

REVIEW

Open Access



More than causing (epi)genomic instability: emerging physiological implications of transposable element modulation

Pu-Sheng Hsu^{1†}, Shu-Han Yu^{1†}, Yi-Tzang Tsai¹, Jen-Yun Chang¹, Li-Kuang Tsai¹, Chih-Hung Ye¹, Ning-Yu Song^{2,3}, Lih-Chiao Yau¹ and Shau-Ping Lin^{1,4,5,6*} 

Abstract

Transposable elements (TEs) initially attracted attention because they comprise a major portion of the genomic sequences in plants and animals. TEs may jump around the genome and disrupt both coding genes as well as regulatory sequences to cause disease. Host cells have therefore evolved various epigenetic and functional RNA-mediated mechanisms to mitigate the disruption of genomic integrity by TEs. TE associated sequences therefore acquire the tendencies of attracting various epigenetic modifiers to induce epigenetic alterations that may spread to the neighboring genes. In addition to posing threats for (epi)genome integrity, emerging evidence suggested the physiological importance of endogenous TEs either as cis-acting control elements for controlling gene regulation or as TE-containing functional transcripts that modulate the transcriptome of the host cells. Recent advances in long-reads sequence analysis technologies, bioinformatics and genetic editing tools have enabled the profiling, precise annotation and functional characterization of TEs despite their challenging repetitive nature. The importance of specific TEs in preimplantation embryonic development, germ cell differentiation and meiosis, cell fate determination and in driving species specific differences in mammals will be discussed.

Keywords: Transposable elements (TEs), Functional RNAs, Epigenetics, Enhancers, Differentiation, Evolution

Introduction

Transposable elements (TEs) were first discovered in maize in the late 1940s [1]. TEs were considered endogenous “junk sequences” or “selfish genomic sequences”. This is mainly due to their virus-based genomic features that are designed to amplify themselves or move around the host genome at the cost of genomic instability of the host cells. In mammals, TEs comprise roughly 45% of the genomic sequences [2]. TEs have been classified into two categories: DNA transposons and retrotransposons. DNA transposons such as *tc1/mariner* cut and

paste themselves to reach transposition. Retrotransposons copy and paste via an RNA intermediate followed by reverse transcription to achieve retrotransposition. Retrotransposons are further classified into long terminal repeats (LTRs), such as endogenous retroviruses (ERVs), and non-LTRs, such as long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). Long since their invasion into an ancient host genome, evolutionarily older TEs have been mutated or truncated and eventually lost their ability to transpose within the genome [3]. On the other hand, evolutionarily younger retrotransposons are the most dangerous threats to genome integrity since they maintain the capability of transposition [4]. The occurrence of TE insertions can cause genomic instability [5–7] and transcriptional deregulation. Under pressure from TE-derived hazards,

*Correspondence: shaupinglin@ntu.edu.tw

[†]Pu-Sheng Hsu and Shu-Han Yu authors contributed equally

¹ Institute of Biotechnology, College of Bio-Resources and Agriculture, National Taiwan University, Taipei, Taiwan

Full list of author information is available at the end of the article



© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

both transcriptional and posttranscriptional defense systems evolve in the host [8]. Although many TEs are neutralized in the host, more than 120 disease-causing TE insertions have been documented in humans [9].

Most TEs contain numerous copies, which makes it difficult for scientists to map each TE sequences to the exact genomic location with second-generation short read sequencing platforms. However, in the last few years, advanced third-generation sequencing technologies (reviewed by Amarasinghe, 2020 [10]) have enabled the detection of long sequencing reads and the identification of location-specific small variations in each TE family member. With the associated optimized bioinformatics packages, getting the precise localization of TEs is now practical. The improvement of technologies helps scientists reveal that these mutated TEs are susceptible to substantial epigenetic modulation, which also spreads to adjacent genomic regions and therefore affects the expression of neighboring genes during the development and physiological function of organisms [11]. TEs actually function as a double-edged sword in host cells. In this review, we introduce not only how TEs are regulated by host organisms and what happens when they are dysregulated but also recent evidence showing the physiological properties of TEs.

The threats associated with TEs in host cells

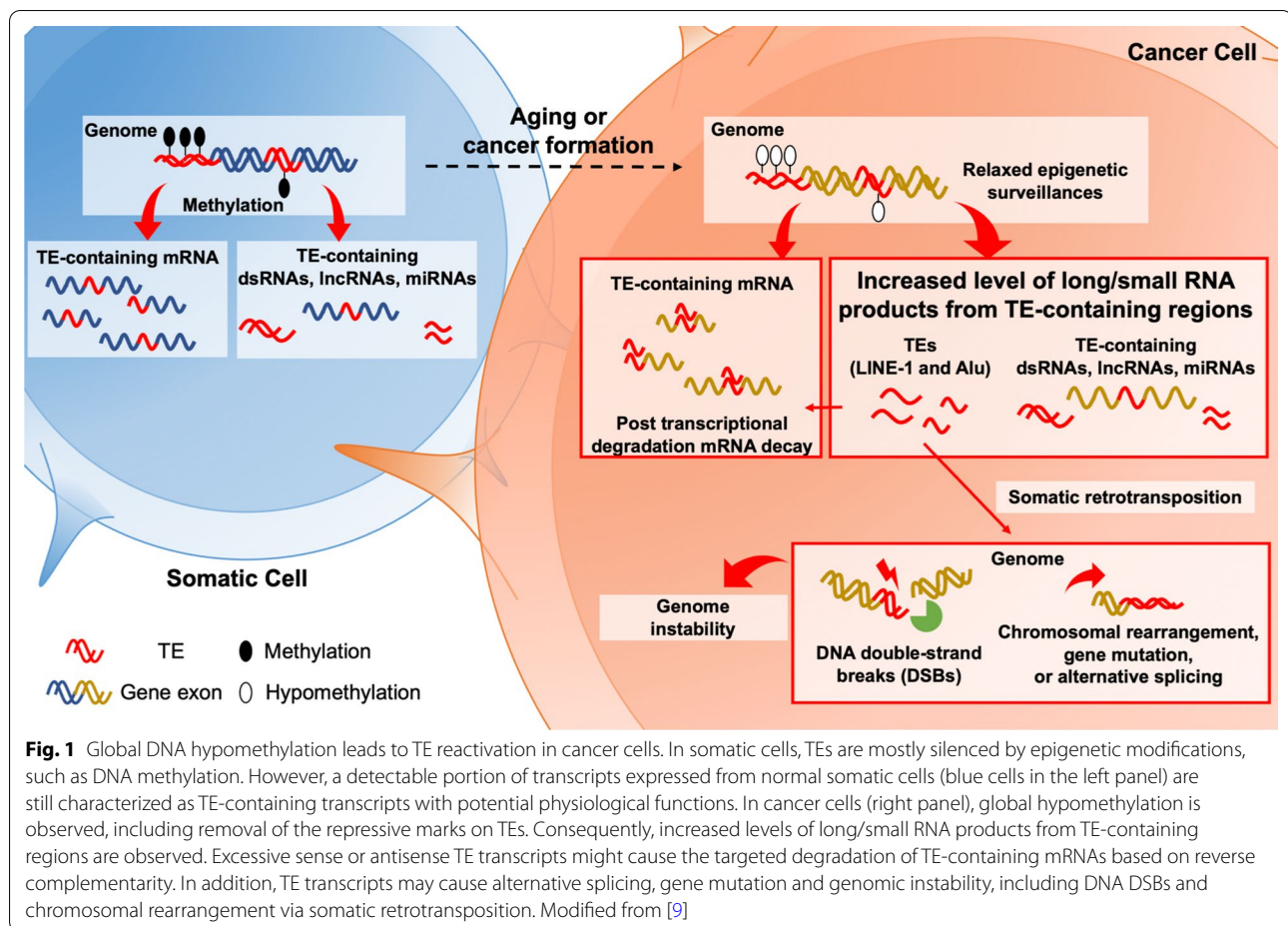
TEs are usually silenced by epigenetic modifications, including DNA methylation and histone modifications [12–14]. Some TEs are packaged into heterochromatin structure associated with nuclear lamins [15]. However, aging-associated or other aberrant micro- or macroenvironment-induced epigenomic defects may cause dysregulation of TEs. These TE deregulations may induce genomic instability [16, 17] and diseases, including neurodevelopmental disorders (reviewed by Lapp, 2019 [18]), neurodegeneration [19, 20], autoimmunity [21] and cancer [22].

Some of the activated TEs have the ability to insert into other genomic sites, the processing of which can cause DNA double-strand breaks (DSBs). Gasior et al. documented that the transfection of LINE-1 into HeLa cells induces DSBs and G2/M cell cycle arrest [23]. Although DNA repair system can fix most of the TE transposition-induced DSBs, this process still disrupts genome stability and may cause chromosomal rearrangement, gene mutation or alternative splicing. For example, LINE-1 insertion usually creates a large genetic deletion [24] and may lead to chromosomal rearrangement [25]. Moreover, scientists have found that more than eighty percent of Alu elements inserted into the exons of mRNAs cause a frameshift or a premature termination codon that affects the expression

of those coding genes [26]. In addition, TEs inserted into introns also cause problems; for example, Alu and LINE-1 induce alternative splicing to affect transcript integrity [27, 28]. If TEs insert into DNA repair genes such as breast cancer type 2 susceptibility protein (*BRCA2*) [29] or tumor suppressor genes such as adenomatous polyposis coli protein (*APC*) [30] and retinoblastoma protein 1 (*RBI*) [31], they may cause genome instability or tumor formation, respectively. Furthermore, several cancers, including lung cancer, renal cancer, breast cancer, etc., are correlated with DNA hypomethylation and TE deregulation [5, 7, 32, 33]. Kong et al. also observed similar phenomena by analyzing the transcriptomes of over twenty different cancers from the Cancer Genome Atlas database [34]. Lee et al. showed the consequences of TE insertion in different cancer samples by performing whole genome sequencing. They observed that LINE-1-inserted genes are usually dysregulated and associated with cancer formation [35].

