REVIEW ARTICLE
https://doi.org/10.1038/s41556-022-00970-4

nature cell biology

Mammalian genome innovation through transposon domestication

Andrew J. Modzelewski[1,2], Johnny Gan Chong[1], Ting Wang[3] and Lin He[1]✉

Since the discovery of transposons, their sheer abundance in host genomes has puzzled many. While historically viewed as largely harmless 'parasitic' DNAs during evolution, transposons are not a mere record of ancient genome invasion. Instead, nearly every element of transposon biology has been integrated into host biology. Here we review how host genome sequences introduced by transposon activities provide raw material for genome innovation and document the distinct evolutionary path of each species.

Barbara McClintock's seminal discovery of transposable elements (TEs) was decades ahead of its time[1]. She postulated the existence of TEs, and speculated their gene regulatory activity long before experimental validation[1]. Similarly, Britten and Davidson's 'gene battery' model, a theoretical framework on how repetitive sequences contribute to coordinated gene regulation[2], was not appreciated until recently. With numerous genomes deciphered[3–5], it becomes evident that TE influence is widespread across metazoan genomes.

Approximately 40% of mammalian genomes originate from TEs[4,6], including DNA transposons (1–2%) and retrotransposons (~40%), both hijacking cellular machineries to spread in host genomes. DNA transposons use a 'cut and paste' mechanism to integrate into the host genome, while retrotransposons use a 'copy and paste' strategy for expansion[7]. In recent evolutionary history, retrotransposon domestication is more frequently observed in mammals than that of DNA transposons. Thus, our Review focuses on retrotransposons and their roles in genome architecture and innovation. Readers can refer to several reviews for DNA transposons[8,9].

Retrotransposons are categorized into two groups: long terminal repeat (LTR) and non-LTR retrotransposons (long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs)). LTR retrotransposons contain two identical LTRs, flanking an internal protein-coding region; they frequently undergo homologous recombination to generate solo-LTRs. Among the non-LTR retrotransposons, LINEs encode proteins for retrotransposition, while the non-autonomous, non-coding SINE elements exploit LINE-encoded proteins for retrotransposition[10].

Given the potential danger associated with rampant transposition, TE abundance in mammals is counterintuitive[11]. TEs and their hosts undergo a constant, on-going arms race. The ability of TEs to colonize, replicate and spread in host genomes is countered by the host's surveillance. Most mammalian TEs have been inactivated via degenerative mutations and/or transcriptional/post-transcriptional silencing. Yet, occasionally, TE–host interactions, which initially serve a selfish purpose in the TE life cycle, can be repurposed for developmental/physiological host functions (Fig. 1). TE fragments could rewire proximal host gene expression by acting as alternative enhancers, promoters, splicing donors/acceptors and polyadenylation signals (Fig. 1). TE elements that encode proteins and/or non-coding RNAs (ncRNAs) could contribute neogenes to the host for novel biological functions (Fig. 1). An intricate balance is struck in TEs between their selfish properties and domesticated functionalities. While host genomes are exposed to risks imposed by TE invasion, they gain opportunities for genome innovation that expands gene regulatory modality, enriches transcript diversity and diversifies functional reservoirs[12].

In this Review, we examine the roles of TEs in mammalian development, physiology and evolution, with a focus on in vivo functional characterization of specific TE elements, as well as the key challenges and opportunities in the field.

Transposons as a functional reservoir of gene regulatory networks

TE–host interactions that mediate TE transcription, splicing and translational regulation are preserved during evolution and wired into host gene regulatory networks. When proximal to host genes, specific TEs can serve as cell-type-specific gene regulatory sequences[13–15] (Fig. 2), often conferring species-specific gene regulation, and ultimately, species-specific biological readouts (Fig. 3).

TE-derived sequences are prevalent in/near protein-coding genes, where 18.4% mouse and 27.4% human Refseq annotations have at least one isoform harbouring a TE-derived sequence in its untranslated regions (UTRs)[16]; 37% mouse and 45% human enhancers are predicted to be TE-derived[17]. Domesticated TEs as gene regulatory elements confer several distinct mechanisms of gene regulation (Fig. 3). Species-specific TEs yield diversification of gene regulation among species or, through convergent evolution, mediate similar gene regulation in different hosts. Additionally, homologous TE loci provide similar/identical gene regulatory sequences to a cohort of host genes, achieving coordinated gene regulation (Fig. 3). These mechanisms greatly enrich host gene regulatory networks.

Successful domestication of TEs as gene regulatory sequences depends on the TEs' gene regulatory capacity, integration sites and selective evolutionary advantages. In mammals, functional characterizations of TE-dependent gene regulation were often described in germ cells[18,19,20] and pre-implantation embryos[21,22], which are characterized by potent TE induction due to extensive epigenetic reprogramming. TE-mediated gene regulation is also observed in other developmental systems that lack strong, global TE induction, including neural, haematopoietic and immune systems[23–27].

[1]Division of Cellular and Developmental Biology, MCB Department, University of California, Berkeley, CA, USA. [2]Department of Biomedical Sciences, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA, USA. [3]Department of Genetics, Edison Family Center for Genome Science and System Biology, McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA. ✉e-mail: lhe@berkeley.edu

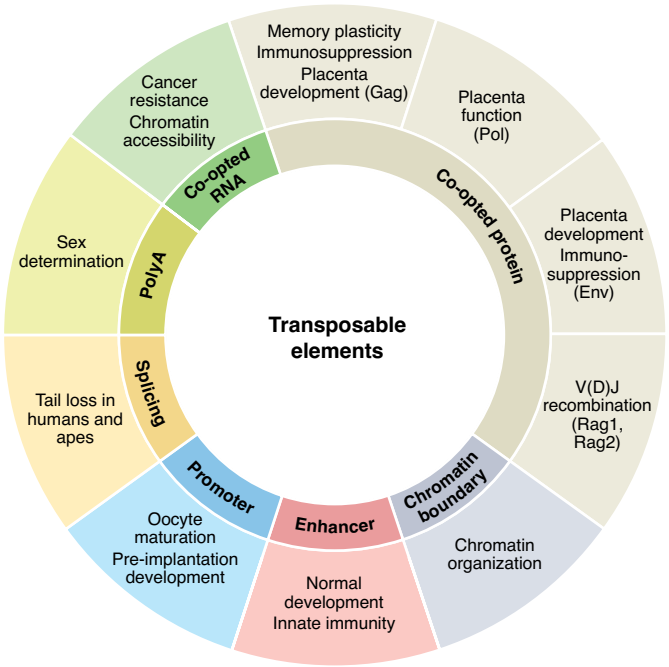

Fig. 1 | Transposon domestication contributes to host biology. Transposon domestication provides new mechanisms for host genome innovation in diverse developmental and physiological processes, generating numerous gene regulatory elements, functional ncRNAs and protein-coding genes[18,19,21,26,30,38,40,46,48–50,53,56,60–62,65,66,69,73,76,77,79,84,89,106–108]. The diagram, while probably representing the tip of an iceberg, summarizes key studies that characterize the in vivo validated transposon functions in host genomes.

Transposons as promoters. Transposon promoters have been co-opted to regulate specific host gene isoforms, expand transcript diversity and enrich gene regulatory networks. Using technologies ranging from complementary DNA library cloning[28] to RNA sequencing[29], hundreds of TE promoters have been identified, generating numerous alternative gene isoforms with distinct expression dynamics and/or altered open reading frames (ORFs). The in vivo importance of a TE promoter was first revealed by a mouse-specific MTC promoter (an LTR retrotransposon), which drives an oocyte-specific, N-terminally truncated DICER isoform, DICER^{O}. DICER^{O} exhibits greater enzymatic activity than the canonical DICER, leading to highly efficient RNA interference (RNAi) during oocyte maturation, thus placing RNAi as the central mechanism for post-transcriptional gene/TE silencing in mouse oocytes[19]. Deletion of this MTC element abolishes *Dicer*^{O} expression, causing meiotic spindle defects in oocytes and, ultimately, female infertility[19,30]. In comparison, many mammals lack *Dicer*^{O}, and use the piRNA pathway instead for TE silencing[31]. These findings reveal the importance of TEs for evolutionary plasticity of species-specific biological processes that are not essential for host viability.

