Genomic Accumulation of Retrotransposons Was Facilitated by Repressive RNA-Binding Proteins: A Hypothesis

Jan Attig* and Jernej Ule*

Retrotransposon-derived elements (RDEs) can disrupt gene expression, but are nevertheless widespread in metazoan genomes. This review presents a hypothesis that repressive RNA-binding proteins (RBPs) facilitated the large-scale accumulation of RDEs. Many RBPs bind RDEs in pre-mRNAs to repress the effects of RDEs on RNA processing, or the formation of inverted repeat RNA structures. RDE-binding RBPs often assemble on extended, multivalent binding sites across the RDE, which ensures repression of cryptic splice or polyA sites. RBPs thereby minimize the effects of RDEs on gene expression, which likely reduces the negative selection against RDEs. While mutations that change splice sites in RDEs act as an off-on switch in exon formation, mutations that decrease the multivalency of RBP binding sites resemble a rheostat that enables a more gradual evolution of new RDE-derived exons. RBPs might also repress aberrant processing of active retrotransposons, thus increasing the chance that full-length copies are made. Taken together, in this review, it is proposed that RBPs facilitate the widespread accumulation of intronic RDEs by repressing RNA processing while chaperoning their potential to gradually evolve into new exons.

1. Introduction

Retrotransposons use reverse transcriptase to copy themselves within the genome through an RNA intermediate, a process known as retrotransposition. Retrotransposons are present in most eukaryotic genomes, including virtually all known mammalian and plant genomes.[1] Retrotransposon-derived sequences contribute to ≈10.9 and ≈44% of the genomes of *Danio rerio* and

Dr. J. Attig, Prof. J. Ule
The Francis Crick Institute
1 Midland Road, London NW1 1AT, UK
E-mail: jan.attig@crick.ac.uk; jernej.ule@crick.ac.uk

Prof. J. Ule
Department of Molecular Neuroscience
UCL Institute of Neurology
Queen Square, London WC1N 3BG, UK

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/bies.201800132.

Homo sapiens, respectively.[2] Many families of retrotransposons have been identified across eukaryotic clades (reviewed in refs. [1,3]). In the human genome, most retrotransposon integrations correspond to the long interspersed nuclear elements (LINEs), including the L1, L2, and CR1 subfamilies, the short interspersed nuclear elements (SINEs), including the primate-specific Alu and SVA elements, and the transposons flanked by long-terminal repeats (LTRs), including single LTRs and endogenous retroviruses (ERVs).

Here, we develop a framework to describe the evolutionary dynamics and consequences of retrotransposons for gene expression at the level of RNA processing, and the protein regulators that bind to them. We propose that active retrotransposons and their inactive genomic progeny share RNA-binding protein (RBP) interaction partners, and discuss at which stages selection might act on them and how these interactions enable retrotransposons to accumulate in the genome without disrupting the fitness of the host.
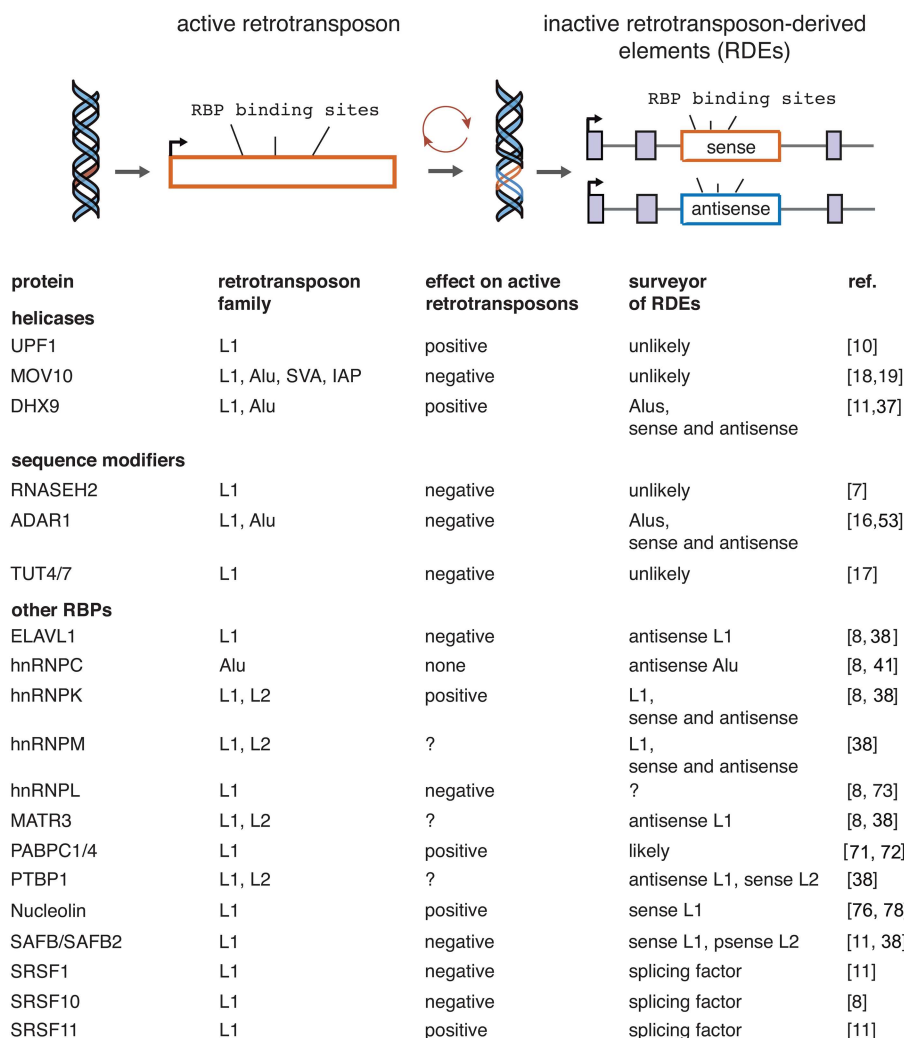
2. Many RNA-Binding Factors Bind Retrotransposons to Regulate Their Retrotransposition

Active retrotransposition poses a threat to genomic stability and cellular function.[4–6] As a consequence, strong transcriptional repressors target active retrotransposon families, as reviewed elsewhere.[7] In addition to these DNA-binding proteins, many host RBPs bind and regulate retrotransposons. Some of the RBPs are co-opted by the retrotransposon and necessary for retrotransposition, while other RBPs act as restriction factors of retrotransposition. Most interaction partners have to date been identified by co-immunoprecipitation with L1 RNPs. L1 encodes two ORFs, ORF1p that is an RBP, and ORF2p, which harbors the reverse transcriptase domain. For reverse transcription, ORF2p uses the polyA tail of the L1, which is annealed as primer:template to a DNA insertion site. Many RBPs are interaction partners of L1-ORF1p, L1-ORF2p, or the L1-ORF1p*ORF2p RNP, and a recent CRISPR screen for factors affecting L1 retrotransposition demonstrated that many of these RBPs have positive or negative effects on L1 retrotransposition (refs. [8–11], listed in **Figure 1**).

For illustration, we discuss as examples MOV10 and TUT4/7; additional restriction factors have already been reviewed in

active retrotransposon

inactive retrotransposon-derived elements (RDEs)

protein	retrotransposon family	effect on active retrotransposons	surveyor of RDEs	ref.
helicases				
UPF1	L1	positive	unlikely	[10]
MOV10	L1, Alu, SVA, IAP	negative	unlikely	[18,19]
DHX9	L1, Alu	positive	Alus, sense and antisense	[11,37]
sequence modifiers				
RNASEH2	L1	negative	unlikely	[7]
ADAR1	L1, Alu	negative	Alus, sense and antisense	[16,53]
TUT4/7	L1	negative	unlikely	[17]
other RBPs				
ELAVL1	L1	negative	antisense L1	[8, 38]
hnRNPC	Alu	none	antisense Alu	[8, 41]
hnRNPK	L1, L2	positive	L1, sense and antisense	[8, 38]
hnRNPM	L1, L2	?	L1, sense and antisense	[38]
hnRNPL	L1	negative	?	[8, 73]
MATR3	L1, L2	?	antisense L1	[8, 38]
PABPC1/4	L1	positive	likely	[71, 72]
PTBP1	L1, L2	?	antisense L1, sense L2	[38]
Nucleolin	L1	positive	sense L1	[76, 78]
SAFB/SAFB2	L1	negative	sense L1, psense L2	[11, 38]
SRSF1	L1	negative	splicing factor	[11]
SRSF10	L1	negative	splicing factor	[8]
SRSF11	L1	positive	splicing factor	[11]

