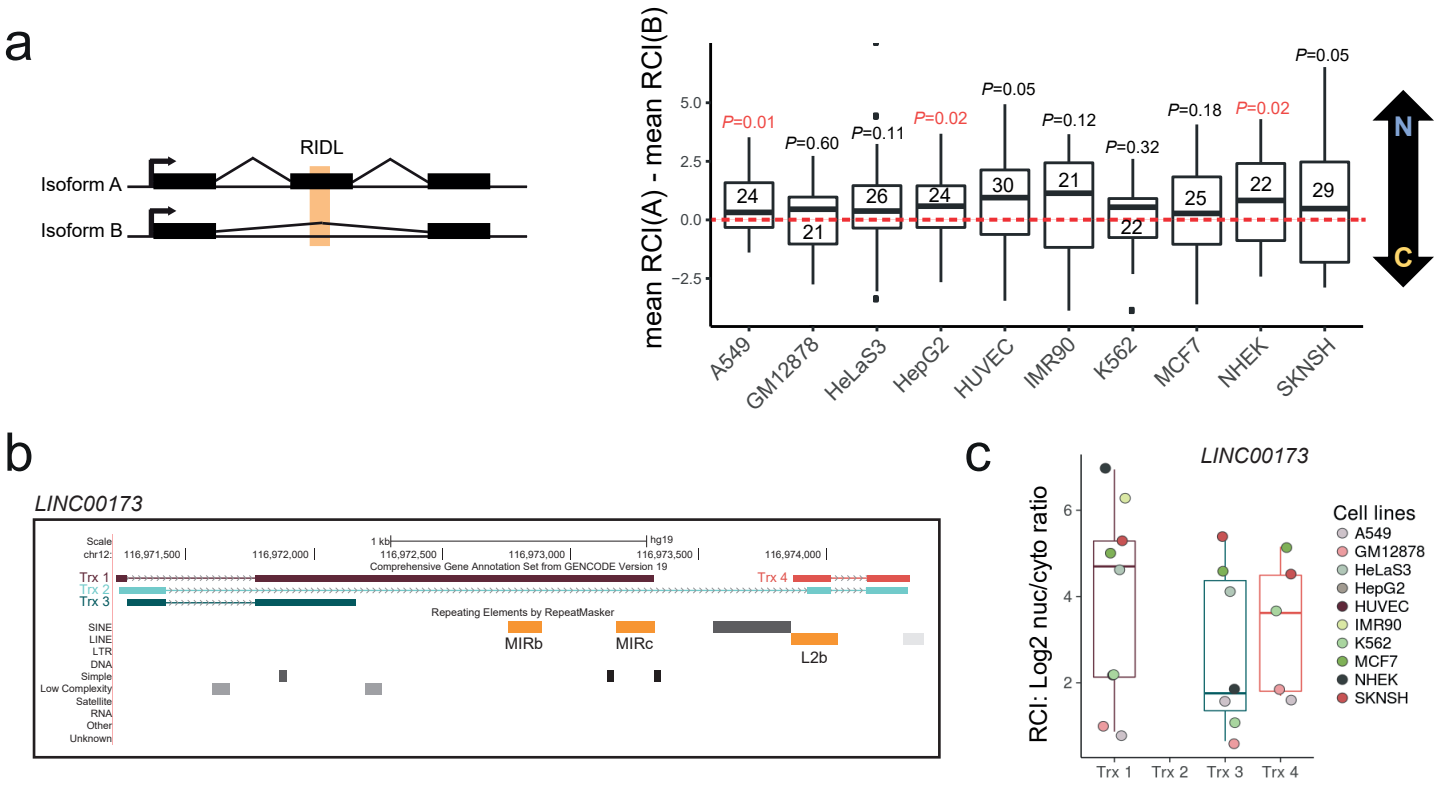

Figure 6: RIDLs correlate with differential localisation of lncRNA transcripts from the same locus.
(a) Distribution of differences between RCI mean of transcripts with nuclear RIDL (mean RCI(A)) and RCI mean of transcripts without nuclear RIDL (mean RCI(B)). A positive value indicates that RIDL-carrying 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. Here "nuclear RIDL" refers to L1AP16, MIRb, MIRc and L2b. *P*-values obtained from one-sided t-test are shown (in red when $P < 0.05$). (b) Example of a lncRNA gene locus containing a mixture of both RIDL- and non-RIDL-containing trancript isoforms: *LINC00173*. Displayed are four isoforms annotated by GENCODE v19 (Trx1-4) and repeat elements from *RepeatMasker*. MIRb, MIRc and L2b RIDLs are highlighted in orange. (c) Boxplot shows for every isoform of *LINC00173* the distribution of log2 nuclear/cytoplasmic ratio in ten cell lines. Horizontal bar indicates the median. Transcripts are named according to (b). Transcript 2 (Trx 2) is not detected in any of the cell lines.

