Ancient exapted transposable elements drive nuclear localisation of lncRNAs

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 localisation, remain largely unknown. One hypothetical source of such domains are neofunctionalised fragments of transposable elements (TEs), otherwise known as RIDLs (Repeat Insertion Domains of Long Noncoding RNA). We here 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 a total of 5374 RIDLs in 3566 gene loci. RIDL-containing lncRNAs are over-represented in functionally-validated and disease-associated genes. By integrating this RIDL map with global lncRNA subcellular localisation data, we uncover a cell-type-invariant and dose-dependent relationship between LINE2 and MIR elements and nuclear-cytoplasmic distribution. Experimental validations establish causality, showing that these RIDLs drive the nuclear localisation of their host lncRNA. Together these data represent the first map of TE-derived functional domains, and suggest a global role for TEs as nuclear localisation signals of lncRNAs.

Introduction

The human genome contains many thousands of long noncoding RNAs, of which at least a fraction are likely to have evolutionarily-selected biological functions (Ulitsky and Bartel 2013). Our current thinking is that, similar to proteins, lncRNAs 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 (Guttman and Rinn 2012; Rory Johnson and Guigó 2014; Mercer and Mattick 2013). 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) (Rory Johnson and Guigó 2014). TEs have been a major contributor 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; R. Johnson et al. 2006), microRNAs (Roberts, Cardin, and Borchert 2014), gene promoters (Faulkner et al. n.d.; Huda et al. 2011), and splice sites (Lev-Maor et al. 2003; Sela et al. 2007).

We recently proposed that TEs may contribute pre-formed functional domains to lncRNAs. We termed these "RIDLs" – *Repeat Insertion Domains of Long noncoding RNAs* (Rory Johnson and Guigó 2014). As RNA, TEs 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. n.d.). Diverse evidence also points to repetitive sequences forming intermolecular Watson-Crick RNA:RNA and RNA:DNA hybrids (Gong and Maquat 2011; Holdt et al. 2013; Rory Johnson and Guigó 2014). However, it is likely that *bona fide* RIDLs represent a small minority of the many exonic TEs, with the remainder being phenotypically-neutral "passengers".

A small but growing number of RIDLs have been described, reviewed in (Rory 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, a number of transcriptome-wide maps of lncRNA-linked TEs have been presented, showing how TEs have contributed extensively to lncRNA gene evolution (Kapusta et al. 2013; D. Kelley and Rinn 2012)(Schmitt et al. 2016)(Hezroni et al. 2015). However, there has been no attempt so far to create passenger-free maps of RIDLs with evidence for functionality in the context of mature lncRNA molecules.

Subcellular localisation, and the domains controlling it, exert a powerful influence over lncRNA functions (reviewed in (L.-L. Chen 2016)). 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 (K. Zhang et al. 2014). This is reflected in diverse patterns of lncRNA localization (Cabili et al. 2015; Mas-Ponte et al. 2017b). If lessons learned from mRNA are also valid for lncRNAs, then short sequence motifs binding to RNA binding proteins (RBPs) will be an important mechanism (Martin and Ephrussi 2009). This was recently demonstrated for the *BORG* lncRNA, where a pentameric motif was

shown to mediate nuclear retention (B. Zhang et al. 2014). Most interestingly, another study implicated an inverted pair of Alu elements in nuclear retention of *lincRNA-P21* (Chillón and Pyle 2016). 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 a potent nuclear localisation on their host transcripts.

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 GTF annotations were downloaded from www.gencodegenes.org (Harrow et al. 2012). 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(Quinlan and Hall 2010). *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. These intronic regions were assigned the strand of the host gene.

The RepeatMasker annotation file was then compared to these exons, and intersections are recorded and then 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 S1).

RIDL identification

Using this Exonic TE Annnotation, we identified the subset of those 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 that has >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 (Lev-Maor et al. 2003; Sela et al. 2007). For all TE instances, the "relative strand" was calculated: positive, if the annotated TE strand matches that of the host transcript; negative, if not. Then for every TE type, the ratio of relative strand sense/antisense was calculated. Statistical significance was calculated empirically: entire gene structures were randomly 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(Siepel et al. 2005)s, while the ECS conserved regions from obtained from Supplementary Data of Smith et al (Smith et al. 2013). 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. 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 evolutionary 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 S2.

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 (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 (G. 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 (Hindorff et al. 2009; Welter et al. 2014) (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. 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(Djebali et al. 2012b; Mas-Ponte et al. 2017a). 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. Cytoplasmic/nuclear ratio expression for each transcript was defined as (cytoplasm polyA+ RPKM)/(nuclear polyA+ RPKM), and only transcripts having non-zero values in both were considered. For each RIDL type and cell type in turn, the cytoplasmic/nuclear ratio distribution of RIDL-containing to non-RIDL-containing lncRNAs was compared using Wilcox test. Only RIDLs having at least 3 host 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).

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, or 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 S3.

Sub-cellular fractionation

Each pair of the wild type and the corresponding 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. The nuclear and cytoplasmic fractionation was carried out as described previously (Suzuki et al. 2010) 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 S4) and GoTaq qPCR master mix (Promega, # A6001). Human *GAPDH* mRNA and *MALAT1* lncRNA were used as cytoplasmic and nuclear markers, respectively.

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).

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 (Rory Johnson and Guigó 2014) (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 TEs in Gencode v21 lncRNAs

Our first aim was to create a comprehensive map of 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 their mature transcript.

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 model 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 (Kapusta et al. 2013). This may reflect selection against disruption of lncRNA transcripts by TEs, and hence is evidence for general lncRNA functionality.

Contribution of TEs to lncRNA gene structures

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

This exonic TE map represents the starting point for the identification of RIDLs, defined as the subset of TEs with evidence for functionality in the context of mature lncRNAs. In this and subsequent analyses, TEs are grouped by type as defined by RepeatMasker. 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 the background represented by intronic sequence (D. Kelley and Rinn 2012). 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 consistent with previous analyses by Kelley and Rinn (D. Kelley and Rinn 2012).

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 (Perepelitsa-Belancio and Deininger 2003). 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 (Perepelitsa-Belancio and Deininger 2003).

In a similar way, we investigated whether exonic TEs display a strand preference relative to host lncRNA (Rory Johnson and Guigó 2014). We took care to avoid a major source of bias: as shown above, many TSS and splice sites of lncRNA are contributed by TE, and such cases would lead to clear 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). We carried out the same operation for lncRNA introns (Supplementary Figure S1). 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 below. Other significantly enriched TE types include LTR78, MLT1B, and MIRc (Figure 3B).

In introns, we observed a widespread depletion of same-strand TE insertions (Supplementary Figure S1). This effect is modest yet statistically significant, and is especially true for LINE1 elements, possibly for reasons mentioned above. In contrast, 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 (Siepel et al. 2005) calculated on primate, placental mammal and vertebrate alignments; second, the more recently published “Evoluationarily Conserved Structures” (ECS) set of conserved RNA structures (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 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 (Figure C3, Supplementary Figure S1B,C,D) and ECS (Supplementary Figure S1) data. The majority of TEs do not exhibit significantly different conservation when found in exons or introns (Grey points). Depending on conservation data used, the method detects significant conservation for 17 to 40 TE types (Figure 3C). We also found a small number of TEs with high evolutionary rates specifically 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).

An annotation of RIDLs

We next combined all TE classes with evidence of functionality into a draft annotation of RIDLs, enriched for functional TE elements. This annotation combined altogether 99 TE types with at least one types 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).

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 S2.

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 (Jurka, Zietkiewicz, and Labuda 1995; Konkel, Walker, and Batzer 2010; Smith et al. 2013).

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 S5). 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 RIDL-carrying lncRNAs. We first asked whether RIDLs are randomly distributed amongst lncRNAs, or else 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 (Quek et al. 2015) (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) (Nissan et al. 2012) 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 (Rory Johnson and Guigó 2014), 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 (Rory 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 of RIDL-carrying lncRNAs can be found in Supplementary File S6.

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. 2017a). We next asked whether RIDLs might define lncRNA localisation (Hacisuleyman et al. 2016; B. Zhang et al. 2014)(Chillón and Pyle 2016) (Figure 5A). Beginning with normalised subcellular RNAseq data from the LncATLAS database (Mas-Ponte et al. 2017a), based on 10 ENCODE cell lines (Djebali et al. 2012a), 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 localisation: 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 difference cell types (Figure 5B,C). Furthermore, the degree of nuclear localisation increases in lncRNAs as a function of the number of RIDLs 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). These data suggest that certain types of RIDL may regulate the subcellular localisation of lncRNAs.