Figure 1. Different classes of RBPs interact with active retrotransposons and retrotransposed genomic copies. Retrotransposons sequences are transcribed into RNA in two different scenarios. On the one hand, for active and autonomously transcribed retrotransposons, the RNA transcript is the template for reverse transcription. On the other hand, past insertions of a retrotransposon in the genome are frequently transcribed as part of another gene, as retrotransposon-derived element or RDE. A number of RBPs are regulators of retrotransposons, and broadly fall in one of three categories: helicases, sequence editing enzymes, and sequence-specific RBPs. Identification of RBPs interacting with active retrotransposons has often been done by co-purification schemes and mass spectrometry.[8–10] Recently, a CRISPR screen has revealed many positive and negative regulators of active L1.[11] Screening of RNA interactome data from crosslinking and immunoprecipitation of RBPs has identified dozens of RBPs that interact with RDEs.[38] A few of the RBPs that restrict L1 retrotransposition, are well-characterized and general mRNA surveillance factors (MOV10, UPF1, TUT4/7, RNASEH2). For these RBPs it seems unlikely that they have specificity toward RDE sequences and regulate intronic RDEs beyond their general role in mRNA surveillance.

detail.[7] MOV10 is an ATP-dependent RNA helicase that was first identified as a restriction factor of Moloney leukemia virus.[12] MOV10 interacts with multiple enzymes that act as restriction factors of retrotransposons, including the adenosine deaminase ADAR1, the cytidine deaminase APOBEC3G, and the terminal uridylyltransferases TUT4/TUT7.[13–17] MOV10 inhibits retrotransposition of all human non-LTR retrotransposons in cultured cells,[18,19] most likely by promoting the activity of these enzymes, for example TUT4/TUT7. When a poly(A) tail is shortened and not protected by poly(A)-binding proteins, TUT4 and TUT7 act as terminal uridylyltransferases to promote mRNA decay.[20] In vitro, L1-ORF1p inhibits the access of TUT4/7 and RNases to L1 RNA. MOV10 displaces L1 ORF1p and thus allows

access of TUT4/7 to L1 RNA to promote its degradation and inhibit reverse transcription.[17]

3. Retrotransposon-Derived Elements, or RDEs, Are Common in Introns

The vast majority of insertions of retrotransposons are silent due to truncations or accumulation of mutations that inactivate their capacity for retrotransposition. For example, >99% of LINEs in the human genome contain truncations or mutations that fully inactivated their capacity for retrotransposition, and therefore can be considered "dead." We refer to such copies as retrotransposon-derived elements, or RDEs, which have lost

all ability to retrotranspose. While active copies of retrotransposons are under negative selection pressure,[21] regulatory evolution of RDEs leads to acquisition of new functions as enhancers, promoters, exons, or polyA sites.[22–24] The acquisition of promoters could be traced to sequence mutations at individual RDEs that took place long after the genomic insertion of the RDE.[25,26] Recent studies made great progress in understanding how the acquisition or loss of binding sites for transcription factors (TFs) and RBPs can be a strong driving force for the regulatory evolution of RDEs.

As a consequence of their sheer abundance, over a million RDEs are transcribed as part of introns within host genes.[3] RDEs often contain cryptic sites for RNA processing, which can perturb gene expression and lead to loss-of-function phenotypes if used in an unregulated manner, a feature that makes them disease alleles.[27,28] Both Alus and LINEs can disrupt host gene expression when cryptic splice or polyA sites within them are erroneously recognized.[29–31] For instance, newly inserted L1 elements are documented to have caused thalassemia due to disruption of *HBB* expression,[32] X-linked retinitis pigmentosa 2 due to disruption of *XRP2*,[33] chronic granulomatous disease due to disruption of *CYBB*,[34] and X-linked dilated cardiomyopathy due to disruption of *DMD*.[35] In these cases, "exonization" of the L1 creates a L1-derived exon within the coding sequence, which introduces premature termination codons (PTCs) that truncate the protein and induce nonsense mediated mRNA decay (*NMD*). Moreover, polymorphic Alu insertions can also affect splicing of nearby alternative exons. For example, an Alu insertion promotes skipping of an exon in the gene CD58, reduced expression of which is associated with risk for developing multiple sclerosis.[36]

The proportion of intronic RDEs affecting RNA processing at first sight seems rather small compared to the overall number of intronic RDEs. Yet, the listed examples represent only the known de novo insertions in human patients, whereas the vast majority of genomic RDEs are fixed in the population. With the deleterious effects of new intronic insertions in mind, a question arises: how do RDEs accumulate in introns at such high proportion, and is there some cellular mechanism in place that controls the majority of them?

4. Many RBPs Bind to Intronic RDEs to Regulate Their RNA Processing

Tens of thousands of intronic RDEs contain cryptic splice sites (**Table 1**) with the potential to disrupt gene expression, yet few seem to do so. It has become evident that many RBPs recognize specific sequence or structural motifs formed by the different RDE families, and thereby efficiently protect host gene expression by reducing recognition of RNA processing sites within them. Prominent examples of RBP:RDE interactions are discussed here in detail, and include control of Alu RDEs by HNRNPC/U2AF65, ADAR/DHX9 and STAU1/PKR, and repression of LINE-derived exons by MATR3/PTBP1.[37–41]

4.1. Repression of RNA Processing in Antisense Alus by hnRNPC

Most intronic Alu elements are 150–300 nt long. Antisense Alus are prone to exonize,[31,42] and exonization events are associated with human diseases.[43] Exonization of sense Alu elements is much rarer due to their lack of pyrimidine tracts and 3′ splice site sequences (Table 1A). Inclusion of Alu-exons frequently introduces a PTC, thus targeting the transcript to NMD. Alu exonization is widely inhibited by hnRNPC, which binds to the U-tracts of antisense *Alu* elements.[41] In sense orientation, the transposing Alu contains an internal A-tract and 3′ polyA-tail, which is necessary for co-option of LINE ORF2p for retrotransposition.[44] Hence, when de novo Alu insertions are transcribed in an antisense orientation as part of the host gene, they contain two U-tracts. The U-tracts are recognized by hnRNPC, which incorporates the Alu sequence into the hnRNP particle and thereby blocks its splicing and 3′ end processing.[41,45] However, if mutations disrupt the U-tract to prevent binding of hnRNPC, it can turn into a binding site for U2AF65 and TIA1/TIAL1, the splicing factors that assist in the recognition of 3′ and 5′ splice sites. We have shown that U2AF65 and hnRNPC directly compete for binding to U-rich motifs in thousands of antisense Alus,[41] and T to C mutations favor U2AF65 binding. As a result, there is a close relationship between the age of Alu insertions, the length of their U-tracts, the relative binding of hnRNPC versus U2AF65, and exonization of Alus.[39]

4.2. ADAR/DHX9 and Editing of RNA Duplexes

When two proximal Alu elements are present within an RNA in opposing orientation, they can form double-stranded RNA structures (dsRNA) if their complementarity has not been disrupted through accumulation of mutations. When such inverted repeats are present within 3′ UTRs, they can become a signal for nuclear retention of the mRNA and its decay in the nucleus.[46,47] Moreover, inverted Alu repeats form the most common substrate of the RNA editing enzyme ADAR1.[48,49] ADAR deaminates adenosine nucleotides to inosines, and interacts with DHX9, an RNA helicase.[37] Editing prevents formation of double-stranded RNA since inosine pairs with cytosine, not uridine, and DHX9 unwinds dsRNA formed by Alus. In the absence of DHX9, the Alu-derived secondary structures can also disrupt RNA processing. Hence, ADAR1 and DHX9 play a similar role as hnRNPC by protecting the transcriptome from misprocessing at Alu.

If inverted Alu repeats are not edited and escape to the cytoplasm, they can be recognized by the viral RNA sensor PKR1, which can lead to partial and unspecific translational shut-down.[50] In mouse, ADAR knockouts are embryonically lethal because of an autoimmune phenotype, triggered by MDA5, a sensor of viral RNA, that erroneously recognizes endogenous RNAs in absence of editing.[51–53] Hence, ADAR is a paradigm example of a restriction factor that also plays a vital role on control of RDEs within the host genome.

4.3. Repression of LINE Exonization by MATR3 and PTBP1

LINEs are prone to be spliced both in sense and antisense orientation.[29] Since LINEs are much longer sequences compared to Alus, more regulators bind to them, and dozens

Table 1. Putative splice sites within mammalian RTEs.