TE promoters can also be repurposed for essential mammalian developmental functions[21]. In pre-implantation embryos, a mouse-specific *MT2B2* (also known as *Rr130*) promoter (an LTR retrotransposon) drives transient, yet potent induction of *Cdk2ap1*$^{\Delta N}$, an N-terminally truncated *Cdk2ap1* isoform[21]. Unlike canonical *Cdk2ap1*, which suppresses cell proliferation, *Cdk2ap1*$^{\Delta N}$ promotes proliferation. Deletion of *MT2B2* abolishes *Cdk2ap1*$^{\Delta N}$ in pre-implantation embryos, causing reduced cell proliferation, embryonic lethality and impaired implantation[21]. The essential role of the *MT2B2* promoter is surprising, as pre-implantation development was presumably normal before its integration into the ancestral mouse genome. The persistence of *MT2B2* in mouse implicates a selective advantage through increased pre-implantation cell proliferation, as induced by *Cdk2ap1*$^{\Delta N}$. Additional changes may have arisen in the mouse genome to adapt to the *MT2B2* integration, ultimately rendering it essential.

Domestication of TE promoters can yield either species-specific or evolutionarily conserved gene regulation (Fig. 3). In the case of *Cdk2ap1* (ref. [21]), nearly all mammals contain a *Cdk2ap1*$^{\Delta N}$ isoform with an evolutionarily conserved ORF[21]. The mouse-specific *MT2B2* promoter drives strong pre-implantation induction of *Cdk2ap1*$^{\Delta N}$ (ref. [21]). In pig and cow, a transposon-independent promoter regulates *Cdk2ap1*$^{\Delta N}$ expression, but yields minimal pre-implantation expression[21]. In primates, another transposon-derived promoter, L2a/Charlie4z, generates a modest pre-implantation expression of *Cdk2ap1*$^{\Delta N}$ (ref. [21]). The L2a/Charlie4z element is upstream of *Cdk2ap1* in many placental mammals, yet is not active in mouse, cow or pig to regulate *Cdk2ap1*$^{\Delta N}$. We speculate that an ancient L2a/Charlie4Z integration yields the *Cdk2ap1*$^{\Delta N}$ isoform in ancestral genomes, and that additional transposon integration and/or L2a/Charlie4z degeneration could reprogram *Cdk2ap1*$^{\Delta N}$ expression in a species-specific manner.

TE-dependent regulation of prolactin (*PRL*) expression in the endometrium tells a different evolutionary story, in which species-specific transposon promoters mediate a conserved gene expression pattern through convergent evolution[32,33]. The acquisition of endometrium *PRL* expression in evolution occurred independently in multiple species by domestication of different transposon promoters, including MER77 in mice, L1-2a in elephants and MER39 in primates (including humans) (Fig. 3), all of which generate a highly conserved expression pattern[32,33].

Transposon-derived promoters provide a powerful mechanism for coordinated gene regulation. As a family of transposons quickly spreads through the host genome, transcription factor-binding sites embedded within TE promoters are rapidly propagated. Given their sequence similarities, related transposon promoters are often coordinately regulated, achieving co-induction of dozens, if not hundreds, of host gene isoforms[21,26,28,29,34] (Fig. 3). The capacity of transposon promoters to generate new transcriptional regulation, to create new host gene isoforms and to rewire gene regulatory networks enables genome innovation, particularly in cell types that are susceptible to transposon induction, such as germ cells, pre-implantation embryos and placenta.

Human specific, transposon-dependent gene regulation probably underlies human-specific biology. For example, humans and great apes maintain fertility for decades, with male fertility being more prolonged than female fertility. Unique to humans and great apes, an ERV9 LTR element was integrated upstream of the *p63* (also known as *TP63*) gene ~10–15 million years ago, acting as a testis-specific promoter to drive a p63 isoform with an altered N-terminus[35]. This ERV9:p63 isoform induces a p53-like pro-apoptotic response to eliminate male germ cells with excessive DNA damage, preserving male fertility in humans and great apes[35]. Germ cell-specific ERV9 expression is desirable for its spread as a selfish element in the host, yet unexpectedly, a specific ERV9 was repurposed as a guardian of germ cell genome integrity, providing an evolutionary advantage[36,37].

Transposons as enhancers, repressors, insulators and chromatin boundaries. TEs have contributed extensively to enhancers, repressors and insulators, as previously reviewed[26,38,39]. On average, ~20% of cell- or tissue-specific, active chromatin elements in human, mouse and zebrafish are within TEs[40–42]. The regulatory potential of TEs as enhancers is largely tied to their sequences for transcription factor (TF) binding and host chromatin factor recognition. Many TFs, including p53 (refs. [43,44]), OCT4 (ref. [45]), CTCF[46] and STAT1 (ref. [26]), have a large repertoire of TE-derived binding sites[47]. Interestingly,

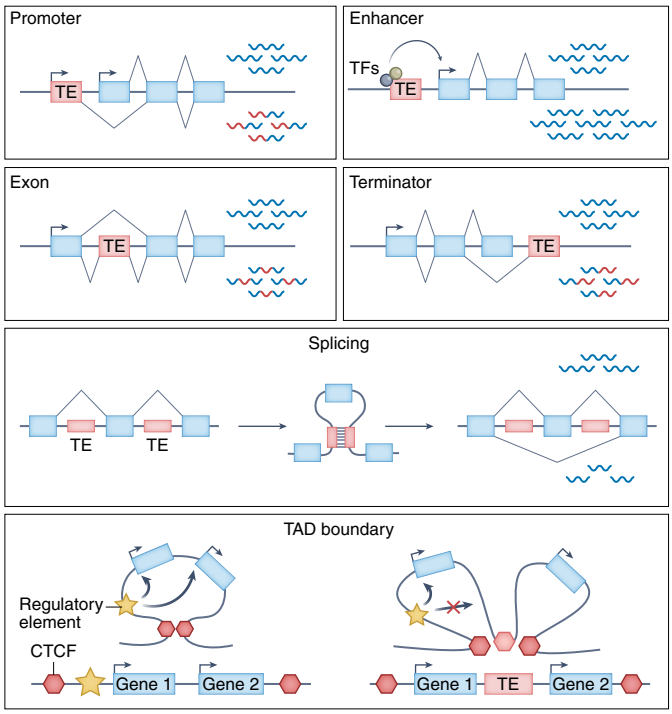

Fig. 2 | Transposon-derived gene regulatory elements diversify host gene isoforms and enrich expression regulation modality. Transposon-derived sequences contribute to gene enhancers, promoters, exons, terminators, splicing donors/acceptors and chromatin boundaries, regulating the structure and expression of proximal host gene isoforms. TE domestication expands gene regulatory modality, enriches transcript diversity and diversifies functional reservoirs in host genomes. Pink rectangles, TE elements; blue rectangles, protein-coding exons or protein-coding genes; red hexagons, CTCF; yellow star, a gene regulatory element.

some TEs bear a whole array of transcription factor-binding sites, collectively functioning as regulatory modules[48]. Hence, TEs are a source of genetic material for enhancer evolution.