RIDLs play a causative role in lncRNA nuclear localisation

To more directly test whether RIDLs play a causative role in nuclear localisation, 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 6A,B and Supplementary Figure S6). Nuclear localisation of these candidates was validated using qRTPCR (Figure 6C and Supplementary Figure S7).

We designed an assay to compare the localisation of transfected lncRNAs carrying wild-type RIDLs, and mutated versions where RIDL sequence has been randomised ("Mutant") (Figure 6D, full sequences available in Supplementary File 3). Wild-type and Mutant lncRNAs were overexpressed in cultured cells and their localisation evaluated by fractionation. Overexpressed copies were many fold greater than endogenous lncRNAs (Supplementary Figure S8). Fractionation quality was verified by Western blotting (Figure 6E) and qRT-PCR (Figure 6F), and we found negligible levels of DNA contamination in cDNA samples (Supplementary Figure S9).

Comparison of Wild-Type and Mutant lncRNAs demonstrated a potent and consistent impact of RIDLs on nuclear-cytoplasmic localisation: for all three candidates, loss of RIDL sequence resulted in relocalisation of the host transcript from nucleus to cytoplasm (Figure 6F). Overexpressed Wild-Type lncRNA copies had slightly lower levels of nuclear enrichment compared to endogenous transcripts, and the two types were distinguished by primers matching a plasmid-encoded region (Supplementary File S4). We repeated these experiments in another cell line, A549, and observed similar effects (Supplementary Figure S7). These results demonstrate a role for L2b, MIRb and MIRc elements in nuclear localisation of 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 that TEs can drive the nuclear localisation of lncRNAs. A global correlation analysis of TEs and RNA localisation data revealed a handful of TEs, most notably LINE2b, MIRb and MIRc, strongly correlate with the degree of nuclear retention 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 transcript localisation.

These data are consistent with our previously-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 (Rory Johnson and Guigó 2014). In the context of localisation, RIDLs could mediate nuclear retention due to their described interactions with nuclear proteins (D. R. Kelley et al. 2014), 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 (Cordaux and Batzer n.d.). Both have previously been associated with acquired roles in the context of genomic DNA (Jjingo et al. 2014)(R. Johnson et al. 2006). Although the evolutionary history of lncRNA 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, Kellis, and Garber 2014). 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(Chuong, Elde, and Feschotte 2017). Given that nuclear retention is at odds with the primary needs of TE RNAs 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. 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.

This study 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)(D. Kelley and Rinn n.d.)(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 the present

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, and the concept of RIDLs in general, may help to address two longstanding questions in the lncRNA field.

How does lncRNA primary sequence evolve new functions so rapidly? The insertion of TE-derived "pre-fabricated" sequence domains, with protein-, RNA- or DNA-binding capacity, confers rapid acquisition of new molecular activities. There is a variety of evidence that TE sequences contain abundant, latent activity of this type (reviewed in (Rory Johnson and Guigó 2014)).

Why do lncRNAs have nuclear-restricted localisation? Although they are readily detected in the cytoplasm, it is nevertheless true that lncRNAs tend 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. 2017a). 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)(B. Zhang et al. 2014). 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 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.

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) 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 1. 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 1. 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 (Quek et al. 2015). 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) We hypothesized that RIDLs may govern the subcellular localization of host lncRNAs.

(B) Results of an *in silico* screen of RIDL types (rows) against lncRNA nuclear/cytoplasmic localisation in human cell lines (columns). For every RIDL type (rows) in every cell type (columns), the localisation of RIDL-containing lncRNAs was compared to all other detected lncRNAs. Significantly-different pairs 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.

(C) The nuclear-cytoplasmic localization of lncRNAs carrying L2b RIDLs in IMR90 cells. Blue indicates lncRNAs carrying $\geq$1 L2b RIDLs, grey indicates all other detected lncRNAs ("not"). Dashed lines represent medians. Significance was calculated using Wilcoxon test.

(D) The nuclear-cytoplasmic ratio of lncRNAs broken down by the number of significant RIDLs that they carry (L1PA16, L2b, MIRb, MIRc). Correlation coefficient (Rho) between number of RIDLs

and nuclear-cytoplasmic ratio of lncRNAs and the corresponding *p*-value (P) are indicated in the plot. In each box, upper value indicates the number of lncRNAs, and lower value the median.

Figure 6: 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) Experimental design

(C) Expression of the three lncRNA candidates as inferred from HeLa RNAseq (Djebali et al. 2012c).

(D) Nuclear/cytoplasmic localisation of endogenous candidate lncRNA copies in wild-type HeLa cells, as measured by qRT-PCR.

(E) The purity of HeLa 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. 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.

Supplementary Figure Legends

Supplementary Figure S1: (A) Figure shows, for every TE type, the ratio of intronic nucleotide coverage in sense vs antisense configuration. "Sense" here is defined as sense of TE annotation relative to the sense of the overlapping intron. 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). (B-D) As for (A), but *y* axis records the ratio of nucleotide overlap in exons vs introns by PhastCons Primate-conserved elements, PhastCons placental-conserved elements, PhastCons vertebrate-conserved elements and Evolutionarily conserved structure, respectively.

Supplementary Figure S2: Evolutionary analysis using 6-mammal alignments. Rows represent human RIDLs. In each column (6 different species) they are coloured in red when they are conserved in the given species and otherwise in light yellow.

Supplementary Figure S3: Density distribution of RIDL-carrying lncRNAs based on the number of RIDLs per kb of exons that they contain. Dotted line indicates median.

Supplementary Figure S4: Creating an exon-length-matched sample of non-RIDL lncRNAs. (A) Exonic length distribution of RIDL-carrying and non-RIDL GENCODE v21 lncRNAs ("others"). (B) Exonic length distribution of RIDL-carrying lncRNAs and a length-matched sample of non-RIDL lncRNAs. The latter were used for all comparisons at gene level.

Supplementary Figure S5: Cytoplasmic-nuclear localization in IMR90. Yellow represents lncRNAs carrying one or more copies of GC-rich RIDL elements, grey represents all other detected lncRNAs. Dashed lines indicate the median of each group.

Supplementary Figure S6: Barplot generated by LncATLAS webserver (Mas-Ponte et al. 2017b). Bars indicate log2-transformed cytoplasmic/nuclear expression ratios (CN RCI) in 15 cell lines for three candidate genes. Colours represent the total nuclear expression of each gene in each cell line.

Supplementary Figure S7: Experimental data for A549 cells. (A) Expression data for three candidate lncRNAs calculated using RNAseq data (Djebali et al. 2012a). (B) Validation of nuclear/cytoplasmic localisation of candidate genes in wild-type cells by qRT-PCR. (C) Validation of subcellular fractionation by Western blot, using GAPDH/Histone H3 proteins as cytoplasmic/nuclear markers, respectively. (D) Nuclear/cytoplasmic localisation of transfected candidate lncRNAs.

Supplementary Figure S8: Estimating overexpression of transfected lncRNA candidates. Data are shown individually for each of four biological replicates. *y* axis represents the fold change of transfected gene level compared to the endogeneous level estimated from wild-type cells.

Supplementary Figure 9: DNA contamination check in RNA Isolated from different subcellular fractions. Total RNA was isolated from different subcellular fractions using RNA isolation kit as described in Materials and Methods. Equal amount of RNAs with (+RT) and without (-RT) reverse transcription, were subjected to PCR (40 cycles) with an exonic primer. The reaction products were electrophoresed on a 1.5 % agarose gel stained with SYBR safe. The T, C, N represent the RNA isolated from total, cytoplasmic and nuclear fractions respectively.

Supplementary Files

Supplementary File S1: File is in BED format with coordinates for human genome assembly GRCh38 / hg38. The file contains the coordinates of genomic nucleotides that overlap both TEs and merged exons of lncRNA genes.

The 4th "name" column contains information on the TE (from *RepeatMasker* annotation) in comma-separated format:

Chromosome of TE annotation, start of TE annotation, end of TE annotation, strand of TE annotation, TE name, TE class, TE family, left in repeat, end in repeat, begin in repeat, hypothetical start if TE were full length, hypothetical end if TE were full length.

In the 5th "score" column is indicated the TE classification:

+1 : tss.plus ; +2 : splice_acc.plus ; +3 : encompass.plus ; +4 : splice_don.plus; +5 : tts.plus ; +6 : inside.plus. The sign (+-) denotes whether the annotated strand of the TE lies on the same/opposite strand with respect to that of the overlapping lncRNA.

Supplementary File S2: File is in BED format with coordinates for human genome assembly GRCh38 / hg38. It was created by merging a subset of the exonic TEs from Supplementary File S1, and thus some lines may be composed of merged features.