(A) Splice site sequences in repeat consensus sequences			(B) Cryptic splice sites within retrotransposons in the human genome				
	# of predicted splice sites in			# of splice sites in intergenic insertions		# of splice sites in intragenic insertions	
Repeat family	Sense [strength]	Antisense [strength]	Repeat family	Sense	Antisense	Sense	Antisense
LIME (L1.3)			LINES	697255		289503	435470
5′ SS	11 [8.0–8.3]	9 [8.0–10.1]	With 5SS	76050	102122	25521	59611
3′ SS	45 [8.3–8.7]	25 [10.3–12.6]	With 3SS	46392	82808	19079	44299
Alus (AluSx, AluJb)			Alus	484786		294626	357352
5′ SS	0 [−4.7 to 4.9]	2* [2.4–4.3]	With 5SS	55107	12083	33638	7744
3′ SS	0 [0.3–0.97]	2* [14.2–6.5]	With 3SS	414	198598	252*	148172
ERV1 (HERVH)			ERVs	439433		102772	178299
5′ SS	7 [8.9–10.5]	35 [10.3–10.7]	With 5SS	54696	78424	12176	31148
3′ SS	43 [10.1–10.8]	36 [7.3–9.8]	With 3SS	41022	41897	8794	17323

(A) Retrotransposons frequently contain putative splice sites. The number of splices sites and their predicted strength is shown for the major families of human RTEs, with L1.3, the sequence of AluSx and AluJb, and the HERVH consensus sequence from repbase as examples. The number of putative splice sites was determined with NetGene2.[99] NetGene2 did not call any of the putative splice sites in Alus, so we used the splice site positions identified by Lev-Maor et al. and Sorek et al. (indicated by *). Splice site strength was predicted with MaxEntScan,[100] and we summarise the strengths of the TOP3 splice sites. The sequences of the TOP3 splice sites are listed in Table S1, Supporting Information. Table S1, also lists splice site analysis for mouse LI and the consensus sequence for mammalian L2. The interquartile range of predicted splice site strengths for constitutive human exons is 6.2–10.0. (B) We show the number of cryptic splice sites within the human genome (assembly hg38) generated through RTE insertions in sense and antisense. To estimate the number for RTEs with 5′ and 3′ splice sites, we searched for NNGGURAG and Y8NBAGR sequences, respectively.

of RBPs are highly enriched on LINEs.[38] We found a number of RBPs show strong enrichment toward targeting of evolutionarily young LINEs, including MATR3, hnRNPM, and SUGP2. While LINE-binding RBPs likely serve a multitude of functions at RDEs, such as resolving RNA duplexes or anchoring RNAs to the chromatin (discussed in more detail below), it was striking that known splicing repressors formed the majority among the RBPs preferentially recognising evolutionarily young LINEs.

This suggested to us that LINEs are pre-marked as intronic sequence at new insertions. We found support for this hypothesis from two angles. Firstly, exons arise about two-times more frequently from evolutionarily old LINEs, and these exons show higher inclusion levels in more tissues, despite comparable strength of their splice site sequences. Secondly, MATR3 and PTBP1 were among the most enriched RBPs on intronic LINE sequences, and depletion of both proteins increased inclusion of cryptic exons derived from antisense L1s.[38] Just as the hnRNPC-repressed Alu-exons, inclusion of such cryptic LINE-exons tends to decrease the stability the mRNA isoform through NMD. Hence, LINEs have a high intrinsic propensity to exonize, but are generally efficiently repressed due to the function of RBPs acting in trans. Interestingly, LINEs are excluded in a wide window from established exons both in mouse and man, suggesting that they interfere with splicing of close-by splice sites.[38]

4.4. Splice Repressors at ERVs?

Due to their LTR promoter structure with a strong 5′SS, ERVs are particularly likely to produce splice-fusion transcripts with downstream genes,[54–56] a feature that might create a need to have splice repressors recognizing ERV sequence to reduce the amount of fusion transcripts. However, so far no RBP has been described to act on ERV transcripts, neither in the context of an ERV RNP particle, nor in the context of insertions transcribed as part of a host gene.

5. RBPs Generally Emerged Earlier in Evolution Than the Regulated RDEs

To assess the timing of evolutionary emergence of RDEs and their regulators, we performed cross-species analysis of the timing of L1 and Alu amplification waves that generated new RDEs at bulk, and the primary RBPs that repress them (**Figure 2**), using both orthoDB[57] and ENSEMBL compara.[58] RBPs that repress Alus, such as hnRNPC, ADAR and DHX9, are conserved between vertebrates and invertebrates, and often their RNA-binding domains are almost perfectly conserved across these lineages. In contrast, Alu elements are primate-specific. Thus, the repressors predate the emergence of intronic Alu repeats, and in the mouse ADAR and DHX9 recognize RNA duplexes formed from other sources than Alus.[59–62]

LINE elements are evolutionarily older than Alus. L1 elements were hyperactive several times early on in mammalian evolution, for instance after the therian clade diversified from monotremes (ref. [63], see also Figure 2). Intronic insertions are bound by repressive RBPs, including MATR3, hnRNPM, and the PTBP family proteins.[38] According to orthoDB,[57] MATR3 likely emerged within an early vertebrate ancestor in the *Osteichthyes* lineage, since no orthologues could be identified in any of the other metazoan classes examined (Figure 2). The PTB protein family and hnRNPM protein have orthologues in all metazoans we investigated (vertebrates and invertebrates). Hence, MATR3, PTBP1, and hnRNPM have predated the major amplification waves of mammalian L1 elements.
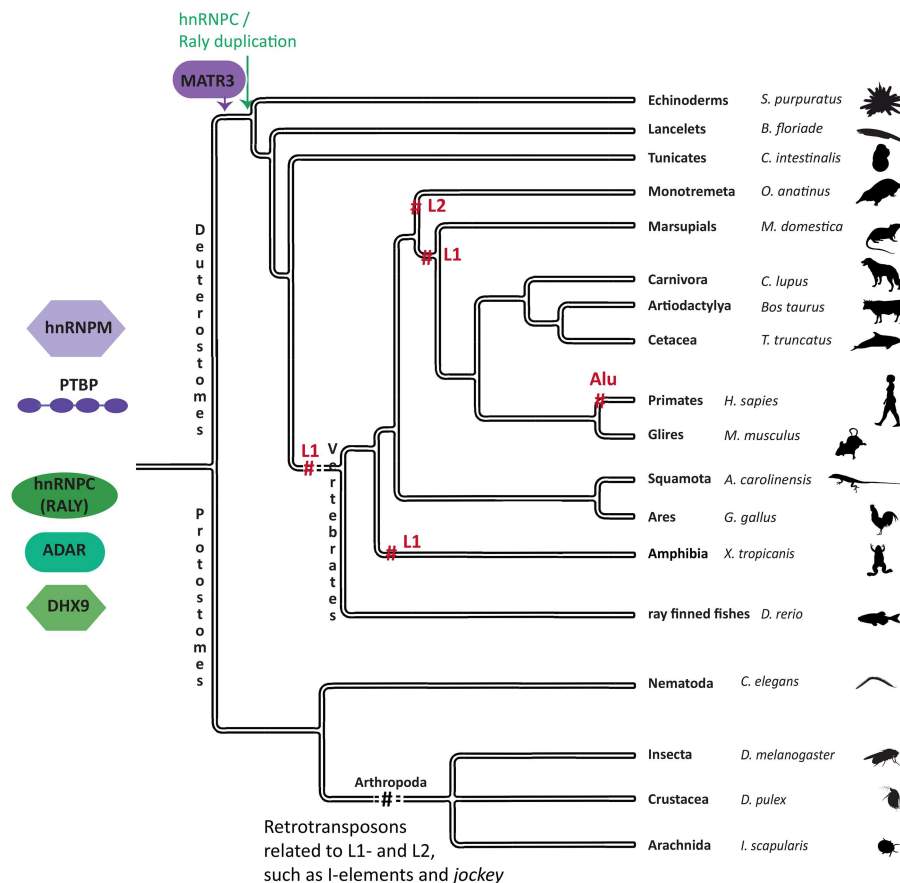
Figure 2. Amplification of RDEs took place in an environment of cognate RNA binding proteins. Phylogenetic comparison of RDE-binding RBPs and known amplification periods of RDEs in reference metazoans. The phylogenetic tree is focused on reference organisms of major clades with increasing evolutionary distance to human and mouse (guided by ref. [101]). The tree is for visualization purposes only and evolutionary distances between clades and species is not to scale. All branches are depicted as accepted in recent literature[101] and major clades within the tree are labeled. Disputed groups within the Arthropoda are shown as multi-nodes but do not affect the analysis of protein orthologues depicted. Known amplification periods of Alu, L1, and L2 elements are marked by hash tags and identified from references.[63,98,102] Kordis et al. established that there was an amplification of L1 elements in the last common ancestor of vertebrates. We focus exclusively on known amplification period, since the initial emergence of LINE families is not resolved. It remains possible that L1s were introduced into a mammalian ancestor through horizontal transfer,[97] instead of being vertically transmitted from the metazoan ancestor. It is worth mentioning that Nematodes and insects are devoid of L1 and L2 elements, and dominated by other LINE families (RTE, jockey) while the simple chordates (*Strongylocentrotus purpuratus, Branchiostoma floridae, Ciona intestinalis*) contain L1 and L2 elements. Orthologues of human RBPs were identified by orthoDB,[57] which groups proteins if they are likely to originate from a distinct common ancestor. hnRNPC and hnRNPCL are closely related to RALY and RALYL proteins in all species and form one group with one common ancestor, so do isoforms of the PTBP proteins.