Even without transposition, the presence of TE sequences may still disrupt (epi)genome stability. TE enriched sequences can be found in the flanking sequences of DSBs in cancer cells. It is suggested that during DNA replication, short inverted repeats such as the Alu element may form a secondary hairpin structure, which can lead to replication stalling and even DNA DSB formation [36, 37]. On the other hand, global hypomethylation of endogenous TEs in cancer cells might further induce alternative promoter activation. For example, Jang et al. investigated 15 cancer types in 7,769 tumor and 625 normal tissue datasets to identify 129 TE-related promoter activation events. The authors showed that the global profile of these TEs is associated with 106 oncogenes across 3,864 tumors, and TEs are the cause of oncogenic activation [38, 39], which may further contribute to tumor initiation and progression [7, 22, 38]. As shown in Fig. 1, DNA damage, chromosomal rearrangement [35, 40–43], alternative splicing and gene expression are induced by abnormal TE activation and, in some cases, subsequent insertion in cancer cells. These TE insertions were shown to further activate aberrant and recombinant gene transcription [44].

Global DNA hypomethylation associated with depression of TEs can be observed in normal aging cells as well [45], but may be reversible. As shown in our recent study, transient ectopic expression of an epigenetic cofactor, DNMT3L, in aging fibroblasts is sufficient to inhibit senescence progression and facilitate epigenetic repression of some aging associated derepressed genes and TEs [46]. It is therefore possible to develop strategies in mitigating aging-associated defects via increasing epigenomic surveillance.



Modulating TEs in host cells

Several strategies for protecting host cells from TE-derived disruption have emerged during evolution. Both transcriptional and posttranscriptional pathways are involved in TE regulation. The consummate management of TEs is presumably performed with the coordination of DNA methylation, histone modifications, and small RNA-mediated RNA degradation in mammalian cells.

DNA methylation

Among the TE modulation machineries, DNA methylation is used for long-term TE surveillance in mammals. Both LTR and non-LTR retrotransposons are inhibited by DNA methylation [13, 47]. In DNA methylation-deficient models, the expression of TEs was increased significantly and caused developmental defects [13, 47, 48].

Thirty years ago, Prof. Timothy Bestor hypothesized that the DNA methylation machinery evolved from immune mechanisms in prokaryotes designed to protect against phage infection into gene expression and genome structural modulation in large-genome plants and vertebrate animals, including silencing of TEs and other repeat

sequences to reduce their exposure to the transcriptional machinery [60]. Recently, Zhou et al. examined the complete DNA sequences of 53 organisms, and the results supported the original hypotheses that DNA methylation enables TE-driven genome expansion. The results of these analysis also indicated that DNA methylation spreads to the flanking host DNA sequences associated with the inserted TEs [59].

DNA methylation in mammals predominantly take place on the 5th carbon of cytosine (5-methylcytosine; 5mC), catalyzed by a family of DNA methyltransferases (DNMTs) [49]. DNMT1 transfer the methyl group from S-adenyl methionine (SAM) mainly to hemimethylated DNA template during DNA replication to maintain DNA methylation mark from mother strand to daughter strand [50–52]. DNMT3A and DNMT3B target unmethylated DNA template to introduce de novo methylation mark during development [53, 54]. Another pluripotent stem cells and developing germ cells enriched DNMT3 family member, DNMT3-Like (DNM3L), lacks functional catalytic domain but serve as an important co-factor to facilitate DNMT3A and DNMT3B for de novo

methylation on TEs and beyond [55, 56]. DNA methylation not only results in the transcriptional repression of TEs but can also result in C-T deamination and inactivate those sequences permanently at a more advanced level [57–59].

Histone modification

Post-translational histone modifications, including acetylation, methylation and ubiquitylation, collaborate with DNA methylation machineries to accomplish multiple layers of epigenetic modulation [61]. The repressive marks H3K9me2/3 and H3K27me3 are responsible for silencing different types of transposons at different differentiation stages and coordinating TE regulation with other silencing strategies [62]. H3K9me2/3 is linked to the maintenance of DNA methylation mediated by ubiquitin-like, containing PHD and RING finger domains protein 1 (UHRF1) in ESCs [63–65]. Furthermore, H3K9me2/3 may be connected to the role of UHRF1 in the maintenance of 5mC levels at intracisternal A-type particles (IAPs) in preimplantation embryos [63].

Krüppel-associated box zinc finger protein (KRAB-ZFP)-KRAB-associated protein-1 (KAP1)-mediated silencing is an interesting system that coevolves with TEs and is critical in early embryogenesis. Its targeting might be adapted to new retroelements through evolutionarily changing the DNA-binding region [8, 66]. KRAB-ZFPs are transcription factors that use the C-terminal ZFP to target TE sequences and the N-terminal KRAB domain to bind tripartite motif-containing 28 protein (TRIM28)/KAP1, which is a scaffold that recruits epigenetic modifiers. The H3K9-specific methylase SET domain bifurcated histone lysine methyltransferase 1 (SETDB1) is recruited to introduce repressive histone modifications, such as H3K9me3 [67, 68]. In addition, ZFP/KAP1 was reported to interact with DNMT3A/3B and to play a role in the maintenance of DNA methylation in embryonic stem cells (ESCs) [69].

Small RNA-mediated TE regulation

The small RNA-mediated pathway is another mechanism that manages gene expression in a sequence-dependent manner. Additionally, it is vital for controlling and counteracting TE transcripts, especially at the developmental stages when repressive DNA methylation and histone marks are modified during epigenetic reprogramming. In germ cells and early embryos, the DNA methylation profile changes dynamically during development [70]. RNA interference (RNAi) was indicated to regulate the expression of the retrotransposons murine endogenous retrovirus-leukemia protein (MuERV-L) and IAPs in preimplantation mouse embryos [71]. P-element-induced wimpy testis protein (PIWI)-interacting RNA (piRNA),

on the other hand, is the best studied small RNA-mediated epigenetic regulator that functions to repress mobile genetic elements in germ cells [72]. The PIWI-piRNA pathway is also highly conserved across the animal kingdom to mitigate the threat of retrotransposition in germ cells [73]. piRNAs are approximately 26–34 nt single-stranded RNAs with 3'-end-2'-O methylation that are processed from long single-stranded transcripts, including TE transcripts. The biogenesis of piRNAs requires PIWI proteins, a specific clade of Argonautes, and other piRNA biogenesis-associated proteins in the nuage/germ granules immediately outside the nuclear envelope of developing germ cells [74–76]. Located in germ granule cement, the PIWI-piRNA pathway is generally considered a posttranscriptional silencing mechanism. With guidance by piRNA, PIWI proteins target sequence-complementary transposon transcripts and destroy RNAs by cleavage. Cleavage degrades TE transcripts during piRNA biogenesis and therefore blocks the reverse transcription and retrotransposition of these TE elements. The “ping-pong cycle” of secondary piRNA biogenesis that involves targeting and cutting sense- and antisense-strand TEs via the PIWI-piRNA complex with mature antisense and sense TE sequence-derived piRNAs, respectively, is particularly efficient at minimizing the expression of full-length TE transcripts [56, 77]. In addition, the PIWI-piRNA complex enters the nucleus, targets nascent TE transcripts and recruits epigenetic silencers, including histone methyltransferases, and even de novo DNA methylation for the long-term maintenance of transcriptional silencing [78–80].

Joint protection by DNA methylation, histone modification, and small RNA-mediated regulation defends against invasion by TEs. The crosstalk between these strategies affords different layers of retroelement regulation and modulates the orchestration of gene expression.

Can't eliminate them, use them: the physiological functions of TEs in host cells

Accumulating evidence suggest that TEs acquire important physiological functions through evolution. Several LTR retrotransposon-derived genes have been discovered in the human genome, domesticated to neogenes of functional proteins. These include sushi/Mart [81], paraneoplastic Ma antigens protein (PNMA), activity-regulated cytoskeleton-associated protein (ARC), skin-specific retroviral-like aspartic protease/aspartic peptidase retroviral like 1 protein (SASPase/ASPRV1) [82], SCAN (SRE-ZBP, CTfin-51, AW-1 and 2 Number 18 cDNA protein) family members [83], recombination activating 1 protein (RAG1) and recombination signal sequences (RSSs) [84]. Those TE containing sequences are important in a myriad of biological processes, such as stem cell properties,

tissue development, inflammation, V(D)J recombination and neurophysiology [84–89]. In addition, through big data analysis, Kong et al. also found that TE expression is correlated with the regulation of cytokine responses and induces the infiltration of some types of immune cells in cancer [34].