The 4th "name" column contains information on the TE (from *RepeatMasker* annotation) in comma-separated format:

Chromosome of TE annotation, start of TE annotation, end of TE annotation, strand of TE annotation, TE name, TE class, TE family, left in repeat, end in repeat, begin in repeat, hypothetical start if TE were full length, hypothetical end if TE were full length.

Supplementary File S3: File contains wild type and mutant synthesized plasmid sequences. Coloured sequences indicate wild-type / scrambled RIDL sequences.

Supplementary File S4: Primer sequences and calculations of their efficiency.

Supplementary File S5: PDF images of insertion profiles of all RIDLs (blue) and intronic insertions (red). *x* axis shows the entire consensus sequence of a given TE type. *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. "Data" / "Reference" indicate the numbers of individual insertions considered for RIDLs / intronic insertions, respectively.

Supplementary File S6: List of GENCODE v21 lncRNAs carrying one or more RIDLs. First column contains gene ID and second column the number of RIDLs present in the lncRNA.

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 and their analysis. Samir Ounzain (CHUV) also contributed suggestions regarding experimental design. Julien Lagarde (CRG) kindly provided help in gene sampling 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

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57: 289–300. https://www.jstor.org/stable/2346101 (August 25, 2017).

Blackwell, Benjamin J et al. "Protein Interactions with piALU RNA Indicates Putative Participation of retroRNA in the Cell Cycle, DNA Repair and Chromatin Assembly." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3383447/pdf/mge-2-26.pdf (August 25, 2017).

Bourque, Guillaume et al. 2008. "Evolution of the Mammalian Transcription Factor Binding Repertoire via Transposable Elements." *Genome research* 18(11): 1752–62. http://www.ncbi.nlm.nih.gov/pubmed/18682548 (August 25, 2017).

———. 2009. "Transposable Elements in Gene Regulation and in the Evolution of Vertebrate Genomes." *Current Opinion in Genetics & Development* 19(6): 607–12. http://www.sciencedirect.com/science/article/pii/S0959437X09001725 (August 25, 2017).

Cabili, Moran N et al. 2015. "Localization and Abundance Analysis of Human lncRNAs at Single-Cell and Single-Molecule Resolution." *Genome biology* 16(1): 20. http://genomebiology.com/2015/16/1/20 (January 30, 2015).

Carrieri, Claudia et al. 2012. "Long Non-Coding Antisense RNA Controls Uchl1 Translation through an Embedded SINEB2 Repeat." *Nature* 491(7424): 454–57. http://www.ncbi.nlm.nih.gov/pubmed/23064229 (February 10, 2015).

Chen, G. et al. 2013. "LncRNADisease: A Database for Long-Non-Coding RNA-Associated Diseases." *Nucleic Acids Research* 41(D1): D983–86. http://www.ncbi.nlm.nih.gov/pubmed/23175614 (December 24, 2016).

Chen, Ling-Ling. 2016. "Linking Long Noncoding RNA Localization and Function." *Trends in biochemical sciences*. http://www.ncbi.nlm.nih.gov/pubmed/27499234 (August 25, 2016).

Chillón, Isabel, and Anna M. Pyle. 2016. "Inverted Repeat *Alu* Elements in the Human lincRNA-p21 Adopt a Conserved Secondary Structure That Regulates RNA Function." *Nucleic Acids Research*: gkw599. http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkw599 (September 5, 2016).

Chuong, Edward B, Nels C Elde, and Cédric Feschotte. 2017. "Regulatory Activities of Transposable Elements: From Conflicts to Benefits." https://www.nature.com/nrg/journal/v18/n2/pdf/nrg.2016.139.pdf (August 25, 2017).

Clark, Michael B et al. 2012. "Genome-Wide Analysis of Long Noncoding RNA Stability." *Genome research* 22(5): 885–98. http://genome.cshlp.org/content/22/5/885.long (June 3, 2014).

Cordaux, Richard, and Mark A Batzer. "The Impact of Retrotransposons on Human Genome Evolution." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2884099/pdf/nihms-201920.pdf (August 29, 2017).

Derrien, Thomas et al. 2012. "The GENCODE v7 Catalog of Human Long Noncoding RNAs: Analysis of Their Gene Structure, Evolution, and Expression." *Genome research* 22(9): 1775–89. http://genome.cshlp.org/content/22/9/1775.long (May 23, 2014).

Djebali, Sarah et al. 2012a. "Landscape of Transcription in Human Cells." *Nature* 489(7414): 101–8. http://www.nature.com/doifinder/10.1038/nature11233 (April 20, 2017).

———. 2012b. "Landscape of Transcription in Human Cells." *Nature* 489(7414): 101–8. http://www.ncbi.nlm.nih.gov/pubmed/22955620 (May 11, 2017).

———. 2012c. “Landscape of Transcription in Human Cells.” *Nature* 489(7414): 101–8. http://www.ncbi.nlm.nih.gov/pubmed/22955620 (July 17, 2017).

Elisaphenko, Eugeny A et al. 2008. “A Dual Origin of the Xist Gene from a Protein-Coding Gene and a Set of Transposable Elements.” http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0002521&type=printable (August 25, 2017).

Faulkner, Geoffrey J et al. “The Regulated Retrotransposon Transcriptome of Mammalian Cells.” https://www.nature.com/ng/journal/v41/n5/pdf/ng.368.pdf (August 25, 2017).

Feschotte, Cédric. 2008. “Transposable Elements and the Evolution of Regulatory Networks.” *Nature reviews. Genetics* 9(5): 397–405. http://www.ncbi.nlm.nih.gov/pubmed/18368054 (September 1, 2017).

Gong, Chenguang, and Lynne E Maquat. 2011. “lncRNAs Transactivate STAU1-Mediated mRNA Decay by Duplexing with 3’ UTRs via Alu Elements.” *Nature* 470(7333): 284–88. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3073508&tool=pmcentrez&rendertype=abstract (March 6, 2015).

Guttman, Mitchell, and John L. Rinn. 2012. “Modular Regulatory Principles of Large Non-Coding RNAs.” *Nature* 482(7385): 339–46. http://www.nature.com/doifinder/10.1038/nature10887 (January 6, 2017).

Hacisuleyman, Ezgi et al. 2016. “Function and Evolution of Local Repeats in the Firre Locus.” *Nature Communications* 7: 11021. http://www.nature.com/doifinder/10.1038/ncomms11021 (November 20, 2016).

Harrow, J. et al. 2012. “GENCODE: The Reference Human Genome Annotation for The ENCODE Project.” *Genome Research* 22(9): 1760–74. http://genome.cshlp.org/cgi/doi/10.1101/gr.135350.111 (November 20, 2016).

Hezroni, Hadas et al. 2015. “Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species.” *Cell Reports* 11(7): 1110–22. http://linkinghub.elsevier.com/retrieve/pii/S2211124715004106 (September 1, 2017).

———. 2017. “A Subset of Conserved Mammalian Long Non-Coding RNAs Are Fossils of Ancestral Protein-Coding Genes.” *Genome Biology* 18(1): 162. http://www.ncbi.nlm.nih.gov/pubmed/28854954 (September 2, 2017).

Hindorff, Lucia A et al. 2009. “Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits.” *Proceedings of the National Academy of Sciences of the United States of America* 106(23): 9362–67. http://www.ncbi.nlm.nih.gov/pubmed/19474294 (August 30, 2017).

Holdt, Lesca M 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(7). http://journals.plos.org/plosgenetics/article/file?id=10.1371/journal.pgen.1003588&type=printable (August 25, 2017).

Hu, Boqin et al. 2017. “POSTAR: A Platform for Exploring Post-Transcriptional Regulation Coordinated by RNA-Binding Proteins.” *Nucleic Acids Research* 45(D1): D104–14. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw888 (September 2, 2017).

Huda, Ahsan, Nathan J. Bowen, Andrew B. Conley, and I. King Jordan. 2011. “Epigenetic Regulation of Transposable Element Derived Human Gene Promoters.” *Gene* 475(1): 39–48.

http://www.sciencedirect.com/science/article/pii/S0378111910004762 (August 27, 2017).

Jjingo, Daudi et al. 2014. “Mammalian-Wide Interspersed Repeat (MIR)-Derived Enhancers and the Regulation of Human Gene Expression.” *Mobile DNA* 5(1): 14. http://mobilednajournal.biomedcentral.com/articles/10.1186/1759-8753-5-14 (September 2, 2017).

Johnson, R. et al. 2006. “Identification of the REST Regulon Reveals Extensive Transposable Element-Mediated Binding Site Duplication.” *Nucleic Acids Research* 34(14): 3862–77. http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkl525 (August 25, 2017).