From such cross-species comparisons it is clear that in most cases, RBPs are evolutionarily more ancient than the RDEs that they repress. This supports the notion that presence of repressive RBPs could increase the capacity of RDEs to accumulate in introns across the genome through minimising the negative selection against them. It would equally be expected that emerging RDE families with pre-existing repressive RBP partners are more likely to spread in the genome.

6. Exons Tend to Emerge From RDEs Through Gradual De-Repression

Since the RBPs are evolutionarily older than the RDEs that they repress, repressive RNPs could have assembled on intronic RDEs since their earliest emergence. We have shown that the consensus RDE sequences are generally rich in multivalent binding sites for repressive RBPs. For instance, antisense Alus contain two separate U-tracts that are often >8 nt long, and a U-tract of 4 nts is sufficient to accommodate 1 RRM domain of hnRNPC.[64,65] Thus, Alus with longer U-tracts can accommodate multivalent binding of the four RRM domains of hnRNPC (two at each U-tract), leading to high-affinity binding. An even more extreme example is the multivalency of the antisense L1 consensus, which is covered by many dozens of binding motifs for PTBP1 and MATR3.[38]

Neutral evolution leads to deviations in the RDEs from the consensus sequence of the founder retrotransposon. As part of this process, some mutations will increase the strength of splice sites or polyA sites within the RDEs, or position 5′ and 3′ splice sites in a conformation that favors inclusion of a new RDE-derived exon into the mRNA product of the gene (**Figure 3**).
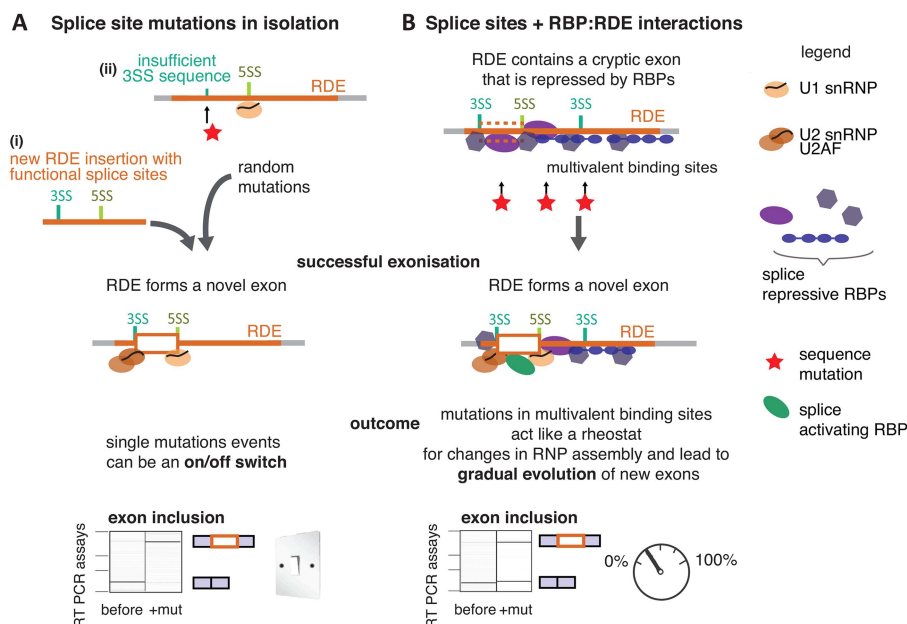
Figure 3. Co-evolution of splice sites and RBP interaction sites in RDEs controls inclusion of RDE-derived exons. RDEs contain both splice sites and binding sites for splice-repressive RBPs. Mutations have different outcomes for inclusion of cryptic exons, depending on if splice sites and the repressive RBPs are coupled or not. The two scenarios are illustrated in (A) and (B), with visualization of exon inclusion levels by exemplary RT-PCR assay images. A) If they evolve independently, changes in splice site conformations will manifest as a binary switch in inclusion or skipping of an RDE-derived exon. This includes scenarios where a new RDE has pre-existing functional splice sites (i) or where random mutations happen to increase splice site strength (ii). B) If splice sites are coupled with splice-repressive proteins, changes to the splice sites determine the capability of the RDE to exonize but inclusion levels remain low. Multiple mutations in the multivalent binding sites will lead to gradual increase in inclusion. In addition, RDEs often lack exonic splice enhancer sequences, which can recruit splice-activating RBPs.[66] The need to accumulate those through mutations further reinforces the gradual emergence of novel exons. Given the random occurence of mutations and position of an emerging RDE-derived exon within transcripts, sudden full inclusion of a novel exon is likely to result in a fitness cost to the organism. As discussed in the text, both Alu and LINE-derived exons are examples of scenario (B). Gradual evolution of a novel exon is likely more suited to create variation while maintaining essential gene functions.

Many RDEs, particularly antisense L1 and Alus, contain sequences that are strong splice sites or are only one base-pair mutation away from the splice site consensus sequence (Table 1, see also refs. [30,38]). If such an RDE inserted but lacks a repressive RBP binding partner, or an existing RDE acquired a mutation creating a strong splice site, that event would likely create an immediately highly included exon (Figure 3A), in effect amounting to a binary on-off switch. However, as we have the shown consensus RDE sequences contain multivalent sequences that are strong binding sites for repressive RBP, and therefore an emerging RNA processing site is likely under direct control of a repressive RBP. In this setting, a novel splice site will first create a cryptic exon that is either not or lowly included into mRNAs (Figure 3B). In addition, RDE consensus sequences in human show a general lack of exonic splice enhancer sequences,[38,66] which could further contribute to low inclusion levels of emerging exons. We have shown that the inclusion level of exons at RDEs correlates with the evolutionary age of the RDE, both for Alu and L1 elements.[38,39] Importantly, changes in splice site sequences and strength distinguished exonising from non-exonising Alus, but did not explain differences in their inclusion levels.[39] Instead, all our findings indicate that upon evolutionary divergence of RTEs, accumulation of mutations tends to gradually decrease the number of multivalent binding motifs for repressive RBPs. Given the large number of repressive multivalent motifs that are coupled with each cryptic splice site, mutations leading to gradual loss of these motifs are likely a major factor in controlling exon inclusion levels. For instance, gradual shortening of the repressive U-tracts at older Alus is accompanied by reduced hnRNPC binding and increased inclusion levels.[39] We also found that the evolutionarily older RDEs, or RDEs with more sequence divergence relative to the RDE consensus, are more accessible to RBPs that enhance splicing or 3′ end processing (U2AF2, TIA1/TIAL1, CSTF2, etc.) and are more likely to contain exonic splice enhancer sequences,[38] in agreement with a previous study from the Eyras lab.[66] We use the term "regulatory evolution" for this process where mutations gradually shift the balance from RBPs that repress to those that promote RNA processing at RDEs. The mutations in the multivalent binding sites could act as a rheostat for regulatory evolution that gradually increases the inclusion of RDE-derived exons through de-repression.

7. Do Repressive RBPs Promote Evolutionary Tinkering at Intronic RDEs?

Francois Jacob has coined the term "evolutionary tinkering" for systems that facilitate evolutionary novelty but are

self-constrained at their origin, in that they have to maintain pre-existing functions.[67] This concept suits the process leading to the exonization of RDEs across the transcriptome. If RDEs were not initially repressed by RBPs, they would undergo rapid negative selection, and be deleted from the gene pool. Since they are repressed, they can diversify and potentially acquire new functions, which can benefit the population and outweigh the risk of aberrant processing at other RDE insertions of the same family.

This risk is likely reduced through efficient repression of intronic RDEs. Misprocessing of RDEs into aberrant exons would impact on the organism's fitness or reproduction (**Figure 4**). RDE-derived exons created through an off-on switch will be most likely removed from the population due to negative selection. Repressive RBPs minimize the likelihood of such a discrete off-on switch and can keep the exons in a cryptic state, thereby minimizing negative selection against the exon and enabling it to persist in the population. We expect that each RDE family attracts its own set of repressive RBPs, which helped with the accumulation of the RDE in the genome, which helped with the accumulation of the RDE in the genome. Our model could explain the observation that surprisingly many RBPs recognize RDEs in antisense orientation, that is, the reverse complement of the retrotransposal sequence. In particular, we identified many more RBPs that recognize L1 elements in antisense orientation than in sense (Figure 1, see also ref. [38]). The human genome has about twice as many cryptic splice sites in antisense L1s compared to sense L1s (Table 1B). Potentially, reduced negative selection against splice sites in antisense L1s favors accumulation of splice sites in antisense L1s over those in sense L1s.