In addition to generating new gene regulatory events, TEs provide redundancy and robustness to existing regulatory networks. CTCF is a chromatin factor with many binding sites derived from species-specific TEs[46]. Roughly 20% species-specific chromatin loop anchors and topologically associated domain (TAD) boundaries are CTCF sites encoded by species-specific TEs[49] (Fig. 2). Notably, around 10% of loop anchors and TAD boundaries that are functionally conserved between human and mouse are derived from species-specific TEs containing CTCF binding sites[49]. This paradox can be explained by TE-mediated CTCF binding site turnover, in which existing CTCF binding sites can be functionally replaced by a new, redundant CTCF site introduced by a proximal species-specific TE insertion that lacks sequence conservation but functionally conserves chromatin organization. Indeed, functional conservation in absence of sequence conservation seems to be the rule rather than the exception in the evolution of gene regulatory networks. It remains to be determined if TE-derived chromatin boundaries contribute primarily to genome innovation or robustness of gene regulation.

Transposons as alternative splicing signals. Transposon-dependent alternative splicing is another widespread phenomenon that contributes to evolutionary innovation on gene structure and function (Fig. 2). In some cases, transposons harbour splicing donors and/or acceptors that mobilize host splicing machineries to generate alternative gene isoforms, enabling the incorporation of transposons as gene exons, contributing to alternative coding sequences and/or UTRs. In other cases, transposon integration into host genes adds unique features of pre-mRNA structure, which alters the canonical splicing pattern to generate gene isoforms with new biological functions.

Among the best examples is an AluY element that integrated into intron 6 of the *TBXT* gene in the hominoid ancestor genome about 25 million years ago[50] (Fig. 1). Adjacent to the AluY element is a more ancient AluSx1 element integrated in the reverse orientation, resulting in a hairpin structure within the *TBXT* pre-mRNA that traps exon 6 and prevents its incorporation in the mRNA. This generates a hominoid-specific, alternative splicing isoform, $TBXT^{\Delta exon6}$, whose emergence during primate evolution coincides with tail loss in hominoid. $TBXT^{\Delta exon6}$ expression in mice results in impaired tail development or complete tail loss, supporting that AluY integration is an evolutionary event that caused/contributed to tail loss in hominids[50]. Tail loss probably confers a selective advantage, possibly by enhancing locomotion and adopting a non-arboreal lifestyle in primates. Hence, a seemingly random event in the transposon–host interaction may have shaped a major event in hominoid evolution.

Transposons as alternative polyadenylation sites. As most transposons mobilize host Pol II machinery for selfish transcription, transposon-derived polyadenylation signals can generate isoforms with altered 3′ UTRs[51]. Since 3′ UTRs regulate mRNA stability, translation, localization, trafficking and protein localization[51], an altered 3′ UTR enables distinct post-transcriptional gene regulation. Alternative polyadenylation can also be coupled with alternative splicing to generate protein isoforms with unique C-termini, and ultimately, a different protein function (Fig. 2).

Mouse sex determination is among the best examples illustrating how TE-dependent alternative polyadenylation yields functional diversity[18]. SRY, a DNA-binding protein, is an essential factor initiating male sex determination in mammals. *Sry* has been considered a single-exon gene for 30 years, until an alternative gene isoform, *Sry-T*, was identified in mice[18]. *Sry-T* is generated by alternative splicing coupled with alternative polyadenylation, in which *Sry* exon 1 splices into a transposon-derived, second exon, consisting of an L3 element and three tandem LTRs. The *Sry-T* transcription terminates at a polyadenylation signal derived from one of the LTR elements. The C-terminal 18-amino-acid sequence of the canonical SRY isoform encodes a degradation motif, which is replaced by a 15-amino-acid, degron-free sequence in the SRY-T isoform. This mechanism renders SRY-T a more stable isoform, reinforcing male specification in mice[18]. Deletion of this transposon-derived *Sry-T* exon 2 in XY mice abolishes *Sry-T* expression, causing male-to-female sex reversal. Hence, the acquisition of an alternative, TE-derived polyadenylation signal for *Sry* confers a mouse-specific functionality in sex determination.

Transposons as a functional reservoir of ncRNAs and proteins

In addition to integrating into the host gene regulatory network, domesticated TEs also generate ncRNAs and/or proteins for the host functional repertoires (Figs. 1 and 4). As such, TEs are often mutated/truncated, retaining minimal sequences for encoding ncRNAs and/or proteins. Through domestication, aspects of their ancestral functions that support TE–host interactions evolved to regulate unique host cellular processes.

Domesticated transposon-encoded functional ncRNAs. In mammals, many ncRNAs contain TE sequences[52]. Pre-implantation-specific LINE1 expression has an important role in chromatin organization during mouse zygotic genome activation. Prolonged transcriptional activation of LINE1 or premature transcriptional silencing of LINE1 in mouse zygotes results in developmental

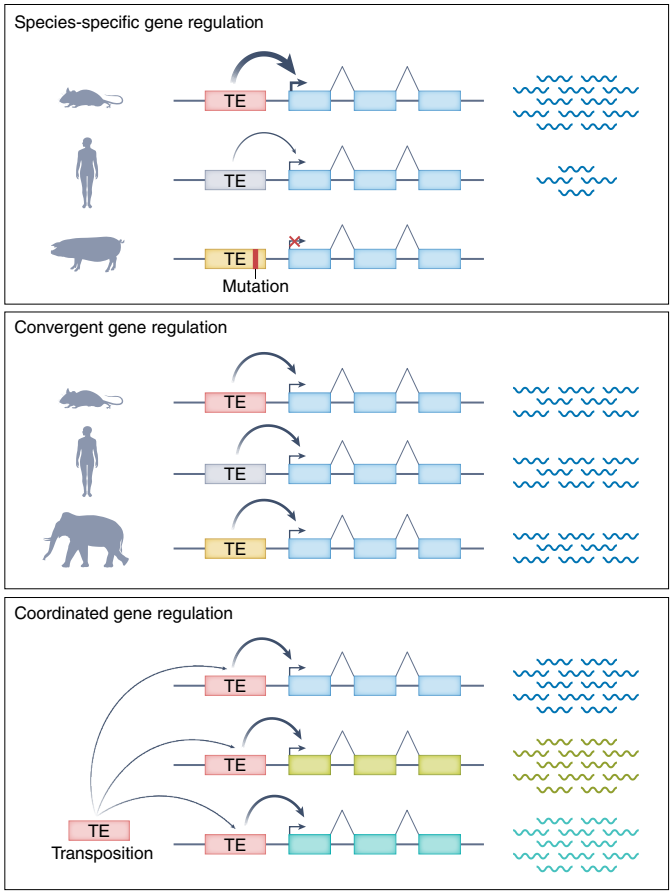

Fig. 3 | Transposons confer unique modes of *cis*-gene regulation in host genomes. Compared with other gene regulatory sequences, transposons have distinct properties, as they are repetitive in nature and frequently species-specific. Top: species-specific gene regulation occurs when distinct TEs integrate proximal to homologous genes across species, generating a unique expression pattern in each species. Middle: convergent gene regulation occurs when distinct TE insertions across species converge on the same regulatory principal to yield similar or identical expression patterns. Bottom: coordinated gene regulation occurs when related transposon elements from the same TE family spread in a given host genome and land proximal to a cohort of host genes to coordinate their expression.

arrest. Surprisingly, this effect is not attributed to LINE1-encoded proteins. Instead, LINE1 ncRNAs regulate the dynamic global chromatin accessibility in early mouse embryos[53]. Hence, TE expression is highly regulated in pre-implantation development, rather than a consequence of extensive epigenetic reprogramming. Similarly, the human HERV-H retrotransposons lose protein-coding capacity, but exhibit strong RNA expression from >100 loci in human embryonic stem cells, where HERV-H long ncRNAs establish or maintain pluripotency[54]. Mechanistically, HERV-H long ncRNAs act as a nuclear scaffold for TFs, transcriptional machineries and chromatin modifiers, promoting the expression of proximal host genes to sustain pluripotency[54]. Likewise, LINE1 ncRNAs act as a nuclear scaffold to recruit nucleolin and KAP1 to silence the *Dux*/MERVL two-cell transcriptional programme and maintain the pluripotency gene network in mouse embryonic stem cells[55].