Johnson, Rory, and Roderic Guigó. 2014. “The RIDL Hypothesis: Transposable Elements as Functional Domains of Long Noncoding RNAs.” *RNA (New York, N.Y.)* 20(7): 959–76. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4114693&tool=pmcentrez&rendertype=abstract (April 29, 2015).

Jurka, J, E Zietkiewicz, and D Labuda. 1995. “Ubiquitous Mammalian-Wide Interspersed Repeats (MIRs) Are Molecular Fossils from the Mesozoic Era.” *Nucleic acids research* 23(1): 170–75. http://www.ncbi.nlm.nih.gov/pubmed/7870583 (August 29, 2017).

Kapusta, Aurélie et al. 2013. “Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs.” ed. Hopi E. Hoekstra. *PLoS genetics* 9(4): e1003470. http://dx.plos.org/10.1371/journal.pgen.1003470 (August 25, 2017).

Kelley, David R, David G Hendrickson, Danielle Tenen, and John L Rinn. 2014. “Transposable Elements Modulate Human RNA Abundance and Splicing via Specific RNA-Protein Interactions.” *Genome Biology* 15(12): 537. http://www.ncbi.nlm.nih.gov/pubmed/25572935 (September 2, 2017).

Kelley, David, and John Rinn. 2012. “Transposable Elements Reveal a Stem Cell-Specific Class of Long Noncoding RNAs.” *Genome Biology* 13(11): R107. http://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-11-r107 (August 25, 2017).

———. “Transposable Elements Reveal a Stem Cell-Specific Class of Long Noncoding RNAs.” https://genomebiology.biomedcentral.com/track/pdf/10.1186/gb-2012-13-11-r107?site=genomebiology.biomedcentral.com (August 25, 2017).

Konkel, Miriam K, Jerilyn A Walker, and Mark A Batzer. 2010. “LINEs and SINEs of Primate Evolution.” *Evolutionary anthropology* 19(6): 236–49. http://www.ncbi.nlm.nih.gov/pubmed/25147443 (August 29, 2017).

Lev-Maor, Galit, Rotem Sorek, Noam Shomron, and Gil Ast. 2003. “The Birth of an Alternatively Spliced Exon: 3’ Splice-Site Selection in Alu Exons.” *Science* 300(5623). http://science.sciencemag.org/content/300/5623/1288/tab-pdf (August 25, 2017).

Li, Jun-Hao et al. 2014. “starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA Interaction Networks from Large-Scale CLIP-Seq Data.” *Nucleic Acids Research* 42(D1): D92–97. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1248 (September 2, 2017).

Martin, Kelsey C, and Anne Ephrussi. 2009. “mRNA Localization: Gene Expression in the Spatial Dimension.” *Cell* 136(4): 719–30. http://www.ncbi.nlm.nih.gov/pubmed/19239891 (August 29, 2017).

Mas-Ponte, David et al. 2017a. “LncATLAS Database for Subcellular Localisation of Long Noncoding RNAs.” *RNA (New York, N.Y.)*: rna.060814.117. http://rnajournal.cshlp.org/lookup/doi/10.1261/rna.060814.117 (April 10, 2017).

———. 2017b. “LncATLAS Database for Subcellular Localization of Long Noncoding RNAs.” *RNA*

23(7): 1080–87. http://rnajournal.cshlp.org/lookup/doi/10.1261/rna.060814.117 (July 19, 2017).

Mercer, Tim R, and John S Mattick. 2013. “Structure and Function of Long Noncoding RNAs in Epigenetic Regulation.” *Nature Publishing Group* 20. https://www.nature.com/nsmb/journal/v20/n3/pdf/nsmb.2480.pdf (August 25, 2017).

Mukherjee, Neelanjan et al. 2017. “Integrative Classification of Human Coding and Noncoding Genes through RNA Metabolism Profiles.” *Nature structural & molecular biology* 24(1): 86–96. http://www.nature.com/doifinder/10.1038/nsmb.3325 (September 2, 2017).

Necsulea, Anamaria et al. 2014. “The Evolution of lncRNA Repertoires and Expression Patterns in Tetrapods.” *Nature* 505(7485): 635–40. http://www.ncbi.nlm.nih.gov/pubmed/24463510 (May 11, 2017).

Ning, Shangwei et al. 2016. “Lnc2Cancer: A Manually Curated Database of Experimentally Supported lncRNAs Associated with Various Human Cancers.” *Nucleic Acids Research* 44(D1): D980–85. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1094 (August 30, 2017).

Nissan, Aviram et al. 2012. “Colon Cancer Associated Transcript-1: A Novel RNA Expressed in Malignant and Pre-Malignant Human Tissues.” *International Journal of Cancer* 130(7): 1598–1606. http://www.ncbi.nlm.nih.gov/pubmed/21547902 (August 29, 2017).

Van Nostrand, Eric L et al. 2016. “Robust Transcriptome-Wide Discovery of RNA-Binding Protein Binding Sites with Enhanced CLIP (eCLIP).” *Nature Methods* 13(6): 508–14. http://www.ncbi.nlm.nih.gov/pubmed/27018577 (September 2, 2017).

Perepelitsa-Belancio, Victoria, and Prescott Deininger. 2003. “RNA Truncation by Premature Polyadenylation Attenuates Human Mobile Element Activity.” *Nature Genetics* 35(4): 363–66. http://www.nature.com/doifinder/10.1038/ng1269 (August 25, 2017).

Quek, Xiu Cheng et al. 2015. “lncRNAdb v2.0: Expanding the Reference Database for Functional Long Noncoding RNAs.” *Nucleic acids research* 43(Database issue): D168-73. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gku988 (December 24, 2016).

Quinlan, Aaron R, and Ira M Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics (Oxford, England)* 26(6): 841–42. http://www.ncbi.nlm.nih.gov/pubmed/20110278 (August 25, 2017).

Roberts, Justin T, Sara E Cardin, and Glen M Borchert. 2014. “Burgeoning Evidence Indicates That microRNAs Were Initially Formed from Transposable Element Sequences.” *Mobile Genetic Elements* 4(3): e29255. http://www.tandfonline.com/doi/abs/10.4161/mge.29255 (August 25, 2017).

Schmitt, Adam M. et al. 2016. “Long Noncoding RNAs in Cancer Pathways.” *Cancer Cell* 29(4): 452–63. http://linkinghub.elsevier.com/retrieve/pii/S1535610816300927 (June 28, 2016).

Seemann, Stefan E. et al. 2017. “The Identification and Functional Annotation of RNA Structures Conserved in Vertebrates.” *Genome Research* 27(8): 1371–83. http://www.ncbi.nlm.nih.gov/pubmed/28487280 (September 2, 2017).

Sela, Noa et al. 2007. “Comparative Analysis of Transposed Element Insertion within Human and Mouse Genomes Reveals Alu’s Unique Role in Shaping the Human Transcriptome.” 8(6). https://genomebiology.biomedcentral.com/track/pdf/10.1186/gb-2007-8-6-r127?site=genomebiology.biomedcentral.com (August 25, 2017).

Siepel, Adam et al. 2005. “Evolutionarily Conserved Elements in Vertebrate, Insect, Worm, and Yeast Genomes.” *Genome research* 15(8): 1034–50. http://www.ncbi.nlm.nih.gov/pubmed/16024819

(August 25, 2017).

Smith, M. A., T. Gesell, P. F. Stadler, and J. S. Mattick. 2013. "Widespread Purifying Selection on RNA Structure in Mammals." *Nucleic Acids Research* 41(17): 8220–36. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3783177&tool=pmcentrez&rendertype=abstract (April 20, 2015).

Suzuki, Keiko et al. 2010. "REAP: A Two Minute Cell Fractionation Method." *BMC Research Notes* 3(1): 294. http://www.ncbi.nlm.nih.gov/pubmed/21067583 (August 28, 2017).

Ulitsky, Igor, and David P. Bartel. 2013. "lincRNAs: Genomics, Evolution, and Mechanisms." *Cell* 154(1): 26–46. http://linkinghub.elsevier.com/retrieve/pii/S0092867413007599 (November 20, 2016).