Due to the capacity of newly created exons to persist in the population in a cryptic, repressed state, gradual accumulation of further mutations can eventually lead to adaptive variation. It has been shown that older RDEs have been increasingly exapted into regulatory roles.[68] This is consistent with the increased density of older RDEs close to exons.[38,69] Moreover inclusion of RDE-derived alternative exons correlates with their evolutionary age, such that both Alu- and LINE-derived exons are included into mRNAs in a highly tissue-specific manner, most frequently in the brain and testis.[38,39] This might be explained by a bias for exonization within tissue-specific genes, or regulation of the cognate repressive RBPs. For example, PTBP1 is not expressed in the brain though its close homologue PTBP2 is, and both hnRNPC and MATR3 contain many sites for post-translational modifications, of yet unknown function. An alternative reason could be the differential strength of selection pressure at different groups of genes during fixation of RDEs. Highly expressed genes are selected for short introns,[70] and essential or ubiquitously expressed genes might have stronger negative selection against insertion of RDEs. Conversely, genes that evolve quickly typically have tissue-specific functions, such as immune tissues or the brain. RDEs could enhance the sequence turnover at such genes, and thus be particularly important for evolutionary tinkering in tissue-specific functions.

8. Do the Same RBPs Act Both on Active Retrotransposons and RDEs?

Active retrotransposons and their intronic progeny will share the same RBP binding sites, if the RDEs are transcribed in sense orientation. Hence, we might predict new regulators of intronic RDEs from factors that regulate active retrotransposons, and vice versa. Reviewing the known interaction partners of Alus and LINEs, we find such a dual role well supported for helicases (Figure 1), but this principle could extend well to other factors, such as splicing repressors.

The DHX9 and UPF1 helicases are known interaction partners of Alu and LINE retrotransposons, among others.[8,10] Depletion of DHX9 or UPF1 reduces L1 retrotransposition.[10,11]
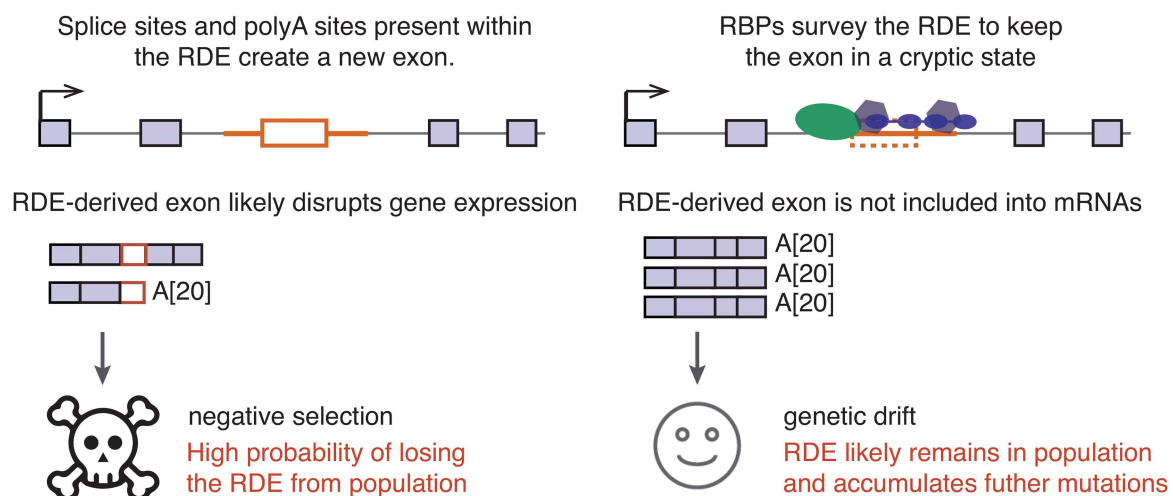
Figure 4. The outcome of a newly emerging exon in an RDE depends on the bound RBPs. Any active RTEs will generate new insertions in the genome, which will be either intergenic or intronic. Intronic RTEs will be transcribed as part of the host gene and exonize if strong splice site sequences are present in the inserting RTE. Exonization of RTEs is prone to reduce expression level of the host gene, and hence has a probability of negatively affecting an organism's fitness. If splice repressors are able to effectively recognize and repress exonization events at novel insertions, an active RTE family will affect fitness at a lower probability. Thereby, the active RTE family has a higher probability to spread in the population.

Here, helicases are likely needed to remove RBPs and secondary RNA structures ahead of ORF2p during reverse transcription. In a genomic context, DHX9 prevents formation of RNA structures derived from inverted repeat RDEs.[37]

Another co-factor of retrotransposons are poly(A) binding proteins PABPC1 and PABPN1 (PABP2), which bind to the polyA-tail of L1 RNA and promote retrotransposition.[71] The polyA-tail of L1 and Alus is crucial for retrotransposition − it increases RNA stability, pairs with T_4 single-strand overhang during target site-primed reverse transcription, and ORF2p has a preference for binding to A-tails.[72] As a result of the need for a polyA tail, LINE RDEs are particularly rich in polyA signals,[73,74] and we expect they need to be repressed in most of them. PABPN1 has been shown to inhibit cryptic or premature polyA sites in the genome,[75] but it is not known if any such sites derive from LINE RDEs.

Finally, nucleolin and SAFB proteins are co-factors of L1, and depletion of either protein reduces retrotransposition.[11,76] At the same time, they also act on nuclear scaffold RNAs derived from or rich in L1 sequence.[77,78] Nucleolin is a ribosome biogenesis factor, and a scaffold component of the nucleolus. Peddigari et al.[76] found nucleolin binds to the inter-ORF spacer of mouse L1, and L1 ORF1p interacts and co-localizes with fibrillarin,[8] another core component of the nucleolus. Hence, the nucleolus plays an as yet undefined role in the L1 retrotransposition cycle. On top, L1-containing RNAs appear to act as scaffold lncRNAs to recruit nucleolin and the transcriptional repressor KAP1 in mouse ES cells, to binding sites important for maintenance of the ES cell transcriptional profile.[78] SAFB, SAFB2, and hnRNPU (also called SAF-A) likely have a similar function in bridging nuclear L1-lncRNAs and transcriptional control. The SAF proteins were identified as nuclear attachment factors, and the DNA-binding SAF-Box domain.[79] Past work has indicated that SAF-Box proteins are scaffolds that link nuclear RNA and active chromatin domains,[80] and a dominant-negative mutant of the SAFB interacting partner hnRNPU (also called SAF-A) leads to dissociation of LINE RDE-containing RNAs from euchromatin.[77] While the precise molecular mechanisms by which hnRNPU/SAFB/SAFB2 complexes regulate the nuclear fate of LINE RDE-containing RNAs remains unclear, the SAF-box proteins are RBPs that again act on active L1 RNP, as well as L1-RDE containing RNAs. It is worth highlighting that both in the work of Percharde et al.[78] and of Hall et al.,[77] the precise sequence and locus of the RNA transcripts remains elusive, but both indicate a pool of highly transcribed lncRNAs that are either L1-derived, or rich in L1 RDEs.

9. Could Repressive RBPs Influence the Efficiency of Active Retrotransposition?

Active L1 retrotransposons in human are replete with cryptic splice and polyA sites.[29,74] Belancio et al.[29] have shown that splicing of an active, human L1 elements results in loss of ORF2p sequence and integration of truncated daughter elements that are dead-on-arrival. Hence, long retrotransposon RNA is vulnerable if splicing precedes the formation of the reverse transcriptase complex, and thus it would benefit from recruiting splice-repressive proteins. We hypothesize that the binding of repressive RBPs might be required to repress the splice and polyA sites in long retrotransposons, thus increasing the chance that they make full-length copies of themselves (**Figure 5**).

Just like active LINEs, the RNA transcript of the full ERV element including the LTRs (called genomic RNA) can only remain active if it remains unspliced until reverse transcription. RBPs binding ERVs are not yet known, but it is likely that specific repressive RBPs act to inhibit splice sites within ERVs. For HIV, SR proteins extensively regulate splicing of the genomic RNA. A change in the abundance of several SR proteins decreased the amount of infectious HIV particles in vitro.[81,82] This is likely a result of changing the ratio of spliced versus genomic RNA, and consequently the ratio of viral proteins. By extrapolation, splicing repression of ERVs might similarly increase the efficiency of their retrotransposition.

