REGULATORY ELEMENTS

Regulatory activities of transposable elements: from conflicts to benefits

Edward B. Chuong, Nels C. Elde and Cédric Feschotte

Abstract | Transposable elements (TEs) are a prolific source of tightly regulated, biochemically active non-coding elements, such as transcription factor-binding sites and non-coding RNAs. Many recent studies reinvigorate the idea that these elements are pervasively co-opted for the regulation of host genes. We argue that the inherent genetic properties of TEs and the conflicting relationships with their hosts facilitate their recruitment for regulatory functions in diverse genomes. We review recent findings supporting the long-standing hypothesis that the waves of TE invasions endured by organisms for eons have catalysed the evolution of gene-regulatory networks. We also discuss the challenges of dissecting and interpreting the phenotypic effect of regulatory activities encoded by TEs in health and disease.

Genetic drift
A process by which mutations become fixed in the population by chance alone.

Transposable elements (TEs) are ubiquitous in eukaryotic genomes and persist through independent replication of their sequences. The idea that TEs have a fundamental role in the evolution of eukaryotic gene regulation reaches back 75 years to the seminal work of Barbara McClintock on 'controlling elements' of maize. She regarded these elements as "normal components of the chromosome responsible for controlling, differentially, the time and type of activity of individual genes" (REF. 1). McClintock's perspective, although scorned by some of her contemporaries, was further developed by other pioneers, most notably Britten and Davidson in the late 1960s. Building on early insights into the complex repetitive nature of eukaryotic genomes[2], which they correctly attributed to transposition activity, Britten and Davidson envisioned a model in which the amplification of diverse repeat families in the genome could spread 'pre-built' regulatory elements to drive the evolution of gene-regulatory networks[3].

Half a century later, we now appreciate that the movement and accumulation of TEs in genomes may be solely explained by their 'selfish' replication activities[4,5] and other non-adaptive forces functioning at the level of the host population, such as genetic drift, inexorably shaping genome architecture[6]. Many studies have documented the disruptive and often deleterious effect of these activities, as well as the more 'constructive' influence of TEs in the evolution of chromosome structure and gene content. But to what extent the pervasive colonization of genomes by TEs has affected the evolution of eukaryotic gene regulation remains a matter of speculation and controversy. At the heart of the debate lies a series of recent large-scale analyses of the genetic regulatory landscape of mammalian cells, revealing the engagement of an unexpectedly large fraction of TE sequences in a wide range of regulatory processes and molecular interactions. These observations, now reported for a diverse range of organisms, have rejuvenated some of the original ideas proposed by McClintock, Britten and Davidson and repositioned transposition as a potent mechanism underlying the evolution of transcriptional gene networks in eukaryotes.

In this Review, we consider emerging evidence that reveals TEs as a genome-wide source of regulatory elements. We also discuss recent advances in our ability to experimentally capture the regulatory activities of TEs, with a primary focus on their contribution as *cis*-regulatory DNA elements. We argue that most of these regulatory activities can be interpreted as relics of strategies used by TEs to spread within genomes and host populations. This simple view bypasses the need to evoke widespread adaptive consequences for the host and recognizes the occasional long-term repercussions of TEs on genome evolution. We weigh current challenges in deciphering the biological effect and evolutionary relevance of TE-derived regulatory activities, drawing examples mostly from studies carried out in mammalian species, but also other eukaryotes, including plants and insects. Finally, we propose how unbalanced control of TE-derived functions might result in transcriptional misregulation, promoting disease states.

What predisposes TEs to *cis*-regulatory activity?

Evolutionary origins of TE cis-regulatory activities. Autonomous TEs encode genes that promote their replication independently of that of host chromosomes. However, as genomic parasites, TEs rely on host cell

Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah 84103, USA.

Correspondence to C.F.
cedric@genetics.utah.edu

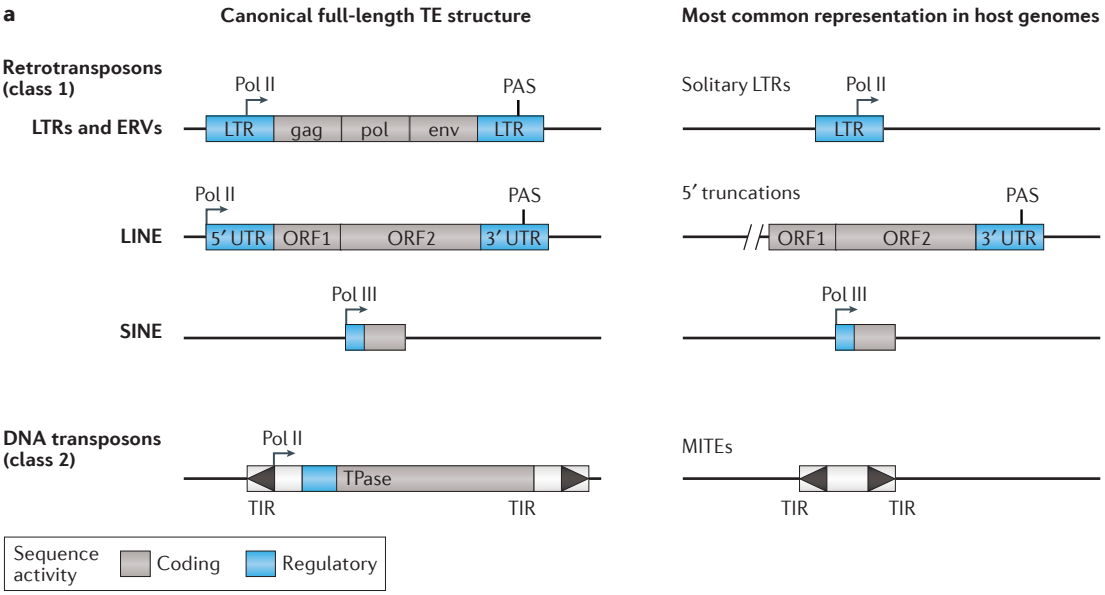
doi:10.1038/nrg.2016.139
Published online 21 Nov 2016

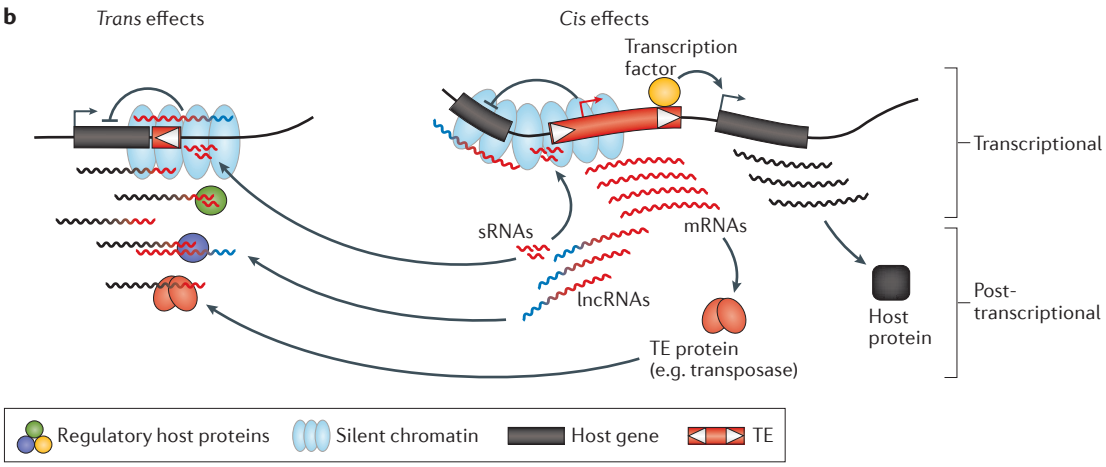
Cis-regulatory sequences
Segments of DNA that regulate the transcription of adjacent genes.

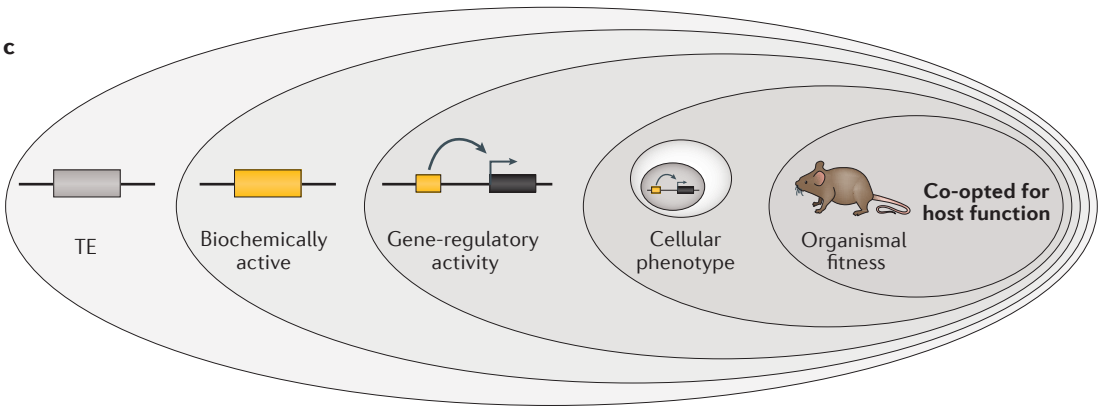
Long terminal repeat elements
(LTR elements). A class of retrotransposons containing direct LTRs flanking the protein-coding sequence.

Long interspersed nuclear elements
(LINEs). Class of non-long terminal repeat retrotransposons that retrotranspose by target-primed reverse transcription.

machinery to express their genes. Thus replication-competent TEs have evolved *cis*-regulatory sequences that function to mimic host promoters (FIG. 1a). In long terminal repeat elements (LTR elements) such as those of endogenous retroviruses (ERVs), *cis*-regulatory sequences and RNA polymerase II (Pol II) promoters are present in duplicate within each of the two LTRs flanking the coding sequence of the elements[7]. By contrast, full-length long interspersed nuclear elements (LINEs) possess an internal Pol II promoter located in their 5′ untranslated region (UTR)[8] as well as an antisense promoter[9]. Other TEs collectively referred to as







Short interspersed nuclear elements
(SINEs). Class of non-autonomous retrotransposons that are copied by the LINE replication machinery.

DNA transposons
Transposable elements that do not generate an RNA intermediate during transposition, which generally occurs through a 'cut-and-paste' mechanism.

short interspersed nuclear elements (SINEs) are derived from cellular genes transcribed by Pol III (for example, transfer RNA (tRNA) or 7SL RNA). Thus, an intact, full-length SINE copy contains internal sequence motifs (A and B boxes) that are capable of recruiting Pol III[8].

The diversity of TEs in their abundance, form and replication mechanism greatly affects the fate of the promoters and the *cis*-regulatory elements they carry. For instance, the mechanism of LTR element replication dictates that each new insertion will introduce two exact copies of the LTR in the host genome[7]. However, following insertion, these elements often undergo ectopic recombination between their LTRs. These events result in the removal of the coding regions of these elements but leave intact solitary LTRs, which contain the original *cis*-regulatory sequences. ERVs occupy ~8% of the human genome but 90% exist as solitary LTRs[10]. By contrast, LINEs account for a larger fraction of the human genome (~20%), but the vast majority suffered 5′ truncations upon insertion that removed their promoter sequences[10]. Similarly, DNA transposons, which generally transpose via a cut-and-paste mechanism, are mostly propagated as miniature inverted repeat TEs (MITEs), which arise from internal deletion derivatives of autonomous elements[11]. Predictably, many MITEs lack the promoter sequences of their parental element. In summary, not all TE insertions are 'created equal' in their ability to retain their original promoters and *cis*-regulatory elements (FIG. 1a).

As for canonical host gene promoters, TE promoters often show spatially or temporally regulated activity that is dependent on cell type or that is in response to environmental cues such as stress or infection[12,13]. Although few TE promoters have been extensively characterized, it is known that a range of TEs have regulated patterns of expression that are established by clusters of *cis*-regulatory sequences. Much like canonical promoters and enhancers, these sequences can recruit and 'integrate' specific combinations of host-encoded transcription factors[8,14–21].

What evolutionary forces shaped the regulatory features of TE promoters? For new TE insertions to persist through vertical inheritance, transposition events must occur in the germline or before germline development, which results in strong selection for TEs to be transcriptionally active in the germline. Indeed, the expression of many TEs seems to be restricted to various stages of gametogenesis or early embryogenesis in both plants[22] and animals[23] (BOX 1). Paradoxically, some TEs also have tightly regulated activity in somatic tissues in a range of organisms[22–25]. Given that somatic transposition events are not *trans*-generationally inherited, activity in these tissues holds no immediately apparent benefit or consequence to the TE. It is tempting to speculate that regulated somatic TE activity may reflect a symbiotic relationship between host and TE; an alternative but not mutually exclusive explanation is that these somatic regulatory activities have been previously moulded by selection on the TE. For example, all ERVs originate from infectious retroviruses that must have infected germ cells (or their progenitor cells) but possibly also somatic cells from diverse tissues. The past viral lifestyle of ERVs may thus explain why their LTRs harbour complex regulatory elements that are capable of driving transcription in a range of tissues and cell types. Regardless of the evolutionary forces driving somatic activity, it is clear that TEs have the potential to affect host gene regulation well beyond the germline and early embryo.