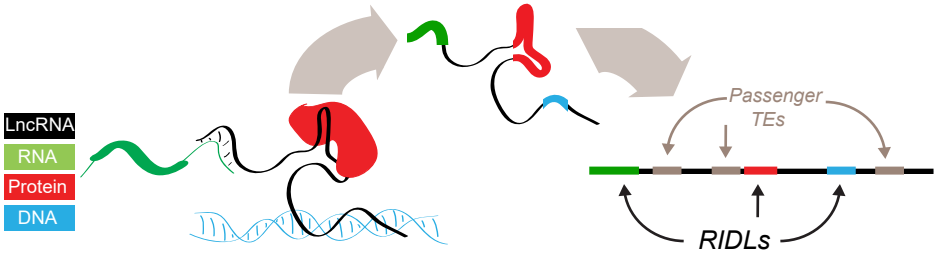
Washietl, Stefan, Manolis Kellis, and Manuel Garber. 2014. "Evolutionary Dynamics and Tissue Specificity of Human Long Noncoding RNAs in Six Mammals." *Genome research* 24(4): 616–28. http://www.ncbi.nlm.nih.gov/pubmed/24429298 (May 25, 2014).

Welter, Danielle et al. 2014. "The NHGRI GWAS Catalog, a Curated Resource of SNP-Trait Associations." *Nucleic Acids Research* 42(D1): D1001–6. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1229 (August 30, 2017).

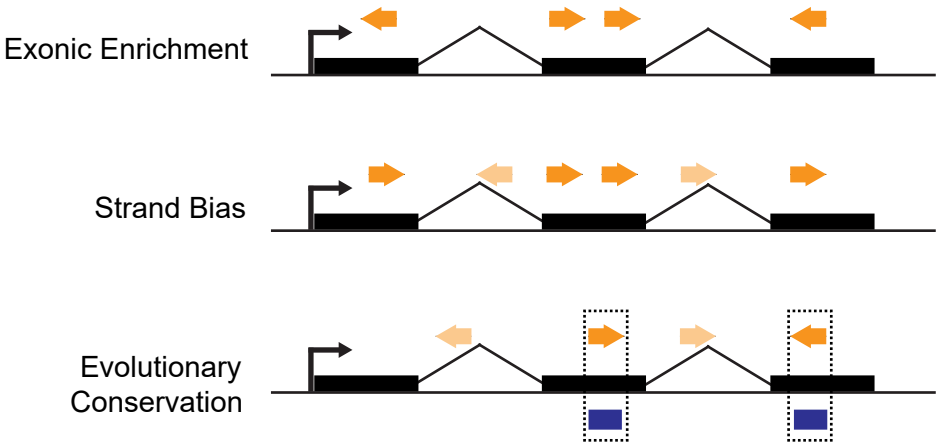
Zhang, B. et al. 2014. "A Novel RNA Motif Mediates the Strict Nuclear Localization of a Long Noncoding RNA." *Molecular and Cellular Biology* 34(12): 2318–29. http://mcb.asm.org/cgi/doi/10.1128/MCB.01673-13 (November 20, 2016).

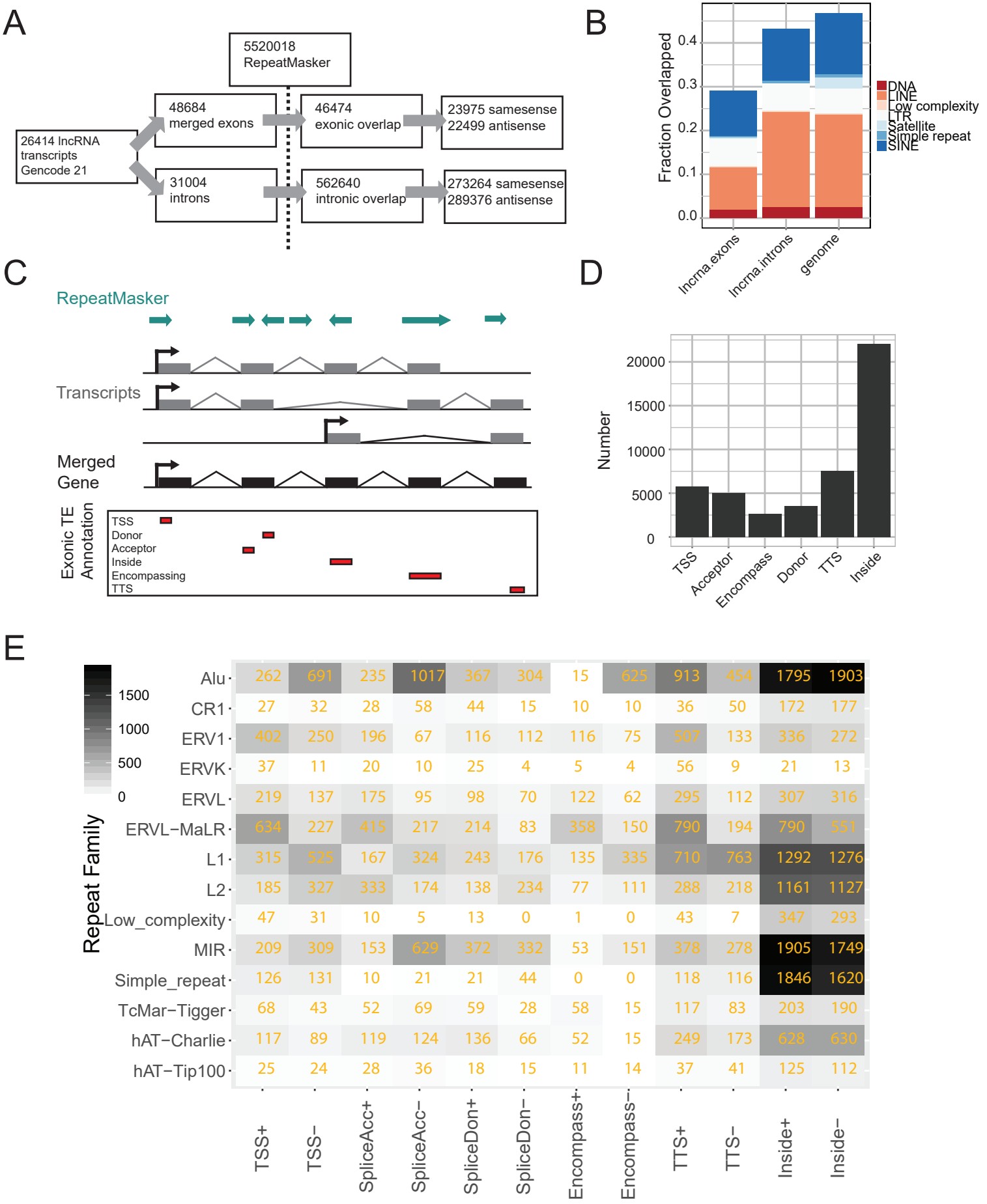
Zhang, Kun et al. 2014. "The Ways of Action of Long Non-Coding RNAs in Cytoplasm and Nucleus." *Gene* 547(1): 1–9. http://www.ncbi.nlm.nih.gov/pubmed/24967943 (March 17, 2015).