The properties of TEs in attracting epigenetic modifiers also enable the inserted TEs to become functionally relevant genomic features. By performing comprehensive chromatin immunoprecipitation (ChIP)-seq analyses in human and mouse leukemia cell lines (K562 and MEL) and lymphoblast cell lines (GM12878 and CH12), researchers have shown that TE sequences are present in 20% of transcription factor binding sites in immune cells in which neighboring areas show open chromatin marks, including DNA hypomethylation, H3K4me1, H3K4me3 and H3K27ac. In addition, LTRs comprise the majority of TE-derived binding peaks in human [90]. Furthermore, full-length retrotransposons are composed of a complete transcription unit, including a strong promoter [91–93]. Newly inserted or endogenously derepressed TEs may drive the transcription of neighboring genes or intergenic regions and evolve into functional RNAs. On the other hand, genome-wide chromatin profiling data and high-throughput sequencing have revealed the expression and lineage-specific distributions of multiple TE subfamilies [91, 94, 95]. Some enhancers are believed to have evolved from TEs [96]. TE-associated enhancers are involved in many developmental processes [97–99]. The functions of transposable elements in mammals are separated into two major categories: TE-containing functional RNAs and TE-containing *cis*-regulatory elements.

TE-containing transcripts functioning in trans

TE-containing transcripts as a miRNA source

TE-containing transcripts can be processed into microRNAs (miRNAs) [100–102]. Many of the TE-derived miRNA loci are located at the 3' untranslated regions (3'UTRs) of protein-coding genes, which is also the target site for miRNA-mediated posttranscriptional regulation. For example, the expression of numerous Argonaute RISC catalytic component 2 protein (AGO2)-associated functional miRNAs and their target sites derived from LINE-2 sequences has been discovered in the brain cortex ([103] and references therein). L2b-derived miR-95 was significantly downregulated in the tumor biopsies of patients with glioblastoma, suggesting that the TE-centered network of miRNA targets might contribute to the normal functions of the brain (Fig. 2A).

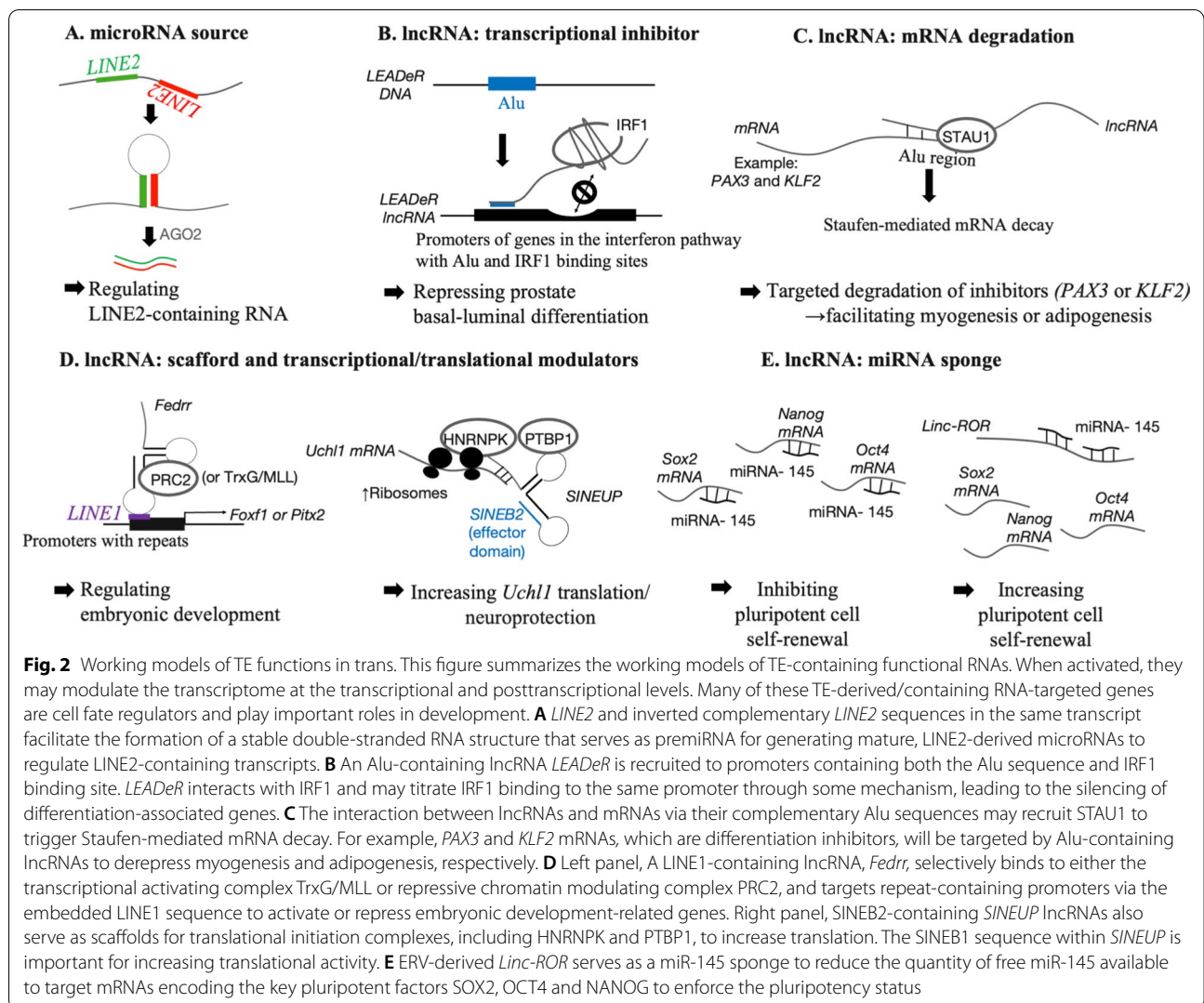
TE-containing long noncoding RNAs (lncRNAs)

TE sequences are prevalent in lncRNAs. Approximately 80% of the lncRNAs identified in several studies contain

at least one TE [104–106]. TEs embedded in lncRNAs may constitute the functional domain or otherwise regulate the expression, processing or localization of the host transcript (reviewed by Fort, 2021 [107]).

The lncRNAs that are involved in development may exploit embedded TEs to interact with the regulatory region of developmental genes for transcriptional modulation. During human prostate development, the canonical miRNA MIR205HG locus alternatively derives a lncRNA, long epithelial Alu-interacting differentiation-related RNA (*LEADeR*). The Alu sequence within *LEADeR* binds to the Alu element present in the regulatory sequences of its target gene, possibly by forming a paired RNA–DNA hybrid, which prevents interferon-regulatory factor 1 (IRF1) from interacting with a binding site proximal to the Alu element, thus leading to transcriptional repression of genes for luminal cell differentiation and subsequently sustaining basal cell identity [108] (Fig. 2B). In addition, LINE1 is embedded within fetal-lethal noncoding developmental regulatory RNA (*Fendrr*), a mouse lateral plate mesoderm-specific lncRNA, functioning as a putative DNA binding domain that binds to low-complexity repeats in the promoter region of target genes. The histone-modifying complexes polycomb repressive complex 2 (PRC2) and mixed lineage leukemia protein (MLL) associated with *Fendrr* are thus recruited to regulate the embryonic development of the heart and body wall by shaping the chromatin signatures of the genes involved [109, 110] (Fig. 2D).

Moreover, TE-containing lncRNAs control the expression of protein-coding genes through diverse posttranscriptional mechanisms. In contrast to the generation of miRNAs, TEs in some lncRNAs function as competing endogenous RNAs (ceRNAs, also called miRNA sponges) that recruit miRNAs with a complementary sequence from their recognition element in an mRNA, thereby stabilizing the target mRNA. For example, long intergenic non-protein-coding RNA, regulator of reprogramming (*Linc-RoR*), which is derived from human endogenous retrovirus H (HERV-H), is reported to protect the mRNAs of pluripotency-associated core transcription factors (octamer-binding transcription factor 4 (*OCT4*), SRY (sex determining region Y)-box 2 (*SOX2*), and *NANOG*) from miR-145-mediated degradation in self-renewing ESCs [111] (Fig. 2E). Recently, a primate-specific transcript isoform of the conserved protein coding gene cytochrome P450, family 20, subfamily A, polypeptide 1 (*CYP20A1*) was found to be untranslated and hence considered an lncRNA (*CYP20A1_Alu-LT*), harboring a stretch of Alu sequences at the 3'UTR to form a potential miRNA sponge [112]. The 9 miRNAs corresponding to *CYP20A1_Alu-LT* expression in primary human neurons are deduced to have mRNA



targets involved in tissue-specific processes of blood coagulation and neuron development. Other cancer-relevant TE-lncRNAs modulate signaling pathways by competing for the miRNAs that target mRNAs encoding the related proteins; for example, the lncRNA hepatocellular carcinoma up-regulated long non-coding RNA (*HULC*, consisting of LTR-mammalian LTR transposon 1 A (MLT1A)) is expressed at high levels in liver cancer, and BRAF-activated nonprotein coding RNA (*BANCR*, consisting of LTR-MER41B) has been shown to act as a sponge for miR-372 to derepress protein kinase cAMP-activated catalytic subunit beta (*PRKACB*) and for miR-338-3p to derepress insulin-like growth factor 1 receptor (*IGF1R*) in esophageal squamous cell carcinoma [113–115].

