Published online 22 April 2022

Nucleic Acids Research, 2022, Vol. 50, No. 9 ***4959–4973***
https://doi.org/10.1093/nar/gkac265

Active endogenous retroviral elements in human pluripotent stem cells play a role in regulating host gene expression

Tianzhe Zhang[1], Ran Zheng[1], Mao Li[1], Chenchao Yan[1], Xianchun Lan[1], Bei Tong[2], Pei Lu[1] and Wei Jiang[1,3,4,*]

[1]Department of Biological Repositories, Frontier Science Center for Immunology and Metabolism, Medical Research Institute, Zhongnan Hospital of Wuhan University, Wuhan University, Wuhan 430071, China, [2]Department of Cardiology, Zhongnan Hospital of Wuhan University, Wuhan 430071, China, [3]Human Genetics Resource Preservation Center of Wuhan University, Wuhan 430071, China and [4]Hubei Provincial Key Laboratory of Developmentally Originated Disease, Wuhan 430071, China

Received October 26, 2021; Revised March 22, 2022; Editorial Decision March 31, 2022; Accepted April 01, 2022

ABSTRACT

Human endogenous retroviruses, also called LTR elements, can be bound by transcription factors and marked by different histone modifications in different biological contexts. Recently, individual LTR or certain subclasses of LTRs such as LTR7/HERVH and LTR5_Hs/HERVK families have been identified as *cis*-regulatory elements. However, there are still many LTR elements with unknown functions. Here, we dissected the landscape of histone modifications and regulatory map of LTRs by integrating 98 ChIP-seq data in human embryonic stem cells (ESCs), and annotated the active LTRs enriching enhancer/promoter-related histone marks. Notably, we found that MER57E3 functionally acted as proximal regulatory element to activate respective ZNF gene. Additionally, HERVK transcript could mainly function in nucleus to activate the adjacent genes. Since LTR5_Hs/LTR5 was bound by many early embryo-specific transcription factors, we further investigated the expression dynamics in different pluripotent states. LTR5_Hs/LTR5/HERVK exhibited higher expression level in naïve ESCs and extended pluripotent stem cells (EPSCs). Functionally, the LTR5_Hs/LTR5 with high activity could serve as a distal enhancer to regulate the host genes. Ultimately, our study not only provides a comprehensive regulatory map of LTRs in human ESCs, but also explores the regulatory models of MER57E3 and LTR5_Hs/LTR5 in host genome.

INTRODUCTION

LTRs contribute to ∼8% of the human genome and are derived from retrovirus infected of the germline. As dominant of transposable elements, the amplification of LTRs is considered to be harmful to genome stability (1), therefore, in most situation, LTRs are silenced by various repressive epigenetic mechanisms, such as DNA methylation, H3K9me3, PRC1/2 complex and piwi/piRNA pathway (2–6). LTRs are involved in many pathological processes, such as neurodegenerative diseases (7–9), cancer progression (10,11) and inflammatory bowel disease (12). In addition, LTRs are also involved in normal biological processes, illustrated by the recent report that species-specific LTRs shape its specific early embryonic development process and are essential for embryonic development (13). A few studies have shown that LTRs can enrich certain active histone modifications and transcription factors, and act as *cis*-regulatory elements to regulate host gene expression, such as serving as species-specific transcriptional insulators or providing transcription factor binding sites (14–16). During species evolution, all transposable elements accumulate mutations and gradually lose their transcription factor binding site motifs and thus reduce the ability of regulating host genes (17,18). In other words, the younger the evolutionary age is, the transposon would be closer to its full length and the less mutations are accumulated, and with a stronger the regulatory ability of transposable elements would be.

It seems unavoidable that LTRs can be reactivated along with the process of epigenetic remodeling, especially in the early embryo and pluripotent stem cells (19–21). For example, HERVL and MERVL elements activate a subset of cleavage stage genes through DUX4/mDux in early embryo development (22,23). In mouse ESCs, MERVL can mark a rare transient cell population with high levels of 2-cell-specific transcripts (24), and is broadly used as a reporter

*To whom correspondence should be addressed. Tel: +86 27 68750208; Fax: +86 27 68759675; Email: jiangw.mri@whu.edu.cn

of 2-cell-like pluripotent cells (25,26). In human pluripotent stem cells, most studies focus on younger LTRs such as LTR7/HERVH and LTR5_Hs/HERVK. LTR7/HERVH transcript is located in nucleus and could function as an lncRNA to determine human ESC identity (27), or establish pluripotency-specific topologically associating domain (TAD) boundary in ESCs and ultimately is critical in maintaining the pluripotency of stem cells (28). LTR5_Hs/HERVK is firstly found to be reactivated and contribute to the innate immunity in human early embryos and pluripotent cells (20). Recently, Pontis and colleagues reported that LTR5_Hs/HERVK could activate blastocyst-specific genes in human naïve pluripotent stem cells and found that the activated ZNFs may feedback regulate the retrotransposon activity (29). Despite these advances, a systematic understanding of the regulatory landscape of LTRs in early embryos or human pluripotent stem cells is still confused and there are still many functional LTRs to be dissected.

In this study, we are aiming to identify new potential functional LTRs in human pluripotent stem cells. We first drew a systematic histone modification landscape of LTRs in human ESCs based on ENCODE dataset (30). We found some LTRs exhibited similar histone modification characteristics to promoter or enhancer. Of note, MER57E3 is mainly located in downstream of the ZNF genes transcription start site (TSS); LTR5_Hs/LTR5/HERVK belongs to the youngest LTRs and still retains the characteristics of provirus and can generate *HERVK* transcripts. To further identify the potential role of LTRs, we undertook an inducible CRISPRi system in human ESCs to suppress the activity of LTRs of interest. We also explored the upstream regulation of these LTRs subclasses through motif analysis and verified it by luciferase reporter assay. Importantly, the activity of LTR5_Hs/LTR5 was correlated with pluripotent states, so we further dissected the role of LTR5_Hs/LTR5 in conventional human ESCs and newly derived feeder-free extended pluripotent stem cells (ffEPSCs).

MATERIALS AND METHODS

Cell cultures and conversion of ffEPSC

Primed human ESC lines HUES8 and H9 were plated and expanded with mTeSR1 medium (Stemcell Technologies, # 1000023391) in Matrigel-coated 6-well plates (MATRIGEL MATRIX HESC-QUALIFIED, BD Bioscience, # 354277). Culture medium was refreshed daily, and the cells were passaged with accutase (Stemcell Technologies, # A1110501) every 4–5 days. The maintenance and conversion of ffEPSCs from ESCs were previously described (31). HEK293T cells were cultured with DMEM containing 10% FBS and 1% penicillin-streptomycin.

ChIP-seq and ChIP-exo data analysis

ChIP-seq and ChIP-exo data for histone modifications, transcription factors, chromatin modifiers were downloaded from ENCODE Project database (https://www.encodeproject.org/). Raw data was filtered by trim_galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) to generate the clean data with the parameter:-q 30, and the reads were aligned to the human genome (hg38) using bowtie2 (32) with the parameters: -p 64 –very-sensitive –end-to-end –no-unal. Reads mapped to the mitochondrial genome were removed using samtools (33), and only the best alignments were kept while multimapped reads were randomly retained once. PCR duplicates were removed using Picard MarkDuplicates (http://broadinstitute.github.io/picard/). The bam alignment files of the same histone modification transcription factors and chromatin modifiers were merged by samtools merge function. The merge bam files were transformed into normalized RPKM (reads per kilobase per million mapped reads) bigwig files and Pearson correlation was calculated using deepTools (34).