Nonrandom integration of TEs in the genome. As TEs evolved regulatory sequences that essentially mimic host *cis*-regulatory elements, it naturally follows that a TE insertion that lands in the vicinity of a host gene has a strong potential to interfere with its expression. Thus, where a TE initially inserts in the genome will often dictate its fate in the population. Accordingly, many TEs seem to have evolved mechanisms that favour integration into genomic regions that maximize their chance of propagation. For instance, some TE families have evolved sophisticated molecular mechanisms to target genomic 'safe havens', such as gene-poor or heterochromatic regions, whereas others favour integration within areas of open, transcriptionally active chromatin[26]. For instance, *Mutator* elements in maize[27], *mPing* MITEs in rice[28], Tf1 retroelements in fission yeast[29] and P elements in *Drosophila melanogaster*[30] all actively target the 5′ region of genes. Some retroviruses, including HIV-1, preferentially integrate within highly transcribed regions of the genome[31]. Presumably, this pattern of propagation ensures that newly inte grated elements reside in a favourable environment for expression and transmission. Such elements are likely to have a stronger propensity for modulating adjacent gene expression and, if tolerated by natural selection, may be more prone to *cis*-regulatory co-option — possibly immediately upon insertion[28]. Finally, it is important to note that as time elapses after TE propagation, forces such as selection and genetic drift are likely to obscure initial integration preferences.

TEs are a rich source of regulatory activities

What is the effect of the widespread dispersal of *cis*-regulatory elements by TEs on genome regulation and cell function? To cope with the parasitic burden of TEs in their genomes, eukaryotes seem to have evolved multilayered mechanisms to prevent TE transcription[22,32,33].

◀ Figure 1 | **Origins of TE regulatory activities and how they may affect host genes.** **a** | Schematic of major transposable element (TE) classes and their typical genetic organization. The left panel depicts the general full-length version of each TE type. Most TEs harbour regulatory sequences that function to promote their own transcription and regulation, such as promoters for RNA polymerase II (Pol II) or Pol III, and polyadenylation signals (PASs). The right panel shows the structures of each type of TE as they most commonly occur in the genome, which can differ substantially depending on the TE (see the main text for details). **b** | Diagram depicting different types of regulatory activities exerted by TEs. These include effects mediated by *cis*-regulatory DNA and RNA elements (right panel) as well as *trans* effects mediated by TE-produced non-coding RNAs and proteins (left panel). **c** | Hierarchy of evidence to consider when determining whether a TE has been co-opted for host functions. Many TEs have biochemical hallmarks of regulatory activity on the basis of genome-wide assays. However, additional evidence is required to determine which of these TEs alter the regulation of host genes and affect organismal phenotypes and fitness. ERV, endogenous retrovirus; LINE, long interspersed nuclear element; lncRNA, long non-coding RNA; LTR, long terminal repeat; MITE, miniature inverted-repeat transposable element; SINE, short interspersed nuclear element; sRNA, small RNA; TIR, terminal inverted repeat; TPase, transposase.

RNA sequencing
(RNA-seq). High-throughput sequencing of complementary DNAs derived from RNAs extracted from cells or tissues.

Cap analysis of gene expression followed by sequencing
(CAGE-seq). A method used to precisely map the transcription start sites of capped RNAs genome-wide.

Consequently, the bulk of TE-derived DNA is generally assumed to be silenced and biochemically inert in most cells. However, examples of host genes driven by TE promoters have been documented in diverse species over the past several decades[14,21,34–36]. Do these cases simply represent rare events in which individual TEs have escaped host silencing or are they hints of pervasive regulatory activity?

A large body of literature has shown the myriad mechanisms by which TEs can alter host gene expression both in *cis* and in *trans*, transcriptionally or post-transcriptionally[14,37] (FIG. 1b). Although most of these mechanisms were initially discovered through the genetic analysis of mutations caused by *de novo* TE insertions, genomic studies have revealed that these activities permeate throughout the genome and often emanate from TEs that have long been fixed in the host population. In particular, genome-wide assays such as those mapping transcriptional activity, open chromatin or binding of transcription factors have provided compelling evidence that a multitude of TE-derived sequences have the biochemical hallmarks of active regulatory elements. Together, these studies indicate that TEs are not as robustly or systematically silenced as is commonly assumed.

Notably, transcriptome analyses have uncovered TEs as an abundant source of tissue-specific and/or alternative promoters. Two principal methods have allowed the mapping of TE-derived promoters at a genome-wide scale. First, deep sequencing of full-length cDNAs (a form of RNA sequencing (RNA-seq)) revealed that chimeric transcripts that initiate within TE-derived promoters constitute a considerable fraction of mammalian transcriptomes, notably during early development[38] (BOX 1) but also in many other mammalian tissues[39]. Second, approaches mapping sites of transcription initiation, such as cap analysis of gene expression followed by sequencing (CAGE-seq) revealed surprising amounts of Pol II initiation within TEs in a range of human and mouse cell types and tissues[12] (for example, up to 30% of all transcription start sites mapped in human embryonic

Box 1 | **A role for ERVs in regulating early mammalian development?**

Genome-wide epigenetic silencing of transposable elements (TEs) is a pivotal step in early mammalian development, but recent studies characterizing the transcriptomes of embryonic stem cells (ESCs) and early embryos have revealed surprisingly high transcriptional activity emanating from endogenous retrovirus (ERV) sequences. ERVs are among the first sequences to be transcribed during zygotic genome activation in mouse two-cell embryos[38]. Similarly, in human embryos, distinct families of primate-specific ERVs are expressed at each stage of pre-implantation development, and this activity ceases as cells differentiate into somatic cells[155,156]. These ERVs are transcriptionally regulated by long terminal repeat (LTR) promoters that contain binding sites for transcription factors controlling early development, such as OCT4 and NANOG[23]. Taken together, these findings indicate that ERVs are seemingly unleashed during pre-implantation development, with potentially widespread regulatory effects on the cellular transcriptome.

A major challenge now lies in understanding the biological consequences of ERV activity in embryonic development. RNA sequencing (RNA-seq) analyses have revealed that hundreds of LTRs drive expression of long non-coding RNAs (lncRNAs)[157] and protein-coding genes[38,158], and chromatin immunoprecipitation followed by sequencing (ChIP-seq) analyses have revealed thousands more solitary LTRs that function as ESC-specific distal enhancers that may also influence host gene expression[17]. The expression of some host genes that are required for pre-implantation development may thus be regulated by ERV-derived promoters or enhancers.

The activity of some of these ERVs, such as the primate-specific HERV-H, is highly correlated with stem cell pluripotency[138,158–160] and knockdown of lncRNAs driven by HERV-H results in rapid differentiation into somatic cell types[158,160–163]. Further molecular investigation in human ESCs suggested that HERV-H transcripts may function as RNA scaffolds recruiting transcriptional activators, including OCT4, Mediator and p300 (REF. 163). Another lncRNA — named human pluripotency-associated transcript 5 (*HPAT5*) and derived from both a primate-specific HUERS-P1 ERV and an *Alu* element — was also discovered to promote pluripotency by functioning as a molecular sponge for the *let-7* family of microRNAs (miRNAs)[164] (FIG. 2a). Together, these results indicate that ERV-derived lncRNAs are capable of modulating stem cell pluripotency, which may be important for proper development. Another study found that the protein Rec encoded by HERV-K binds to a multitude of host mRNAs and potentiates antiviral resistance in cell culture, raising the provocative hypothesis that HERV-K expression protects pre-implantation embryos from exogenous viral infection[156].

These data provide mounting evidence that ERVs have been co-opted as essential regulators of human development. However, an important limitation to these studies is that they are typically carried out in cultured ESCs, thus it remains possible that functional phenotypes linked to ERV-regulatory activity may not be required for organismal development. One pioneering study used small interfering RNA (siRNA) injections to simultaneously knock down expression of three ERV-derived lncRNAs (*HPAT2*, *HPAT3* and *HPAT5*) in human two-cell embryos and found that cells depleted of these lncRNAs were no longer capable of contributing to the inner cell mass of the blastocyst[164]. This result suggests, but does not prove, that ERVs have been co-opted to regulate blastocyst development. Given the limited sample size and potential for off-target or transcriptional effects of siRNA silencing, further studies are necessary to conclusively establish a developmental role for human ERVs. Indeed it is possible that ERV activity in ESCs is not essential for development but instead merely reflects the recent selfish exploitation of this developmental niche by retroviruses[165].

Although the necessary experiments would be challenging with human material, mice similarly show dynamic expression of ERVs during early embryonic development[23]. Despite all ERVs with regulatory activity in human embryos being primate specific, mouse embryos similarly show rodent-specific ERVs that are expressed throughout early development[38,157]. If ERV activity is truly required for both human and mouse early development, it would imply a remarkable scenario in which ERVs were independently co-opted in primates and rodents to regulate embryogenesis. Note that there are precedents for such convergent TE co-option events[166,167] (BOX 3).

Box 2 | **TEs as a source of non-coding regulatory RNAs**

In addition to providing *cis*-regulatory DNA elements, transposable elements (TEs) have also been documented to contribute non-coding regulatory sequences that modulate gene expression post-transcriptionally. The best documented are TE-derived regulatory sequences embedded within the untranslated regions (UTRs) of protein-coding genes. For instance, *Alu* elements inserted within 3′ UTRs can provide recognition sequences for Staufen 1-mediated mRNA decay[168] or can influence mRNA localization[169], whereas *Alu* elements incorporated within 5′ UTRs often modulate mRNA translation efficiency[170].

TEs are also major contributors to the evolutionary origination and biogenesis of various regulatory RNAs that regulate gene expression either transcriptionally or post-transcriptionally, including microRNAs (miRNAs)[171], long non-coding RNAs (lncRNAs)[138,139] and circular RNAs (circRNAs)[172,173]. In human, mouse and zebrafish, more than two-thirds of lncRNAs were found to contain exonic TE sequences, whereas TEs seldom occur in protein-coding transcripts[138,139]. Of course, many of these lncRNAs may have no biological function and could merely reflect the pervasive transcriptional activity of TEs. They could also reflect the greater capacity of lncRNAs to tolerate and to assimilate TE insertions over evolutionary time[14,69,174]. Nevertheless, there is a growing number of examples in which the sequences conferring regulatory activities to non-coding RNAs are directly derived from TEs[69,163,164,174–176] (BOX 1).

Many of the conceptual ideas developed in this Review for TE-derived *cis*-regulatory DNA elements can be applied to those functioning at the RNA level. In particular, a *cis*-acting sequence (for example, a recognition motif for an RNA-binding protein) that is present within an ancestral TE sequence will be seeded as the TE amplifies throughout the genome, opening the door for the co-option of multiple TE copies to assemble complex regulatory circuits modulating the expression of a vast number of genes[14,175,177]. Further details on the role of TEs in the evolution of regulatory RNAs and post-transcriptional gene regulation have been reviewed elsewhere[37,69,174].

cells) as well as in *D. melanogaster* embryonic development[40]. There is also growing evidence that TEs are a substantial source of promoter activity in plants[36]. Thus TEs seem to have dispersed vast numbers of developmentally regulated promoters in a wide range of species that often remain active and drive the transcription of adjacent DNA in a tissue- or stage-specific manner.

Evidence is mounting that TEs provide a profusion of other types of *cis*-regulatory elements including enhancers, insulators and repressive elements. In mammals, chromatin immunoprecipitation followed by sequencing (ChIP-seq) studies have revealed that for any given transcription factor and cell type examined, TEs contribute a substantial fraction of binding sites across the genome (5–40%; average ~20%)[15,19]. LTR elements tend to contribute more than other TE types[14,21,41], probably because LTRs are more likely to possess and to retain their ancestral *cis*-regulatory activities as explained above (FIG. 1a). Importantly, for a given transcription factor the majority of TE-derived binding sites are contributed by a fairly small number of specific TE families that are highly enriched for transcription factor-binding events compared with what would be expected on the basis of the density of these TE families in the genome[17,19,42,43]. In most cases for which the origin and the sequence of the binding sites have been examined, it could be inferred that a canonical binding motif pre-existed within the ancestral TE sequence and was subsequently dispersed through the genome via transposition — a scenario consistent with a 'copy-and-paste' model of transcription factor-binding site dispersion[14].