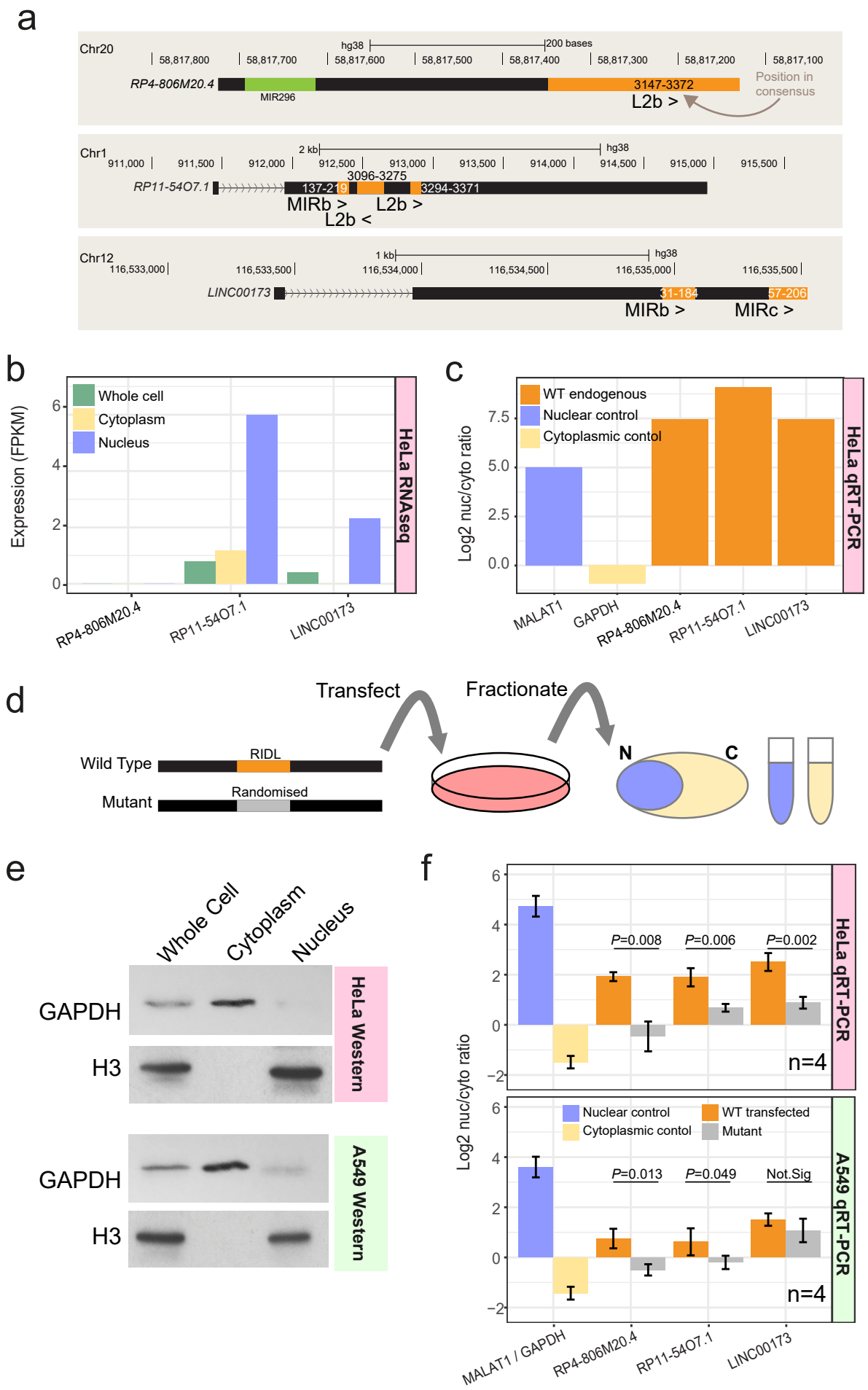

Figure 7: Disruption of RIDLs results in lncRNA relocalisation from nucleus to cytoplasm.
(a) Structures of candidate RIDL-lncRNAs. Orange indicates RIDL positions. For each RIDL, numbers indicate the position within the TE consensus, and its orientation with respect to the lncRNA is indicated by arrows (“>” for same strand, “<” for opposite strand). (b) Expression of the three lncRNA candidates as inferred from HeLa RNAseq (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, and error bars represent standard error of the mean. *P*-values for paired t-test (1 tail) are shown.