The annotations of LTR elements were obtained from UCSC Genome Browser RepeatMasker. We filtered out annotations < 300 bp, and then size-matched annotations corresponding to each LTR elements were generated by bedtools random functions. For LTR elements enrichment analysis, we calculated the total counts of each family element in the annotations of LTR elements and random annotations, and calculated their $\log_2$(fold-change) ratio. The coverage signals in LTR elements were generated by deeptools. For individual LTR elements visual track views are based on uniquely mapped reads.

ChIP-exo data of ZNF730 was downloaded from GSE78099 (35) and the binding peaks of ZNF730 were called by MACS2 (36) with the threshold: $q < 0.0001$ and fold-change > 4.

Transcriptome sequencing and analysis

Total RNA was isolated using HiPure total RNA mini kit (Magen, # R4111-03), and RNA sequencing was performed by illumina novaseq 6000 PE150.

For RNA-seq analysis, the reads were aligned to the human genome (hg38) using hisat2 (37) with the default parameters. For LTR elements, only the best alignments were kept while multimapped reads were randomly retained once. Bedtools (38) was used for the counting of individual LTR sites and only uniquely mapped reads were kept for individual LTR. For the count of other genes, we used featureCounts (39) to generate raw counts, and the TPM (transcript per million) of protein-coding genes and the CPM (counts per million) of LTR elements were calculated. Differential gene expression analysis was performed using DESeq2. Genes with TPM < 1 in all samples were filtered out, and the remaining genes with abs|$\log_2$(fold-change)| > 1, *P*-adjust < 0.05 were defined as differentially expressed genes.

Motif enrichment

We used the random bed file of LTR5_Hs/LTR5 and MER57E3 to generate the background sequence of the corresponding position, and then used the findMotifs tool in HOMER (40) to calculate the enrichment level of motif.

Principal component analysis

Principal component analysis was performed using prcomp function in the R stats package. Covariance matrix was ex-

tracted to detect the contribution of histone modifications to LTR elements.

Generation of inducible CRISPRi system

The Gen1 (pAAVS1-NDi-CRISPRi, a gift from Bruce Conklin, Addgene plasmid # 73497) and pX459 (pSpCas9(BB)-2A-Puro, a gift from Feng Zhang, Addgene plasmid # 48139) containing sgRNA targeting AAVS1 locus were electroporated into HUES8 cells using Nucleofector (Lonza) and cells were selected with 1 μg/μl G418. After 14 days of selection, clonal lines were generated by dilution in 96-well plates, and further selected based on doxycycline induction of mCherry expression.

For the repression of MER57E3 activity, sgRNA was designed by CRISPOR (http://crispor.tefor.net/), and cloned to the LentiGuide-Puro vector (a gift from Feng Zhang, Addgene plasmid # 52963). As for LTR5_Hs/LTR5, 6 sgRNAs were selected from previous reports, and sgRNA pool was cloned to the LentiGuide-Puro plasmid backbone. All sgRNA sequences were provided in Supplementary Table S5. The above plasmids were co-transfected into HEK293T cells together with psPAX2 (a gift from Didier Trono, Addgene plasmid # 12260), pMD2.G (a gift from Didier Trono, Addgene plasmid # 12259) to generate lentivirus. Concentrated lentiviral particles were added to the culture medium of KRAB-dCas9 HUES8 line. After 4 days of selection with 2 ug/ml puromycin, stable cell line was established.

Generation of CRISPR knock out cell line

The two sgRNAs targeting ZNF730-MER57E3 were cloned into pX459. The plasmids were electroporated into HUES8 cells using Nucleofector (Lonza). Clonal lines were generated by dilution in 96-well plates. Genotype identification primer and sgRNA sequences are shown in Supplementary files.

Generation of overexpression cell line

The coding sequence of TEF was cloned into PB-CAG-BGHpA vector (a gift from Xiaohua Shen, Addgene plasmid # 92161). The plasmid was electroporated into HUES8 cells using Nucleofector (Lonza). Clonal lines were selected with 250 ug/ml Hygromycin B.

Dual-luciferase reporter assays

MER57E3 sequence and mini promoter were cloned into pGL4.17 vector while GFP-coding sequence served as a control, and then the TEF coding sequence was cloned into pCMV vector. 1×10^5 HEK293T cells were seeded into each well of a tissue culture-treated 24-well plate. Firefly and Renilla (Control background) luciferase and TEF were co-transfected into HEK293T cells through LIPO-plus (Sage, # Q03004). The luciferase luminescence was measured after 48 hours. The ratio of Firefly/Renella (Nluc/Fluc) was calculated and normalized to the negative control.

Nuclear and cytoplasmic extract

Approximately 1×10^7 cells were collected and incubated on ice for 30 minutes in CE buffer (10 mM HEPES, 60 mM KCl, 1 mM EDTA, 0.075% NP40, 1 mM DTT and 1 mM PMSF, adjusted to pH 7.6). After centrifuge for 15 minutes at 4°C, the precipitate and supernatant were separated as the nucleus and cytoplasm respectively, to prepare for RNA extraction.

The RT-qPCR data is converted into cytoplasmic-nuclear relative concentration index (CN-RCI), using the following formula:

$$\mathrm{RCI} = \log_2\left(\frac{\text{Cytoplasmic expression } 2\text{^}(-\text{cq})}{\text{Nuclear expression } 2\text{^}(-\text{cq})}\right)$$

RT-qPCR

Total RNA was reversely transcribed into cDNA using cDNA Synthesis SuperMix (Bimake, # B24408). qPCR reactions were performed with the SYBR Green qPCR Master Mix (Biomake, # B21203) on the CFX384 Touch Real-Time PCR Detection System (All primers used in qPCR are shown in Supplementary Table S5).

Western blot

RIPA buffer (10 mM Tris–HCl, pH 8.0, 150 mM NaCl, 1% NP-40, 0.1% sodium deoxycholate) with protease inhibitors was used to lyse cells. Protein lysates were loaded on SDS-PAGE gels. After transfer to nitrocellulose membranes, membranes were blocked with TBST (TBS with 0.1% Tween-20) and 5% milk and then incubated with antibodies overnight (all antibodies used in western blot are shown in Supplementary Table S5). Signals were detected using HRP-conjugated secondary antibodies and Western Lightning Plus-ECL.

Immunofluorescence microscopy

Cells were grown on Matrigel-coated plate and fixed using 4% PFA for 30 minutes at room temperature followed by blocking and permeablizing with 10% donkey serum, 0.3% Triton-X 100 in PBS (antibody buffer) supplemented with 10% serum for species-matched secondary antibody. Primary antibodies (1:200) were resuspended in antibody buffer and incubated at 4°C overnight. After wash, secondary antibodies (1:200) were added in the dark (All antibodies used in Immunofluorescence are shown in Supplementary Table S5). Finally, the cells were incubated with DAPI for 10 minutes and then imaged on fluorescence microscope (Olympus).