TE-derived ncRNAs have been associated with cancer resistance by promoting innate immune surveillance (Fig. 1). In blind mole rats, pre-malignant cells experience a global loss of DNA methylation, which triggers retrotransposon induction, generates cytoplasmic RNA/DNA hybrids and activates the cGAS–STING pathway to induce cell death[56]. Similarly, treating human cancer cells with the DNMT inhibitor 5-azacytidine yields retrotransposon induction, which generates cytoplasmic double-stranded RNAs and triggers the RNA-sensing pathway to promote type I interferon response[57]. In both cases, the pathogenic properties of TE ncRNAs serve as a sensor for disease state, triggering an innate immune response to eliminate cells with inappropriate TE induction. It is unclear whether such benefit is co-opted by the host, or a side effect of harbouring transposons by the host.

Domestication of transposon-encoded proteins. Both DNA transposon- and retrotransposon-encoded proteins are co-opted in mammalian genomes, yet annotated retrotransposon proteins are greater in number owing to their recent domestication. Ancestral LTR retrotransposons and LINEs express proteins to mediate retrotransposition, most of which undergo deleterious mutations and/or epigenetic silencing. Nevertheless, a subset of LTR retrotransposons, particularly endogenous retroviruses (ERVs), retain protein-coding capacity. Among 19 mammalian species examined, 0.05–0.15% of ERVs retain protein-coding capacity of retroviral origin[58]. Since the origin of anciently domesticated transposon-derived proteins may not be easily recognizable, both DNA transposon- and retrotransposon-derived protein-coding genes could be underestimated in numbers.

Some retrotransposons retain the protein-coding capacity of Gag, Pol and Env proteins of their retroviral origin. The domestication of retrotransposon-encoded proteins possibly enriches host cellular functions and empowers the host to resist invasion by similar TEs[8]. A recent genome analysis in 700 vertebrate genomes uncovered 177 independent co-option events for retroviral protein-coding genes, with the majority being Gag and Env[58]. Many of these events are retained for a short evolutionary timeframe. Similar functionality of ERV proteins can be repeatedly adopted by different mammals from different ancient retroviruses[58] (Fig. 3). Intriguingly, some protein-coding retrotransposons evolve into essential genes, supporting that their invasion provides novel ORFs to fulfil new host functions with a selective advantage (Fig. 4).

ERV-encoded Gag proteins. Retrotransposon-encoded Gag was once essential for retroviral packaging and budding. Gag contains three key domains: the N-terminal matrix (MA) domain for plasma membrane binding and virion assembly, the central capsid (CA) domain for viral capsid core formation, and the nuclear capsid (NC) domain for viral RNA packaging. Analysis of all annotated human protein-coding genes reveals dozens of Gag-like genes[59]. In addition to annotated individual cellular genes with a retrotransposon origin, mammalian genomes also harbour ERV loci with partial or complete protein-coding capacity[58]. Limited functional studies so far suggest that the molecular functions of domesticated Gag-like proteins all have their roots in those of viral Gag in the retrovirus life cycle[59].

Arc. Activity-regulated cytoskeleton-associated protein (Arc) is a key regulator of synaptic plasticity, long-term learning and memory consolidation. Arc originates from the Ty3/gypsy family *gag* gene[60–62]. Arc self-assembles into a virus-like capsid that encapsulates its own mRNA in extracellular vesicles that are released from active synapses. This mechanism transfers *Arc* mRNAs into the dendrites of neighbouring neurons for localized translation[61,62] (Fig. 4). Despite analogous functions, cellular Arc and bona fide retroviral Gag exhibit mechanistic differences. Arc originates from truncated retrotransposons lacking Env, and therefore relies on a different mechanism of uptake[63]. Unlike retroviral Gag, which binds specifically to its own retroviral RNAs, Arc binds its own mRNA, and associates with other cellular mRNAs with a lower affinity[61]. Arc also evolves a synaptic function that is atypical of a retroviral

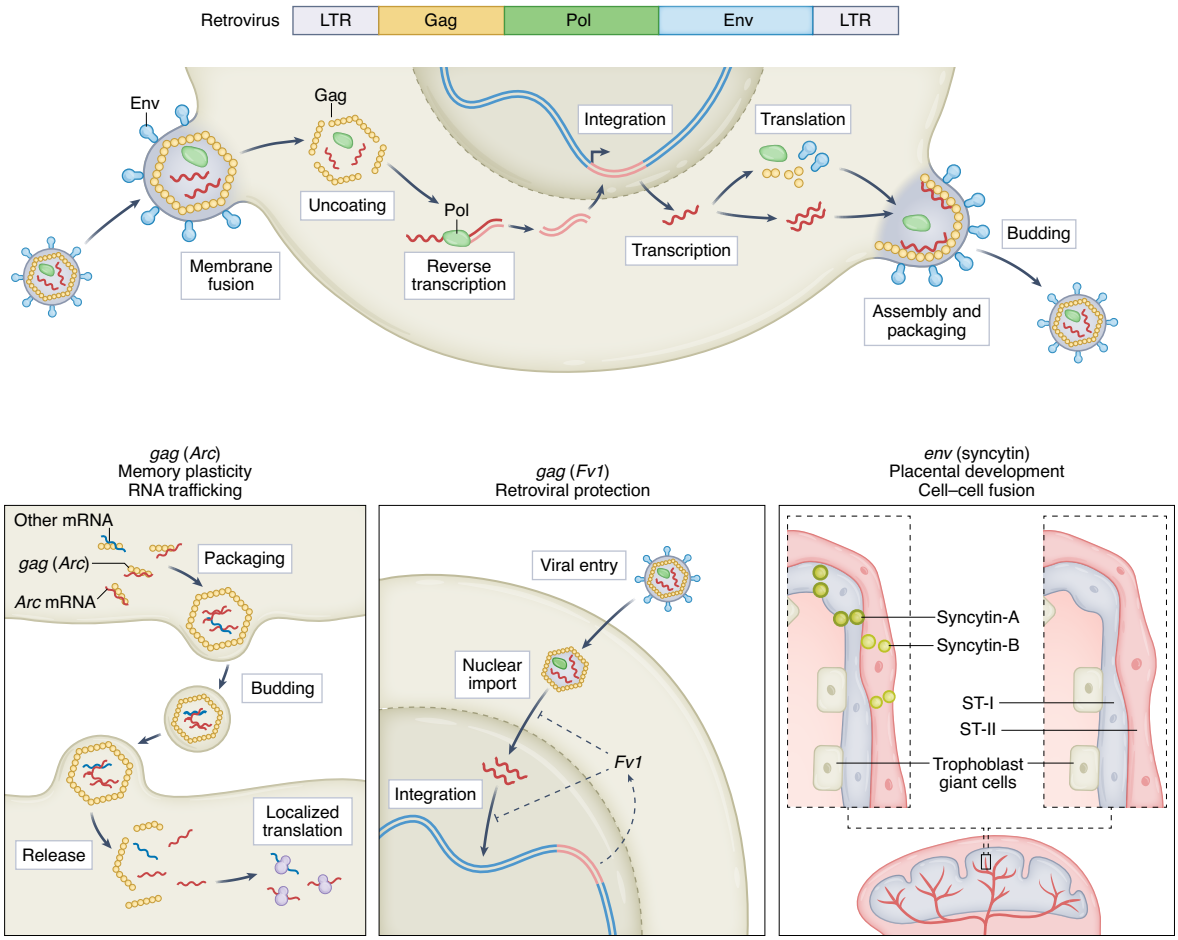

Fig. 4 | Co-option of transposon-encoded proteins contributes to new host biology. A diagram illustrates the functional parallel between the retroviral Gag, Pol and Env proteins and their domesticated counterparts encoded by retrotransposons. Top: retroviral life cycle begins when retroviruses infect the host cells and integrate into the host genome. Subsequently, the host machineries drives the expression of viral Gag, Pol and Env, allowing the retrovirus to mature before released from the host cells. Bottom: here we show examples of domesticated *gag* and *env* genes, which are repurposed for neuronal functions, host defence and placenta development. The remarkable modern innovations conferred by retrotransposon-encoded proteins can be traced back to their proviral functions.