10. How to Test the Hypothetical Role of RBPs in the Genomic Accumulation of RDEs?

We propose that repressive RBPs have facilitated the spread of RDEs in metazoan genomes through a combination of two factors: by reducing the negative selection against intronic REs, and by facilitating retrotransposition of full-length retrotransposons. We expect that an active retrotransposon would invade the genome at a reduced rate in the absence of the cognate RDE. A possible experiment to test this model could involve expression of the RBP in cells of a species naturally devoid of it, but with an active cognate retrotransposon. However, no such species is known to date, since we find that the RBPs appear to have emerged in evolution earlier than their cognate RDEs (Figure 5). Nevertheless, it is possible that the RBP binding partners of LINEs co-evolved, and that clade-specific properties of RNPs contribute to efficient repression of subfamilies of LINE retrotransposons. Therefore, one could infect for instance a chicken cell line with a mammal-specific L1 family, to test if the trans-acting RBPs in chicken are less efficient in promoting the accumulation of the mammalian L1. Another approach is creating knockout cells of an RBP-of-interest, and testing the retrotransposition rate of L1. Here, one would need to keep in mind the other functions of each RBP, which might affect the knockout cells capacity for long-term propagation. Instead we suggest exploiting synthetic L1s, which have been established to monitor the rate of retrotransposition and recovery of insertion sites.[83,84] These synthetic L1 sequences could be designed to lack the binding sites of repressive RBPs, hence allowing them to be tested for their role in retrotransposition. The effect of the RBP depletion on new insertions could then be measured after a few dozen cell divisions.

The regulation of RDEs has been studied almost exclusively in human cells, hence limiting the inferences that we can make about co-evolution of their sequence with their cellular environment. We need to learn more about the impact of RDEs on gene expression in other genomes, such as Drosophila and monotremes. Here, LINEs are particularly interesting because they are present across vertebrates, but different families of LINE elements dominate the genomes of different lineages, and similar retrotransposon families are present across arthropoda.
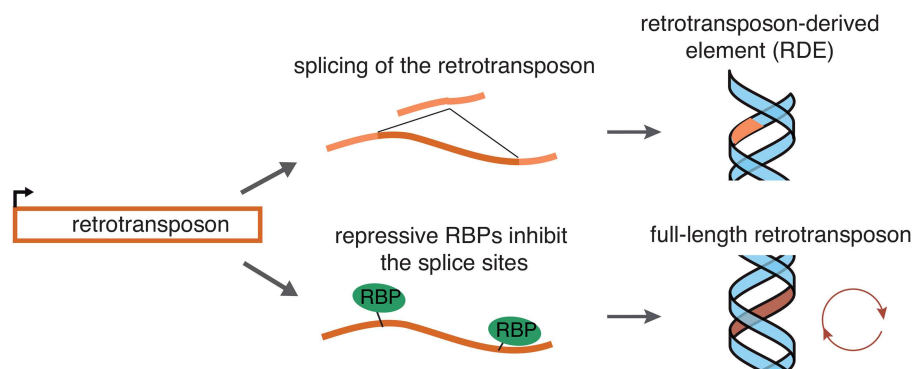
Figure 5. Splicing repressors could promote full-length retrotransposition. Any retrotransposon relies on generating full-length transcripts to copy itself faithfully at the insertion site. Active LINEs can undergo splicing.[29] This generates shortened LINE transcripts, some of which can still insert into the genome but then are "dead on arrival," that is, can not seed further amplification. This could be prevented by splicing repressors, which thereby would contribute to survival of active retrotransposing elements.

If different families of RDEs are kept in check in these species by the orthologues of the human RBPs, then these independently evolved families all contain the binding sites of some shared RBPs, it would provide further support for our hypothesis of the importance of splice-repressive RBPs in regulating retrotransposons. An example of a shared role of an RBP is the capacity of DHX9 to resolve inverted repeat sequences formed by Alus in human cells, and by the structurally related B2 elements in mouse cells. Many RDEs contain U-rich motifs in one of their orientations, which can be bound by a number of RBPs, including hnRNPC, MATR3 and PTBP1, and could thus act as a shared sequence feature. Experimental studies of these RBPs in non-mammalian species will be required to understand if their roles in RDE control are more ancient than so far anticipated. This would open the avenue for further comparisons of the adaptive evolution of diverse RDE families, and the sequences that enable their regulation by RBPs.

11. Similarities in the Evolution of DNA- and RNA-Binding Sites Within RDEs

Actively amplifying RDEs are kept in control by transcriptional silencing, primarily through recognition by the KRAB family of zinc finger transcription factors and piRNAs. Retrotransposition is effectively limited to the germline in many organisms, and it has been argued this is mutually beneficial: somatic transposition does not propagate the retrotransposon to the next generation, but could be lethal to the host. Hence, there is an equilibrium with the host repressing retrotransposon transcription globally versus selective escape by the retrotransposons targeting the germline. This dynamic is best embodied in Drosophila, where the I-element retrotransposon propagates via the oocyte nursing cells,[85] which effectively excludes any fitness cost to the individuum itself; as well as the mammalian testis, in which activation of IAV elements in the male mouse occurs in spermatogonia.[86]

While piRNAs are changeable at the sequence level to adapt to novel transposon families, the relationship between KRAB transcription factors and their cognate RDE has significant parallels to the dynamics we propose for RDE:RBP interactions. KRAB transcription factors recruit the KAP1 transcriptional repressor to a number of retrotransposons families, and depletion of KAP1 activates their transcription.[87–89] The KRAB family of genes appears to co-evolve with newly arising active elements in each evolutionary lineage, as evident by the increased number of the KRAB family of genes in parallel with amplification of retrotransposons in mammalian genomes.[90,91]

Similar to the evolution of RBP binding sites within transcribed RDEs, regulatory evolution also shapes the landscape of transcription factor binding sites within RDEs. RDEs in vicinity of promoters and enhancers and recruitment of a KRAB protein can lead to repression of the close-by gene.[92] It has been proposed early on that this seeds transcriptional modules in the genome.[93] Indeed some KRAB proteins do no longer target active retrotransposons but exclusively inactive RDEs, are expressed in a tissue-specific manner themselves, and control transcription of target genes through RDE-binding.[94] At the level of transcriptional control, multiple enhancers and promoters mirror in part the multivalent binding sites within RNA — again, a combinatorial regulation confers robustness,[95,96] and can ensure gradual incorporation of novel, functional binding sites into an existing promoter structure. Comparisons of sequence mutation rate and activity of evolutionarily young transcription start sites showed that, while they mutate quite quickly, their activity and inclusion as regulatory elements is gradual.[26] Hence, new RDEs have similar features at the level of DNA and RNA binding sites — initial strong repression allows accumulation of new, but silenced genetic elements, while step-wise derepression allows gradual evolution of functionally important elements.

12. Conclusion and Outlook

Most RDEs contain cryptic splice sites and polyA sites, in both the sense and antisense orientation, but the potentially disruptive inclusion of RDE-derived exons into host transcripts is mitigated by repressive RBPs. We present the hypothesis that RDE:RBP interactions are beneficial for the host to protect transcriptome integrity, but also to the retrotransposon to maintain its full-length retrotransposition and to minimize negative selection against RDEs. A key feature of this model is that the repressive RBPs must

evolutionarily predate the emergence of the retrotransposons, and be able to recognize new insertion sites based on the sequence features of the active retrotransposon. Our analysis of LINE and Alu retrotransposons confirmed this to be the case for the most abundant retrotransposons in mammals.

The rapid accumulation of genomic data from many individuals and species will be an excellent resource to further analyse how the mutation rate at RDEs shapes variations in gene expression. So far, most studies have focused on mechanisms involving piRNAs, TFs, and RBPs that restrict retrotransposition.[7,16] More work is needed to understand the complexity of regulatory forces that shape the gradual evolution of RDEs toward functional elements that contribute to the regulation and expression of the host genes.

Abbreviations

LINE, retrotransposons of the long interspersed nuclear element family; LTR, retrotransposon family flanked by long terminal repeats; NMD, nonsense mediated mRNA decay; RBP, RNA-binding protein; RDE, retrotransposed element, the sequence from past retrotransposon insertions; SINE, retrotransposons of the short interspersed nuclear element family; TF, transcription factor.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

Conflict of Interest

The authors declare no conflict of interest.