RESULTS

More than 10% LTR elements are marked with active histone modifications in human ESCs

In order to profile the landscape of histone modifications and transcription factors enrichment on LTR elements in

human ESCs, we downloaded and reanalyzed the ChIP-seq data of H1 cell line from ENCODE Project (30). For the different dataset about the same modifications or transcription factors, we merged the clean data after reads-mapping and removal of PCR duplicates. Given that the sequence of LTRs is conservative, and most of the obtained ChIP-seq data are from single-end sequencing with length generally less than 100 bp, we applied the high score alignment and randomly assigned to an LTR element in order to get a more realistic enrichment, while the uniquely mapped reads were used for the visualization of ChIP-seq signal of LTRs (41). Then, we generated a random size-matched genome region for each LTR, and calculated the enrichment score using $\log_2$(ratio) (the total counts of each family element over the genomic size of respective random region) to characterize the enrichment of histone modifications on LTRs (41). We also calculated the Pearson correlation coefficients of different ChIP-seq samples on all LTR elements, and the regulation/modifications with similar annotation were indeed clustered together, such as the maintenance of TAD boundaries (CTCF and RAD21), the maintenance of heterochromatin (CBX5 and H3K9me3), enhancer/promoter characters (H3K4me1/2, H3K27ac and POLR2A) and others (Supplementary Figure S1A), suggesting the reliability of this analysis. Importantly, we found that more than one quarter of LTR families (163/584) exhibited at least one enriched modification in human ESCs. Since H3K9me3 is the mark of heterochromatin and is considered as the most common silencing mechanism of transposon elements (3,4), in our analysis we indeed found nearly one tenth of LTR families (54/584) were significantly enriched by H3K9me3 signal, comparable with the result in mouse ESCs (41). Actually, the LTR elements like MER9a2 occupied by H3K9me3 were hardly enriched by other histone modifications (Figure 1A and D) and with pool enrichment of DNaseI (Figure 1E), representing a subset of silent LTRs. In addition, almost no LTR elements were significantly enriched by H3K27me3, H3K36me3, H4K20me1 and H3K79me1/2 (Figure 1A).

In order to explore the role of LTR elements in human ESCs, we performed principal component analysis, showing these LTRs distributed discretely (Figure 1B). We focused on the histone modifications with strongest contributions to PC1 and PC2, respectively. PC1 was composed of enhancer-related modifications such as H3K4me1/2 and H3K27ac while PC2 contained promoter-related H3K4me2/3 and H3K27ac (Figure 1C). These results suggest those active LTR elements may act as *cis*-regulator in human ESCs. This comprehensive landscape revealed the regulatory potential of LTRs in human ESCs. We noticed that MER57E3, LTR5_Hs/LTR5/HERVK, LTR7/HERVH were highly enriched in active histone modifications in human ESCs. Among them, MER57E3 was significantly distinguished by promoter-related modifications, while LTR5_Hs/LTR5 and LTR7 were marked with enhancer-related modifications (Figure 1B, D and E). Since LTR7 has been extensively studied in previous reports (27–28,42–44), hence we next focused on the characterization of MER57E3 and LTR5_Hs/LTR5 in this study.

MER57E3 is bound and regulated by transcription factor PAR bZIP family

As an ancient LTR, MER57E3 was previously identified as active in acute myeloid leukemia (45). In addition, recent reports found MER57E3 were significantly activated in human pachytene spermatocytes and they suggested the evolutionary or regulatory links between ZNF and MER57E3 because of their location association, but its function is yet largely unknown (46,47). To gain insights into the functional mode of MER57E3, we first looked into the genomic location and found MER57E3 is usually located non-randomly at about 1 kb downstream of the TSS of certain coding genes. Given that this location might cause the spread of histone modification signal from the TSS, particularly H3K4me3 (48), we defined the active and silent loci of MER57E3 by chromatin accessibility (Supplementary Figure S2A), and counted the length and Smith-Waterman alignment score (swScores, to describe the level of conservation with the MER57E3 sequence in Dfam database) (49,50) of MER57E3. Interestingly, we noticed that MER57E3 closer to full length and with higher swScores exhibited higher chromatin accessibility (Figure 2A). In addition, we found that MER57E3 in the promoter region is with higher integrity and fewer mutations (Figure 2B) compared to the other MER57E3, indicating an active role because the integrity is one of the key conditions for functional repetitive elements (51). One example was shown for TSS-related MER57E3 and TSS-unrelated MER57E3 respectively (Figure 2C). We found that most of MER57E3 elements are associated with gene regulation function and most interestingly, the majority are related to genes encoding ZNF protein family (Figure 2D). Certain ZNF proteins have been reported that can recruit TRIM28 to establish heterochromatin at its binding site (52), or interact with some transcriptional activators to regulate target genes (53–56). Therefore, MER57E3 might be involved in regulating the expression of ZNF genes to regulate host gene expression.

We further investigated the upstream regulation of MER57E3 elements by transcription factor binding motif analysis. The most significantly enriched motif is the PAR bZIP family, including HLF, TEF and DBP (Figure 2E). Previous studies have shown that both HLF and TEF can recognize the extremely similar motif in the hematopoietic cells of leukemia and play an important role in the proliferation of hematolymphoid progenitors (57,58). In addition, another study performed the ChIP-seq experiment of HLF, TEF, DBP in HepG2 cells (30), therefore we downloaded the raw data and reanalyzed the binding on MER57E3 elements. As expected, HLF and TEF had a significant binding peak on MER57E3, and the peak summit was directly located in MER57E3 (Supplementary Figure S2B), illustrated by visualized location of two MER57E3 sites (Supplementary Figure S2C). From RNA-seq dataset, we found TEF exhibited the highest expression level in human ESCs (Figure 2F). To further demonstrate the regulation of TEF over MER57E3 elements, we cloned the sequences of two individual MER57E3 sites into a luciferase report vector, and co-transfected with TEF coding sequence as well. The results showed that MER57E3 sequence could strongly ac-