Interestingly, TE-derived transcription factor-binding sites tend to be lineage specific. For instance, in an analysis of binding events for 26 pairs of orthologous transcription factors across two comparable human and mouse cell lines, >98% of >130,000 TE-derived peaks identified in each species were species specific[19]. This result can be partly explained by the fact that the majority of binding events occur within TEs that have amplified after the split of the two species examined; that is, primate- or rodent-specific TE families. Another factor is that more ancient TEs have accumulated many more neutral substitutions, leading to degradation of their ancestral transcription factor-binding sites. The differential decay of these ancestral TE sequences across species may also result in species-specific transcription factor binding events. In any case, these data suggest that transposition represents a common mechanism for the gain of novel transcription factor-binding sites during mammalian evolution.

TEs have also been documented to function as insulator and/or boundary elements. Such sequences function to partition the genome into domains of active or inactive transcription ranging in size from 100 kb to 1 Mb, often by preventing the spread of heterochromatin[44]. Several studies showed that many TEs, particularly SINEs, harbour binding sites for factors such as CTCF or TFIIIC (also known as GTF3C1) that confer insulator activity and organize nuclear architecture[45–47]. A subset of these TEs seem to have roles in the three-dimensional organization of the genome by functioning as 'anchors' that isolate regions of active transcription. Indeed, studies investigating intra- or inter-chromosomal interactions underlying these topologies at a genome-wide scale using chromatin-interaction profiles have found that SINEs are enriched at the borders of these domains[48–50]. By providing a fertile source of binding sites for architectural factors, TEs may be important contributors to high-order genomic organization, which controls the transcriptional regulation of large chromosomal regions containing many genes.

In addition to providing *cis*-regulatory DNA elements, TEs have also been documented to contribute a wealth of non-coding regulatory RNA transcripts, such as microRNAs (miRNAs) and long non-coding RNAs (lncRNAs), which can modulate gene expression in *cis* or in *trans* (BOX 2; FIG. 1b).

Finally, although we primarily focus this Review on regulatory activities encoded within TE sequences, TEs can also alter gene expression through many other effects, including the disruptive effects of their chromosomal integration. An interesting example is the adaptive insertion of a *POGON1* DNA transposon within the 3′ UTR of the gene *CG11699* in *D. melanogaster*[51]. The insertion disrupts the 'normal' (ancestral) polyadenylation signal of the gene resulting in a shorter 3′ UTR, elevated mRNA levels and increased resistance to xenobiotic stress[51]. TEs may also exert a broad influence on genome regulation as a result of being targeted by host silencing pathways[32] (FIG. 1b). In particular, the repressive chromatin nucleated at TE sequences may 'spread' to adjacent regions and silence the expression of nearby genes, a phenomenon that has been observed in several organisms including plants[52] and mammals[53].

Chromatin immunoprecipitation followed by sequencing (ChIP-seq). A method for identifying protein–DNA interactions genome-wide. Following crosslinking, a protein of interest is immunoprecipitated and its binding sites in the genome are identified by high-throughput sequencing of the co-purified DNA fragments.

The biological effect of TE regulatory activities

The work summarized above firmly anchors the idea that TEs are a prolific source of biochemical regulatory activity in host cells. These observations raise the question as to whether this phenomenon has exerted a considerable influence on the biology and the phenotypic evolution of species. Although it may be tempting to interpret the tightly controlled regulatory activity originating from TEs as indicative of a widespread role in physiology or development, the biochemical manifestation of this activity alone does not signify that it is important for proper genome function[54,55] (FIG. 1c). The terms 'function', 'activity' or 'co-option' are widely used but can have different meanings depending on the field of study and we caution that imprecise use of these terms might confuse our understanding of the influence of TE regulatory activity on biological outcomes. Coming from an evolutionary perspective, we use the term 'co-option', which is also referred to as exaptation[56] or domestication[57], when a TE-derived sequence acquired a cellular function conferring a selectable phenotype that predictably increases the fitness of the host organism (FIG. 1c). Given that this Review focuses on regulatory activity of TEs rather than on their protein products, we will use the term 'co-opted TEs' to refer specifically to the co-option of DNA sequences within the TE that are responsible for *cis*-regulatory activity.

With the goal of advancing our current understanding of the biological importance of TE regulatory activities, we discuss the major challenges involved in disentangling effects that are simply relics of the selfish 'behaviour' of TEs with no benefit to the host from those that reflect bona fide co-option events resulting in adaptive cellular innovations.

Co-option of TEs revealed by evolutionary sequence conservation. How can evolutionarily co-opted, or 'domesticated', TEs be distinguished from TEs with regulatory activity that is non-adaptive and simply tolerated by host cells? One of the most direct lines of evidence that a DNA sequence has exerted biological function is to show that it has evolved under a regime of purifying selection. Selective constraint on non-coding DNA is most effectively revealed by detecting a signature of evolutionary sequence conservation when compared across distantly related species, relative to DNA sequences that are assumed to be unconstrained and neutrally evolving. Such comparative genomic studies have revealed tens of thousands of non-coding TE fragments in the human genome that are orthologous and highly conserved across species and that show clear signatures of purifying selection[14,55,58]. Although most of these sequences derive from relatively ancient TEs (which are old enough to be compared orthologously across distant species), they are of various ages and types. They are also unevenly distributed in the genome and tend to be enriched near transcription factor genes and other developmental genes. Together, these data strongly suggest that TEs have indeed provided a rich source of non-coding sequence material fuelling regulatory innovation during vertebrate evolution.

Striking examples of TEs co-opted for gene regulation identified by virtue of evolutionary conservation are those of the so-called 'living fossil' SINE (LF-SINE) family[59]. Multiple LF-SINE copies are highly conserved across tetrapods and one was shown to possess tissue-specific enhancer activity in a mouse reporter assay (TABLE 1). Another interesting example is a highly conserved mammalian-wide interspersed repeat (MIR) element that in mice functions as an intronic enhancer for the forkhead box P3 (*Foxp3*) gene required for extrathymic generation of regulatory T cells, which reinforces maternal–fetal immunotolerance during pregnancy[60]. In addition, two ancient TEs, a mammalian apparent LTR retrotransposon (MaLR) and a SINE, were independently co-opted during mammalian evolution to function as hypothalamus-specific enhancers that are required to regulate the proopiomelanocortin (*POMC*) gene, which functions in the brain to control food intake[61]. Examples such as these highlight how evolutionary conservation of TEs can guide experiments validating regulatory functions.

Although tens of thousands of TEs in the human genome contain sequences that have evolved under purifying selection[58], as well as sequence motifs and biochemical signatures that are consistent with some form of *cis*-regulatory activity[62,63], it is important to emphasize that, thus far, very few of these elements have been characterized using functional assays. Interestingly, a recent study[64] identified a DNA segment, which is highly conserved across eutherian mammals and is derived from three distinct juxtaposed TEs (AmnSINE1, X6b_DNA and MER117), that appears to function as an enhancer driving frontonasal expression of the crucial development gene *Wnt5* in the mouse embryo. However, mice carrying a homozygous deletion of the TE-derived enhancer did not show any obvious phenotypic defects, raising the paradox — not unprecedented[65] — that even sequences with a strong signature of evolutionary constraint and demonstrated *cis*-regulatory activity may only exert subtle effects on organismal development.

Uncovering recently co-opted TEs. Although sequence conservation is a strong predictor of biological function, there are several caveats that may preclude comparative genomics from identifying TEs that have been co-opted for host gene regulation. First, only relatively ancient TEs that have been co-opted for a sufficiently long evolutionary period will show a robust signature of sequence constraint[66]. These ancient TEs are technically challenging to recognize as such, especially in species with rapid evolution rate such as fruitflies (BOX 3). In human and mouse, the vast majority of TEs showing regulatory activity in functional genomics assays are non-orthologous and derive from relatively young insertions[19,41]. The recent origins of these elements make it difficult to assess whether they have evolved under functional constraint, although some methods have been developed in an attempt to address this issue[67]. The matter is worsened by the intrinsic shortness and degeneracy of the sequence motifs underlying regulatory activity, which increases the difficulty of detecting purifying selection acting on these motifs through sequence analysis alone[68,69].

Purifying selection
Selection against mutations that are deleterious to the fitness of the individual.

Reporter assay
A putative *cis*-regulatory DNA sequence is cloned upstream of a reporter gene (such as luciferase) either in an episomal vector or as a chromosomally integrated construct and tested for its ability to enhance transcription of the reporter gene.

Retrotransposon
A type of transposable element that replicates through an RNA intermediate in a 'copy-and-paste' mechanism.

Table 1 | **Case examples of TEs functioning as *cis*-regulatory DNA elements***

Species	Gene product	TE	*Cis*-regulatory activity	Function	Evidence	Refs
Tetrapod vertebrates	ISL1	LF-SINE (vertebrate SINE)	Neural enhancer	Brain development (inferred)	Sequence conservation and mouse reporter assay	59
Human	AIM2	MER41 (primate ERV)	Interferon-inducible enhancer	Regulates inflammatory response	CRISPR knockout in human cell culture	43
Human	β-Globin	ERV9 (primate ERV)	Erythroid enhancer	Controls developmental switch from fetal to adult globin	Cre–*loxP* knockout in transgenic mouse BAC and chromatin looping (3C)	181
Mouse	Dicer	MT-C (rodent LTR retrotransposon)	Oocyte-specific promoter	Necessary for female oocyte function and fertility	TALEN knockout in mouse	74
Mouse	Growth hormone	B2 SINE (mouse retrotransposon)	Insulator	Developmental regulation of growth hormone (inferred)	Enhancer blocking assay	182
Human	Prolactin	MER39 (ERV)	Endometrium-specific promoter	Prolactin expression during pregnancy	5′ RACE on endometrial tissue	183
Fly (*Drosophila melanogaster*)	Cyp6g1	Accord (LTR retrotransposon)	Enhancer	Increased resistance to DDT insecticide	Selective sweep, genetic mapping and reporter assay	184
Fly (*D. melanogaster*)	Jheh1 and Jheh2	Bari1 (DNA transposon)	Antioxidant response elements (enhancer)	Increased resistance to oxidative stress	Selective sweep and phenotypic assays	185
Fly (*Drosophila simulans*)	Slowpoke	Shellder (LTR retrotransposon)	Altered splicing	Courtship song variation	Trait mapping and *in vivo* CRISPR knockout	186
Maize	TB1	Hopscotch (teosinte LTR retrotransposon)	Enhancer	Apical dominance branching pattern	Trait mapping and reporter assay in maize leaf protoplast	76
Sicilian blood orange	Ruby	Tcs1 (LTR retrotransposon)	Cold-responsive promoter	Required for cold-induced accumulation of anthocyanin	Trait mapping and 5′ RACE	77

3C, chromatin conformation capture; AIM2, absent in melanoma 2; BAC, bacterial artificial chromosome; CRISPR, clustered regularly interspaced short palindromic repeats; DDT, dichlorodiphenyltrichloroethane; ERV, endogenous retrovirus; LF-SINE, living-fossil SINE; LTR, long terminal repeat; RACE, rapid amplification of cDNA ends; SINE, short interspersed nuclear element; TALEN, transcription activator-like effector nuclease. *Further examples can be found in the catalogue of genes affected by transposable elements (TEs; C-GATE)[35].

Despite these hurdles, there is a growing set of recent TEs for which solid evidence has been gathered to support the idea that they were co-opted for important regulatory innovation[35] (some examples are provided in FIG. 2 and TABLE 1). These are often cases in which the TE functions as a tissue-specific promoter for a gene with crucial function in that tissue. Indeed, these cases can be readily discerned by the detection of a tissue-specific gene transcript initiating within a TE sequence[34]. A classic example in mammals is the acquisition of salivary expression of the digestive enzyme amylase from a retroviral LTR inserted in the common ancestor of anthropoid primates[70]. Another interesting case is the endometrial expression of the hormone prolactin, which is crucial for pregnancy in mammals[71,72] (BOX 3).

In contrast to TEs with promoter activity, the function of TEs located distal to genes is more challenging to discern and validate. Therefore, far fewer examples of TE-derived enhancers with clear biological roles have been documented[35]. A traditional experimental approach to delineate the *cis*-regulatory effects of a TE is a reporter assay. If the reporter expression pattern driven by the TE recapitulates that of the endogenous gene associated with the TE, it may be reasonable to conclude that the TE contributes to the regulation of the endogenous gene *in vivo*[59,73]. However, these experiments are still limited by the fact that they dissociate the TE from its native chromosomal context and cannot establish a direct causal link between the *cis*-regulatory activity of the TE and the endogenous gene expression.