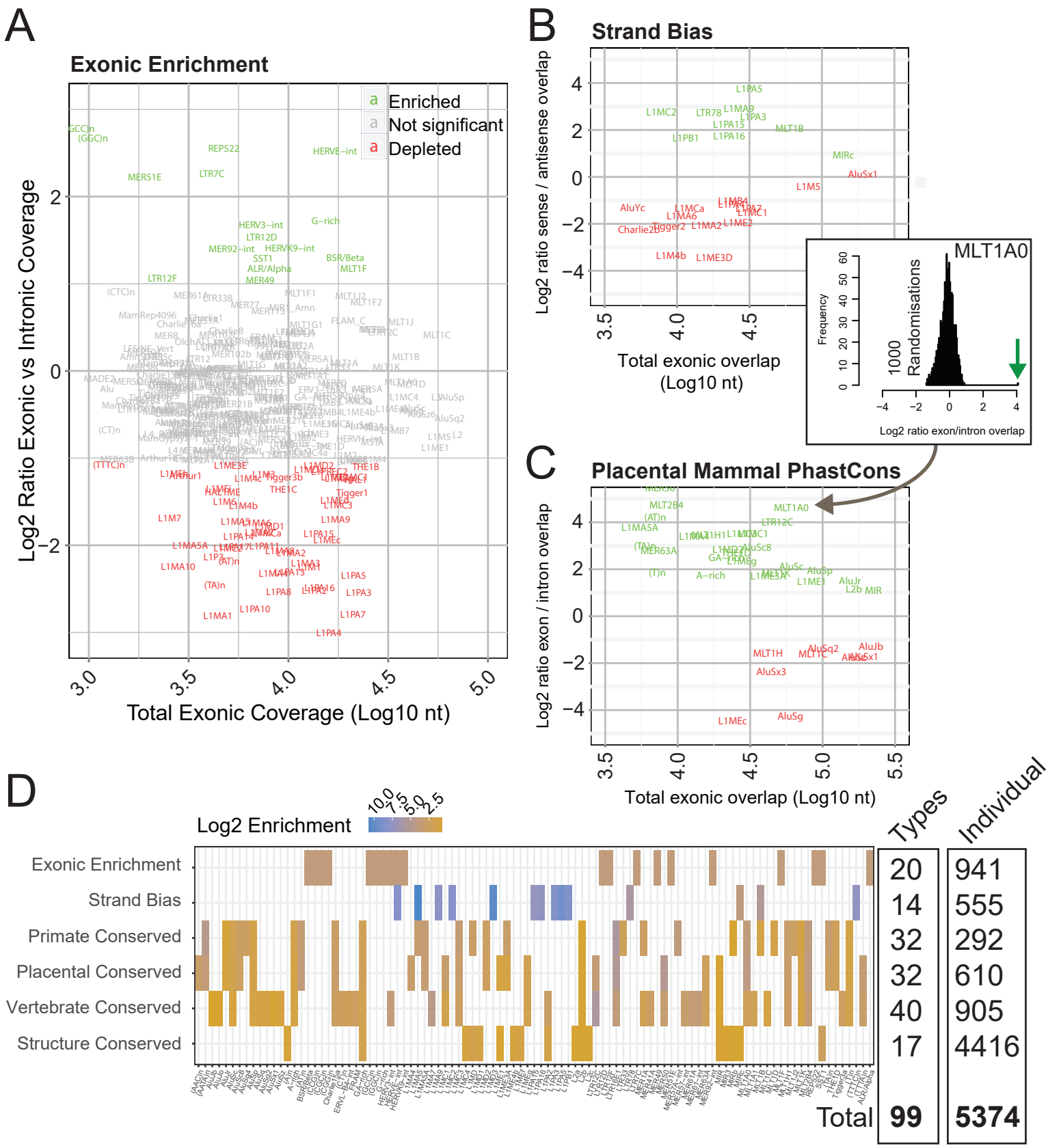
A

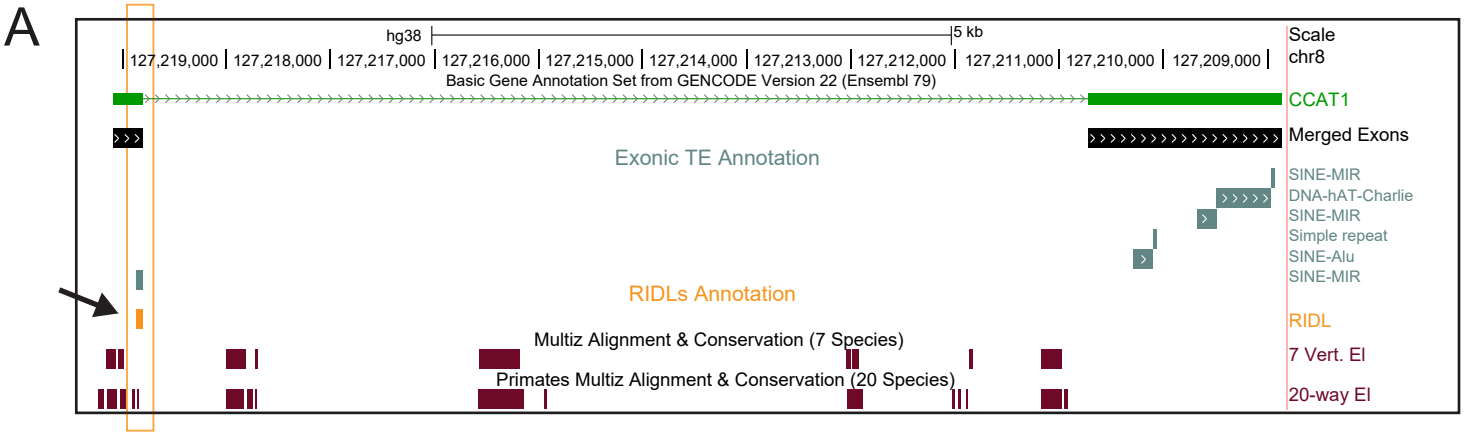

LncRNA
RNA
Protein
DNA
Passenger
TEs
RIDLs

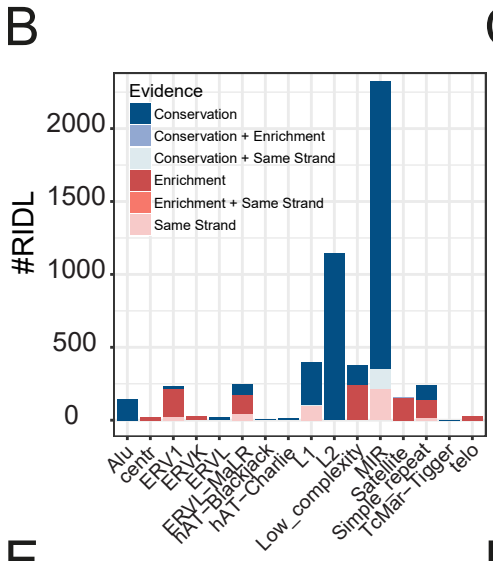
B

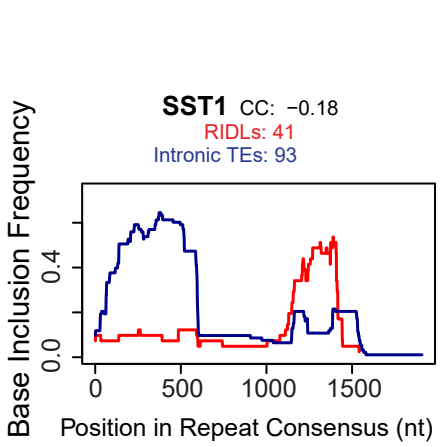

Exonic Enrichment
Strand Bias
Evolutionary
Conservation

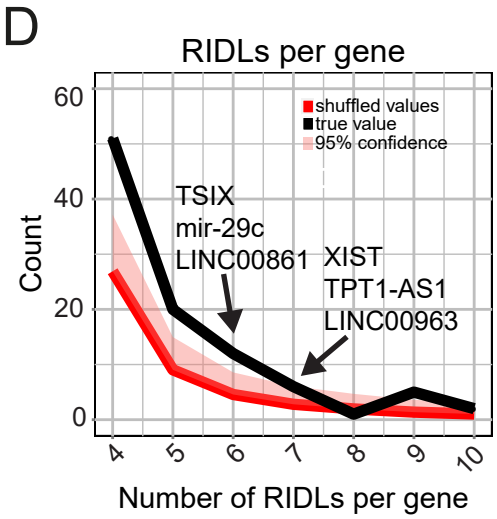

A
5520018
RepeatMasker
26414 lncRNA
transcripts
Gencode 21
48684
merged exons
31004
introns
46474
exonic overlap
562640
intronic overlap
23975 samesense
22499 antisense
273264 samesense
289376 antisense
B
Fraction Overlapped
0.0
0.1
0.2
0.3
0.4
lncrna.exons
lncrna.introns
genome
DNA
LINE
Low complexity
LTR
Satellite
Simple repeat
SINE
C
RepeatMasker
Transcripts
Merged
Gene
Exonic TE
Annotation
TSS
Donor
Acceptor
Inside
Encompassing
TTS
D
Number
0
5000
10000
15000
20000
TSS
Acceptor
Encompass
Donor
TTS
Inside
E
Repeat Family
1500
1000
500
0
Alu 262 691 235 1017 367 304 15 625 913 454 1795 1903
CR1 27 32 28 58 44 15 10 10 36 50 172 177
ERV1 402 250 196 67 116 112 116 75 507 133 336 272
ERVK 37 11 20 10 25 4 5 4 56 9 21 13
ERVL 219 137 175 95 98 70 122 62 295 112 307 316
ERVL−MaLR 634 227 415 217 214 83 358 150 790 194 790 551
L1 315 525 167 324 243 176 135 335 710 763 1292 1276
L2 185 327 333 174 138 234 77 111 288 218 1161 1127
Low_complexity 47 31 10 5 13 0 1 0 43 7 347 293
MIR 209 309 153 629 372 332 53 151 378 278 1905 1749
Simple_repeat 126 131 10 21 21 44 0 0 118 116 1846 1620
TcMar−Tigger 68 43 52 69 59 28 58 15 117 83 203 190
hAT−Charlie 117 89 119 124 136 66 52 15 249 173 628 630
hAT−Tip100 25 24 28 36 18 15 11 14 37 41 125 112
TSS+
TSS−
SpliceAcc+
SpliceAcc−
SpliceDon+
SpliceDon−
Encompass+
Encompass−
TTS+
TTS−
Inside+
Inside−