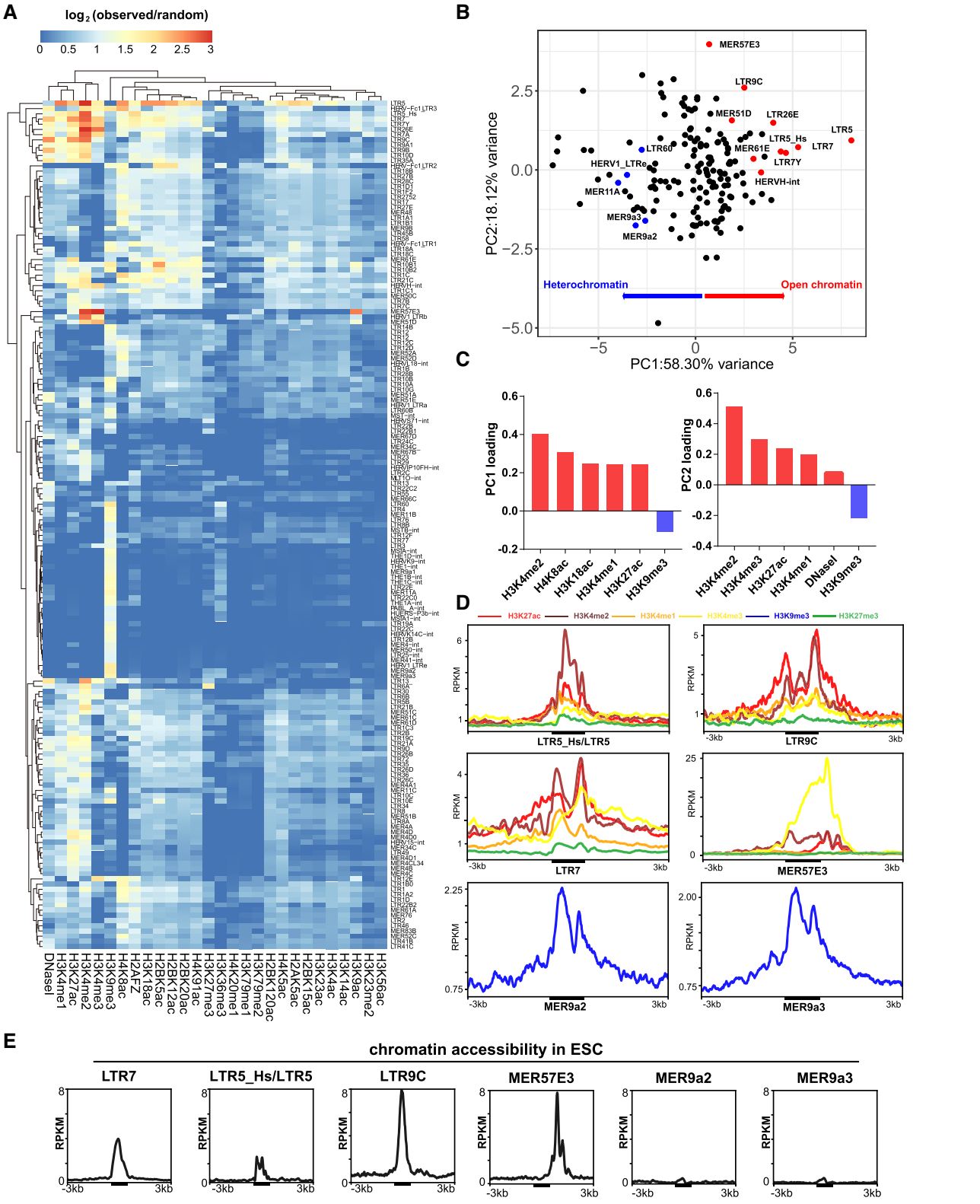

Figure 1. Identification of LTRs with regulatory potential in human ESCs. (**A**) Heatmap of the $\log_2$(observed/random) enrichment for DNaseI and histone modifications in at least one modified LTRs ($\log_2$(observed/random)>1). (**B**) Principal component analysis of enrichment of histone modifications and DNaseI in all LTRs. Red (active): high H3K27ac, or H3K4me1/2/3. Blue (inactive): high H3K9me3. (**C**) Principal component top loadings on PC1 and PC2, respectively. (**D**) Plot of average ChIP-seq signal profiles for selected LTRs within active and inactive clusters. (**E**) Plot of average DNaseI-seq signal profiles for selected LTRs within active and inactive clusters defined in (D).

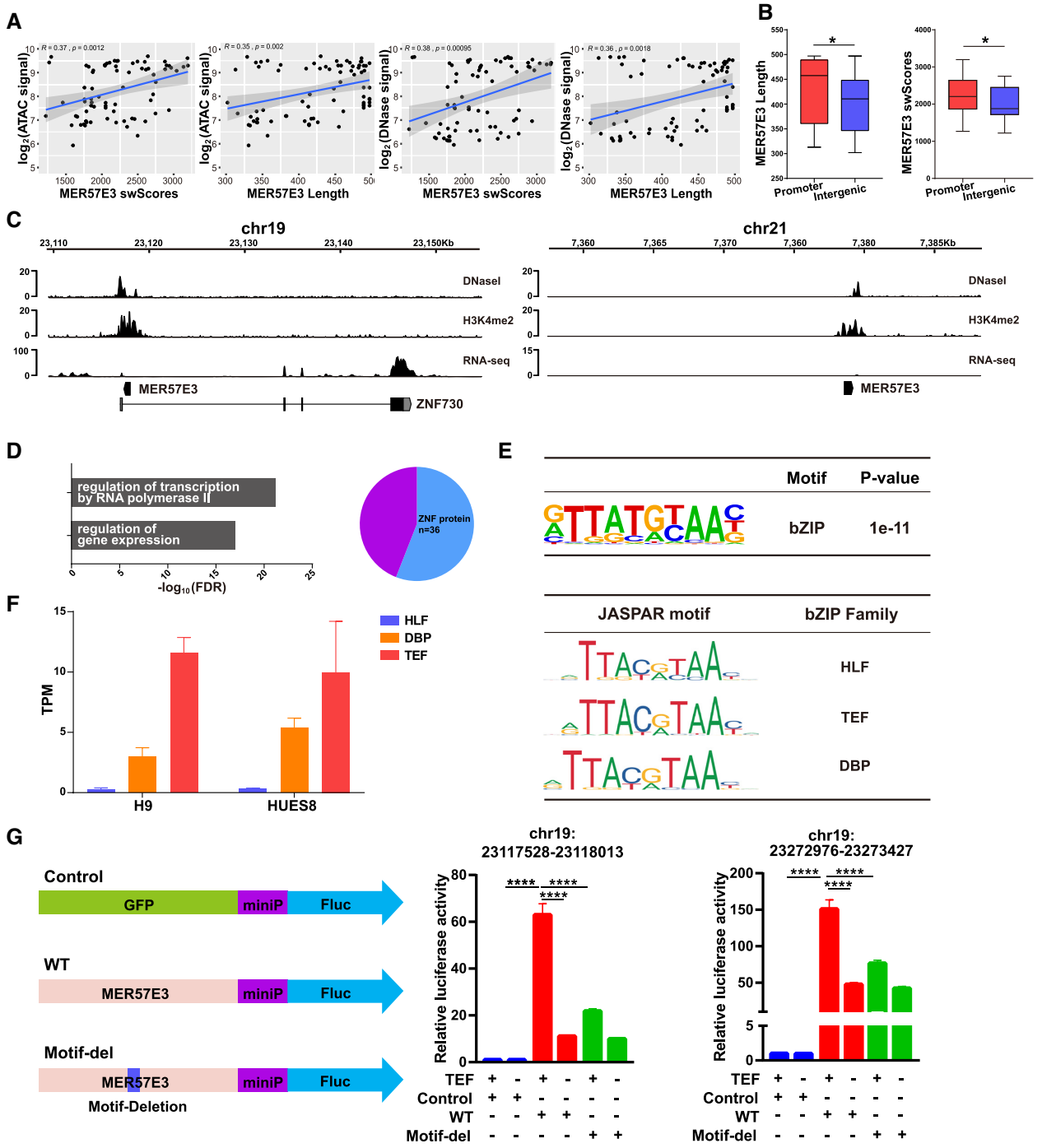

Figure 2. TEF binds on MER57E3 to drive the expression of ZNF genes. (**A**) DNase-seq/ATAC-seq signal within 2kb up/downstream of MER57E3; Pearson correlation generated by DNase-seq/ATAC-seq signals and swScores or MER57E3 length. (**B**) Box plot showing Smith-Waterman alignment score and MER57E3 length of promoter/non-promoter MER57E3 subtypes. Error bars represent mean ± s.e.m. $^{*}P < 0.05$; unpaired Student's *t*-test. (**C**) Track view of the Promoter related MER57E3 and non-promoter related MER57E3 in human ESC. (**D**) Left: bar plot showing the statistically significant GO terms for genes adjacent to MER57E3. Right: pie chart showing the proportion of ZNF genes adjacent to MER57E3. (**E**) Motif enrichment analysis of MER57E3 for putative TF-binding sites. (**F**) PAR bZIP genes TPM expression in human ESC line H9 and HUES8. (**G**) Dual-luciferase reporter assays. Relative fold changes in luciferase activity were shown. Error bars represent mean ± s.e.m. $^{****}P < 0.0001$; unpaired Student's *t*-test. At least three biological replicates were examined.

tivate the luciferase signal upon TEF overexpression (Figure 2G). More importantly, the mutated MER57E3 version with the bZIP family motif deletion exhibited much lower transcription activity even with TEF co-transfection (Figure 2G), supporting the view that the MER57E3 element functioned through the TEF protein. Furthermore, these data together illustrate that MER57E3 has the potential to act as a proximal regulatory element and then regulate host gene expression.