SINEUP, a type of lncRNA containing SINE elements that upregulate the translation of target mRNAs, is a bipartite antisense RNA with one effector domain

containing the SINE element and an RNA-binding domain to recognize the target mRNA by complementary pairing with the 5' end sequence surrounding the AUG start codon. The underlying mechanism was recently suggested: the SINE element domain contributes to the recruitment of the RNA binding proteins polypyrimidine tract-binding protein 1 (PTBP1) and heterogeneous nuclear ribonucleoprotein K (HNRNPK), leading to the translocation of the paired lncRNA and target mRNA into the cytoplasm to facilitate the assembly of translational initiation complexes [116]. A lncRNA antisense to mouse ubiquitin carboxy-terminal hydrolase L1 (*Uchl1*) containing the *SINEUP* feature has been identified. It increases the translation of *Uchl1*/Park5, which is essential for brain function and particularly for neuron maintenance [117] (Fig. 2D, right panel). Moreover, a human *SINEUP* lncRNA discovered in the brain transcriptome was shown to upregulate the translation of protein

phosphatase 1 regulatory subunit 12A (PPP1R12A), a downstream effector of inhibitory glutamate receptor delta-1 (GluD1), in postsynaptic cortical pyramidal neurons [118]. This unique regulatory function of TE-containing lncRNAs has recently prompted scientists to design and apply synthetic *SINEUP* to increase the translation of proteins of interest [119].

One type of TE-containing lncRNA is involved in mRNA degradation, specifically through the Staufen-mediated mRNA decay (SMD) mechanism [120, 121]. STAU (Staufen protein) binds to double-stranded RNA that is formed by imperfect base pairing between an Alu element in the 3'UTR of the target mRNA and another Alu element within a lncRNA (named half-STAU1-binding site RNA, *½-sbsRNA*) to elicit the SMD mechanism by recruiting up-frameshift suppressor 1 (UPF1) and UPF2, the core factors in the mRNA degradation pathway. Examples of development-relevant *½-sbsRNA* lncRNAs in humans include lncRNAs that target the mRNA of *PAX3* (encoding the myogenesis inhibitor paired box gene 3), which is implicated in myogenesis [122], lncRNAs that target the Krüppel-like factor 2 (*KLF2*) mRNA (the *KLF2* protein, in turn, negatively regulates the adipogenic gene *PPAR γ*) in adipogenesis [123], and lncRNAs involved in mouse myogenesis [124], with *SINE B1*, *B2*, and *B4* subfamilies (except for the primate-specific Alu) among the putative mouse *½-sbsRNAs* targeting the 3'UTRs of several mRNAs for degradation (Fig. 2C).

The gray zone between *cis* and *trans*: native elongating TE-containing functional RNAs modulate X chromosome inactivation and reactivation

Xist is a well-known lncRNA that is critical for initiating and was recently shown to also be important for maintaining X chromosome inactivation (Xi) in female cells. It consists of various tandem repeats (A–F), possibly originating from a variety of TE families, including ERVs, LINES and SINES. When *Xist* “coats” the inactivating X chromosome, these TE components within the RNA sequence are required for the recruitment of several transcriptional silencers, polycomb repressive histone modifiers, and other factors related to the establishment and maintenance of X chromosome inactivation ([125]; reviewed by Pintacuda, 2017 [126]).

Additionally, relevant to the X chromosome dosage compensation process but having the opposite function, primate-specific *Xact* competes with *Xist* and enables erosion of the Xi chromosome, which results in X chromosome reactivation (XCR) [112–114]. When we studied *Xact* lincRNA sequences in the UCSC genome browser on Human Feb. 2009 (GRCh37/hg19) Assembly, we identified its embedded TE elements, including LTR9B, AluY, and MLT1J (unpublished observation).

Despite the presence of TEs in *Xact*, the function and interacting proteins of TEs have yet to be determined. The ERV1 LTR9B-derived sequence binds to OCT4 and SOX2 proteins, which subsequently modulate the expression of various pluripotent genes [127]. Further investigations of whether LTR9B-containing *Xact* is also involved in the initiation or maintenance of pluripotency, in addition to its X chromosome reactivation function, would be interesting.

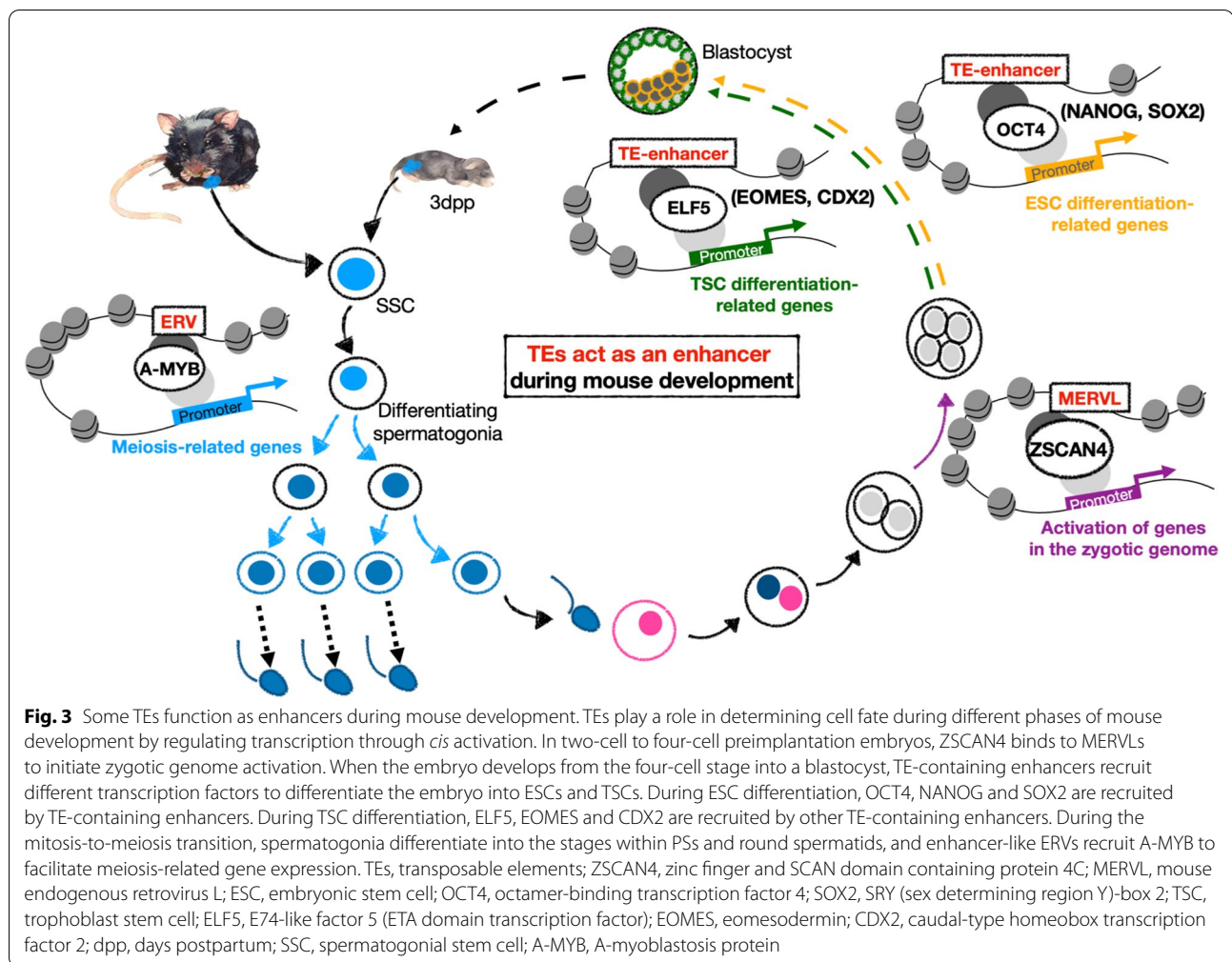
TE-containing *cis*-regulatory elements

Apart from being incorporated into functional RNAs to execute trans-acting functions, TE DNA sequences are also substantially involved in modulating gene expression by serving as binding sites for heavily weighted transcription factors, epigenetic modifiers or insulator binding proteins. For example, thousands of ERVs carry functional tumor suppressor protein P53 binding sites and regulate nearby genes, especially in the event of DNA damage [128, 129]. The expansion of these mobile carriers of transcriptional modulators and chromatin looping factors also provided opportunities for strain-specific or species-specific transcriptional networks and phenotypes [130].

TE-containing enhancers

After fertilization, different ERVs are activated at different stages of preimplantation embryo development [131]. For instance, oocytes/zygotes express the ERVK family member RLTR40, while zygotes/2-cell stage embryos express the ERVL-MaLR family member MTA. When 2-cell embryos differentiate to the 4-cell stage, the embryo faces zygotic genome activation (ZGA), the stage in which the embryos produce necessary RNA and protein from their own genome and gradually wean from those inherited from the oocytes. The expression and enhancer activities of MERVL, as documented by increased chromatin accessibility, are both critical for ZGA. Embryonic development is arrested upon MERVL deficiency [132]. The expression of 3' downstream proximal genes, including many cleavage-stage specific genes (cleavage genes), is modulated by a mechanism depending on MERVL accessibility. [133]. Upon activation by DUX4 and Zscan4c, key factors in the ZGA stage, MERVL plays an important role in cleavage gene regulation [131] (Fig. 3). Moreover, MERVL is also involved in translational modulation at the ZGA stage [134]. In summary, MERVL is an ERV that is specifically expressed and serves as an active enhancer during ZGA to modulate cleavage genes crucial for zygote development [132].