Keywords

[1] C. R. Huang, K. H. Burns, J. D. Boeke, *Ann. Rev. Genet.* **2012**, *46*, 651.
[2] A. Smit, R. Hubley, P. Green, http://www.repeatmasker.org, **1996–2010**.
[3] A. F. Smit, *Curr. Opin. Genet. Dev.* **1999**, *9*, 657.
[4] F. Cammas, M. Mark, P. Dolle, A. Dierich, P. Chambon, R. Losson, *Development* **2000**, *127*, 2955.
[5] D. Bourc'his, T. H. Bestor, *Nature* **2004**, *431*, 96.
[6] M. A. Carmell, A. Girard, H. J. van de Kant, D. Bourc'his, T. H. Bestor, D. G. de Rooij, G. J. Hannon, *Dev. Cell* **2007**, *12*, 503.
[7] J. L. Goodier, *Mob. DNA* **2016**, *7*, 16.
[8] J. L. Goodier, L. E. Cheung, H. H. Kazazian, Jr., *Nucleic Acids Res.* **2013**, *41*, 7401.
[9] M. S. Taylor, I. Altukhov, K. R. Molloy, P. Mita, H. Jiang, E. M. Adney, A. Wudzinska, S. Badri, D. Ischenko, G. Eng, K. H. Burns, D. Fenyo, B. T. Chait, D. Alexeev, M. P. Rout, J. D. Boeke, J. LaCava, *eLife* **2018**, *7*, e30094.
[10] M. S. Taylor, J. Lacava, P. Mita, K. R. Molloy, C. R. Huang, D. Li, E. M. Adney, H. Jiang, K. H. Burns, B. T. Chait, M. P. Rout, J. D. Boeke, L. Dai, *Cell* **2013**, *155*, 1034.
[11] N. Liu, C. H. Lee, T. Swigut, E. Grow, B. Gu, M. Bassik, J. Wysocka, *Nature* **2017**, *553*, 228.
[12] K. Mooslehner, U. Muller, U. Karls, L. Hamann, K. Harbers, *Mol. Cell. Biol.* **1991**, *11*, 886.
[13] L. H. Gregersen, M. Schueler, M. Munschauer, G. Mastrobuoni, W. Chen, S. Kempa, C. Dieterich, M. Landthaler, *Mol. Cell* **2014**, *54*, 573.
[14] C. Liu, X. Zhang, F. Huang, B. Yang, J. Li, B. Liu, H. Luo, P. Zhang, H. Zhang, *J. Biol. Chem.* **2012**, *287*, 29373.
[15] C. Esnault, J. Millet, O. Schwartz, T. Heidmann, *Nucleic Acids Res.* **2006**, *34*, 1522.
[16] E. Orecchini, M. Doria, A. Antonioni, S. Galardi, S. A. Ciafre, L. Frassinelli, C. Mancone, C. Montaldo, M. Tripodi, A. Michienzi, *Nucleic Acids Res.* **2017**, *45*, 155.
[17] Z. Warkocki, P. S. Krawczyk, D. Adamska, K. Bijata, J. L. Garcia-Perez, A. Dziembowski, *Cell* **2018**, *174*, 1537.
[18] S. Arjan-Odedra, C. M. Swanson, N. M. Sherer, S. M. Wolinsky, M. H. Malim, *Retrovirology* **2012**, *9*, 53.
[19] J. L. Goodier, L. E. Cheung, H. H. Kazazian, Jr., *PLoS Genet.* **2012**, *8*, e1002941.
[20] J. Lim, M. Ha, H. Chang, S. C. Kwon, D. K. Simanshu, D. J. Patel, V. N. Kim, *Cell* **2014**, *159*, 1365.
[21] S. Boissinot, J. Davis, A. Entezam, D. Petrov, A. V. Furano, *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 9590.
[22] M. Cowley, R. J. Oakey, *PLoS Genet.* **2013**, *9*, e1003234.
[23] G. Bourque, *Curr. Opin. Genet. Dev.* **2009**, *19*, 607.
[24] E. B. Chuong, N. C. Elde, C. Feschotte, *Nat. Rev. Genet.* **2017**, *18*, 71.
[25] E. B. Chuong, N. C. Elde, C. Feschotte, *Science* **2016**, *351*, 1083.
[26] C. Li, B. Lenhard, N. M. Luscombe, *Genome Res.* **2018**, *28*, 676.
[27] E. Sukarova, A. J. Dimovski, P. Tchacarova, G. H. Petkov, G. D. Efremov, *Acta Haematol.* **2001**, *106*, 126.
[28] J. Chen, A. Rattner, J. Nathans, *Hum. Mol. Genet.* **2006**, *15*, 2146.
[29] V. P. Belancio, D. J. Hedges, P. Deininger, *Nucleic Acids Res.* **2006**, *34*, 1512.
[30] G. Lev-Maor, R. Sorek, N. Shomron, G. Ast, *Science* **2003**, *300*, 1288.
[31] R. Sorek, G. Lev-Maor, M. Reznik, T. Dagan, F. Belinky, D. Graur, G. Ast, *Mol. Cell* **2004**, *14*, 221.
[32] L. Lanikova, J. Kucerova, K. Indrak, M. Divoka, J. P. Issa, T. Papayannopoulou, J. T. Prchal, V. Divoky, *Hum. Mutat.* **2013**, *34*, 1361.
[33] U. Schwahn, S. Lenzner, J. Dong, S. Feil, B. Hinzmann, G. van Duijnhoven, R. Kirschner, M. Hemberger, A. A. Bergen, T. Rosenberg, A. J. Pinckers, R. Fundele, A. Rosenthal, F. P. Cremers, H. H. Ropers, W. Berger, *Nat. Genet.* **1998**, *19*, 327.
[34] C. Meischl, M. Boer, A. Ahlin, D. Roos, *Eur. J. Hum. Genet.* **2000**, *8*, 697.

[35] K. Yoshida, A. Nakamura, M. Yazaki, S. Ikeda, S. Takeda, *Hum. Mol. Genet.* **1998**, *7*, 1129.

[36] L. M. Payer, J. P. Steranka, D. Ardeljan, J. Walker, K. C. Fitzgerald, P. A. Calabresi, T. A. Cooper, K. H. Burns, *Nucleic Acids Res.* **2019**, *47*, 421.

[37] T. Aktas, I. Avsar Ilik, D. Maticzka, V. Bhardwaj, C. Pessoa Rodrigues, G. Mittler, T. Manke, R. Backofen, A. Akhtar, *Nature* **2017**, *544*, 115.

[38] J. Attig, F. Agostini, C. Gooding, A. M. Chakrabarti, A. Singh, N. Haberman, J. A. Zagalak, W. Emmett, C. W. Smith, N. M. Luscombe, J. Ule, *Cell* **2018**, *174*, 1.

[39] J. Attig, I. Ruiz de Los Mozos, N. Haberman, Z. Wang, W. Emmett, K. Zarnack, J. Konig, J. Ule, *eLife* **2016**, *5*, e19545.

[40] R. A. Elbarbary, L. E. Maquat, *Cell Cycle* **2014**, *13*, 345.

[41] K. Zarnack, J. König, M. Tajnik, I. Martincorena, S. Eustermann, I. Stevant, A. Reyes, S. Anders, N. M. Luscombe, J. Ule, *Cell* **2013**, *152*, 453.

[42] R. Sorek, G. Ast, D. Graur, *Genome Res.* **2002**, *12*, 1060.

[43] I. Vorechovsky, *Hum. Genet.* **2010**, *127*, 135.

[44] V. Ahl, H. Keller, S. Schmidt, O. Weichenrieder, *Mol. Cell* **2015**, *60*, 715.

[45] M. Tajnik, A. Vigilante, S. Braun, H. Hanel, N. M. Luscombe, J. Ule, K. Zarnack, J. Konig, *Nucleic Acids Res.* **2015**, *43*, 10492.

[46] L. L. Chen, J. N. DeCerbo, G. G. Carmichael, *EMBO J.* **2008**, *27*, 1694.

[47] L. L. Chen, G. G. Carmichael, *Mol. Cell* **2009**, *35*, 467.

[48] L. Bazak, E. Y. Levanon, E. Eisenberg, *Nucleic Acids Res.* **2014**, *42*, 6876.

[49] I. X. Wang, E. So, J. L. Devlin, Y. Zhao, M. Wu, V. G. Cheung, *Cell Rep.* **2013**, *5*, 849.

[50] R. A. Elbarbary, W. Li, B. Tian, L. E. Maquat, *Genes Dev.* **2013**, *27*, 1495.

[51] K. Pestal, C. C. Funk, J. M. Snyder, N. D. Price, P. M. Treuting, D. B. Stetson, *Immunity* **2015**, *43*, 933.

[52] B. J. Liddicoat, R. Piskol, A. M. Chalk, G. Ramaswami, M. Higuchi, J. C. Hartner, J. B. Li, P. H. Seeburg, C. R. Walkley, *Science* **2015**, *349*, 1115.

[53] N. M. Mannion, S. M. Greenwood, R. Young, S. Cox, J. Brindle, D. Read, C. Nellaker, C. Vesely, C. P. Ponting, P. J. McLaughlin, M. F. Jantsch, J. Dorin, I. R. Adams, A. D. Scadden, M. Ohman, L. P. Keegan, M. A. O'Connell, *Cell Rep.* **2014**, *9*, 1482.