Gag function. The key Arc function in regulating AMPA receptor trafficking and membrane density probably stems from unique interactions between its ancestral Ty3/gypsy retrotransposon and host cellular proteins[64].

During evolution, Gag proteins from two lineages of Ty3/gypsy retrotransposons were independently domesticated, leading to convergent evolution of *Arc* genes in both tetrapods and Diptera phyla of the animal kingdom[60]. While vertebrate Arc proteins in mouse, human and rat contain only a predicted MA domain and a CA domain to mediate intercellular mRNA transfer between neurons[62], insect Arc in *Drosophila* contains MA, CA and NC motifs and mediates mRNA transfer among neuromuscular junctions[61]. In both cases, the 3′ UTR of *Arc* mRNA is necessary and sufficient for binding to the Arc protein. The evolutionary origin of Arc provides important insights into mechanisms governing synaptic function and highlights the potential of ancient Gag-derived cellular genes for mRNA trafficking.

Peg10. Paternally expressed 10 (*Peg10*), derived from a Ty3/gypsy LTR retrotransposon, is an evolutionarily conserved, imprinted gene in all eutherian mammals[65]. *Peg10* is paternally expressed in placenta, in which deletion in mice caused lethality at embryonic day 9.5, largely due to impaired placental development[65]. Interestingly, *Peg10* retains the retroviral-derived overlapping ORFs, generating two ORFs from the same transcript[66]. Peg10-ORF1 encodes a Gag-like protein containing the CA and NC domains, while Peg10-ORF1/2 encodes a fusion of Gag and Pol generated by a programmed −1 frameshift during the translation of Peg10-ORF1 (ref. [66]). This mechanism resembles the translation of Gag–Pol in retroviruses, supporting a bona fide retrotransposon origin for *Peg10*.

Similar to Arc, Peg10 encapsulates its own mRNA to form capsid-like particles that are secreted in budding vesicles[67]. The ability of human PEG10 to encapsulate and transport its own mRNA has been exploited to generate a modular platform for mRNA delivery by fusing the *PEG10* 3′ UTR motif to cargo mRNA[68]. Pseudotyped PEG10 virus-like particles encapsulate such chimeric RNAs in extracellular vesicles to mediate efficient intercellular transfer, thus using endogenous proteins to minimize immunogenicity in nucleic acid therapy[68], and providing an innovative method to complement existing viral delivery systems.

ERV-encoded Env proteins. Retroviral Env proteins bind to cell surface receptors to mediate fusion between host and viral membranes, thus determining tissue tropism for infection. Syncytin proteins, derived from *env* genes of multiple ancestral ERVs, provide a similar function in placenta by promoting cell–cell fusion of

mononucleated cytotrophoblasts to establish the multi-nucleated syncytiotrophoblast layer[69]. Syncytiotrophoblast layers are formed during implantation, and maintained throughout gestation to mediate exchange of nutrient, gas and waste between maternal and fetal blood, and shield the fetus from the maternal immune response. Syncytin-mediated fusion of cytotrophoblasts is essential for syncitiotrophoblast maturation and placenta development in mammals (Fig. 4).

During mammalian evolution, syncytin genes emerged through at least nine independent domestication events from distinct, species-specific ERVs[70]. While different mammalian syncytin genes are not conserved in protein sequences owing to their distinct retroviral origins, they all exhibit placenta-specific expression, retain fusogenic activity and persist in evolution for extended periods (>10 million years)[70]. Although many placental mammals, including humans, have domesticated an Env protein as syncytin to mediate cell–cell fusion[69], functional studies have been performed only in mice[71]. Most mammals have one syncytin gene and one syncitiotrophoblast layer, yet mice have two syncytin genes (*Syna* and *Synb*) and two syncytiotrophoblast layers (ST-I and ST-II), adding to functional redundancy, complexity and robustness (Fig. 4) (ref. [69]). *Syna* and *Synb* entered the rodent genome approximately 20 million years ago, regulating the formation of ST-I and ST-II, respectively[71,72]. Deletion of *Syna* disrupts cell fusion of the ST-I layer in placenta, causing aberrant cell expansion, apoptosis and impaired fetal vascularization and, ultimately, embryonic lethality[71]. By contrast, *Synb* null placenta displays impaired cell fusion of the ST-II layer, yet the embryos are viable with only limited late-onset growth defects.

Synb also exhibits immune suppressive activity, an innate property of retroviral Env proteins, probably conferring maternal–fetal tolerance[73]. It is tempting to speculate that the consecutive retroviral gene capture by the rodent genome provides a biological innovation that generates a multi-layered placental structure with functional redundancy. The domestication of Syncytin in the ancestral mammals could be a pivotal event for the emergence of placental mammals. The replacement of the ancestral syncytin gene with a new *env* gene in each species probably contributes to a species-specific mechanism for placentation.

ERV-encoded Pol and Gag–Pol proteins. Retroviral Pol protein contains several important domains, including the protease that self-cleaves the polyprotein, the reverse transcriptase domain that converts the RNA genome into cDNA, and the integrase domain that integrates the retrotransposon genome into the host genome. Pol domestication occurs at a much lower frequency compared with that of Gag and Env, possibly due to the difficulty in taming reverse transcriptase activity that renders detrimental effects.

Bioinformatic analyses have identified two evolutionarily conserved *pol* genes, the *GIN1* gene harbouring an integrase domain, and *CGIN1* containing an RNase H and an integrase domain[74,75]. *GIN1* and *CGIN1* are evolutionarily conserved in mammals, implicating a potential host function. Another example of Pol domestication is *Peg10*, which encodes both Gag and Gag–Pol[66]. While *Peg10* deletion leads to mid-gestation lethality in mice[65], mutation of its Pol protease motif causes perinatal lethality, with fetal and placental growth defects due to impaired fetal vasculature[76]. *Peg10* is expressed in the three trophoblast layers, but not the surrounding fetal capillary epithelial cells[76]. Interestingly, *Peg11*, presumably derived from the same retrotransposon family as *Peg10*, is specifically expressed in fetal endothelial cells, but not trophoblasts. *Peg11* contains Gag and Pol regions, and its deficiency in mice leads to impaired fetal capillaries in placenta during mid to late gestation, resembling the phenotype caused by *Peg10*-protease-motif mutant mice[76,77]. While the exact molecular basis remains elusive, the protease activity of *Peg10* in trophoblasts and the Pol-like activity of *Peg11* in fetal endothelial cells act at the fetal–maternal interface to safeguard the development of fetal vasculature.