A more conclusive approach to assess the effects of an individual TE on host gene regulation is to carry out a loss-of-function experiment. The recent development of precise genome-editing technologies has begun to facilitate such experimental knockout studies. In one

Box 3 | **Evolutionary dynamics of TE regulatory activities**

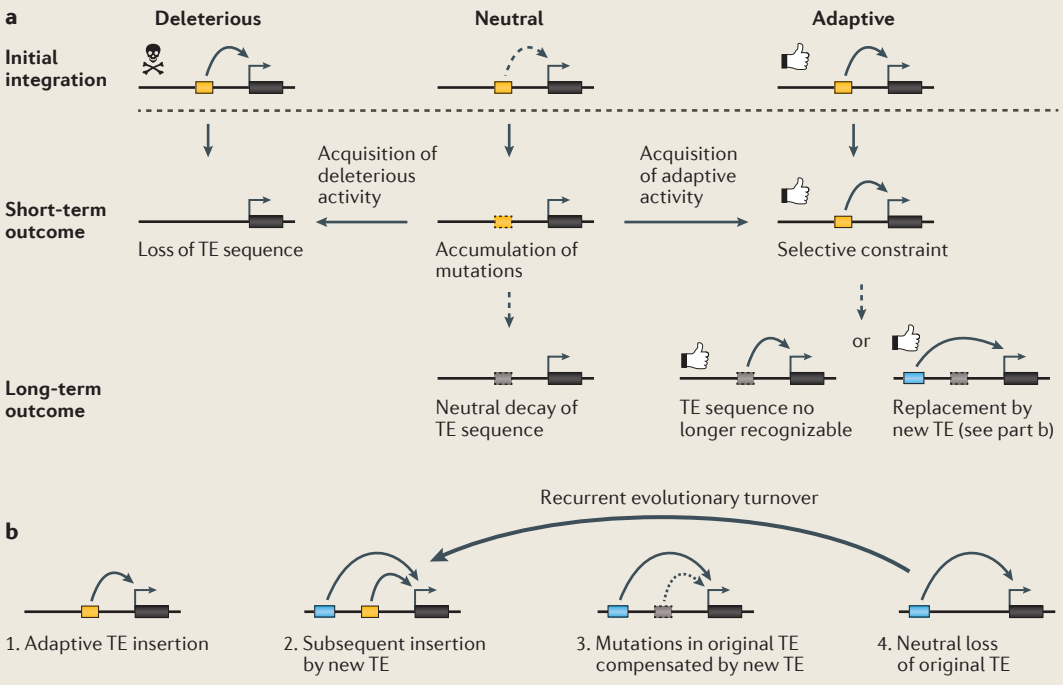
What is the mode and tempo of transposable element (TE) co-option for *cis*-regulatory function? Does TE co-option occur in major bursts that coincide with the invasion of TEs, or is it a more gradual process tapping into a reservoir of decaying TEs with various levels of predisposition?

The most straightforward path to co-option is when a TE confers an adaptive regulatory function immediately upon insertion (see the figure, part **a**, right). This scenario may conceivably be common because the *cis*-regulatory elements mapped within adaptive TEs are often inferred to have pre-existed at the time of their insertion in the genome[43]. Some of the adaptive TE insertions that recently swept in the *Drosophila melanogaster* population may represent examples of new *cis*-regulatory sequences gained by transposition[178]. It could also be that some TEs have immediate *cis*-regulatory effects that are independent of their own sequences, simply through their disruptive[51] or epigenetic activities.

There is also evidence that a TE may become 'latently' co-opted long after it inserts into the genome (see the figure, part **a**, middle). Although many TEs possess built-in *cis*-regulatory sequences, these may remain silent and/or inconsequential for adjacent gene expression until additional mutations unmask or bolster their regulatory activity. Along those lines, Emera and Wagner[71] proposed a model of 'epistatic capture' to describe the series of mutational events by which a TE was transformed into a strong promoter for decidual prolactin. It is also conceivable that such epistatic mutational events occurring outside of the TE sequence but within its genomic neighbourhood could promote latent co-option of the TE. A variety of mechanisms can be envisaged, such as point mutations introducing or removing other transcription factor-binding sites, as well as insertion or deletion events of *cis*-regulatory elements, genes or even other TEs.

What is the evolutionary fate of a TE that has become co-opted for host function? As discussed in the main text, a recent study in *Drosophila miranda* found that dosage compensation on a newly evolved 'neo-X' chromosome (~1 million years old) was mediated by the recent spread of an ISX transposon, which harbours a 'recognition element' motif for the male-specific lethal (MSL) dosage compensation complex[102]. Remarkably, the authors also found evidence that dosage compensation on the older XR chromosome (~15 million years old) was partly established by an older expansion of a related TE family, named ISXR[102]. The older ISXR elements were characterized by stronger MSL recognition elements and decaying TE sequences. The authors also found evidence that the MSL recognition elements in the younger ISX elements have already undergone additional evolutionary 'fine-tuning' to optimize binding of the MSL dosage complex, facilitated by a gene-conversion-like process between copies[179]. These data suggest a two-step model of TE domestication, whereby elements with a regulatory predisposition upon insertion are further refined by post-insertional mutation and selection. Eventually, co-opted TEs will lose signatures of their TE origins as all but the most essential nucleotides are eroded by neutral substitutions[102].

Another emerging insight is that the co-option of TEs as regulatory elements may be a surprisingly volatile process. Several cases have now been described whereby multiple species have independently, but convergently co-opted lineage-specific TEs to regulate the same gene. In mammals, this is exemplified by the prolactin gene[71] and also the NLR family apoptosis-inhibitory protein (*NAIP*) gene[180]. One potential explanation is that similar selective pressures drove independent co-option of TEs at the same locus in different species. However, it is also possible that the ancestral locus was also regulated by a TE, which has since been replaced by a new TE in each lineage. Thus, analogous to a gene duplication event, the insertion of a TE with similar regulatory properties might result in relaxed selection and eventual decay of a nearby co-opted TE, and over time this would result in a cycle of co-opted TEs being replaced by newer TEs (see the figure, part **b**). Such a turnover model could explain the prevalence of lineage-specific TEs associated with cellular processes that have deep evolutionary origins, such as pregnancy[89] or stem cell development[17].



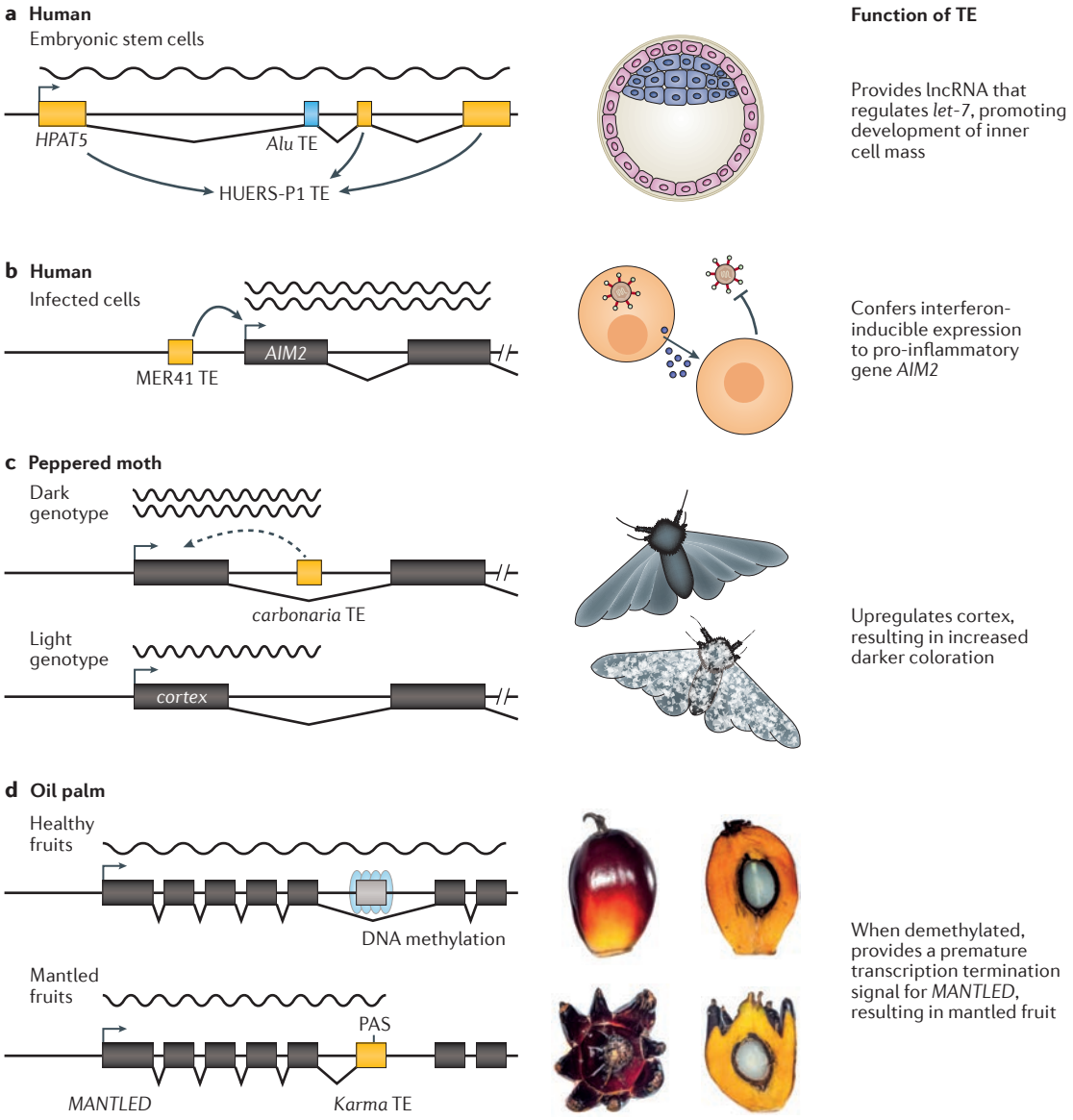


Figure 2 | **Examples of phenotypes driven by TE regulatory activity. a** | The human pluripotency-associated transcript 5 (*HPAT5*) is a long non-coding RNA (lncRNA) that is derived from a composite of a HUERS-P1 endogenous retrovirus (ERV) element and an *Alu* short interspersed nuclear element (SINE). In all figure parts, wavy lines indicate pre-mRNA transcripts and angled lines indicate spliced introns. *HPAT5* regulates the *let-7* family of microRNAs through *let-7*-binding sites carried by the *Alu* element and was shown to be essential for inner cell mass formation during early embryonic development[164]. **b** | A MER41 transposable element (TE) provides an interferon-inducible enhancer upstream of the human absent in melanoma 2 (*AIM2*) gene, which regulates inflammation in response to infection[43]. **c** | In the peppered moth, a polymorphic *carbonaria* TE insertion within an intron of the *cortex* gene enhances *cortex* expression levels (dotted line indicates an uncharacterized regulatory mechanism), which underlie the adaptive cryptic colouration that occurred during the industrial revolution[82]. **d** | In oil palm, sporadic demethylation of a *Karma* TE within an intron of the *MANTLED* gene results in unmasking of a cryptic splice acceptor site and a premature termination signal, causing the mantled fruit phenotype[143]. The left panel depicts the genomic locus (not drawn to scale). PAS, polyadenylation signal. Images in part **d** are reprinted from REF. 143, Nature Publishing Group.

study, transcription activator-like effector nuclease (TALEN)-mediated deletion of a mouse transcript (MT) retroelement providing an alternative promoter for the *Dicer1* gene led to oocyte malfunction and female sterility[74]. In another study, CRISPR-mediated deletion of a MER41 ERV-derived interferon-inducible enhancer associated with the absent in melanoma 2 (*AIM2*) gene in human cells impaired the innate immune response to viral infection[43] (FIG. 2b). In a third example, a screen for *cis* elements regulating the pregnancy-specific histocompatibility gene *HLA-G* combining massively parallel reporter assays and CRISPR genome editing identified a LTR71-derived distal enhancer required for *HLA-G* expression in the placenta[75].

TE-mediated cis-regulatory variation and the domestication of species. TEs were first discovered in maize and have since been characterized as a major component of other crops. Intriguingly, there is mounting evidence that many traits associated with the domestication of crop plants evolved through artificial selection of TEs with *cis*-regulatory effects on adjacent host genes[13,36]. In maize for instance, a *Hopscotch* retrotransposon insertion functions as an enhancer for the *teosinte branched1* gene underlying increased apical dominance compared with its ancestral wild relative teosinte[76]. Another study examining blood oranges identified a *copia*-like retrotransposon that functions as a cold temperature-inducible enhancer of the *ruby* locus that modulates fruit colour[77] (TABLE 1).