Under in vitro culture conditions, ESCs and trophoblast stem cells (TSCs) can be derived from blastocyst-stage embryos. ESCs and TSCs are responsible for the



development of fetal and extraembryonic tissue, respectively. Researchers performed assay for transposase-accessible chromatin using sequencing (ATAC-seq), ChIP-seq and promoter capture Hi-C (PChI-C) to study ESCs and TSCs and suggested that they contained distinct TE subfamilies that functioned as enhancers to regulate gene expression and determine cell differentiation. By performing ATAC-seq and H3K27ac enrichment analyses and confirming binding by at least one of the three key transcription factors (NANOG, OCT4 and CTCF in ESCs; ELF5, EOMES and CDX2 in TSCs), a subset of TEs were defined as “ESC/TSC TE enhancers” [98, 99] (Fig. 3). The activity of these TE enhancers is more restricted in ESCs/TSCs than that of non-TE enhancers. Using PChI-C to analyze the correlation between enhancers and gene expression, researchers showed that TE enhancer-interacting genes displayed higher expression levels in both ESCs and TSCs than genes interacting with non-TE enhancers [98]. Furthermore, the analysis of gene expression levels across a wide

array of tissues indicated that genes interacting with TE enhancers were almost exclusively expressed in ESCs or TSCs [135]. Genetic and chromatin analyses suggested that TE enhancers may be used to support lineage-specific expression of a subset of genes in early embryonic development [136]. Thus, TE enhancers play a critical role in early embryonic development and differentiation.

In male germ cells, dramatic reorganization of epigenomic modifications occurs during the transition from mitosis to meiosis in spermatogenesis. Enhancer-like ERVs such as RLTR10 also recruit A-myoblastosis protein (A-MYB) to facilitate germ cell differentiation (Fig. 3). In vitro dual-luciferase assays indicated that A-MYB dramatically increased the enhancer activity of ERVs. Interestingly, A-MYB depletion led to a decrease in the H3K27ac level, suggesting that A-MYB plays a role in the activation of enhancer-like ERVs. Human ERVs also exhibit this enhancer-like function in spermatogenesis. MER57E3 (ERV1) and LTR5B (ERVK) are enriched with H3K27ac in pachytene spermatocytes (PSs).

Moreover, ERV1 and ERVK also contain binding motifs for A-MYB, which is expressed at high levels in human spermatocytes, suggesting that enhancer-like ERVs have similar activation mechanisms in human spermatogenesis. A small fraction of super-enhancers required for the mitosis-to-meiosis transition are also ERV-containing, A-MYB-binding enhancers, associated with the activation of meiosis-associated genes [137, 138] (personal communication between Prof. Satoshi Namekawa and Prof. Shau-Ping Lin).

Enhancer-like ERVs are also associated with the diversity of gene expression in different species during evolution. Forty-eight mouse-specific genes were identified among 381 enhancer-like ERV-adjacent genes. Moreover, by analyzing rodent-specific ERVKs, researchers found that enhancer-like ERVs show differences in both copy numbers and genome distributions between mice and rats. In humans, 52 of 66 enhancer-like MER57E3 sequences are located at the first intron of a zinc finger protein. Among them, 47 of 52 were KRAB-ZFPs, suggesting a coevolutionary mechanism. This phenomenon was also observed in neuronal differentiation—briefly, KRAB-ZFPs partner with ERVs, regulating gene expression in human neurons [11]. The levels of enhancer-like MER57E3 and ERVK-adjacent genes are upregulated during the mitosis-to-meiosis transition in humans. Similar to those in mice, 61 of 138 enhancer-like ERV-adjacent genes were identified as primate-specific genes. Thus, ERVs have rapidly evolved in mammals to regulate several function-specific genes in the host genome [137, 138].

In the immune system, Chuong et al. showed that ERVs are significantly enriched in numerous interferon (IFN) regulatory elements in different mammalian genomes [139]. As a result, ERVs are considered IFN-inducible enhancers, and ERVs are strongly correlated with the innate immunity-associated IFN response [139]. Additionally, TEs are more highly enriched near immune genes in cytotoxic T cells and CD8⁺ cells than in nonimmune cells, supporting the hypothesis that the immune response may depend on the function of TEs as enhancers to rapidly activate immune pathways. In adaptive immune cells, TEs have been reported to function as enhancers that regulate putative CD8⁺ T cell immunity [140]. Ye et al. [140] employed genome-wide chromatin analysis and ATAC-seq to assess the contribution of TEs to T lymphocyte development. Researchers divided the T cell enhancer region into three distinct domains, an accessible core, proximal flanking region, and distal flanking region, to further elucidate the regulatory functions of different TEs. The authors proposed that different TEs may be predisposed to contribute distinct regulatory functions. For example, ERVs enriched at enhancer cores

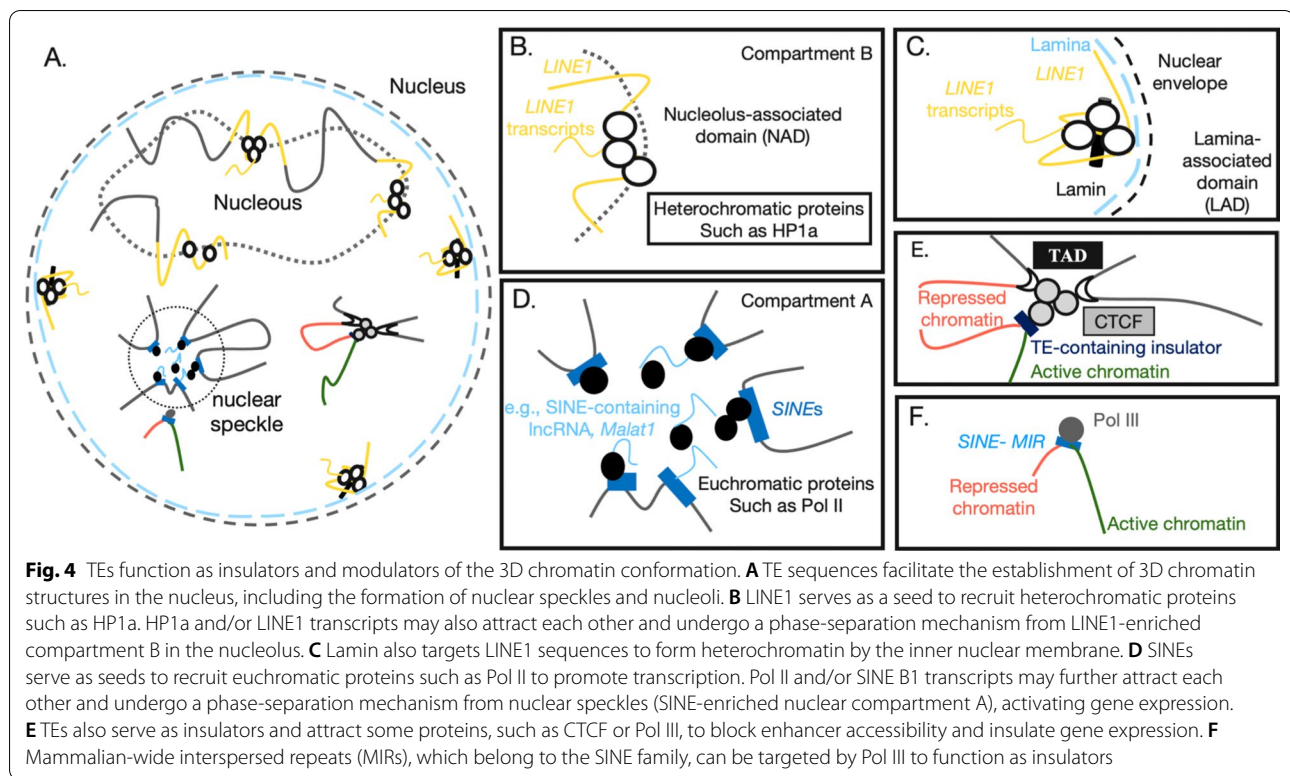
may provide transcription factor binding sites, while B1 SINEs enriched in enhancer flanks are more likely to facilitate chromatin organization. Moreover, SINEs are associated with high levels of the histone mark H3K4me1 [140], which is thought to serve as an enhancer [141]. Ye et al. [140] further suggested that epigenetic dysregulation of TE-derived enhancers may result in inappropriate activation of CD8⁺ T cells, which further shows a high correlation with TEs, especially those in the ERV subfamily [140].

Tes function as insulators and modulate the 3D chromatin conformation

Apart from attracting epigenetic modifiers to spread histone modifications and sometimes DNA methylation and therefore affecting the transcriptional activities of their neighboring genes, the involvement of TE sequences as insulators or factors contributing to the modulation of 3D chromatin architecture has been revealed over the last 2 decades (reviewed by Nishihara, 2019 [142]) (Fig. 4A–F). For example, the on-site transcripts of retrotransposon SINE B2 repeats function as insulators to prevent enhancer access and thus provide domain boundaries during organogenesis [143]. The 11-zinc-finger CCCTC-binding factor (CTCF) is a well-known *trans*-acting transcriptional repressor and a critical mediator of chromatin looping. Some retrotransposons contain CTCF binding sites, and with their expansion in a particular host genome, the 3D chromatin looping structures change with them, sometimes generating species-specific chromatin looping structures [144] (Fig. 4E). In addition to CTCF-dependent modulation of the 3D genome architecture, a recent study also identified another transposable element, mammalian-wide interspersed repeats (MIRs), serving as insulator elements in immune cells via a CTCF-independent pathway [145] (Fig. 4F). Homotypic clustering of L1 and B1/Alu transcripts reorganizes the 3D genome into higher-order compartments [146] (Fig. 4B–D). From the perspective of evolution, evolutionarily young TE subfamilies such as L1PA and AluY are significantly enriched at topological associating domain (TAD) boundaries in the developing cortex of human brains, while older TE subfamilies such as MIR, LINE-2, Charlie, and MaLR are enriched at TAD boundaries conserved across species [130].