ZNF genes adjacent to MER57E3 exhibit a comprehensive effect on the regulation of LTR elements and host genes

We annotated 84 MER57E3 elements over 300bp in length in the human genome, of which 47 were located near the promoter region. Among them, 36 MER57E3 elements are annotated in the first intron region of the ZNF genes (Figure 2C and D). These MER57E3-related ZNF genes exhibited a high expression level in the ESC line HUES8, and were broadly expressed from 8-cell stage of the early embryo and maintained high expression in the pluripotent stage (Figure 3A).

We collected the publicly available ChIP-exo data of the above-mentioned 12 ZNF genes (35). We found 11 of these ZNF proteins have a strong enrichment signal on different classes of LTR elements (Supplementary Figure S3A, Supplementary Table S4). Of note, we annotated the 1405 bound peaks of ZNF730 which exhibited the highest expression level in human ESCs (Figure 3B), and found that 10% of the peaks bound in the promoter region (Supplementary Figure S3B). Taken together, these data indicated that these MER57E3-associated ZNF genes might regulate host gene transcription level and other LTR elements as well.

MER57E3 elements mediate the transcription of ZNF genes by proximal regulatory function

Next, to validate whether these MER57E3 elements indeed mediate the transcription of adjacent ZNF genes in human pluripotent stem cells, we constructed a CRISPRi system in ESC line HUES8 (Figure 3C) (59). We confirmed the success of transfection by doxycycline-induced expression of mCherry (Supplementary Figure S4A) and dCas9 protein (Supplementary Figure S4B). We first designed two uniquely matched sgRNAs targeting the MER57E3-ZNF730 locus (892bp and 934bp downstream from the TSS of ZNF730, respectively) (Figure 3D). After 7 days of treatment with doxycycline, both sgRNAs resulted in a 55–75% reduction in mRNA level of ZNF730 detected by RT-qPCR (Figure 3D). To further verify the universal role of MER57E3 as a proximal regulatory element, we selected another MER57E3-associated ZNF gene, ZNF678, for CRISPRi experiment (Figure 3D). We designed two sgRNA sequences (588bp and 613bp downstream from the TSS of ZNF678, respectively) and both sgRNAs of ZNF678-MER57E3 achieved a 75–80% reduction in mRNA level of ZNF678 (Figure 3D).

In order to explore the broad effect of the MER57E3-ZNF targets, we performed the RNA-seq of sgMER57E3-ZNF730. We calculated the expression of 515 ZNF genes, and found only ZNF730 was down-regulated, suggesting the specificity of individual MER57E3-mediated regulation (Supplementary Figure S4D). Meanwhile, according to the genome-wide differential expression analysis, we identified 11 down-regulated genes, of which 5 having the binding site at the TSS of ZNF730 according to the previous dataset (Supplementary Figure S4E and F).

Previous reporter assay indicated TEF was responsible to MER57E3-mediated regulation of respective ZNF genes. To further verify that the transcription factor TEF can directly activate MER57E3-related genes, we over-expressed TEF in human ESCs (Figure 3E), and then we examined the expression of several MER57E3-related genes. Compared with EmptyVector control, TEF-overexpression significantly upregulated the expression of these MER57E3-related genes (Figure 3F). Since the CRISPRi by targeting MER57E3 element might also cause the spreading of repression to the TSS, we further knocked out the ZNF730-MER57E3 element. Using the generated homozygote and heterozygote clones with the loss of ZNF730-MER57E3 (Supplementary Figure S4C), we next examined the mRNA level of ZNF730 and found ZNF730 was significantly downregulated after knocking out MER57E3 (Figure 3G). Using the ZNF730-MER57E3 knockout ESC line, we again performed TEF overexpression experiment (Figure 3H). The upregulation of ZNF730 observed in wild-type ESCs upon overexpressing TEF was not detected any more in the ZNF730-MER57E3 knockout ESC line (Figure 3I). Taken together, our results suggest that the transcription factor TEF interacts with MER57E3 and regulates the adjacent gene expression.

LTR5_Hs/LTR5 /HERVK is epigenetically regulated and exhibits moderate *cis*-regulatory ability in human ESCs

The LTR5_Hs/LTR5 element is the youngest family of LTR elements in human and could produce HERVK provirus transcripts and proteins (60). In previous reports (20,61), HERVK has been indicated associated with the pluripotency. Thenussein *et al.* reported that the HERVK-related LTR5 and LTR5_Hs, recognized as a marker of naive pluripotency, were more active in naive pluripotent stem cells and drove the expression of genes in early embryos (62), herein we surveyed the enrichment of pluripotency-related transcription factors on LTR5_Hs/LTR5 based on the available ChIP-seq data. We observed significant enrichment of OCT4, SOX2, KLF4, MYC, NANOG and EP300 in LTR5_Hs/LTR5 loci (Figure 4A). In conventional human ESCs, LTR5_Hs/LTR5 was mainly enriched with H3K27ac and H3K4me1/2, without H3K4me3 (Figure 1A). We separately detected the solo LTR5_Hs/LTR5 or proviral LTR5_Hs/LTR5 (HERVK) and found no difference in terms of H3K27ac signal enrichment (Figure 4B), which is consistent with previous reports that the expression of HERVK in human ESCs is at a moderate level (20). HERVK transcripts can be transferred to the cytoplasm and translated into proteins that can make up virus-like particles (20), so we separated the nuclear and cytoplasmic fractions and compared the transcript abundance by qRT-PCR experiments in three pluripotent stem cell lines. Intriguingly, we found that the HERVK transcripts were mainly located