Furthermore, there is evidence that regulatory variation introduced by TEs has facilitated the domestication of animals. A study investigating the domesticated silkworm found that the insertion of a *Taguchi* LINE enhances expression of the ecdysone oxidase (*EO*) gene, which inhibits premature metamorphosis of silk-producing pupae[78]. In addition, several TE insertions with *cis*-regulatory effects have been associated with traits selected in domestic dog and cat breeds, such as small body size[79] and coat colour pattern[80,81].

A recent report[82] identified an intronic insertion of a TE that enhances the expression of the *cortex* gene, and this increased expression was found to underlie industrial melanism in the English peppered moth (FIG. 2c). In this textbook example of adaptation, peppered moths were not domesticated per se, but were subject to strong artificial selection by coal pollution during the industrial revolution.

Together, these findings suggest that TEs frequently have pivotal roles in facilitating plant and animal domestication, which is characterized by artificial selection for specific traits often modulated through changes in gene expression[83].

TEs and the evolution of gene-regulatory networks. Whereas most previous studies focused on individual cases of a particular TE regulating a single host gene, it is theoretically possible that TEs exert a more extensive influence in regulatory evolution by providing the 'wiring' connecting large gene 'batteries'. Such regulatory networks coordinate the expression of multiple gene products that function in concert to control entire pathways and complex biological processes. The assembly of new gene-regulatory networks is thought to underlie major evolutionary innovations including the emergence of new morphological structures and cell types[84]. There is mounting evidence that modification of the *cis*-regulatory architecture underlying the transcriptional control of genes is an important force driving the evolution of new networks[85]. Transposition has long been proposed as an attractive mechanism to facilitate the concurrent mutational events that are required to deeply remodel *cis*-regulatory architecture[3,14].

This model has gained support from recent studies linking the expansion of certain TE families with the dispersal of a regulatory module that is important for the execution of a specific developmental programme. For instance, MER20 transposons seem to have deposited numerous cyclic AMP (cAMP)-inducible enhancers specifying endometrial gene expression in the eutherian mammal ancestor, which coincided with the emergence of mammalian pregnancy more than 100 million years ago[62]. Another example is the dispersal of hundreds of enhancers with forebrain-specific activity in the mouse through the ancient expansion of the MER130 family, which may have been associated with the evolution of the mammalian neocortex[63]. There is also evidence supporting a role for lineage-specific TEs in driving more recent adaptive evolution of gene-regulatory networks. For example, the mouse-specific RLTR13 family of ERVs dispersed hundreds of placenta-specific enhancers within the past 15–25 million years[86], which may reflect and possibly directly contribute to the rapid morphological diversification of the mammalian placenta[87].

Functional evidence for network rewiring by TEs. These recent findings support a model in which waves of TE activity deposited the raw material for large-scale *cis*-regulatory changes underlying major evolutionary innovations. However, current evidence is mostly limited to correlative observations on the basis of chromosomal association of TEs with nearby genes encoding functions with biologically plausible links to the phenotypes or pathways considered[62,63,86,88,89]. Although these associations are interesting, they should be interpreted with caution. An alternative, but not mutually exclusive, interpretation is that TEs inserted within or near genes that are highly transcribed in a given cell type are more likely to be accessible to regulatory proteins (for example, transcription factors) expressed in those cells. As TEs may have an intrinsic preference for integration within 'open' transcriptionally permissive chromatin[90,91] and must also replicate in the germline for new insertions to be inherited, these properties may introduce a biased association with genes that are highly expressed in the germline and early embryonic cell types[92–94]. Second, there is probably some level of selection against insertions that considerably alter host gene expression patterns upon insertion, meaning that most TEs retained in the genome might cause minimal changes in host gene expression[95–101]. This bias would also favour the retention of TEs near or within genes that share similar expression profiles, as these insertions are less likely to significantly alter existing gene expression patterns. Thus, although genomic analyses often seem to support a causal role for TEs in shaping regulatory networks, these data alone cannot falsify the null hypothesis that a given TE has no regulatory effect on a nearby gene, even when showing biochemical signatures that are consistent with *cis*-regulatory activity[55].

A few recent studies have gone a step further in testing more directly the role of TE co-option in the evolution of gene-regulatory networks by using experimental genetic manipulation. In a study investigating the interferon-inducible gene-regulatory network, CRISPR-mediated deletion of four separate primate-specific MER41 elements impaired the expression of adjacent genes with

known innate immune functions[43]. Thus the co-option of multiple MER41 elements with the same regulatory properties at several genomic loci seems to have facilitated the establishment of a coordinated transcriptional response to infection during primate evolution.

Another genetic network in which TEs have been implicated is the *cis*-regulatory circuit enabling dosage compensation on sex chromosomes. Studies of the emergence of dosage compensation on the newly evolved (<1 million years old) 'neo-X' chromosome of the fruit-fly *Drosophila miranda* showed that an ISX transposon was responsible for spreading dozens of binding sites for the dosage compensation machinery[102] (BOX 3). These ISX elements are strikingly enriched on the X chromosome and experimental insertion of an ISX element into an autosomal chromosome was sufficient to recruit the dosage compensation complex to this ectopic site. Together with data indicative of the involvement of TEs in mammalian dosage compensation[103,104], these findings suggest that transposition had a recurring role in the evolution of sex chromosomes by enabling the rapid copying-and-pasting of *cis*-regulatory sequences along entire chromosomes.

A growing number of studies also suggest a potential role for TEs in rewiring gene-regulatory networks that control early mammalian development[23]. Although the biological relevance of these findings remains to be clarified, these reports raise the exciting possibility that ERVs have been repeatedly co-opted during mammalian evolution to remodel the complex genetic circuits underlying early embryonic development (BOX 1).

Pathogenic effects of TE regulatory activities

In line with their parasitic origins and selfish behaviour, TEs have long been associated with mutant phenotypes and disease. TEs are well documented to cause disease through two primary mechanisms: insertional mutagenesis and chromosomal rearrangements. For example, *de novo* germline TE insertions disrupting normal gene function have been implicated in more than one hundred human inherited diseases[105]. Furthermore, both transposition and TE-mediated chromosomal rearrangements in somatic cells have been causally linked to several types of cancer[106]. In this section, we explore how the regulatory activities of TEs may also represent an underappreciated source of disease phenotypes.

TEs as pathogenic regulatory variants. *Cis*-regulatory variation is increasingly recognized as an important factor influencing disease susceptibility[107]. As TE insertions represent a common form of structural variation in human genomes[108], it is plausible that some of these polymorphic insertions contribute to disease risk by modulating the expression of adjacent genes. This idea has yet to be investigated systematically, but a plausible example recently came to light following the discovery of a link between the complement C4 system and schizophrenia risk[109]. The authors found that individuals carrying a polymorphic ERV intronic insertion in the C4 gene have elevated C4 expression, which in turn was found to cause synapse over-pruning, which is a phenotype that is associated with schizophrenia[109]. Although the evidence linking ERV-regulatory activity to disease remains indirect, this case is intriguing in light of previous observations of an association between schizophrenia and elevated ERV transcriptional activity[110].

Causes and consequences of TE reactivation. TEs that are insertionally fixed in the human population but encode dormant regulatory sequences may also contribute to pathogenesis. It has been widely observed that TE transcript levels are significantly increased in numerous cancers and other disease states[111,112]. The reasons why particular TEs seem to be upregulated in certain disease conditions remain poorly understood. Recent data suggest that environmental stimuli, including infection[113] and cellular stress[114], as well as natural cellular processes such as senescence[115–117], destabilize epigenetic marks that normally silence the bulk of TEs in the genome, thereby triggering their sporadic transcriptional activation.

Does the derepression of TEs have a major role in driving disease states or is it simply a side effect of pathogenesis[112,118,119]? There are several routes by which inappropriate transcriptional activation of TEs might have pathogenic consequences for the host (FIG. 3a). First, this activity may result in increased rates of propagation through transposition. Indeed, DNA hypomethylation and transcriptional reactivation of replication-competent LINE-1 copies seems to explain why some tumours, particularly epithelial ones, may have an increased rate of transposition relative to matched healthy tissues[120–122]. The role of somatic insertional mutagenesis and its contribution to tumorigenesis and malignancy is an area of considerable interest[118]. Indeed, recent studies have showed that *de novo* LINE-1 insertions can activate oncogenic pathways in hepatocellular carcinoma[123] and colorectal cancer[124]. Activation of the LINE-1 transposition machinery may have additional pathological consequences, such as the infliction of double-strand breaks in the genome through LINE-1-encoded endonuclease activity[125].

It is important to keep in mind that the vast majority of TEs that are transcriptionally activated in disease (such as ERVs) are no longer replication competent and as such are unable to generate new transposition events. However, there are other potential routes for an excess of TE-encoded transcripts to lead to pathogenesis (FIG. 3a). For example, some of these transcripts may encode peptides that have adverse cellular activities. For instance, overexpression of ERV envelope proteins, as seen in the brain of patients affected with some neurodegenerative and autoimmune disorders, can induce a wide range of cellular processes and abnormalities associated with these pathologies, such as neurodegeneration[126], autoinflammation[127], demyelination[128] and superantigen activity[129]. Furthermore, the cytoplasmic accumulation of nucleic acids derived from activated TEs, including double-stranded RNA, reverse transcribed cDNA or RNA–DNA hybrids, are increasingly regarded as potent immunological 'adjuvants' that may trigger autoimmune responses[130,131] and other toxicities when present in elevated quantities[132,133].

Structural variation
Genomic variation resulting from large-scale DNA mutations such as deletions, insertions or rearrangements.

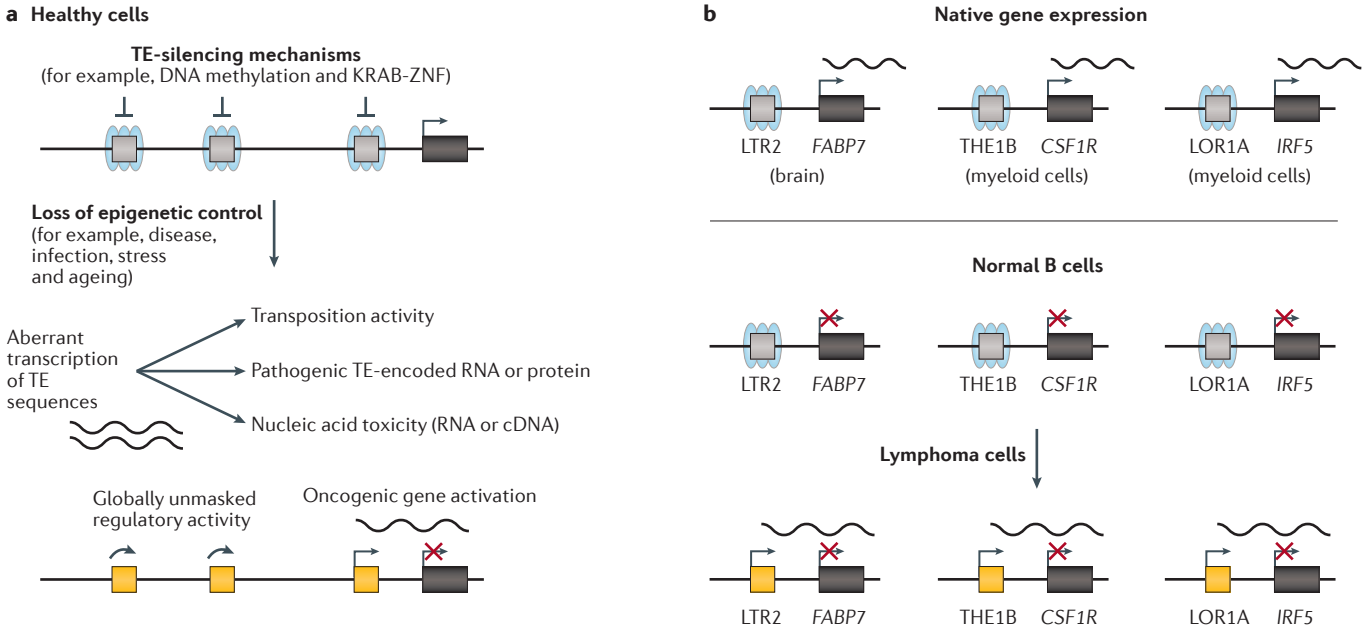


Figure 3 | **TEs can be aberrantly unmasked to promote disease states. a** | Epigenetic perturbations can result in global transposable element (TE) reactivation and pathogenic consequences. The repressive epigenetic marks that normally silence TE transcription include DNA methylation and binding by Krüppel associated box zinc finger (KRAB-ZNF) transcription factors, and these can be depleted upon various stresses. The reactivation of TE sequences across the genome can result in a wide range of pathogenic consequences, including genome instability via transposition, other pathogenic activities of TE-encoded peptides or non-coding RNAs, and cellular toxicity due to build-up of RNA or cDNA intermediate molecules. **b** | Studies examining B cell lymphomas have revealed oncogenic activation of colony-stimulating factor 1 receptor (*CSF1R;* driven by a THE1B promoter in Hodgkin lymphoma)[134], fatty acid-binding protein 7 gene (*FABP7*; driven by a LTR2 promoter in diffuse large B cell lymphoma)[135] and interferon-regulatory factor 5 (*IRF5*; driven by a LOR1A promoter in Hodgkin lymphoma)[136]. All of these examples involve long terminal repeat (LTR) or endogenous retrovirus (ERV) elements, highlighting again the proclivity of this class of TEs to retain potent but generally repressed *cis*-regulatory activity in the human genome.