Conclusions

TEs can cause genomic instability via transposition dependent and independent mechanisms, potentially resulting in cell death or cancer formation. DNA methylation, histone modifications and functional RNA machineries are evolved to modulate TE at the cell type dependent and developmental stage-dependent manner.



Cumulative evidence also suggested physiological significance of TE sequence-dependent mechanisms that provide novel layers of transcriptome modulation in epigenetic, nuclear architecture and post-transcriptional levels. These include TE-containing transcripts serving as miRNA sources, miRNA sponges and functional RNAs for guiding DNA binding proteins; TE-containing cis regulatory element as enhancers, promoters and insulator to regulate gene expression. In addition, TEs also act as genetic accelerators of evolution, contributing to the genome size, species-specific gene regulatory network rewiring, morphological innovation. Further understanding of TE related physiological functions and pathological etiology could lead to novel therapeutic opportunities.

Abbreviations

AGO2: Argonaute RISC catalytic component 2 protein; APC: Adenomatous polyposis coli protein; ARC: Activity-regulated cytoskeleton-associated protein; A-MYB: A-myoblastosis protein; ATAC-seq: Assay for transposase-accessible chromatin using sequencing; *BANCR*: BRAF-activated nonprotein coding RNA; BRCA2: Breast cancer type 2 susceptibility protein; CDX2: Caudal-type homeobox transcription factor 2; CD8: Cluster of differentiation 8; ceRNA: Competing endogenous RNAs; ChIP: Chromatin immunoprecipitation; CTCF: The 11-zinc-finger CCCTC-binding factor; *CYP20A1*: Cytochrome P450, family 20, subfamily A, polypeptide 1; DNMTs: DNA methyltransferases; DNMT3L: DNA methyltransferases 3 like; DSBs: Double-strand breaks; DUX4: Double homeobox protein 4; ELF5: E74-like factor 5 (ETA domain transcription factor); EOMES: Eomesodermin; ERVs: Endogenous retrovirus; ESCs: Embryonic stem cells; *Fendrr*: Fetal-lethal noncoding developmental regulatory RNA; GluR1: Glutamate receptor delta-1; HERV-H: Human endogenous retrovirus H;

HNRNPk: Heterogeneous nuclear ribonucleoprotein K; *HULC*: Hepatocellular carcinoma up-regulated long non-coding RNA; H3K4me1: Histone H3 lysine 4 monomethylation; H3K4me3: Trimethylation of lysine 4 on histone H3 protein subunit; H3K9me2/3: Dimethylation/trimethylation of lysine 9 on histone H3 protein subunit; H3K27ac: Histone H3 lysine 27 acetylation; H3K27me3: Histone H3 lysine 27 trimethylation; IAPs: Intracisternal A-type particles; IGF1R: Insulin-like growth factor 1 receptor; IFN: Interferon; IRF1: Interferon-regulatory factor 1; KAP1: KRAB-associated protein-1 (also known as TRIM28); *KLF2*: Krüppel-like factor 2; KRAB-ZFPs: Krüppel-associated box zinc finger proteins; KRAB-ZNF-KAP1: KRAB-zinc finger and KAP1 protein complex; *LEADER*: Long epithelial Alu-interacting differentiation-related RNA; *Linc-RoR*: Long intergenic non-protein-coding RNA, regulator of reprogramming; LINES: Long interspersed nuclear elements; lncRNAs: Long noncoding RNAs; LTRs: Long terminal repeats; L2b: LINE-2 b; MEL: Murine erythroleukemia cells; MERV: Murine endogenous retrovirus; miR: MicroRNA; miRNAs: MicroRNAs; MIRs: Mammalian-wide interspersed repeats; MIR205HG: MIR205 host gene microRNA; MLL: Mixed lineage leukemia protein; MLT1A/MLT1J: Mammalian LTR transposon 1 A/J; mRNA: Messenger RNA; MuERV-L: Murine endogenous retrovirus-leukemia protein; OCT4: Octamer-binding transcription factor 4; *PAX3*: Paired box gene 3; PCHI-C: Promoter capture Hi-C; piRNA: PIWI-interacting RNA; PIWI: P-element-induced wimpy testis protein; PNMA: Paraneoplastic Ma antigens protein; *PPAR γ* : Peroxisome proliferator-activated receptor gamma; PPP1R12A: Protein phosphatase 1 regulatory subunit 12A; PRC2: Polycomb repressive complex 2; PRKACB: Protein kinase cAMP-activated catalytic subunit beta; PSs: Pachytene spermatocytes; PTBP1: Polypyrimidine tract-binding protein 1; RAG1: Recombination activating 1 protein; RB1: Retinoblastoma protein 1; RNAi: RNA interference; RSSs: Recombination signal sequences; SAM: S-adenosyl methionine; SASPase/ASPRV1: Skin-specific retroviral-like aspartic protease/aspartic peptidase retroviral like 1 protein; SCAN: SRE-ZBP, CTfn-51, AW-1, 2 Number 18 cDNA protein; SEs: Superenhancers; SETDB1: SET domain bifurcated histone lysine methyltransferase 1; SINEs: Short interspersed nuclear elements; SMD: Staufen-mediated mRNA decay; SOX2: SRY (sex determining region Y)-box 2; TAD: Topological associating domain; TEs: Transposable elements; TRIM28: Tripartite motif-containing 28 protein; TSCs: Trophoblast stem cells; Uchl1/Park5: Ubiquitin carboxy-terminal hydrolase L1; UHRF1: Ubiquitin-like, containing PHD and RING finger domains protein 1; UPF1/2:

Up-frameshift suppressor 1/2; XCR: X chromosome reactivation; Xi: X chromosome inactivation; ZGA: Zygotic genome activation; Zscan4c: Zinc finger and SCAN domain containing protein 4C; *1/2-sbsRNA*: Half-STAU1-binding site RNA; 3'UTR: 3' Untranslated region; 5mC: 5-Methylcytosine.

Acknowledgements

The authors would like to express their sincere gratitude to Elva Chia-Chen Lu, Han-Chung Yeh, Yan-Han Lin and Hui-Ping Cheng for the critical discussions and some original writing material.

Authors' contributions

SPL conceived the original idea, set the scope and organized the study group and periodic discussions to develop and consolidate materials in this review manuscript. SHY contributed to the part related to cancer and immunology with assistance from CHY and NYS. CHY also drew the first draft of Fig. 1 which was further modified by PSH. PSH drew the other figures with assistance from SPL and JYC, commented by all co-authors. PSH, JYC, YTT, LKT, LCY and SPL wrote different parts of the original manuscript. PSH, SHY, and SPL led the many rounds of reconstruction and editing of the developing manuscript with major assistance from YTT and JYC. All authors were involved in editing and proofreading. All authors read and approved the final manuscript.

Funding

This study was supported by grants from the Ministry of Science and Technology, Taiwan to SPL (MOST 107-2313-B-002-054-MY3, 109-2311-B-002-024 and 110-2311-B-002-021-MY3), the Young Scholar Fellowship Program to SHY (MOST 109-2636-B-002-012), and grants from National Taiwan University to SPL (110L893305).

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author details

¹Institute of Biotechnology, College of Bio-Resources and Agriculture, National Taiwan University, Taipei, Taiwan. ²Department of Biomedical Engineering, Stony Brook University, Stony Brook, NY, USA. ³Department of Biomedical Engineering, National Taiwan University, Taipei, Taiwan. ⁴Agricultural Biotechnology Research Center, Academia Sinica, Taipei, Taiwan. ⁵Center of Systems Biology, National Taiwan University, Taipei, Taiwan. ⁶The Research Center of Developmental Biology and Regenerative Medicine, National Taiwan University, Taipei, Taiwan.