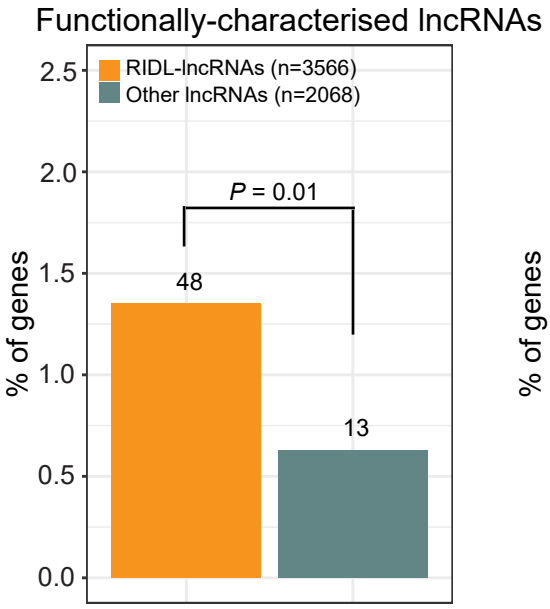

A
Exonic Enrichment
Enriched
Not significant
Depleted
Log2 Ratio Exonic vs Intronic Coverage
Total Exonic Coverage (Log10 nt)
B
Strand Bias
Log2 ratio sense / antisense overlap
Total exonic overlap (Log10 nt)
MLT1A0
1000 Randomisations
Frequency
Log2 ratio exon/intron overlap
C
Placental Mammal PhastCons
Log2 ratio exon / intron overlap
Total exonic overlap (Log10 nt)
D
Log2 Enrichment
Exonic Enrichment
Strand Bias
Primate Conserved
Placental Conserved
Vertebrate Conserved
Structure Conserved
Types
Individual
20 941
14 555
32 292
32 610
40 905
17 4416
Total 99 5374

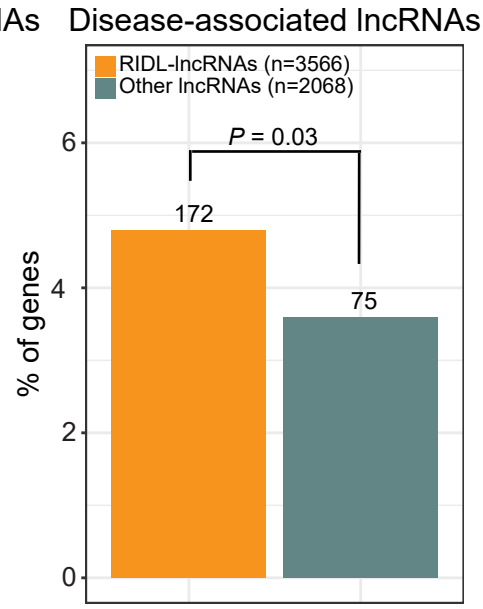

A
hg38
5 kb
Scale
chr8
127,219,000
127,218,000
127,217,000
127,216,000
127,215,000
127,214,000
127,213,000
127,212,000
127,211,000
127,210,000
127,209,000
Basic Gene Annotation Set from GENCODE Version 22 (Ensembl 79)
CCAT1
Merged Exons
Exonic TE Annotation
SINE-MIR
DNA-hAT-Charlie
SINE-MIR
Simple repeat
SINE-Alu
SINE-MIR
RIDLs Annotation
RIDL
Multiz Alignment & Conservation (7 Species)
7 Vert. EI
Primates Multiz Alignment & Conservation (20 Species)
20-way EI

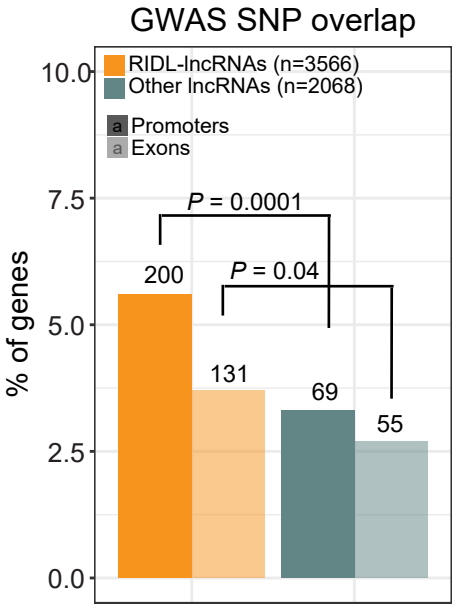

B
Evidence
Conservation
Conservation + Enrichment
Conservation + Same Strand
Enrichment
Enrichment + Same Strand
Same Strand
#RIDL
2000
1500
1000
500
0
Alu
centr
ERV1
ERVK
ERVL
ERVL-MaLR
hAT-Blackjack
hAT-Charlie
L1
L2
Low_complexity
MIR
Satellite
Simple repeat
TcMar-Tigger
telo

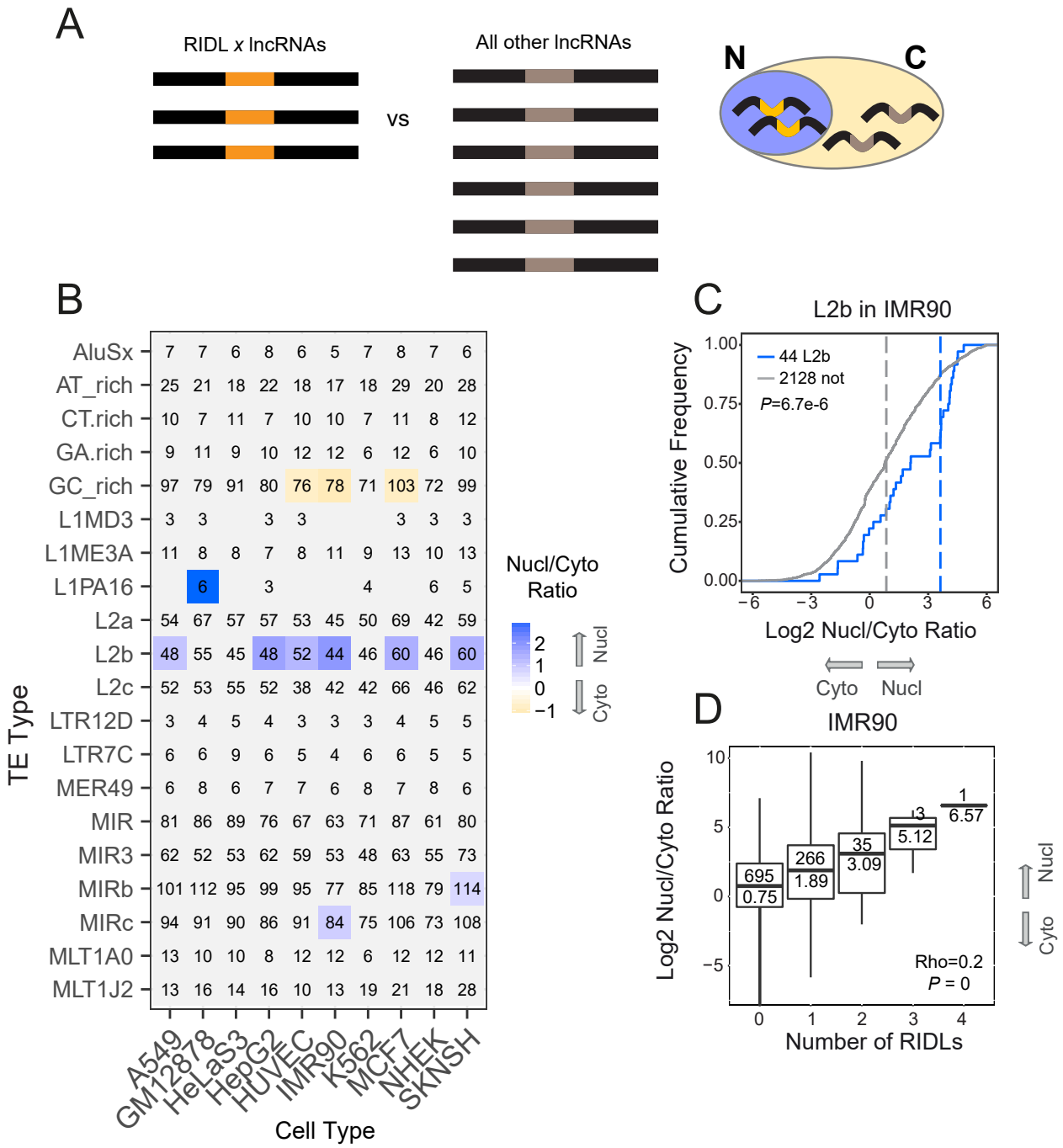

C
SST1 CC: −0.18
RIDLs: 41
Intronic TEs: 93
Base Inclusion Frequency
0.0
0.4
0
500
1000
1500
Position in Repeat Consensus (nt)