Misregulation of host genes by TEs. Perhaps most importantly, and in line with the major theme of this Review, the reactivation of TEs may promote disease states indirectly by altering host gene expression. TEs that are normally silenced by DNA methylation may show *cis*- or *trans*-regulatory activity that could cause global dysregulation of host genes in *cis* or *trans* (through the various mechanisms illustrated in FIG. 1b). Thus ectopic activation of dormant TE *cis*-regulatory sequences may result in the pathogenic activation of genes or pathways in some cells. Although evidence supporting this model remains limited, there is a growing number of studies showing how derepression of a particular TE copy activates transcription of an adjacent proto-oncogene[134–136] (FIG. 3b). Whereas the loss of regulatory control at these specific elements may be a rare stochastic event occurring in a small subset of cells, it is possible that this process could be favoured by selection during tumour evolution. In this model termed 'onco-exaptation', individual cells where an oncogenic TE is aberrantly unmasked acquire a fitness advantage over other cells as a result of altered oncogene expression[137]. This process would favour clonal propagation of cells in which the TE is unmasked and which perpetuate tumour growth.

In addition to altering the expression of adjacent protein-coding genes, reactivated TEs can also drive widespread expression of non-coding RNAs, which themselves are mainly derived from TE sequences[138–140]. Many of these transcripts are likely to be non-functional, but there is evidence that some have oncogenic properties. For example, the lncRNA BRAF-activated non-protein-coding RNA (*BANCR*) is specifically expressed in melanoma cells and promotes the proliferation and migration of these cells in culture[141]. Interestingly, *BANCR* exons are mainly derived from a MER41 ERV insertion and its promoter is derived from the MER41 LTR[139]. A similar example is *EVADR*, which is a lncRNA derived from the ERV MER48 that is recurrently overexpressed in adenocarcinomas[142]. Further studies are warranted to determine the mechanisms by which these TE-derived transcripts might contribute to tumorigenesis.

The regulatory activities of TEs are also emerging as drivers of pathogenesis in non-human species. In the oil palm, for example, a fixed *Karma* TE insertion within an intron of the gene *MANTLED* is normally methylated, but sporadic demethylation of the region provides an alternative splice site and a premature termination signal for *MANTLED*[143] (FIG. 2d). This epigenetic dysregulation was found to underlie the spontaneous production of deformed oil palm fruits by genetic clones[143].

The double-edged sword of co-opting TEs. How can potentially pathogenic TEs persist in host genomes and not be purged by negative selection? In some cases it might just be a matter of time until TE insertions segregating at low frequency in the population are eventually eliminated. It is also possible that the pathogenic activity of certain TEs is only unmasked in a small number of individuals, or exert only weak or post-reproductive effects, and therefore impose an insufficient fitness cost to be purged by natural selection. Another possibility is that some TEs have an adaptive role, but occasional misregulation represents a negative, disease-causing side effect of co-option. Under such a scenario co-opted TEs might be viewed as 'double-edged swords' in host evolution: beneficial in some situations or individuals, but detrimental in others.

Our recent discovery of a MER41 element conferring interferon-inducibility to the *AIM2* gene[43] (FIG. 2b) may represent an example of a double-edged TE. *AIM2* encodes an important immunity factor that also functions as a potent tumour suppressor[144]. Constitutive transcriptional dysregulation of *AIM2* (either up or down) has been recurrently observed in cancer and autoimmune diseases, although the mechanisms underlying this misregulation are not well understood[145]. In one study examining colorectal colon cancer cells, in which *AIM2* is constitutively silent and unresponsive to interferon, the upstream promoter region (~700 bp upstream of the *AIM2* gene) was found to be consistently hypermethylated in cancer samples[146]. Intriguingly, this region coincides with the location of the MER41 element, which suggests that aberrant methylation of this TE might account for silencing of *AIM2* in the cancer cells.

Conclusions and future perspectives

The past decade was marked by tremendous progress in our understanding of how TEs shape genome evolution. Major advances were mainly driven by advances in technologies enabling genome-scale analysis. Genome-wide surveys revealed TEs as a substantial source of *cis*-regulatory elements in diverse eukaryotic species, lending credence to ideas pioneered decades ago by McClintock, Britten and Davidson, among others. It is now apparent that TEs evolved many complex mechanisms and biochemical activities that, to various extents, predispose them to transitions from parasitic elements to integral components of host gene regulation; 'from conflicts to benefits'. A pressing challenge is to obtain more direct assessments of the biological consequences of the regulatory activity of TEs, which in turn will provide a better understanding of the long-term evolutionary implications of TE dispersal. Despite decades of genetic analyses of mutant and disease phenotypes, showing how TE insertions and rearrangements can alter host gene expression patterns in many ways, there is still fairly little experimental evidence that TEs have promoted evolutionary innovations in organismal development and physiology.

Spurred by new tools for direct genetic manipulation such as the CRISPR–Cas9 system, recent studies add to the small but growing list of TEs that have been unambiguously co-opted for regulating cellular functions. A major outstanding task is to gain a better grasp of the role of TE co-option in driving the evolution of gene-regulatory networks on a broader scale. Recent work suggests that TEs can extensively remodel regulatory networks that are involved in specific processes including dosage compensation[102], immunity[43] and early embryonic development[23]. Other comparative studies of enhancer evolution in the liver[147] and the neocortex[148] among mammals found only a minor contribution for host co-option of TE activity. These observations raise the question of whether TE-mediated regulatory evolution may be inherently biased for certain biological processes or whether it is a more general mechanism for large-scale genetic innovation.

Using phylogenetic and other retrospective approaches for assessing the function of ancient TEs will also provide insights into how more recent TEs have shaped gene-regulatory networks in modern species. One of the major hurdles is to reach back to events of the distant past to study specific steps leading to TE co-option, which are obscured by millions of years of evolution (BOX 2). Therefore, in addition to these types of retrospective studies, investigations of ongoing adaptation in populations[149,150] or using protocols of real-time experimental evolution[151,152] will shed light on how TEs can facilitate regulatory evolution in response to defined selective pressures during the earliest stages of adaptation.

There is also growing evidence linking aberrant TE regulatory activity and disease states. Although only a few examples of pathogenic TE misregulation have been documented so far that contribute to cancer, TE reactivation is clearly associated with other pathological conditions including ageing[153], neurological disorders[154] and autoimmunity[132]. Our mechanistic understanding of the cause and consequence of such ectopic TE activation for disease aetiology or progression is still in its infancy. As technological advances continue to increase our ability to functionally test the involvement of non-coding regulatory processes in pathogenesis, it will open new avenues to assess the role of TE-mediated gene dysregulation in disease.

1. McClintock, B. Intranuclear systems controlling gene action and mutation. *Brookhaven Symp. Biol.* **8**, 58–74 (1956).
2. Britten, R. J. & Kohne, D. E. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* **161**, 529–540 (1968).
3. Britten, R. J. & Davidson, E. H. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q. Rev. Biol.* **46**, 111–138 (1971).
4. Orgel, L. E., Crick, F. H. & Sapienza, C. Selfish DNA. *Nature* **288**, 645–646 (1980).
5. Doolittle, W. F. & Sapienza, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**, 601–603 (1980).
6. Lynch, M. *The Origins of Genome Architecture* (Sinauer Associates Incorporated, 2007).
7. Mager, D. L. & Stoye, J. P. Mammalian endogenous retroviruses. *Microbiol. Spectr.* **3**, MDNA3-0009-2014 (2015).
8. Richardson, S. R. *et al.* The influence of LINE-1 and SINE retrotransposons on mammalian genomes. *Microbiol. Spectr.* **3**, MDNA3-0061-2014 (2015).
9. Speek, M. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol. Cell. Biol.* **21**, 1973–1985 (2001).
10. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
11. Feschotte, C., Zhang, X. & Wessler, S. R. *Mobile DNA II* (eds Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M.) 1147–1158 (ASM Press, 2002).

12. Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* **41**, 563–571 (2009).
13. Lisch, D. How important are transposons for plant evolution? *Nat. Rev. Genet.* **14**, 49–61 (2013).
14. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9**, 397–405 (2008).
15. Bourque, G. *et al.* Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.* **18**, 1752–1762 (2008).
16. Roman, A. C., Benitez, D. A., Carvajal-Gonzalez, J. M. & Fernandez-Salguero, P. M. Genome-wide B1 retrotransposon binds the transcription factors dioxin receptor and Slug and regulates gene expression *in vivo*. *Proc. Natl Acad. Sci. USA* **105**, 1632–1637 (2008).
17. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42**, 631–634 (2010).
18. Testori, A. *et al.* The role of transposable elements in shaping the combinatorial interaction of transcription factors. *BMC Genomics* **13**, 400 (2012).
19. Sundaram, V. *et al.* Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.* **24**, 1963–1976 (2014).
This is the most comprehensive assessment so far of the contribution of TEs to the genomic landscape of transcription factor binding in human and mouse cells.
20. Teng, L., He, B., Gao, P., Gao, L. & Tan, K. Discover context-specific combinatorial transcription factor interactions by integrating diverse ChIP-Seq data sets. *Nucleic Acids Res.* **42**, e24 (2014).
21. Thompson, P. J., Macfarlan, T. S. & Lorincz, M. C. Long terminal repeats: from parasitic elements to building blocks of the transcriptional regulatory repertoire. *Mol. Cell* **62**, 766–776 (2016).
22. Lisch, D. & Damon, L. Regulation of transposable elements in maize. *Curr. Opin. Plant Biol.* **15**, 511–516 (2012).
23. Gerdes, P., Richardson, S. R., Mager, D. L. & Faulkner, G. J. Transposable elements in the mammalian embryo: pioneers surviving through stealth and service. *Genome Biol.* **17**, 100 (2016).
24. Kazazian, H. H. Jr. Mobile DNA transposition in somatic cells. *BMC Biol.* **9**, 62 (2011).
25. Perrat, P. N. *et al.* Transposition-driven genomic heterogeneity in the *Drosophila* brain. *Science* **340**, 91–95 (2013).
26. Levin, H. L. & Moran, J. V. Dynamic interactions between transposable elements and their hosts. *Nat. Rev. Genet.* **12**, 615–627 (2011).
27. Liu, S. *et al. Mu* transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *PLoS Genet.* **5**, e1000733 (2009).
28. Naito, K. *et al.* Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* **461**, 1130–1134 (2009).
29. Leem, Y.-E. *et al.* Retrotransposon Tf1 is targeted to Pol II promoters by transcription activators. *Mol. Cell* **30**, 98–107 (2008).
30. Liao, G.-C., Rehm, E. J. & Rubin, G. M. Insertion site preferences of the P transposable element in *Drosophila melanogaster*. *Proc. Natl Acad. Sci.* **97**, 3347–3351 (2000).
31. Schröder, A. R. W. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521–529 (2002).
32. Friedli, M. & Trono, D. The developmental control of transposable elements and the evolution of higher species. *Annu. Rev. Cell Dev. Biol.* **31**, 429–451 (2015).
33. Wolf, G., Greenberg, D. & Macfarlan, T. S. Spotting the enemy within: targeted silencing of foreign DNA in mammalian genomes by the Krüppel-associated box zinc finger protein family. *Mob. DNA* **6**, 17 (2015).
34. Cohen, C. J., Lock, W. M. & Mager, D. L. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene* **448**, 105–114 (2009).
35. Rebollo, R., Farivar, S. & Mager, D. L. C-GATE — catalogue of genes affected by transposable elements. *Mob. DNA* **3**, 9 (2012).
36. Hirsch, C. D. & Springer, N. M. Transposable element influences on gene expression in plants. *Biochim. Biophys. Acta* http://dx.doi.org/10.1016/j.bbagrm.2016.05.010 (2016).
37. Elbarbary, R. A., Lucas, B. A. & Maquat, L. E. Retrotransposons as regulators of gene expression. *Science* **351**, aac7247 (2016).
38. Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57–63 (2012).
References 17 and 38 are seminal studies hinting at the recurrent regulatory co-option of ERVs in mammalian early development.
39. Melé, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
40. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T. R. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res.* **23**, 169–180 (2013).
41. Jacques, P.-É., Jeyakani, J. & Bourque, G. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet.* **9**, e1003504 (2013).
42. Wang, T. *et al.* Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl Acad. Sci. USA* **104**, 18613–18618 (2007).
43. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).
This is the first application of the CRISPR–Cas9 system to establish the functional importance of TE-derived *cis*-regulatory elements, here, shown to be co-opted for the human innate immune response.
44. Van Bortle, K. & Corces, V. G. The role of chromatin insulators in nuclear architecture and genome function. *Curr. Opin. Genet. Dev.* **23**, 212–218 (2013).
45. Schmidt, D. *et al.* Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335–348 (2012).
46. Crepaldi, L. *et al.* Binding of TFIIIC to sine elements controls the relocation of activity-dependent neuronal genes to transcription factories. *PLoS Genet.* **9**, e1003699 (2013).
47. Wang, J. *et al.* MIR retrotransposon sequences provide insulators to the human genome. *Proc. Natl Acad. Sci. USA* **112**, E4428–E4437 (2015).
References 45–47 show that SINEs frequently have properties of architectural elements partitioning the mammalian genome into distinct topological domains.
48. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
49. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
50. Cournac, A., Koszul, R. & Mozziconacci, J. The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. *Nucleic Acids Res.* **44**, 245–255 (2016).
51. Mateo, L., Ullastres, A. & Gonzalez, J. A transposable element insertion confers xenobiotic resistance in *Drosophila*. *PLoS Genet.* **10**, e1004560 (2014).
This is a meticulous investigation of the physiological effect of a TE insertion with adaptive regulatory effects in *D. melanogaster*.
52. Lippman, Z. *et al.* Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**, 471–476 (2004).
53. Rebollo, R. *et al.* Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. *PLoS Genet.* **7**, e1002301 (2011).
54. de Souza, F. S. J., Franchini, L. F. & Rubinstein, M. Exaptation of transposable elements into novel *cis*-regulatory elements: is the evidence always strong? *Mol. Biol. Evol.* **30**, 1239–1251 (2013).
55. del Rosario, R. C. H., Rayan, N. A. & Shyam, P. Noncoding origins of anthropoid traits and a new null model of transposon functionalization. *Genome Res.* **24**, 1469–1484 (2014).
56. Brosius, J. & Gould, S. J. On 'genomenclature': a comprehensive (and respectful) taxonomy for pseudogenes and other 'junk DNA'. *Proc. Natl Acad. Sci. USA* **89**, 10706–10710 (1992).
57. Miller, W. J., McDonald, J. F. & Pinsker, W. Molecular domestication of mobile elements. *Genetica* **100**, 261–270 (1997).
58. Lowe, C. B. & Haussler, D. 29 mammalian genomes reveal novel exaptations of mobile elements for likely regulatory functions in the human genome. *PLoS ONE* **7**, e43128 (2012).
59. Bejerano, G. *et al.* A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**, 87–90 (2006).
60. Samstein, R. M., Josefowicz, S. Z., Arvey, A., Treuting, P. M. & Rudensky, A. Y. Extrathymic generation of regulatory T cells in placental mammals mitigates maternal–fetal conflict. *Cell* **150**, 29–38 (2012).
This is a careful study tracing the origination of an enhancer that is essential for extrathymic regulatory T cell differentiation to an ancient TE insertion that was co-opted in the common ancestor of placental mammals.
61. Lam, D. D. *et al.* Partially redundant enhancers cooperatively maintain mammalian *Pomc* expression above a critical functional threshold. *PLoS Genet.* **11**, e1004935 (2015).
62. Lynch, V. J., Leclerc, R. D., May, G. & Wagner, G. P. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. **43**, 1154–1159 (2011).
63. Notwell, J. H., Chung, T., Heavner, W. & Bejerano, G. A family of transposable elements co-opted into developmental enhancers in the mouse neocortex. *Nat. Commun.* **6**, 6644 (2015).
64. Nishihara, H. *et al.* Coordinately co-opted multiple transposable elements constitute an enhancer for wnt5a expression in the mammalian secondary palate. *PLOS Genet.* **12**, e1006380 (2016).
65. Ahituv, N. *et al.* Deletion of ultraconserved elements yields viable mice. *PLoS Biol.* **5**, e234 (2007).
66. Stone, E. A., Cooper, G. M. & Sidow, A. Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **6**, 143–164 (2005).
67. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394 (2003).
68. Weirauch, M. T. & Hughes, T. R. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet.* **26**, 66–74 (2010).
69. Kapusta, A. & Feschotte, C. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends Genet.* **30**, 439–452 (2014).
70. Ting, C.-N., Rosenberg, M. P., Snow, C. M., Samuelson, L. C. & Meisler, M. H. Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. *Genes Dev.* **6**, 1457–1465 (1992).
71. Emera, D. & Wagner, G. P. Transformation of a transposon into a derived prolactin promoter with function during human pregnancy. *Proc. Natl Acad. Sci. USA* **109**, 11246–11251 (2012).
72. Gerlo, S., Davis, J. R. E., Mager, D. L. & Kooijman, R. Prolactin in man: a tale of two promoters. *Bioessays* **28**, 1051–1055 (2006).
73. Sasaki, T. *et al.* Possible involvement of SINEs in mammalian-specific brain formation. *Proc. Natl Acad. Sci. USA* **105**, 4220–4225 (2008).
74. Flemr, M. *et al.* A retrotransposon-driven dicer isoform directs endogenous small interfering RNA production in mouse oocytes. *Cell* **155**, 807–816 (2013).
This is a compelling study using TALEN editing to show the phenotypic importance of a lineage-specific TE insertion driving the emergence of an alternative protein isoform with a novel function.
75. Ferreira, L. M. R. *et al.* A distant trophoblast-specific enhancer controls HLA-G expression at the maternal–fetal interface. *Proc. Natl Acad. Sci. USA* **113**, 5364–5369 (2016).
76. Studer, A., Zhao, Q., Ross-Ibarra, J. & Doebley, J. Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nat. Genet.* **43**, 1160–1163 (2011).
77. Butelli, E. *et al.* Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell* **24**, 1242–1255 (2012).
78. Sun, W., Shen, Y.-H., Han, M.-J., Cao, Y.-F. & Zhang, Z. An adaptive transposable element insertion in the regulatory region of the *EO* gene in the domesticated silkworm, *Bombyx mori*. *Mol. Biol. Evol.* **31**, 3302–3313 (2014).
79. Klütsch, C. F. C. & de Caprona, M. D. C. The IGF1 small dog haplotype is derived from Middle Eastern grey wolves: a closer look at statistics, sampling, and the alleged Middle Eastern origin of small dogs. *BMC Biol.* **8**, 119 (2010).