Received: 11 April 2021 Accepted: 19 July 2021

Published online: 07 August 2021

References

- Mc CB. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A*. 1950;36(6):344–55.
- Mills RE, et al. Which transposable elements are active in the human genome? *Trends Genet*. 2007;23(4):183–91.
- Kokosar J, Kordis D. Genesis and regulatory wiring of retroelement-derived domesticated genes: a phylogenomic perspective. *Mol Biol Evol*. 2013;30(5):1015–31.
- Franke V, et al. Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes. *Genome Res*. 2017;27(8):1384–94.
- Anwar SL, Wulaningsih W, Lehmann U. Transposable elements in human cancer: causes and consequences of deregulation. *Int J Mol Sci*. 2017;18(5):974.
- Liu J, et al. LINE-1 element insertion at the t(11;22) translocation breakpoint of a desmoplastic small round cell tumor. *Genes Chromosomes Cancer*. 1997;18(3):232–9.
- Daskalos A, et al. Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer. *Int J Cancer*. 2009;124(1):81–7.
- Molaro A, Malik HS. Hide and seek: how chromatin-based pathways silence retroelements in the mammalian germline. *Curr Opin Genet Dev*. 2016;37:51–8.
- Hancks DC, Kazazian HH Jr. Roles for retrotransposon insertions in human disease. *Mob DNA*. 2016;7:9.
- Amarasinghe SL, et al. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol*. 2020;21(1):30.
- Turelli P, et al. Primate-restricted KRAB zinc finger proteins and target retrotransposons control gene expression in human neurons. *Sci Adv*. 2020;6(35): eaba3200.
- Woodcock DM, et al. Asymmetric methylation in the hypermethylated CpG promoter region of the human L1 retrotransposon. *J Biol Chem*. 1997;272(12):7810–6.
- Walsh CP, Chaillet JR, Bestor TH. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet*. 1998;20(2):116–7.
- Martens JH, et al. The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *EMBO J*. 2005;24(4):800–12.
- Guelen L, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*. 2008;453(7197):948–51.
- Wallace MR, et al. A de novo Alu insertion results in neurofibromatosis type 1. *Nature*. 1991;353(6347):864–6.
- Miki Y, et al. Mutation analysis in the BRCA2 gene in primary breast cancers. *Nat Genet*. 1996;13(2):245–7.
- Lapp HE, Hunter RG. Early life exposures, neurodevelopmental disorders, and transposable elements. *Neurobiol Stress*. 2019;11: 100174.
- Tam OH, et al. Postmortem cortex samples identify distinct molecular subtypes of ALS: retrotransposon activation, oxidative stress, and activated glia. *Cell Rep*. 2019;29(5):1164–1177 e5.
- Liu EY, et al. Loss of nuclear TDP-43 is associated with decondensation of LINE retrotransposons. *Cell Rep*. 2019;27(5):1409–1421 e6.
- Thomas CA, et al. Modeling of TREX1-dependent autoimmune disease using human stem cells highlights L1 accumulation as a source of neuroinflammation. *Cell Stem Cell*. 2017;21(3):319–331 e8.
- Jang HS, et al. Transposable elements drive widespread expression of oncogenes in human cancers. *Nat Genet*. 2019;51(4):611–7.
- Gasior SL, et al. The human LINE-1 retrotransposon creates DNA double-strand breaks. *J Mol Biol*. 2006;357(5):1383–93.
- Gilbert N, Lutz-Prigge S, Moran JV. Genomic deletions created upon LINE-1 retrotransposition. *Cell*. 2002;110(3):315–25.
- Han K, et al. Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic Acids Res*. 2005;33(13):4040–52.
- Sorek R, Ast G, Graur D. Alu-containing exons are alternatively spliced. *Genome Res*. 2002;12(7):1060–7.
- Belancio VP, Hedges DJ, Deininger P. LINE-1 RNA splicing and influences on mammalian gene expression. *Nucleic Acids Res*. 2006;34(5):1512–21.
- Lev-Maor G, et al. Intronic Alu influence alternative splicing. *PLoS Genet*. 2008;4(9): e1000204.
- Teugels E, et al. De novo Alu element insertions targeted to a sequence common to the BRCA1 and BRCA2 genes. *Hum Mutat*. 2005;26(3):284.
- Miki Y, et al. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res*. 1992;52(3):643–5.
- Rodriguez-Martin C, et al. Familial retinoblastoma due to intronic LINE-1 insertion causes aberrant and noncanonical mRNA splicing of the RB1 gene. *J Hum Genet*. 2016;61(5):463–6.
- Park SY, et al. Alu and LINE-1 hypomethylation is associated with HER2 enriched subtype of breast cancer. *PLoS ONE*. 2014;9(6): e100429.

33. de Cubas AA, et al. DNA hypomethylation promotes transposable element expression and activation of immune signaling in renal cell cancer. *JCI Insight*. 2020. <https://doi.org/10.1172/jci.insight.137569>.
34. Kong Y, et al. Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. *Nat Commun*. 2019;10(1):5228.
35. Lee E, et al. Landscape of somatic retrotransposition in human cancers. *Science*. 2012;337(6097):967–71.
36. Voineagu I, et al. Replication stalling at unstable inverted repeats: interplay between DNA hairpins and fork stabilizing proteins. *Proc Natl Acad Sci U S A*. 2008;105(29):9936–41.
37. Lu S, et al. Short inverted repeats are hotspots for genetic instability: relevance to cancer genomes. *Cell Rep*. 2015;10(10):1674–80.
38. Wolff EM, et al. Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. *PLoS Genet*. 2010;6(4): e1000917.
39. Hur K, et al. Hypomethylation of long interspersed nuclear element-1 (LINE-1) leads to activation of proto-oncogenes in human colorectal cancer metastasis. *Gut*. 2014;63(4):635–46.
40. Ade C, Roy-Engel AM, Deininger PL. Alu elements: an intrinsic source of human genome instability. *Curr Opin Virol*. 2013;3(6):639–45.
41. Zhang W, et al. Alu distribution and mutation types of cancer genes. *BMC Genomics*. 2011;12:157.
42. Elliott B, Richardson C, Jasin M. Chromosomal translocation mechanisms at intronic alu elements in mammalian cells. *Mol Cell*. 2005;17(6):885–94.
43. Jeffs AR, et al. The BCR gene recombines preferentially with Alu elements in complex BCR-ABL translocations of chronic myeloid leukaemia. *Hum Mol Genet*. 1998;7(5):767–76.
44. Cui F, Sirotnin MV, Zhurkin VB. Impact of Alu repeats on the evolution of human p53 binding sites. *Biol Direct*. 2011;6:2.
45. Cruickshanks HA, et al. Senescent cells harbour features of the cancer epigenome. *Nat Cell Biol*. 2013;15(12):1495–506.
46. Yu YC, et al. Transient DNMT3L expression reinforces chromatin surveillance to halt senescence progression in mouse embryonic fibroblast. *Front Cell Dev Biol*. 2020;8:103.
47. Bourc'his D, Bestor TH. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature*. 2004;431(7004):96–9.
48. Zamudio N, et al. DNA methylation restrains transposons from adopting a chromatin signature permissive for meiotic recombination. *Genes Dev*. 2015;29(12):1256–70.
49. Robertson KD. DNA methylation and human disease. *Nat Rev Genet*. 2005;6(8):597–610.
50. Egger G, et al. Identification of DNMT1 (DNA methyltransferase 1) hypomorphs in somatic knockouts suggests an essential role for DNMT1 in cell survival. *Proc Natl Acad Sci U S A*. 2006;103(38):14080–5.
51. Robert MF, et al. DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells. *Nat Genet*. 2003;33(1):61–5.
52. Li Y, et al. Stella safeguards the oocyte methylome by preventing de novo methylation mediated by DNMT1. *Nature*. 2018;564(7734):136–40.
53. Okano M, et al. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*. 1999;99(3):247–57.
54. Chedin F. The DNMT3 family of mammalian de novo DNA methyltransferases. *Prog Mol Biol Transl Sci*. 2011;101:255–85.
55. Karetka MS, et al. Reconstitution and mechanism of the stimulation of de novo methylation by human DNMT3L. *J Biol Chem*. 2006;281(36):25893–902.
56. Liao HF, et al. Functions of DNA methyltransferase 3-like in germ cells and beyond. *Biol Cell*. 2012;104(10):571–87.
57. Zamudio N, Bourc'his D. Transposable elements in the mammalian germline: a comfortable niche or a deadly trap? *Heredity (Edinb)*. 2010;105(1):92–104.
58. Rollins RA, et al. Large-scale structure of genomic methylation patterns. *Genome Res*. 2006;16(2):157–63.
59. Zhou W, et al. DNA methylation enables transposable element-driven genome expansion. *Proc Natl Acad Sci U S A*. 2020;117(32):19359–66.
60. Bestor TH. DNA methylation: evolution of a bacterial immune function into a regulator of gene expression and genome structure in higher eukaryotes. *Philos Trans R Soc Lond B Biol Sci*. 1990;326(1235):179–87.
61. Rose NR, Klose RJ. Understanding the relationship between DNA methylation and histone lysine methylation. *Biochim Biophys Acta*. 2014;1839(12):1362–72.
62. Walter M, et al. An epigenetic switch ensures transposon repression upon dynamic loss of DNA methylation in embryonic stem cells. *Elife*. 2016. <https://doi.org/10.7554/eLife.11418>.
63. Maenohara S, et al. Role of UHRF1 in de novo DNA methylation in oocytes and maintenance methylation in preimplantation embryos. *PLoS Genet*. 2017;13(10): e1007042.
64. Liu X, et al. UHRF1 targets DNMT1 for DNA methylation through cooperative binding of hemi-methylated DNA and methylated H3K9. *Nat Commun*. 2013;4:1563.
65. Rothbart SB, et al. Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation. *Nat Struct Mol Biol*. 2012;19(11):1155–60.
66. Jacobs FM, et al. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature*. 2014;516(7530):242–5.
67. Karimi MM, et al. DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell Stem Cell*. 2011;8(6):676–87.
68. Schultz DC, et al. SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev*. 2002;16(8):919–32.
69. Quenneville S, et al. In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Mol Cell*. 2011;44(3):361–72.
70. Sasaki H, Matsui Y. Epigenetic events in mammalian germ-cell development: reprogramming and beyond. *Nat Rev Genet*. 2008;9(2):129–40.
71. Svoboda P, et al. RNAi and expression of retrotransposons MuERV-L and IAP in preimplantation mouse embryos. *Dev Biol*. 2004;269(1):276–85.
72. Kabayama Y, et al. Roles of MIWI, MILI and PLD6 in small RNA regulation in mouse growing oocytes. *Nucleic Acids Res*. 2017;45(9):5387–98.
73. Houwing S, et al. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell*. 2007;129(1):69–82.
74. Ku HY, Lin H. PIWI proteins and their interactors in piRNA biogenesis, germline development and gene expression. *Natl Sci Rev*. 2014;1(2):205–18.
75. Voronina E, et al. RNA granules in germ cells. *Cold Spring Harb Perspect Biol*. 2011. <https://doi.org/10.1101/cshperspect.a002774>.
76. Chang KW, et al. Stage-dependent piRNAs in chicken implicated roles in modulating male germ cell development. *BMC Genomics*. 2018;19(1):425.
77. Ernst C, Odom DT, Kutter C. The emergence of piRNAs against transposon invasion to preserve mammalian genome integrity. *Nat Commun*. 2017;8(1):1411.
78. Sienski G, Donertas D, Brennecke J. Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. *Cell*. 2012;151(5):964–80.
79. Watanabe T, et al. IWI2 targets RNAs transcribed from piRNA-dependent regions to drive DNA methylation in mouse prospermatogonia. *EMBO J*. 2018. <https://doi.org/10.15252/emboj.201695329>.
80. Zoch A, et al. SPOCD1 is an essential executor of piRNA-directed de novo DNA methylation. *Nature*. 2020;584(7822):635–9.
81. Brandt J, et al. Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. *Gene*. 2005;345(1):101–11.
82. Campillos M, et al. Computational characterization of multiple Gag-like human proteins. *Trends Genet*. 2006;22(11):585–9.
83. Emerson RO, Thomas JH. Gypsy and the birth of the SCAN domain. *J Virol*. 2011;85(22):12043–52.
84. Kapitonov VV, Jurka J. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol*. 2005;3(6): e181.
85. Nikolaienko O, et al. Arc protein: a flexible hub for synaptic plasticity and cognition. *Semin Cell Dev Biol*. 2018;77:33–42.
86. Matsui T, et al. SASPase regulates stratum corneum hydration through profilaggrin-to-filaggrin processing. *EMBO Mol Med*. 2011;3(6):320–33.