Non-LTR retrotransposon-encoded proteins. Non-LTR retrotransposons, such as LINEs, have protein-coding capacity, yet their ORFs are domesticated less frequently in mammals. L1TD1 is perhaps the best-known example in human. L1TD1 originates from a co-opted LINE1 element that was initially integrated into the common ancestor of eutherian mammals, but subsequently lost or pseudogenized multiple times in some species during mammalian evolution[78]. It has been speculated that L1TD1 confers genome defence against LINE1 and may have later evolved other functions such as pluripotency maintenance[78].

DNA transposon-encoded proteins. DNA transposons are less abundant and active in modern mammalian genomes compared with their retrotransposon counterparts, yet their domestications have also shaped important developmental/physiological processes in evolution. In jawed vertebrates, RAG1 and RAG2, the key enzymes for V(D)J recombination essential for humoral immunity, are derived from transposase genes of ancient, eukaryotic *Transib* DNA transposons[79]. *Thap1*, *Thap9* and *Thap11* represent a family of zinc-finger TFs with a DNA-binding domain homologous to *Drosophila* P-element transposase. Mutations in *Thap1* causes DYT6 dystonia in mouse and human[80]; *Thap11* deletion causes peri-implantation lethality and defects in the inner cell mass in mice[81]; human THAP9 exhibits an active P-element transposase activity, yet its function is unknown[82].

Intriguingly, several transposon–host fusion genes are evolved because of exon shuffling, which contain a transposase DNA-binding domain and a host-derived KRAB domain[83]. These KRAB–transposase fusions functionally combine DNA-binding specificity with transcriptional repression to repress expression of specific genes[83]. Thus, transposase capture is a recurrent mechanism for gene evolution, providing not only DNA-binding specificity, but also splicing sites for novel fusions.

Transposon-encoded proteins as an evolutionary adaption for host defence

A reoccurring theme in TE domestication is their adaptation to provide host defence against similar pathogens. As divergent as prokaryotes and vertebrates, their key enzymes for genome defence could all be traced back to ancient DNA transposons that had once invaded the host genome[84]. In addition to RAG1/RAG2 where ancient transposases are repurposed for humoral immunity[79], multiple CRISPR–Cas components are probably co-opted from DNA transposons[8]. Cas1, a key component of the class I CRISPR–Cas system, is derived from the transposase of a *Casposon* DNA transposon[8]. Cas9, the key component of the class II CRISPR–Cas system is derived from IscB, an RNA-guided DNA nuclease encoded by the IS200/IS605 family of DNA transposons[85,86]. This family of transposons also encode TnpB, an endonuclease distantly related to IscB and a possible ancestral protein for Cas12 (ref. [86]). Thus, RNA-guided DNA nucleases encoded by transposons are probably ancestors for key enzymatic components of the CRISPR–Cas system.

Retrotransposons have also been co-opted for host defence against pathogens. Env proteins can act as restriction factors against infection from related retroviruses. The Env of a retrotransposon could block the activity of a related Env receptor in infected host cells, a process termed receptor interference[87]. Another interesting example is the HERV-T Env protein, which directly binds the cell surface receptor, monocarboxylate transporter-1 (MCT-1), to block its activity, hence protecting the cells from additional infection by HERV-T. Domesticated HERV-T Env probably contributed to the extinction of HERV-T that circulated in primate genomes for ~25 million years before going extinct ~8 million years ago[88].

Env-mediated host defence also occurs in human pre-implantation embryos. HERVK is transiently induced at zygotic genome activation, followed by the translation of its ORFs and assembly of virus-like particles[89]. HERVK encodes multiple ORFs, including Rec, a homologue of HIV Rev. Rec expression leads to induction of the interferon-induced viral restriction factor IFITM1, thereby trigging an innate antiviral response to protect embryos from repeated infection[89]. Similarly, the *env* gene suppressyn (also known as *ERVH48-1*), which originates from an HERV-fb insertion, acts as a potential restriction factor against retroviruses in pre-implantation embryos of humans and other hominoids[90].

Gag can also act as a restriction factor. The mouse *Fv1* gene probably originates from an ancient MuERV-L *gag* gene given their sequence similarity[91]. Fv1 protects the host from a variety of retroviruses, particularly murine leukaemia virus (MLV)[91]. The exact antiviral mechanism of Fv1 is unclear, yet Fv1 is shown to target capsid proteins of exogenous MLV, blocking MLV infection after viral entry but before viral integration and provirus formation[91,92]. It is intriguing that an MLV-unrelated Gag protein restricts MLV infection, implicating an unexpected interaction between these two retrotransposon *gag* genes[92].

Challenges and opportunities for transposon research

Limited read length of genomic sequencing data, underdeveloped computational tools and suboptimal TE annotations all contribute to analytical challenges associated with the repetitive nature of TEs. Many adopt a strategy that relies on uniquely mapped TE reads, thus underestimating TE abundance by ignoring numerous multiply mapped reads[93–95]. TE functional characterization is also complicated by its repetitiveness. While CRISPR-, TALEN- or RNAi-based technologies could target some TE families if the number of loci is optimal, it is difficult to attribute phenotypes to a specific locus. Conversely, genetic disruption of a single TE locus is technically feasible, yet selecting a single TE locus for functional studies is challenging owing to ambiguity in TE mapping. Finally, investigating the evolutionary history of a TE family can be hampered by inaccurate TE annotations, particularly in genomes assembled from short sequencing reads. Renewed efforts to sequence complete mammalian genomes and transcriptomes with long reads will undoubtedly advance the field[96–98].

The integration, spreading and fixation/elimination of TEs in a host genome document the unique evolutionary history of that species. Once selfish elements, TEs that are domesticated, co-opted and repurposed during evolution have contributed a substantial amount of raw material for host genome innovation. Modern-day koalas present a unique experimental system to investigate TE endogenization, TE co-option and TE evolution[99], as they are undergoing genomic colonization by an exogenous retrovirus, KoRV[99], which has begun transitioning into an ERV.

Understanding TE–host interactions will yield powerful strategies for gene delivery, gene manipulation and genome engineering. Gene delivery mediated by DNA TEs has long been harnessed for genetic studies[100,101]. More recently, components of domesticated retrotransposons, such as *PEG10*, have been engineered as gene delivery tools for RNA therapy, utilizing their efficient RNA packaging ability and capacity to infect a variety of host cell types without eliciting immune response[68]. The innate host mechanisms that silence TEs, including RNAi[102] and CRISPR[103–105], can be reprogrammed to silence or engineer endogenous host genes for therapeutic purposes. These approaches have created numerous possibilities to treat a spectrum of human diseases.

Altogether, TE domestication reveals the evolutionary history of genes, gene regulation and genome organization, and substantially contributes to the molecular basis for species-specific, phenotypic diversity. TE biology enriches our understanding of disease mechanisms and empowers us with new therapeutic strategies. Friends or foes, our intimate relationship with TEs may have shaped who we are as a species and will likely continue to do so as long as we co-evolve with our TEs.