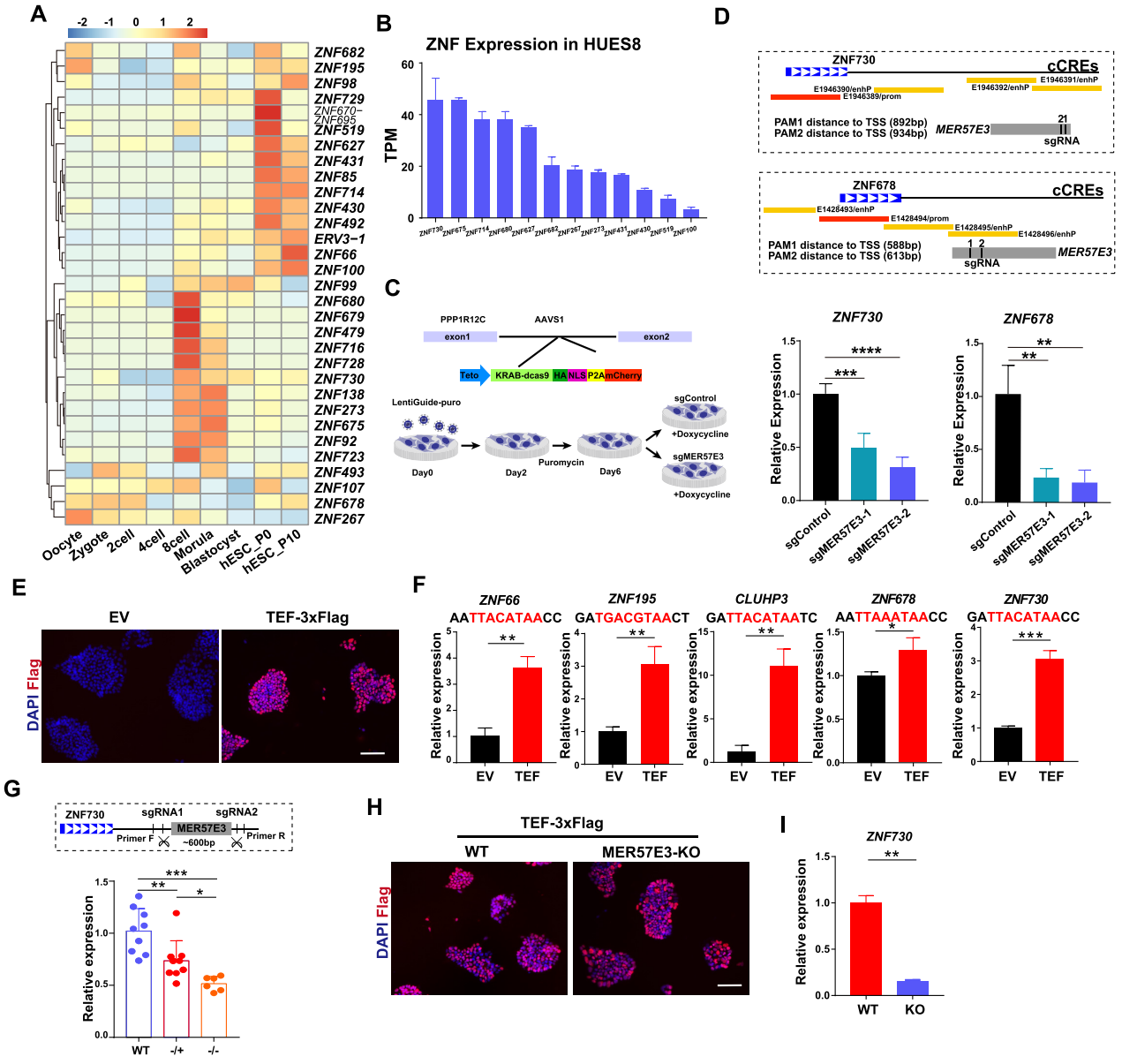

Figure 3. MER57E3 regulates ZNF gene expression in human ESCs. (**A**) Heatmap of hierarchical clustering (Euclidean distance) using *Z*-score of scaled TPM of MER57E3-related ZNF gene expression in early embryonic development. (**B**) Bar plot showing the TPM expression of MER57E3-related ZNF genes in HUES8 cell line. (**C**) Diagram of CRISPRi cell line construction. (**D**) Top: Schematic diagram of the positional relationship between cCREs (candidate *cis*-Regulatory Elements), MER57E3 and CRISPRi experiment sgRNA. Bottom: RT-qPCR analysis of ZNF730 and ZNF678 expression in HUES8 cells induced with dCas9-KRAB (CRISPRi). Error bars represent mean ± s.e.m. $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$; unpaired Student's *t*-test. At least three biological replicates were examined. (**E**) Immunostaining for the TEF overexpression experiment in HUES8 cell line. Scale bars = 100 μm. (**F**) RT-qPCR analysis of MER57E3-related genes. Error bars represent mean ± s.e.m. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$; unpaired Student *t*-test. At least three biological replicates were examined. (**G**) RT-qPCR analysis of ZNF730 expression in MER57E3 knockout cell lines. Error bars represent mean ± s.e.m. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$; unpaired Student's *t*-test. At least three biological replicates were examined. (**H**) Immunostaining for the TEF overexpression experiment in WT and ZNF730-MER57E3 knockout HUES8 cell line. (**I**) RT-qPCR analysis of ZNF730 expression in WT/KO TEF overexpression cell lines. Error bars represent mean ± s.e.m. $^{**}P < 0.01$; unpaired Student's *t*-test. At least three biological replicates were examined.

in the nucleus (Figure 4C), implying that the HERVK transcript might play a role in regulating chromatin accessibility in human pluripotent stem cells.

In order to verify the *cis*-regulatory ability of LTR5_Hs/LTR5 in primed ESCs, we used a previously reported strategy to simultaneously incorporate 6 consecutive sgRNAs in CRISPRi system (63,64). We adapted this system in the KRAB-dCas9 HUES8 cell line (Figure 4D). After 7 days of treatment with doxycycline, the transcription level of *HERVK* decreased by about 70%, and the RNA expression of the pluripotency marker OCT4, NANOG and SOX2 did not change (Figure 4E). The immunofluorescence staining also displayed the unaltered protein levels of OCT4 and NANOG (Figure 4F).