80. Clark, L. A., Wahl, J. M., Rees, C. A. & Murphy, K. E. Retrotransposon insertion in *SILV* is responsible for merle patterning of the domestic dog. *Proc. Natl Acad. Sci. USA* **103**, 1376–1381 (2006).
81. David, V. A. *et al.* Endogenous retrovirus insertion in the *KIT* oncogene determines *White* and *White spotting* in domestic cats. *G3 (Bethesda)* **4**, 1881–1891 (2014).
82. Van't Hof, A. E. *et al.* The industrial melanism mutation in British peppered moths is a transposable element. *Nature* **534**, 102–105 (2016).
This is a classic example of environmental adaptation traced to a TE insertion with *cis*-regulatory effect.
83. Swinnen, G., Goossens, A. & Pauwels, L. Lessons from domestication: targeting *cis*-regulatory elements for crop improvement. *Trends Plant Sci.* **21**, 506–515 (2016).
84. Shubin, N., Tabin, C. & Carroll, S. Deep homology and the origins of evolutionary novelty. *Nature* **457**, 818–823 (2009).
85. Sorrells, T. R. & Johnson, A. D. Making sense of transcription networks. *Cell* **161**, 714–723 (2015).
86. Chuong, E. B., Rumi, M. A. K., Soares, M. J. & Baker, J. C. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat. Genet.* **45**, 325–329 (2013).
87. Chuong, E. B. Retroviruses facilitate the rapid evolution of the mammalian placenta. *Bioessays* **35**, 853–861 (2013).
88. Xie, M. *et al.* DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nat. Genet.* **45**, 836–841 (2013).
89. Lynch, V. J. *et al.* Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Rep.* **10**, 551–561 (2015).
90. Dewannieux, M., Dupressoir, A., Harper, F., Pierron, G. & Heidmann, T. Identification of autonomous IAP LTR retrotransposons mobile in mammalian cells. *Nat. Genet.* **36**, 534–539 (2004).
91. Brady, T. *et al.* Integration target site selection by a resurrected human endogenous retrovirus. *Genes Dev.* **23**, 633–642 (2009).
92. Fontanillas, P., Hartl, D. L. & Reuter, M. Genome organization and gene expression shape the transposable element distribution in the *Drosophila melanogaster* euchromatin. *PLoS Genet.* **3**, e210 (2007).
93. McVicker, G. & Green, P. Genomic signatures of germline gene expression. *Genome Res.* **20**, 1503–1511 (2010).
94. Zhang, Y., Ying, Z. & Mager, D. L. Gene properties and chromatin state influence the accumulation of transposable elements in genes. *PLoS ONE* **7**, e30158 (2012).
95. Han, J. S., Szak, S. T. & Boeke, J. D. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* **429**, 268–274 (2004).
96. Sironi, M. *et al.* Gene function and expression level influence the insertion/fixation dynamics of distinct transposon families in mammalian introns. *Genome Biol.* **7**, R120 (2006).
97. Simons, C., Pheasant, M., Makunin, I. V. & Mattick, J. S. Transposon-free regions in mammalian genomes. *Genome Res.* **16**, 164–172 (2006).
98. Kvikstad, E. M. & Makova, K. D. The (r)evolution of SINE versus LINE distributions in primate genomes: sex chromosomes are important. *Genome Res.* **20**, 600–613 (2010).
99. Zhang, Y., Romanish, M. T. & Mager, D. L. Distributions of transposable elements reveal hazardous zones in mammalian introns. *PLoS Comput. Biol.* **7**, e1002046 (2011).
100. Nellåker, C. *et al.* The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biol.* **13**, R45 (2012).
101. Campos-Sánchez, R., Cremona, M. A., Pini, A., Chiaromonte, F. & Makova, K. D. Integration and fixation preferences of human and mouse endogenous retroviruses uncovered with functional data analysis. *PLoS Comput. Biol.* **12**, e1004956 (2016).
102. Ellison, C. E. & Bachtrog, D. Dosage compensation via transposable element mediated rewiring of a regulatory network. *Science* **342**, 846–850 (2013).
103. Lyon, M. F. X-Chromosome inactivation: a repeat hypothesis. *Cytogenet. Cell Genet.* **80**, 133–137 (1998).
104. Chow, J. C. *et al.* LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. *Cell* **141**, 956–969 (2010).
105. Hancks, D. C. & Kazazian, H. H. Jr. Roles for retrotransposon insertions in human disease. *Mob. DNA* **7**, 9 (2016).
106. Belancio, V. P., Roy-Engel, A. M. & Deininger, P. L. All y'all need to know 'bout retroelements in cancer. *Semin. Cancer Biol.* **20**, 200–210 (2010).
107. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015).
108. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
109. Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
110. Leboyer, M., Tamouza, R., Charron, D., Faucard, R. & Perron, H. Human endogenous retrovirus type W (HERV-W) in schizophrenia: a new avenue of research at the gene–environment interface. *World J. Biol. Psychiatry* **14**, 80–90 (2013).
111. Rodić, N. *et al.* Long interspersed element-1 protein expression is a hallmark of many human cancers. *Am. J. Pathol.* **184**, 1280–1286 (2014).
112. Kassiotis, G. Endogenous retroviruses and the development of cancer. *J. Immunol.* **192**, 1343–1349 (2014).
113. van der Kuyl, A. C. HIV infection and HERV expression: a review. *Retrovirology* **9**, 6 (2012).
114. Mourier, T., Nielsen, L. P., Hansen, A. J. & Willerslev, E. Transposable elements in cancer as a by-product of stress-induced evolvability. *Front. Genet.* **5**, 156 (2014).
115. De Cecco, M. *et al.* Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues. *Aging* **5**, 867–883 (2013).
116. Van Meter, M. *et al.* SIRT6 represses LINE1 retrotransposons by ribosylating KAP1 but this repression fails with stress and age. *Nat. Commun.* **5**, 5011 (2014).
117. Wood, J. G. *et al.* Chromatin-modifying genetic interventions suppress age-associated transposable element activation and extend life span in *Drosophila*. *Proc. Natl Acad. Sci. USA* **113**, 11277–11282 (2016).
118. Rodić, N. & Burns, K. H. Long interspersed element-1 (LINE-1): passenger or driver in human neoplasms? *PLoS Genet.* **9**, e1003402 (2013).
119. Magiorkinis, G., Belshaw, R. & Katzourakis, A. 'There and back again': revisiting the pathophysiological roles of human endogenous retroviruses in the post-genomic era. *Philos. Trans. R. Soc. B* **368**, 20120504 (2013).
120. Lee, E. *et al.* Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967–971 (2012).
121. Rodić, N. *et al.* Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. *Nat. Med.* **21**, 1060–1064 (2015).
122. Doucet-O'Hare, T. T. *et al.* LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma. *Proc. Natl Acad. Sci. USA* **112**, E4894–E4900 (2015).
123. Shukla, R. *et al.* Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell* **153**, 101–111 (2013).
124. Scott, E. C. *et al.* A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res.* **26**, 745–755 (2016).
125. Gasior, S. L., Wakeman, T. P., Xu, B. & Deininger, P. L. The human LINE-1 retrotransposon creates DNA double-strand breaks. *J. Mol. Biol.* **357**, 1383–1393 (2006).
126. Li, W. *et al.* Human endogenous retrovirus-K contributes to motor neuron disease. *Sci. Transl Med.* **7**, 307ra153 (2015).
127. Duperray, A. *et al.* Inflammatory response of endothelial cells to a human endogenous retrovirus associated with multiple sclerosis is mediated by TLR4. *Int. Immunol.* **27**, 545–553 (2015).
128. Kremer, D. *et al.* Human endogenous retrovirus type W envelope protein inhibits oligodendroglial precursor cell differentiation. *Ann. Neurol.* **74**, 721–732 (2013).
129. Conrad, B. *et al.* A human endogenous retroviral superantigen as candidate autoimmune gene in type I diabetes. *Cell* **90**, 303–313 (1997).
130. Yu, P. The potential role of retroviruses in autoimmunity. *Immunol. Rev.* **269**, 85–99 (2015).
131. Hurst, T. P. & Magiorkinis, G. Activation of the innate immune response by endogenous retroviruses. *J. Gen. Virol.* **96**, 1207–1218 (2015).
132. Volkman, H. E. & Stetson, D. B. The enemy within: endogenous retroelements and autoimmune disease. *Nat. Immunol.* **15**, 415–422 (2014).
133. Chiappinelli, K. B. *et al.* Inhibiting DNA methylation causes an interferon response in cancer via dsRNA including endogenous retroviruses. *Cell* **164**, 1073 (2016).
134. Lamprecht, B. *et al.* Derepression of an endogenous long terminal repeat activates the *CSF1R* proto-oncogene in human lymphoma. *Nat. Med.* **16**, 571–579 (2010).
135. Lock, F. E. *et al.* Distinct isoform of FABP7 revealed by screening for retroelement-activated genes in diffuse large B-cell lymphoma. *Proc. Natl Acad. Sci. USA* **111**, E3534–E3543 (2014).
136. Babaian, A. *et al.* Onco-exaptation of an endogenous retroviral LTR drives IRF5 expression in Hodgkin lymphoma. *Oncogene* **35**, 2542–2546 (2016).
References 134–136 provide clear instances in which the derepression of LTR elements leads to the activation of proto-oncogenes in human lymphoma.
137. Babaian, A. & Mager, D. L. Endogenous retroviral promoter exaptation in human cancer. *Mob. DNA* in press.
138. Kelley, D. & Rinn, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.* **13**, R107 (2012).
139. Kapusta, A. *et al.* Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.* **9**, e1003470 (2013).
140. Hashimoto, K. *et al.* CAGE profiling of ncRNAs in hepatocellular carcinoma reveals widespread activation of retroviral LTR promoters in virus-induced tumors. *Genome Res.* **25**, 1812–1824 (2015).
141. Flockhart, R. J. *et al.* $BRAF^{V600E}$ remodels the melanocyte transcriptome and induces *BANCR* to regulate melanoma cell migration. *Genome Res.* **22**, 1006–1014 (2012).
142. Gibb, E. A. *et al.* Activation of an endogenous retrovirus-associated long non-coding RNA in human adenocarcinoma. *Genome Med.* **7**, 22 (2015).
143. Ong-Abdullah, M. *et al.* Loss of *Karma* transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature* **525**, 533–537 (2015).
This is a striking case of epigenetic derepression of a TE associated with a deleterious phenotype that has long plagued the palm oil industry.
144. Man, S. M. *et al.* Critical role for the DNA sensor AIM2 in stem cell proliferation and cancer. *Cell* **162**, 45–58 (2015).
145. Man, S. M., Karki, R. & Kanneganti, T.-D. AIM2 inflammasome in infection, cancer, and autoimmunity: role in DNA sensing, inflammation, and innate immunity. *Eur. J. Immunol.* **46**, 269–280 (2016).
146. Woerner, S. M. *et al.* The putative tumor suppressor *AIM2* is frequently affected by different genetic alterations in microsatellite unstable colon cancers. *Genes Chromosomes Cancer* **46**, 1080–1089 (2007).
147. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
148. Emera, D., Yin, J., Reilly, S. K., Gockley, J. & Noonan, J. P. Origin and evolution of developmental enhancers in the mammalian neocortex. *Proc. Natl Acad. Sci. USA* **113**, E2617–E2626 (2016).
149. Casacuberta, E. & González, J. The impact of transposable elements in environmental adaptation. *Mol. Ecol.* **22**, 1503–1517 (2013).
150. Quadrana, L. *et al.* The *Arabidopsis thaliana* mobilome and its impact at the species level. *eLife* **5**, e15716 (2016).
151. Feng, G., Leem, Y.-E. & Levin, H. L. Transposon integration enhances expression of stress response genes. *Nucleic Acids Res.* **41**, 775–789 (2013).
152. Barroso-Batista, J. *et al.* The first steps of adaptation of *Escherichia coli* to the gut are dominated by soft sweeps. *PLoS Genet.* **10**, e1004182 (2014).
153. Li, W. *et al.* Activation of transposable elements during aging and neuronal decline in *Drosophila*. *Nat. Neurosci.* **16**, 529–531 (2013).
154. Reilly, M. T., Faulkner, G. J., Dubnau, J., Ponomarev, I. & Gage, F. H. The role of transposable elements in health and diseases of the central nervous system. *J. Neurosci.* **33**, 17577–17586 (2013).
155. Goke, J. *et al.* Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell* **16**, 135–141 (2015).

156. Grow, E. J. *et al.* Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**, 221–225 (2015).
This study reports that the HERV-K family is precisely activated during early human development and expresses proteins promoting the formation of retroviral particles and potentially modulating embryonic gene expression.
157. Fort, A. *et al.* Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat. Genet.* **46**, 558–566 (2014).
158. Wang, J. *et al.* Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* **516**, 405–409 (2014).
159. Santoni, F. A., Guerra, J. & Luban, J. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology* **9**, 111 (2012).
160. Ohnuki, M. *et al.* Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proc. Natl Acad. Sci. USA* **111**, 12426–12431 (2014).
161. Loewer, S. *et al.* Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat. Genet.* **42**, 1113–1117 (2010).
162. Wang, Y. *et al.* Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. *Dev. Cell* **25**, 69–80 (2013).
163. Lu, X. *et al.* The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat. Struct. Mol. Biol.* **21**, 423–425 (2014).
164. Durruthy-Durruthy, J. *et al.* The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. *Nat. Genet.* **48**, 44–52 (2015).
165. Izsvák, Z., Wang, J., Singh, M., Mager, D. L. & Hurst, L. D. Pluripotency and the endogenous retrovirus HERVH: conflict or serendipity? *Bioessays* **38**, 109–117 (2016).
166. Casola, C., Hucks, D. & Feschotte, C. Convergent domestication of *pogo*-like transposases into centromere-binding proteins in fission yeast and mammals. *Mol. Biol. Evol.* **25**, 29–41 (2007).
167. Cornelis, G. *et al.* Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. *Proc. Natl Acad. Sci. USA* **112**, E487–E496 (2015).
168. Gong, C. & Maquat, L. E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature* **470**, 284–288 (2011).
169. Hu, S.-B. *et al.* Protein arginine methyltransferase CARM1 attenuates the paraspeckle-mediated nuclear retention of mRNAs containing IRA*lus*. *Genes Dev.* **29**, 630–645 (2015).
170. Shen, S. *et al.* Widespread establishment and regulatory impact of *Alu* exons in human genes. *Proc. Natl Acad. Sci. USA* **108**, 2837–2842 (2011).
171. Roberts, J. T., Cardin, S. E. & Borchert, G. M. Burgeoning evidence indicates that microRNAs were initially formed from transposable element sequences. *Mob. Genet. Elements* **4**, e29255 (2014).
172. Zhang, X.-O. *et al.* Complementary sequence-mediated exon circularization. *Cell* **159**, 134–147 (2014).
173. Liang, D. & Wilusz, J. E. Short intronic repeat sequences facilitate circular RNA production. *Genes Dev.* **28**, 2233–2247 (2014).
174. Johnson, R. & Guigó, R. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA* **20**, 959–976 (2014).
175. Spengler, R. M., Oakley, C. K. & Davidson, B. L. Functional microRNAs and target sites are created by lineage-specific transposition. *Hum. Mol. Genet.* **23**, 1783–1793 (2014).
176. Creasey, K. M. *et al.* miRNAs trigger widespread epigenetically activated siRNAs from transposons in *Arabidopsis*. *Nature* **508**, 411–415 (2014).
177. Kelley, D. R., Hendrickson, D. G., Tenen, D. & Rinn, J. L. Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. *Genome Biol.* **15**, 537 (2014).
178. González, J., Lenkov, K., Lipatov, M., Michael Macpherson, J. & Petrov, D. A. High rate of recent transposable element-induced adaptation in *Drosophila melanogaster*. *PLoS Biol.* **6**, e251 (2008).
179. Ellison, C. E. & Bachtrog, D. Non-allelic gene conversion enables rapid evolutionary change at multiple regulatory sites encoded by transposable elements. *eLife* **4**, e05899 (2015).
This study reveals a gene conversion-like process by which *cis*-regulatory sequences derived from related TE copies may be rapidly ameliorated to fine-tune their binding affinity for a *trans*-regulatory protein.
180. Romanish, M. T., Lock, W. M., van de Lagemaat, L. N., Dunn, C. A. & Mager, D. L. Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian mvolution. *PLoS Genet.* **3**, e10 (2007).
181. Tuan, D. & Pi, W. In human β-globin gene locus, ERV-9 LTR retrotransposon interacts with and activates β-but not γ-globin gene. *Blood* **124**, 2686 (2014).
182. Lunyak, V. V. *et al.* Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science* **317**, 248–251 (2007).
183. Emera, D. *et al.* Convergent evolution of endometrial prolactin expression in primates, mice, and elephants through the independent recruitment of transposable elements. *Mol. Biol. Evol.* **29**, 239–247 (2012).
184. Chung, H. *et al.* *Cis*-regulatory elements in the *Accord* retrotransposon result in tissue-specific expression of the *Drosophila melanogaster* insecticide resistance gene *Cyp6g1*. *Genetics* **175**, 1071–1077 (2007).
185. Guio, L., Barrón, M. G. & González, J. The transposable element *Bari-Jheh* mediates oxidative stress response in *Drosophila*. *Mol. Ecol.* **23**, 2020–2030 (2014).
186. Ding, Y. *et al.* Natural courtship song variation caused by an intronic retroelement in an ion channel gene. *Nature* **536**, 329–332 (2016).
This study uses an elegant mix of genetic mapping and genome editing to show that a polymorphic TE insertion within an intron of an ion channel gene is responsible for driving courtship song variation in *Drosophila simulans*.

Acknowledgements
The authors apologize to many colleagues who have produced primary research on the topic but who could not be cited or discussed owing to space limitations. This work was supported by funds from the US National Institutes of Health (GM77582, GM112972, GM059290 to C.F. and GM114514 to N.C.E). E.B.C. was supported by a Howard Hughes Medical Institute postdoctoral fellowship from the Jane Coffin Childs Memorial Fund. N.C.E. was supported by the Biomedical Scholars Program of the Pew Charitable Trusts.

Competing interests statement
The authors declare no competing interests.

FURTHER INFORMATION
C-GATE https://sites.google.com/site/tecatalog/

ALL LINKS ARE ACTIVE IN THE ONLINE PDF