Received: 13 March 2022; Accepted: 27 June 2022;
Published online: 25 August 2022

References

1. McClintock, B. The origin and behavior of mutable loci in maize. *Proc. Natl Acad. Sci. USA* **36**, 344–355 (1950).
2. Britten, R. J. & Davidson, E. H. Gene regulation for higher cells: a theory. *Science* **165**, 349–357 (1969).
3. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
4. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
5. Craig Venter, J. et al. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
6. Waterston, R. H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
7. Wicker, T. et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
8. Jangam, D., Feschotte, C. & Betrán, E. Transposable element domestication as an adaptation to evolutionary conflicts. *Trends Genet.* **33**, 817–831 (2017).
9. Feschotte, C. & Pritham, E. J. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev. Genet.* **41**, 331–368 (2007).
10. Levin, H. L. & Moran, J. V. Dynamic interactions between transposable elements and their hosts. *Nat. Rev. Genet.* **12**, 615–627 (2011).
11. Doolittle, W. F. & Sapienza, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**, 601–603 (1980).
12. Capy, P. Taming, domestication and exaptation: trajectories of transposable elements in genomes. *Cells* **10**, 3590 (2021).
13. Polak, P. & Domany, E. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics* **7**, 133–148 (2006).
14. Gifford, W. D., Pfaff, S. L. & Macfarlan, T. S. Transposable elements as genetic regulatory substrates in early development. *Trends Cell Biol.* **23**, 218–226 (2013).
15. Garcia-Perez, J. R., Widmann, T. J. & Adams, I. R. The impact of transposable elements on mammalian development. *Development* **143**, 4101–4114 (2016).
16. van de Lagemaat, L. N., Landry, J.-R., Mager, D. L. & Medstrand, P. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.* **19**, 530–536 (2003).
17. Simonti, C. N., Pavličev, M. & Capra, J. A. Transposable element exaptation into regulatory regions is rare, influenced by evolutionary age, and subject to pleiotropic constraints. *Mol. Biol. Evol.* **34**, 2856–2869 (2017).
18. Miyawaki, S. et al. The mouse Sry locus harbors a cryptic exon that is essential for male sex determination. *Science* **370**, 121–124 (2020).
19. Flemr, M. et al. A retrotransposon-driven dicer isoform directs endogenous small interfering RNA production in mouse oocytes. *Cell* **155**, 807–816 (2013).
20. Sakashita, A. et al. Endogenous retroviruses drive species-specific germline transcriptomes in mammals. *Nat. Struct. Mol. Biol.* **27**, 967–977 (2020).
21. Modzelewski, A. J. et al. A mouse-specific retrotransposon drives a conserved Cdk2ap1 isoform essential for development. *Cell* **184**, 5541–5558.e22 (2021).
22. Senft, A. D. & Macfarlan, T. S. Transposable elements shape the evolution of mammalian development. *Nat. Rev. Genet.* **22**, 691–711 (2021).
23. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).
24. Ito, J. et al. Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *PLoS Genet.* **13**, e1006883 (2017).
25. 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).
26. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).
27. Ye, M. et al. Specific subfamilies of transposable elements contribute to different domains of T lymphocyte enhancers. *Proc. Natl Acad. Sci. USA* **117**, 7905–7916 (2020).
28. Peaston, A. E. et al. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev. Cell* **7**, 597–606 (2004).

29. Macfarlan, T. S. et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57–63 (2012).
30. Franke, V. et al. Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes. *Genome Res.* **27**, 1384–1394 (2017).
31. Hasuwa, H. et al. Production of functional oocytes requires maternally expressed PIWI genes and piRNAs in golden hamsters. *Nat. Cell Biol.* **23**, 1002–1012 (2021).
32. Gerlo, S., Davis, J. R. E., Mager, D. L. & Kooijman, R. Prolactin in man: a tale of two promoters. *BioEssays* **28**, 1051–1055 (2006).
33. 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).
34. Davis, M. P. et al. Transposon-driven transcription is a conserved feature of vertebrate spermatogenesis and transcript evolution. *EMBO Rep.* **18**, 1231–1247 (2017).
35. Beyer, U., Moll-Rocek, J., Moll, U. M. & Dobbelstein, M. Endogenous retrovirus drives hitherto unknown proapoptotic p63 isoforms in the male germ line of humans and great apes. *Proc. Natl Acad. Sci. USA* **108**, 3624–3629 (2011).
36. Pi, W. et al. The LTR enhancer of ERV-9 human endogenous retrovirus is active in oocytes and progenitor cells in transgenic zebrafish and humans. *Proc. Natl Acad. Sci. USA* **101**, 805–810 (2004).
37. Hu, T. et al. Long non-coding RNAs transcribed by ERV-9 LTR retrotransposon act in cis to modulate long-range LTR enhancer function. *Nucleic Acids Res.* **45**, 4479–4492 (2017).
38. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9**, 397–405 (2008).
39. Sundaram, V. & Wysocka, J. Transposable elements as a potent source of diverse *cis*-regulatory sequences in mammalian genomes. *Philos. Trans. R. Soc. Lond. B* **375**, 20190347 (2020).
40. Xie, M. et al. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nat. Genet.* **45**, 836–841 (2013).
41. Pehrsson, E. C., Choudhary, M. N. K., Sundaram, V. & Wang, T. The epigenomic landscape of transposable elements across normal human development and anatomy. *Nat. Commun.* **10**, 5640 (2019).
42. Miao, B. et al. Tissue-specific usage of transposable element-derived promoters in mouse development. *Genome Biol.* **21**, 255–280 (2020).
43. Bourque, G. et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.* **18**, 1752–1762 (2008).
44. 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).
45. Kunarso, G. et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42**, 631–634 (2010).
46. Schmidt, D. et al. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335–348 (2012).
47. Sundaram, V. et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res* **24**, 1963–1976 (2014).
48. Sundaram, V. et al. Functional *cis*-regulatory modules encoded by mouse-specific endogenous retrovirus. *Nat. Commun.* **8**, 14550 (2017).
49. Choudhary, M. N. K. et al. Co-opted transposons help perpetuate conserved higher-order chromosomal structures. *Genome Biol.* **21**, 16 (2020).
50. Xia, B. et al. The genetic basis of tail-loss evolution in humans and apes. Preprint at *bioRxiv* https://doi.org/10.1101/2021.09.14.460388 (2021).
51. Mayr, C. Regulation by 3′-untranslated regions. *Annu. Rev. Genet.* **51**, 171–194 (2017).
52. Kelley, D. & Rinn, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.* **13**, R107 (2012).
53. Jachowicz, J. W. et al. LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nat. Genet.* **49**, 1502–1510 (2017).
54. 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).
55. Percharde, M. et al. A LINE1–nucleolin partnership regulates early development and ESC identity. *Cell* **174**, 391–405.e19 (2018).
56. Zhao, Y. et al. Transposon-triggered innate immune response confers cancer resistance to the blind mole rat. *Nat. Immunol.* **22**, 1219–1230 (2021).
57. Chiappinelli, K. B. et al. Inhibiting DNA methylation causes an interferon response in cancer via dsRNA including endogenous retroviruses. *Cell* **162**, 974–986 (2015).
58. Ueda, M. T. et al. Comprehensive genomic analysis reveals dynamic evolution of endogenous retroviruses that code for retroviral-like protein domains. *Mob. DNA* **11**, 29–46 (2020).
59. Campillos, M., Doerks, T., Shah, P. K. & Bork, P. Computational characterization of multiple Gag-like human proteins. *Trends Genet.* **22**, 585–589 (2006).
60. Zhang, W. et al. Structural basis of Arc binding to synaptic proteins: implications for cognitive disease. *Neuron* **86**, 490–500 (2015).
61. Pastuzyn, E. D. et al. The neuronal gene *Arc* encodes a repurposed retrotransposon Gag protein that mediates intercellular RNA transfer. *Cell* **172**, 275–288.e18 (2018).
62. Ashley, J. et al. Retrovirus-like Gag protein Arc1 binds RNA and traffics across synaptic boutons. *Cell* **172**, 262–274.e11 (2018).
63. Kedrov, A. V., Durymanov, M. & Anokhin, K. V. The Arc gene: retroviral heritage in cognitive functions. *Neurosci. Biobehav Rev.* **99**, 275–281 (2019).
64. Okuno, H. et al. Inverse synaptic tagging of inactive synapses via dynamic interaction of Arc/Arg3.1 with CaMKIIβ. *Cell* **149**, 886–898 (2012).
65. Ono, R. et al. Deletion of *Peg10*, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. *Nat. Genet.* **38**, 101–106 (2005).
66. Clark, M. B. et al. Mammalian gene *PEG10* expresses two reading frames by high efficiency –1 frameshifting in embryonic-associated tissues. *J. Biol. Chem.* **282**, 37359–37369 (2007).
67. Abed, M. et al. The Gag protein PEG10 binds to RNA and regulates trophoblast stem cell lineage specification. *PLoS ONE* **14**, e0214110 (2019).
68. Segel, M. et al. Mammalian retrovirus-like protein PEG10 packages its own mRNA and can be pseudotyped for mRNA delivery. *Science* **373**, 882–889 (2021).
69. Sha, M. et al. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403**, 785–789 (2000).
70. Esnault, C., Cornelis, G., Heidmann, O. & Heidmann, T. Differential evolutionary fate of an ancestral primate endogenous retrovirus envelope gene, the EnvV syncytin, captured for a function in placentation. *PLoS Genet.* **9**, e1003400 (2013).
71. Dupressoir, A. et al. Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. *Proc. Natl Acad. Sci. USA* **106**, 12127–12132 (2009).
72. Dupressoir, A. et al. A pair of co-opted retroviral envelope syncytin genes is required for formation of the two-layered murine placental syncytiotrophoblast. *Proc. Natl Acad. Sci. USA* **108**, E1164–E1173 (2011).
73. Mangeney, M. et al. Placental syncytins: genetic disjunction between the fusogenic and immunosuppressive activity of retroviral envelope proteins. *Proc. Natl Acad. Sci. USA* **104**, 20534–20539 (2007).
74. Marco, A. & Marín, I. CGIN1: a retroviral contribution to mammalian genomes. *Mol. Biol. Evol.* **26**, 2167–2170 (2009).
75. Lloréns, C. & Marín, I. A mammalian gene evolved from the integrase domain of an LTR retrotransposon. *Mol. Biol. Evol.* **18**, 1597–1600 (2001).
76. Shiura, H. et al. PEG10 viral aspartic protease domain is essential for the maintenance of fetal capillary structure in the mouse placenta. *Development* **148**, dev199564 (2021).
77. Kitazawa, M., Tamura, M., Kaneko-Ishino, T. & Ishino, F. Severe damage to the placental fetal capillary network causes mid- to late fetal lethality and reduction in placental size in *Peg11/Rtl1* KO mice. *Genes Cells* **22**, 174–188 (2017).
78. McLaughlin, R. N. et al. Positive selection and multiple losses of the LINE-1-derived *L1TD1* gene in mammals suggest a dual role in genome defense and pluripotency. *PLoS Genet.* **10**, e1004531 (2014).
79. Huang, S. et al. Discovery of an active RAG transposon illuminates the origins of V(D)J recombination. *Cell* **166**, 102–114 (2016).
80. Ruiz, M. et al. Abnormalities of motor function, transcription and cerebellar structure in mouse models of THAP1 dystonia. *Hum. Mol. Genet.* **24**, 7159–7170 (2015).
81. Dejosez, M. et al. Ronin is essential for embryogenesis and the pluripotency of mouse embryonic stem cells. *Cell* **133**, 1162–1174 (2008).
82. Majumdar, S., Singh, A. & Rio, D. C. The human *THAP9* gene encodes an active P-element DNA transposase. *Science* **339**, 446–448 (2013).
83. Cosby, R. L. et al. Recurrent evolution of vertebrate transcription factors by transposase capture. *Science* **371**, eabc6405 (2021).
84. Koonin, E. V. & Krupovic, M. Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat. Rev. Genet.***16**, 184–192 (2014).
85. Kapitonov, V. V., Makarova, K. S. & Koonin, E. V. ISC, a novel group of bacterial and archaeal DNA transposons that encode Cas9 homologs. *J. Bacteriol.* **198**, 797–807 (2015).
86. Altae-Tran, H. et al. The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases. *Science* **374**, 57–65 (2021).
87. Malfavon-Borja, R. & Feschotte, C. Fighting fire with fire: endogenous retrovirus envelopes as restriction factors. *J. Virol.* **89**, 4047–4050 (2015).
88. Blanco-Melo, D., Gifford, R. J. & Bieniasz, P. D. Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors. *eLife* **6**, e22519 (2017).