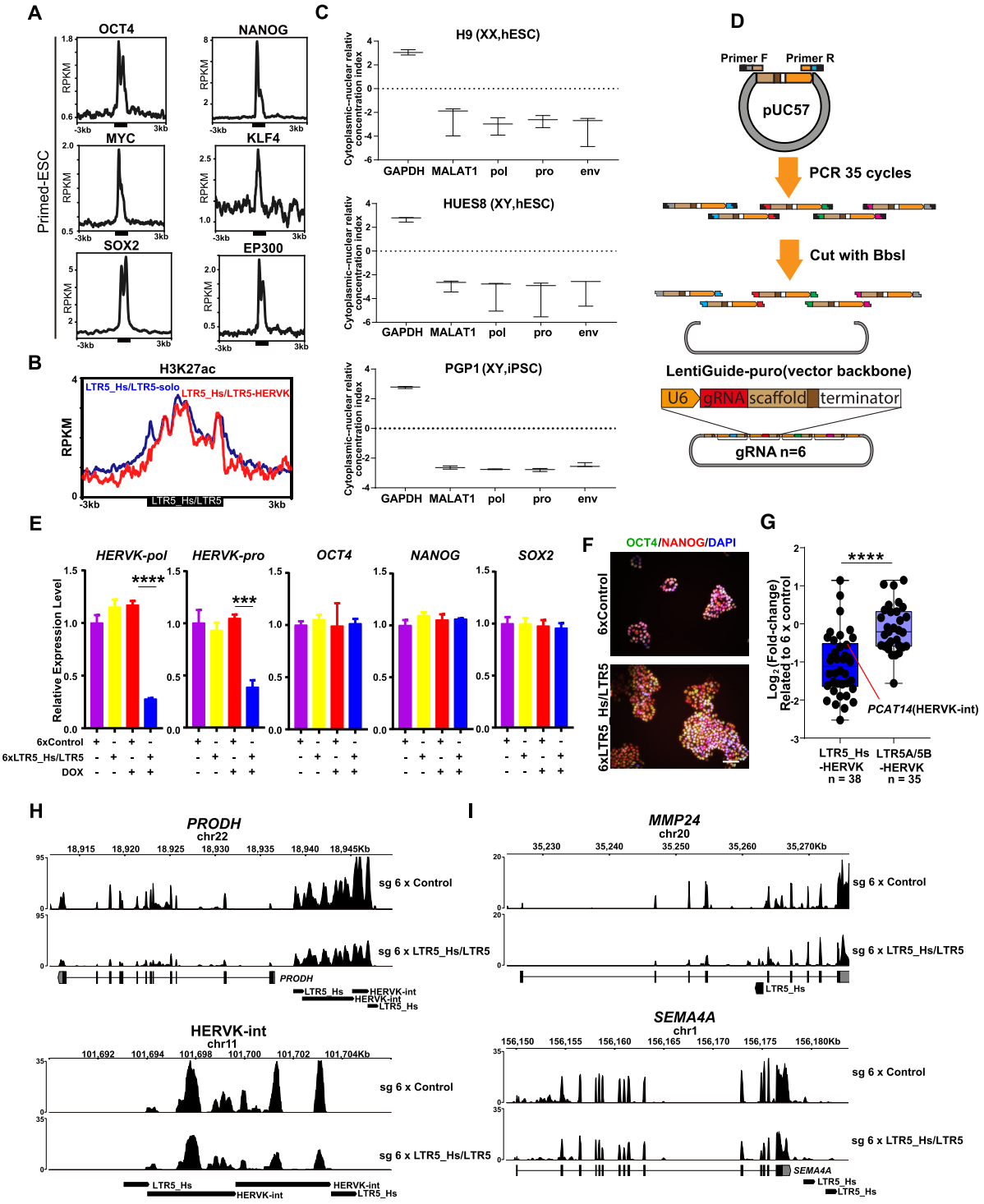

Figure 4. Identification of LTR5_Hs/LTR5/HERVK characteristic in human ESCs. (**A**) Plot of average ChIP-seq signal profiles for LTR5_Hs/LTR5. MYC, OCT4, SOX2, NANOG, KLF4, EP300 are directly related to pluripotency. (**B**) Plot of average H3K27ac ChIP-seq signal profiles for LTR5_Hs/LTR5. Red for HERVK associated LTR5_Hs/LTR5 and blue for solo LTR5_Hs/LTR5. (**C**) RT-qPCR analysis of nuclear and cytoplasmic *HERVK* transcript level. The RT-qPCR data is converted into Cytoplasmic–nuclear relative concentration index (CN-RCI). Error bars represent mean ± s.e.m. At least three biological replicates were examined. (**D**) Schematic of experimental assembly strategy of sgRNA multiplexed array. (**E**) RT-qPCR analysis of *HERVK* and pluripotent markers expression in HUES8 cells induced with dCas9-KRAB (CRISPRi). Error bars represent mean ± s.e.m. $^{***}P < 0.001$, $^{****}P < 0.0001$; unpaired Student's *t*-test. At least three biological replicates were examined. (**F**) Immunostaining for the pluripotent markers NANOG and OCT4 in CRISPRi experiment. Scale bars = 100 μm (**G**) HERVK loci expression level based on RNA-seq data in CRISPRi experiment. Left: LTR5_Hs/LTR5 associated HERVK loci. Right: HERVK loci not associated with LTR5_Hs/LTR5. (**H**) Track view of the highest expressed HERVK loci, showing knockdown efficiency in human ESC based on RNA-seq data. (**I**) Track view of the LTR5_Hs/LTR5 target genes in human ESC based on RNA-seq data.

This indicates that LTR5_Hs/LTR5 does not affect the expression of pluripotent markers in human ESCs. We next wondered whether LTR5_Hs/LTR5 has a *cis*-regulatory effect on human ESCs. We performed RNA-seq of CRISPRi with sgLTR5_Hs/LTR5 and control ESCs, and observed that most of the *HERVK* related to LTR5_Hs/LTR5 were suppressed (Figure 4G), indicating the successful interference of LTR5_Hs/LTR5. For individual locus, the two with the highest expression of HERVK were visualized as example (Figure 4H). However, we only found fewer than 10 differentially expressed genes; among them, *PRODH* has been identified in a previous report (65), and MMP24 and SEMA4A were adjacent to LTR5_Hs (Figure 4I), with significant enrichment of enhancer-associated histone modifications (Supplementary Figure S5A). These results indicated that LTR5_Hs/LTR5 has limited regulatory ability in human ESCs.

In addition, by ChIP-seq analysis we found that activating epigenetic factors (ASH2L, TAF1) and inhibitory epigenetic factors (E2F6, MAX, HDAC2) co-localized on LTR5_Hs/LTR5 in human ESCs (Supplementary Figure S6A). Albeit E2F6 and MAX are directly related to PRC complex (66), we did not observe either obvious signals of H3K27me3 and H2AK119ub in LTR5_Hs/LTR5 (Supplementary Figure S6B and C), or other PRC1/2 members binding to LTR5_Hs/LTR5 (Supplementary Figure S6D).

LTR5_Hs/LTR5/HERVK is more active in the early stages of pluripotency

We performed motif enrichment analysis on LTR5_Hs/LTR5 and identified a series of pluripotency factors, including MYC, SOX2 and KLF5 (Figure 5A), consistent with the ChIP data that most of these transcriptional factors indeed bind to LTR5_Hs/LTR5 loci (Figure 4A). LTR5_Hs/LTR5 is expressed from the 8-cell stage of human embryos and continues to the inner cell mass in blastocyst stage (Figure 5B), and this process is correlated with the up-regulation of pluripotent factors (Figure 5C). Since conventional human ESCs are considered as epiblast stage (late blastocyst) (31,67), and LTR5_Hs/LTR5-bound transcription factors and the HERVK transcript exhibited the higher expression levels in early 8-cell or morula stage, we wondered whether the activity of LTR5_Hs/LTR5 elements are higher in pluripotent cell with early embryonic characteristics, such as naïve ESCs and extended pluripotent stem cells (EPSCs) (68). We detected the activity of LTR5_Hs/LTR5/HERVK in naïve ESCs and EPSCs, truly higher than conventional human ESCs (Figure 5D, Supplementary Table S7). Consistently, we also observed that LTR5_Hs/LTR5-bound transcription factors are significantly up-regulated in both EPSCs and naïve ESCs, such as KLF4, KLF5 and TFAP2C (Figure 5E, Supplementary Table S8).

To evaluate whether LTR5_Hs/LTR5 activity is associated with early pluripotency, we calculated the $\log_2$(fold-change) of differential expressed genes in naive ESC and EPSC relative to primed ESC, then we found that genes closer to LTR5_Hs/LTR5 were more significantly up-regulated (Figure 5F, Supplementary Tables S8 and S9), together indicating that LTR5_Hs/LTR5 can act as a *cis*-regulatory element and correlate with pluripotency.

LTR5_Hs/LTR5/HERVK regulates the gene network of EPSCs through long-range effects

LTR5_Hs/LTR5 elements exhibit higher activity in EPSCs, so we focused on the regulation in feeder-free EPSCs (ffEPSCs) previously established in our laboratory (31). We first validated the expression pattern by RT-qPCR showing HERVK is much higher in ffEPSCs than the parental ESCs (Figure 6A). Next, we established a conventional ESC line H9 (H9-ESCs) with a LTR5_Hs/LTR5-EGFP reporter using the LTR5_Hs/LTR5 conservative sequence. By converting H9-ESCs into ffEPSCs (H9-ffEPSCs), we found that the fluorescence intensity of EGFP in ffESPCs is much brighter than ESCs (Figure 6B), confirming that LTR5_Hs/LTR5 activity is correlated with different pluripotency states.

To address the role of LTR5_Hs/LTR5 in the extended pluripotent stage, we converted the ESCs with inducible targeting of LTR5_Hs/LTR5 (Figure 6C) to the extended pluripotent state. After 7 days of doxycycline treatment, mCherry arose, representing the unsilenced vector. The morphology of ffEPSCs with repressive LTR5_Hs/LTR5/HERVK was comparable with the control (Figure 6C), indicating that the activity of LTR5_Hs/LTR5/HERVK was not necessary for the conversion or maintenance of ffEPSCs. The LTR5_Hs/LTR5 still played a role in regulating the target genes in ffEPSCs (Figure 6D), so we further performed the RNA-seq of LTR5_Hs/LTR5-targeted ffEPSCs to check the global regulation. Interestingly, we identified 80 down-regulated genes and 14 up-regulated genes (Figure 6E and G, Supplementary Table S6), much more than in primed ESCs. We calculated the distance between transcription start site of the down-regulated genes and LTR5_Hs/LTR5, and the result showed that most of the down-regulated genes are located in the range of 25kb around LTR5_Hs/LTR5 (Figure 6F).