[54] A. Fort, K. Hashimoto, D. Yamada, M. Salimullah, C. A. Keya, A. Saxena, A. Bonetti, I. Voineagu, N. Bertin, A. Kratz, Y. Noro, C. H. Wong, M. de Hoon, R. Andersson, A. Sandelin, H. Suzuki, C. L. Wei, H. Koseki, F. Consortium, Y. Hasegawa, A. R. Forrest, P. Carninci, *Nat. Genet.* **2014**, *46*, 558.

[55] C. J. Cohen, W. M. Lock, D. L. Mager, *Gene* **2009**, *448*, 105.

[56] A. B. Conley, J. Piriyapongsa, I. K. Jordan, *Bioinformatics* **2008**, *24*, 1563.

[57] E. M. Zdobnov, F. Tegenfeldt, D. Kuznetsov, R. M. Waterhouse, F. A. Simao, P. Ioannidis, M. Seppey, A. Loetscher, E. V. Kriventseva, *Nucleic Acids Res.* **2016**.

[58] ENSEMBL compara 1999-2018. http://www.ensembl.org/info/genome/compara/index.html

[59] L. L. Chen, G. G. Carmichael, *Cell Cycle* **2008**, *7*, 3294.

[60] M. G. Blango, B. L. Bass, *Genome Res.* **2016**, *26*, 852.

[61] J. B. LeGendre, Z. T. Campbell, P. Kroll-Conner, P. Anderson, J. Kimble, M. Wickens, *J. Biol. Chem.* **2013**, *288*, 2532.

[62] J. D. Laver, X. Li, K. Ancevicius, J. T. Westwood, C. A. Smibert, Q. D. Morris, H. D. Lipshitz, *Nucleic Acids Res.* **2013**, *41*, 9438.

[63] D. Kordis, N. Lovsin, F. Gubensek, *Syst. Biol.* **2006**, *55*, 886.

[64] J. König, K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D. J. Turner, N. M. Luscombe, J. Ule, *Nat. Struct. Mol. Biol.* **2010**, *17*, 909.

[65] Z. Cienikova, S. Jayne, F. F. Damberger, F. H. Allain, C. Maris, *RNA* **2015**, *21*, 1931.

[66] A. Corvelo, E. Eyras, *Genome Biol.* **2008**, *9*, R141.

[67] F. Jacob, *Science* **1977**, *196*, 1161.

[68] C. B. Lowe, G. Bejerano, D. Haussler, *Proc. Natl. Acad. Sci.* **2007**, *104*, 8005.

[69] R. M. Buckley, R. D. Kortschak, J. M. Raison, D. L. Adelson, *Genome Biol. Evol.* **2017**, *9*, 2336.

[70] C. I. Castillo-Davis, S. L. Mekhedov, D. L. Hartl, E. V. Koonin, F. A. Kondrashov, *Nat. Genet.* **2002**, *31*, 415.

[71] L. Dai, M. S. Taylor, K. A. O'Donnell, J. D. Boeke, *Mol. Cell. Biol.* **2012**, *32*, 4323.

[72] A. J. Doucet, J. E. Wilusz, T. Miyoshi, Y. Liu, J. V. Moran, *Mol. Cell* **2015**, *60*, 728.

[73] J. Y. Lee, Z. Ji, B. Tian, *Nucleic Acids Res.* **2008**, *36*, 5581.

[74] V. Perepelitsa-Belancio, P. Deininger, *Nat. Genet.* **2003**, *35*, 363.

[75] M. Jenal, R. Elkon, F. Loayza-Puch, G. van Haaften, U. Kuhn, F. M. Menzies, J. A. Vrielink, A. J. Bos, J. Drost, K. Rooijers, D. C. Rubinsztein, R. Agami, *Cell* **2012**, *149*, 538.

[76] S. Peddigari, P. W. Li, J. L. Rabe, S. L. Martin, *Nucleic Acids Res.* **2013**, *41*, 575.

[77] L. L. Hall, D. M. Carone, A. V. Gomez, H. J. Kolpa, M. Byron, N. Mehta, F. O. Fackelmayer, J. B. Lawrence, *Cell* **2014**, *156*, 907.

[78] M. Percharde, C. J. Lin, Y. Yin, J. Guan, G. A. Peixoto, A. Bulut-Karslioglu, S. Biechele, B. Huang, X. Shen, M. Ramalho-Santos, *Cell* **2018**, *174*, 391.

[79] M. Kipp, F. Gohring, T. Ostendorp, C. M. van Drunen, R. van Driel, M. Przybylski, F. O. Fackelmayer, *Mol. Cell. Biol.* **2000**, *20*, 7480.

[80] O. Nayler, W. Stratling, J. P. Bourquin, I. Stagljar, L. Lindemann, H. Jasper, A. M. Hartmann, F. O. Fackelmayer, A. Ullrich, S. Stamm, *Nucleic Acids Res.* **1998**, *26*, 3542.

[81] C. Mahiet, C. M. Swanson, *Biochem. Soc. Trans.* **2016**, *44*, 1417.

[82] C. M. Stoltzfus, J. M. Madsen, *Curr. HIV Res.* **2006**, *4*, 43.

[83] N. Gilbert, S. Lutz-Prigge, J. V. Moran, *Cell* **2002**, *110*, 315.

[84] D. E. Symer, C. Connelly, S. T. Szak, E. M. Caputo, G. J. Cost, G. Parmigiani, J. D. Boeke, *Cell* **2002**, *110*, 327.

[85] L. Wang, K. Dou, S. Moon, F. J. Tan, Z. Z. Zhang, *Cell* **2018**, *174*, 1082.

[86] A. Dupressoir, T. Heidmann, *Mol. Cell. Biol.* **1996**, *16*, 4495.

[87] H. M. Rowe, J. Jakobsson, D. Mesnard, J. Rougemont, S. Reynard, T. Aktas, P. V. Maillard, H. Layard-Liesching, S. Verp, J. Marquis, F. Spitz, D. B. Constam, D. Trono, *Nature* **2010**, *463*, 237.

[88] D. Wolf, S. P. Goff, *Cell* **2007**, *131*, 46.

[89] N. Castro-Diaz, G. Ecco, A. Coluccio, A. Kapopoulou, B. Yazdanpanah, M. Friedli, J. Duc, S. M. Jang, P. Turelli, D. Trono, *Genes Dev.* **2014**, *28*, 1397.

[90] F. M. Jacobs, D. Greenberg, N. Nguyen, M. Haeussler, A. D. Ewing, S. Katzman, B. Paten, S. R. Salama, D. Haussler, *Nature* **2014**, *516*, 242.

[91] J. H. Thomas, S. Schneider, *Genome Res.* **2011**, *21*, 1800.

[92] H. M. Rowe, A. Kapopoulou, A. Corsinotti, L. Fasching, T. S. Macfarlan, Y. Tarabay, S. Viville, J. Jakobsson, S. L. Pfaff, D. Trono, *Genome Res.* **2013**, *23*, 452.

[93] C. Feschotte, *Nat. Rev. Genet.* **2008**, *9*, 397.

[94] M. Imbeault, P. Y. Helleboid, D. Trono, *Nature* **2017**, *543*, 550.

[95] C. Berthelot, D. Villar, J. E. Horvath, D. T. Odom, P. Flicek, *Nat. Ecol. Evol.* **2018**, *2*, 152.

[96] C. G. Danko, L. A. Choate, B. A. Marks, E. J. Rice, Z. Wang, T. Chu, A. L. Martins, N. Dukler, S. A. Coonrod, E. D. Tait Wojno, J. T. Lis, W. L. Kraus, A. Siepel, *Nat. Ecol. Evol.* **2018**, *2*, 537.

[97] A. M. Ivancevic, R. D. Kortschak, T. Bertozzi, D. L. Adelson, *Genome Biol.* **2018**, *19*, 85.

[98] H. S. Malik, W. D. Burke, T. H. Eickbush, *Mol. Biol. Evol.* **1999**, *16*, 793.

[99] S. M. Hebsgaard, P. G. Korning, N. Tolstrup, J. Engelbrecht, P. Rouze, S. Brunak, *Nucleic Acids Res.* **1996**, *24*, 3439.

[100] G. Yeo, C. B. Burge, *J. Comput. Biol.* **2004**, *11*, 377.

[101] M. A. O'Leary, J. I. Bloch, J. J. Flynn, T. J. Gaudin, A. Giallombardo, N. P. Giannini, S. L. Goldberg, B. P. Kraatz, Z. X. Luo, J. Meng, X. Ni, M. J. Novacek, F. A. Perini, Z. S. Randall, G. W. Rougier, E. J. Sargis, M. T. Silcox, N. B. Simmons, M. Spaulding, P. M. Velazco, M. Weksler, J. R. Wible, A. L. Cirranello, *Science* **2013**, *339*, 662.

[102] N. Lovsin, F. Gubensek, D. Kordis, *Mol. Biol. Evol.* **2001**, *18*, 2213.