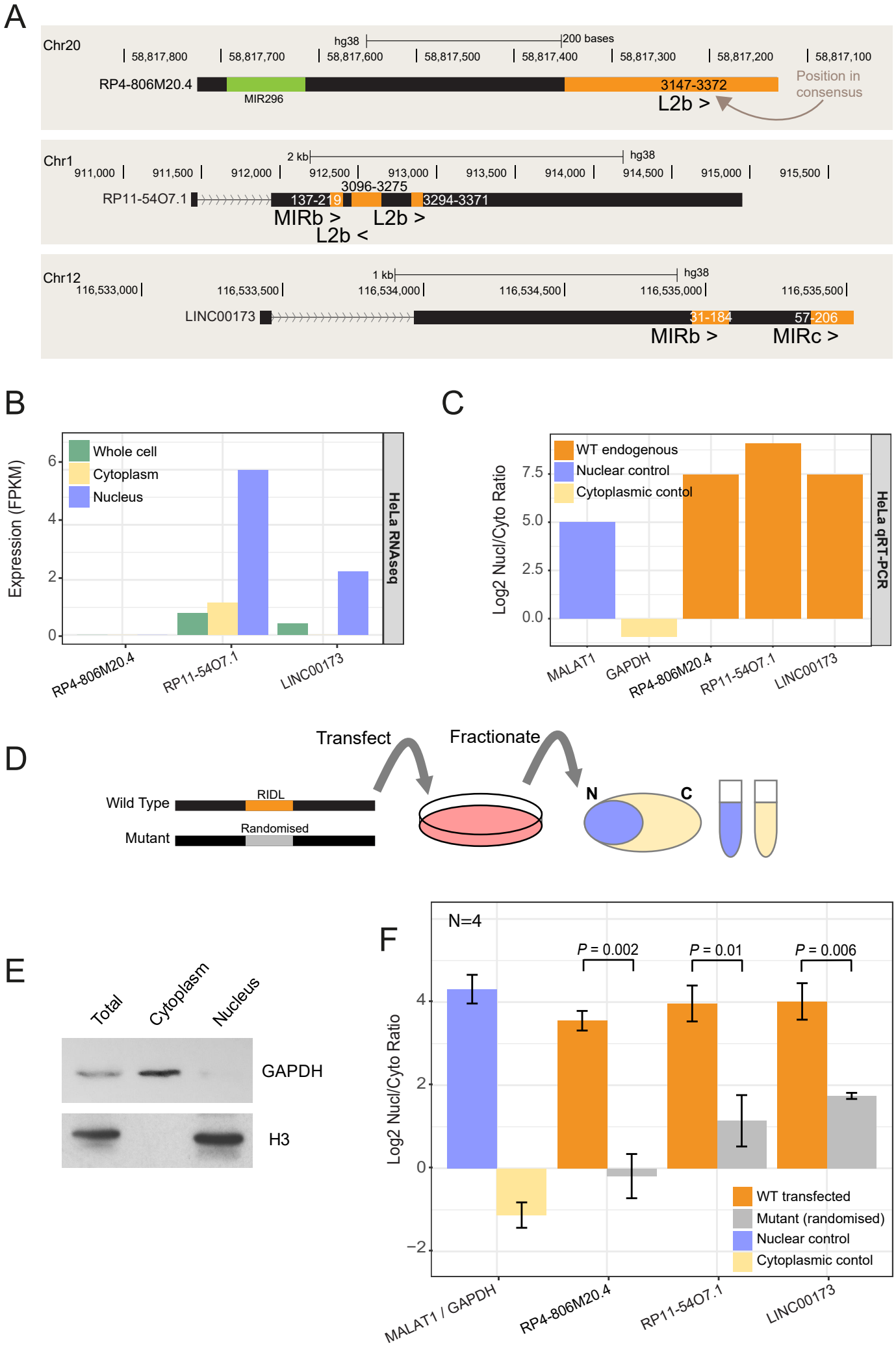

D
RIDLs per gene
shuffled values
true value
95% confidence
Count
60
40
20
0
TSIX
mir-29c
LINC00861
XIST
TPT1-AS1
LINC00963
4
5
6
7
8
9
10
Number of RIDLs per gene


E
Functionally-characterised lncRNAs
RIDL-lncRNAs (n=3566)
Other lncRNAs (n=2068)
% of genes
2.5
2.0
1.5
1.0
0.5
0.0
P = 0.01
48
13


F
Disease-associated lncRNAs
RIDL-lncRNAs (n=3566)
Other lncRNAs (n=2068)
% of genes
6
4
2
0
P = 0.03
172
75


G
GWAS SNP overlap
RIDL-lncRNAs (n=3566)
Other lncRNAs (n=2068)
Promoters
Exons
% of genes
10.0
7.5
5.0
2.5
0.0
P = 0.0001
P = 0.04
200
131
69
55


A
RIDL x lncRNAs
vs
All other lncRNAs
N
C
B
TE Type
AluSx 7 7 6 8 6 5 7 8 7 6
AT_rich 25 21 18 22 18 17 18 29 20 28
CT.rich 10 7 11 7 10 10 7 11 8 12
GA.rich 9 11 9 10 12 12 6 12 6 10
GC_rich 97 79 91 80 76 78 71 103 72 99
L1MD3 3 3 3 3 3 3 3 3
L1ME3A 11 8 8 7 8 11 9 13 10 13
L1PA16 6 3 4 6 5
L2a 54 67 57 57 53 45 50 69 42 59
L2b 48 55 45 48 52 44 46 60 46 60
L2c 52 53 55 52 38 42 42 66 46 62
LTR12D 3 4 5 4 3 3 3 4 5 5
LTR7C 6 6 9 6 5 4 6 6 5 5
MER49 6 8 6 7 7 6 8 7 8 6
MIR 81 86 89 76 67 63 71 87 61 80
MIR3 62 52 53 62 59 53 48 63 55 73
MIRb 101 112 95 99 95 77 85 118 79 114
MIRc 94 91 90 86 91 84 75 106 73 108
MLT1A0 13 10 10 8 12 12 6 12 12 11
MLT1J2 13 16 14 16 10 13 19 21 18 28
A549
GM12878
HeLaS3
HepG2
HUVEC
IMR90
K562
MCF7
NHEK
SKNSH
Cell Type
Nucl/Cyto Ratio
2
1
0
−1
Nucl
Cyto
C
L2b in IMR90
44 L2b
2128 not
P=6.7e-6
Cumulative Frequency
1.00
0.75
0.50
0.25
0.00
−6
−3
0
3
6
Log2 Nucl/Cyto Ratio
Cyto
Nucl
D
IMR90
Log2 Nucl/Cyto Ratio
10
5
0
−5
695
0.75
266
1.89
35
3.09
3
5.12
1
6.57
Rho=0.2
P = 0
0
1
2
3
4
Number of RIDLs
Nucl
Cyto


A
Chr20
hg38
200 bases
58,817,800
58,817,700
58,817,600
58,817,500
58,817,400
58,817,300
58,817,200
58,817,100
RP4-806M20.4
MIR296
3147-3372
L2b >
Position in consensus
Chr1
2 kb
hg38
911,000
911,500
912,000
912,500
913,000
913,500
914,000
914,500
915,000
915,500
RP11-54O7.1
3096-3275
137-219
3294-3371
MIRb >
L2b >
L2b <
Chr12
1 kb
hg38
116,533,000
116,533,500
116,534,000
116,534,500
116,535,000
116,535,500
LINC00173
31-184
57-206
MIRb >
MIRc >
B
Expression (FPKM)
Whole cell
Cytoplasm
Nucleus
HeLa RNAseq
RP4-806M20.4
RP11-54O7.1
LINC00173
C
Log2 Nucl/Cyto Ratio
WT endogenous
Nuclear control
Cytoplasmic contol
HeLa qRT-PCR
MALAT1
GAPDH
RP4-806M20.4
RP11-54O7.1
LINC00173
D
Transfect
Fractionate
Wild Type
RIDL
Mutant
Randomised
N
C
E
Total
Cytoplasm
Nucleus
GAPDH
H3
F
N=4
P = 0.002
P = 0.01
P = 0.006
Log2 Nucl/Cyto Ratio
WT transfected
Mutant (randomised)
Nuclear control
Cytoplasmic contol
MALAT1 / GAPDH
RP4-806M20.4
RP11-54O7.1
LINC00173