89. Grow, E. J. et al. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**, 221–225 (2015).
90. Frank, J. A. et al. Antiviral activity of a human placental protein of retroviral origin. Preprint at *bioRxiv* https://doi.org/10.1101/2020.08.23.263665 (2020).
91. Yap, M. W., Colbeck, E., Ellis, S. A. & Stoye, J. P. Evolution of the retroviral restriction gene Fv1: inhibition of non-MLV retroviruses. *PLoS Pathog.* **10**, e1003968 (2014).
92. Horikoshi, M. et al. Positional cloning of the mouse retrovirus restriction gene Fvl. *Nature* **382**, 826–829 (1996).
93. Jin, Y., Tam, O. H., Paniagua, E. & Hammell, M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593–3599 (2015).
94. Yang, W. R., Ardeljan, D., Pacyna, C. N., Payer, L. M. & Burns, K. H. SQuIRE reveals locus-specific regulation of interspersed repeat expression. *Nucleic Acids Res.* **47**, e27 (2019).
95. Bendall, M. L. et al. Telescope: characterization of the retrotranscriptome by accurate estimation of transposable element expression. *PLoS Comput. Biol.* **15**, e1006453 (2019).
96. Miga, K. H. et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
97. Miga, K. H. & Wang, T. The need for a human pangenome reference sequence. *Annu Rev. Genomics Hum. Genet* **22**, 81–102 (2021).
98. Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
99. Stoye, J. P. Koala retrovirus: a genome invasion in real time. *Genome Biol.* **7**, 241 (2006).
100. Dupuy, A. J., Akagi, K., Largaespada, D. A., Copeland, N. G. & Jenkins, N. A. Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. *Nature* **436**, 221–226 (2005).
101. Ding, S. et al. Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell* **122**, 473–483 (2005).
102. Zamore, P. D., Tuschl, T., Sharp, P. A. & Bartel, D. P. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**, 25–33 (2000).
103. Jinek, M. et al. RNA-programmed genome editing in human cells. *eLife* **2013**, e00471 (2013).
104. Mali, P. et al. RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
105. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
106. Casacuberta, E. *Drosophila*: retrotransposons making up telomeres. *Viruses* **9**, 192–208 (2017).
107. Abad, J. P. et al. TAHRE, a novel telomeric retrotransposon from *Drosophila melanogaster*, reveals the origin of *Drosophila* telomeres. *Mol. Biol. Evol.* **21**, 1620–1624 (2004).
108. Levis, R. W., Ganesan, R., Houtchens, K., Tolar, L. A. & Sheen, F. Miin Transposons in place of telomeric repeats at a *Drosophila* telomere. *Cell* **75**, 1083–1093 (1993).

Acknowledgements

We are grateful to L. B. King for editing and proofreading this review. A.J.M. is supported by NIH (R00HD096108) and the Siebel Stem Cell Institute. T.W. is supported by NIH (R01HG007175, U24ES026699, U01CA200060, U01HG009391, U41HG010972 and U24HG012070). L.H. is a Thomas and Stacey Siebel Distinguished Chair Professor, and a Chan-Zuckerberg Biohub Investigator, supported by an HHMI Faculty Scholar award, a Bakar Fellow award and NIH grants (1R01GM114414, R01CA139067, 1R21OD027053, GRANT12095758, 1R01HD106809 and R01NS120287).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Lin He.

Peer review information *Nature Cell Biology* thanks Todd Macfarlan and William Theurkauf for their contribution to the peer review of this work.