87. Pang SW, et al. PNMA family: protein interaction network and cell signalling pathways implicated in cancer and apoptosis. *Cell Signal*. 2018;45:54–62.
88. Edelman LC, Collins T. The SCAN domain family of zinc finger transcription factors. *Gene*. 2005;359:1–17.
89. Henke C, et al. Selective expression of sense and antisense transcripts of the sushi-ichi-related retrotransposon-derived family during mouse placentogenesis. *Retrovirology*. 2015;12:9.
90. Sundaram V, et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res*. 2014;24(12):1963–76.
91. Faulkner GJ, et al. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet*. 2009;41(5):563–71.
92. Liang D, et al. Genomic analysis revealed a convergent evolution of LINE-1 in coat color: a case study in water buffaloes (*Bubalus bubalis*). *Mol Biol Evol*. 2021;38(3):1122–36.
93. Medstrand P, Landry JR, Mager DL. Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. *J Biol Chem*. 2001;276(3):1896–903.
94. Goke J, et al. Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell*. 2015;16(2):135–41.
95. Fort A, et al. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat Genet*. 2014;46(6):558–66.
96. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet*. 2017;18(2):71–86.
97. Lynch VJ, et al. Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Rep*. 2015;10(4):551–61.
98. Chuong EB, et al. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat Genet*. 2013;45(3):325–9.
99. Kurnarso G, et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet*. 2010;42(7):631–4.
100. Spengler RM, Oakley CK, Davidson BL. Functional microRNAs and target sites are created by lineage-specific transposition. *Hum Mol Genet*. 2014;23(7):1783–93.
101. Borchert GM, et al. Comprehensive analysis of microRNA genomic loci identifies pervasive repetitive-element origins. *Mob Genet Elements*. 2011;1(1):8–17.
102. Piriyaopongsa J, Marino-Ramirez L, Jordan IK. Origin and evolution of human microRNAs from transposable elements. *Genetics*. 2007;176(2):1323–37.
103. Petri R, et al. LINE-2 transposable elements are a source of functional human microRNAs and target sites. *PLoS Genet*. 2019;15(3): e1008036.
104. Kang D, et al. TE composition of human long noncoding RNAs and their expression patterns in human tissues. *Genes Genomics*. 2015;37(1):87–95.
105. Kapusta A, et al. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet*. 2013;9(4): e1003470.
106. Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol*. 2012;13(11):R107.
107. Fort V, Khelifi G, Hussein SMI. Long non-coding RNAs and transposable elements: a functional relationship. *Biochim Biophys Acta Mol Cell Res*. 2021;1868(1): 118837.
108. Profumo V, et al. LEADeR role of miR-205 host gene as long noncoding RNA in prostate basal cell differentiation. *Nat Commun*. 2019;10(1):307.
109. Grote P, Herrmann BG. The long non-coding RNA Fendrr links epigenetic control mechanisms to gene regulatory networks in mammalian embryogenesis. *RNA Biol*. 2013;10(10):1579–85.
110. Grote P, et al. The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Dev Cell*. 2013;24(2):206–14.
111. Wang Y, et al. Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. *Dev Cell*. 2013;25(1):69–80.
112. Bhattacharya A, et al. Multiple Alu exonization in 3'UTR of a primate-specific isoform of CYP20A1 creates a potential miRNA sponge. *Genome Biol Evol*. 2021. <https://doi.org/10.1093/gbe/evaa233>.
113. Song W, et al. Long noncoding RNA BANCR mediates esophageal squamous cell carcinoma progression by regulating the IGF1R/Raf/MEK/ERK pathway via miR3383p. *Int J Mol Med*. 2020;46(4):1377–88.
114. Wang J, et al. CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. *Nucleic Acids Res*. 2010;38(16):5366–83.
115. Panzitt K, et al. Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA. *Gastroenterology*. 2007;132(1):330–42.
116. Toki N, et al. SINEUP long non-coding RNA acts via PTBP1 and HNRNPK to promote translational initiation assemblies. *Nucleic Acids Res*. 2020;48(20):11626–44.
117. Carrieri C, et al. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature*. 2012;491(7424):454–7.
118. Schein A, et al. Identification of antisense long noncoding RNAs that function as SINEUPs in human cells. *Sci Rep*. 2016;6:33605.
119. Indrieri A, et al. Synthetic long non-coding RNAs [SINEUPs] rescue defective gene expression in vivo. *Sci Rep*. 2016;6:27315.
120. Gowravaram M, et al. Insights into the assembly and architecture of a Staufen-mediated mRNA decay (SMD)-competent mRNP. *Nat Commun*. 2019;10(1):5054.
121. Gong C, Maquat LE. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature*. 2011;470(7333):284–8.
122. Gong C, et al. SMD and NMD are competitive pathways that contribute to myogenesis: effects on PAX3 and myogenin mRNAs. *Genes Dev*. 2009;23(1):54–66.
123. Cho H, et al. Staufen1-mediated mRNA decay functions in adipogenesis. *Mol Cell*. 2012;46(4):495–506.
124. Wang J, Gong C, Maquat LE. Control of myogenesis by rodent SINE-containing lncRNAs. *Genes Dev*. 2013;27(7):793–804.
125. Elisaphenko EA, et al. A dual origin of the Xist gene from a protein-coding gene and a set of transposable elements. *PLoS ONE*. 2008;3(6): e2521.
126. Pintacuda G, Young AN, Cerase A. Function by structure: spotlights on Xist long non-coding RNA. *Front Mol Biosci*. 2017;4:90.
127. Jacques PE, Jeyakani J, Bourque G. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet*. 2013;9(5): e1003504.
128. Wang T, et al. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc Natl Acad Sci U S A*. 2007;104(47):18613–8.
129. Wei CL, et al. A global map of p53 transcription-factor binding sites in the human genome. *Cell*. 2006;124(1):207–19.
130. Luo X, et al. 3D Genome of macaque fetal brain reveals evolutionary innovations during primate corticogenesis. *Cell*. 2021;184(3):723–740 e21.
131. Zhang W, et al. Zscan4c activates endogenous retrovirus MERVL and cleavage embryo genes. *Nucleic Acids Res*. 2019;47(16):8485–501.
132. Huang Y, et al. Stella modulates transcriptional and endogenous retrovirus programs during maternal-to-zygotic transition. *Elife*. 2017. <https://doi.org/10.7554/eLife.22345>.
133. Wu J, et al. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature*. 2016;534(7609):652–7.
134. Yang F, et al. DUX-miR-344-ZMYM2-mediated activation of MERVL LTRs induces a totipotent 2C-like state. *Cell Stem Cell*. 2020;26(2):234–250 e7.
135. Todd CD, et al. Functional evaluation of transposable elements as enhancers in mouse embryonic and trophoblast stem cells. *Elife*. 2019. <https://doi.org/10.7554/eLife.44344>.
136. Macfarlan TS, et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*. 2012;487(7405):57–63.
137. Sakashita A, et al. Endogenous retroviruses drive species-specific germline transcriptomes in mammals. *Nat Struct Mol Biol*. 2020;27(10):967–77.
138. Maezawa S, et al. Super-enhancer switching drives a burst in gene expression at the mitosis-to-meiosis transition. *Nat Struct Mol Biol*. 2020;27(10):978–88.
139. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science*. 2016;351(6277):1083–7.

140. Ye M, et al. Specific subfamilies of transposable elements contribute to different domains of T lymphocyte enhancers. *Proc Natl Acad Sci U S A*. 2020;117(14):7905–16.
141. Heintzman ND, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*. 2007;39(3):311–8.
142. Nishihara H. Transposable elements as genetic accelerators of evolution: contribution to genome size, gene regulatory network rewiring and morphological innovation. *Genes Genet Syst*. 2020;94(6):269–81.
143. Lunyak VV, et al. Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science*. 2007;317(5835):248–51.
144. Diehl AG, Ouyang N, Boyle AP. Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. *Nat Commun*. 2020;11(1):1796.
145. Wang J, et al. MIR retrotransposon sequences provide insulators to the human genome. *Proc Natl Acad Sci U S A*. 2015;112(32):E4428–37.
146. Lu JY, et al. Homotypic clustering of L1 and B1/Alu repeats compartmentalizes the 3D genome. *Cell Res*. 2021. <https://doi.org/10.1038/s41422-020-00466-6>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