The protein interaction network analysis based on these LTR5_Hs/LTR5-related genes highlighted GSTA3/4 and GSTP1 (Supplementary Figure S7A), two important genes related to oxidative metabolism process. Given that LTR5_Hs elements contain the binding site of SOX2, a key transcription factor function in pluripotent and neural cells, and HERVK is reported with high expression in neurons of amyotrophic lateral sclerosis (ALS) patients (69), we were wondering whether LTR5_Hs/LTR5 would be activated in the nerves besides of ffEPSCs. We analyzed the H3K27ac levels in different fetal tissues, and we indeed found a specific enrichment of H3K27ac at LTR5_Hs/LTR5 elements only in brain and retina pigmented epithelium (Supplementary Figure S7B). Consistent with this notion, many neural-related genes were reactivated in ffEPSCs (Supplementary Figure S7C), possibly due to the higher activity of LTR5_Hs/LTR5.

DISCUSSION

LTR elements with regulatory capabilities have attracted more and more attentions in developmental biology and cancer studies (45,70–71). Typically, many LTRs are more

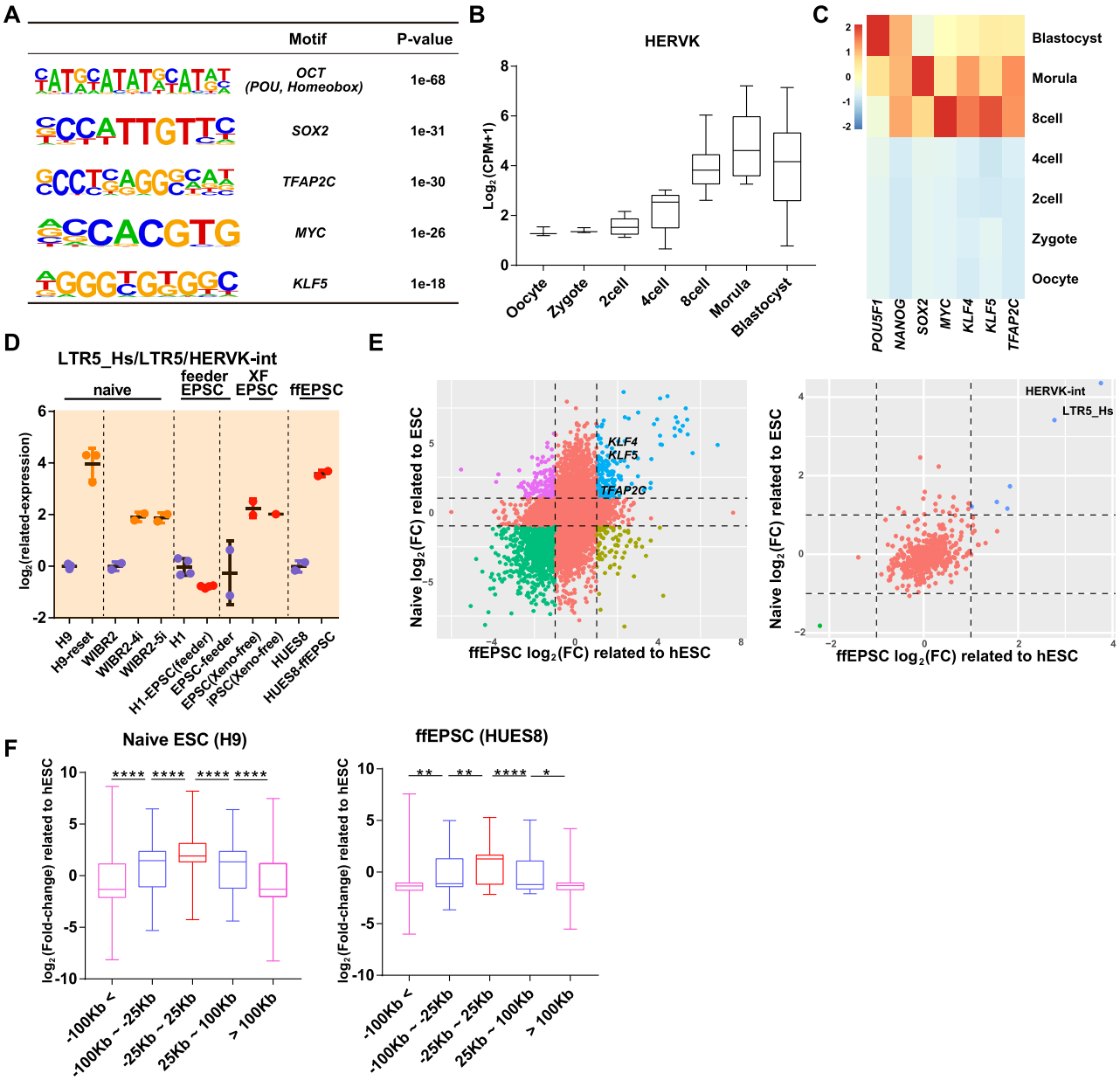

Figure 5. Expression dynamics of LTR5_Hs/LTR5/HERVK in early embryos and different pluripotent stem cells. (**A**) Motif enrichment analysis of LTR5_Hs/LTR5 for putative TF-binding sites. (**B**) Expression level (counts per million) dynamics of HERVK in early embryos based on smart-seq2 dataset. (**C**) Heatmap of hierarchical clustering (Euclidean distance) using Z-score of scaled TPM of regulatory factors of HERVK. (**D**) The expression level (CPM) of LTR5_Hs/LTR5/HERVK is normalized using the same batch RNA-seq data of primed human ESC, and the expression of HERVK in xenofree-EPSC (XF-EPSC) is normalized to feeder-EPSC. (**E**) Left: the expression levels of all protein coding genes in ffEPSC and naïve ESC relative to primed human ESC. Right: the expression levels of all LTRs in ffEPSC and naïve ESC relative to primed human ESC. (**F**) Box plots of $\log_2$(fold-change) relative to primed human ESC in differentially expressed genes between naive ESC and ffEPSC. Error bars represent mean ± s.e.m. $^*P < 0.05$, $^{**}P < 0.01$, $^{****}P < 0.0001$; unpaired Student's *t*-test.

active in ESCs than somatic cells (72,73), offering a good system to explore the functions of LTRs. However, only certain individual LTRs have been studied in human ESCs. Here, we systematically annotated the epigenetic modification map of LTRs in human pluripotent stem cells based on the comprehensive histone modifications and transcription factor ChIP-seq data of the H1 cell line (Supplementary Table S2 and S3). We defined the active LTRs in the pluripotent state, and found that MER57E3 (promoter-related LTR) could act as a proximal regulatory element together with promoter and regulated ZNF gene expression to regulate host genes (Figure 3 and Supplementary Figure S4D–F), while LTR5_Hs/LTR5 (enhancer-related LTR) could act as enhancer in conventional human ESCs and ffEPSCs (Figure 4I, Figure 6E and F). Our results further illustrated the diversity of the pathways in which LTRs

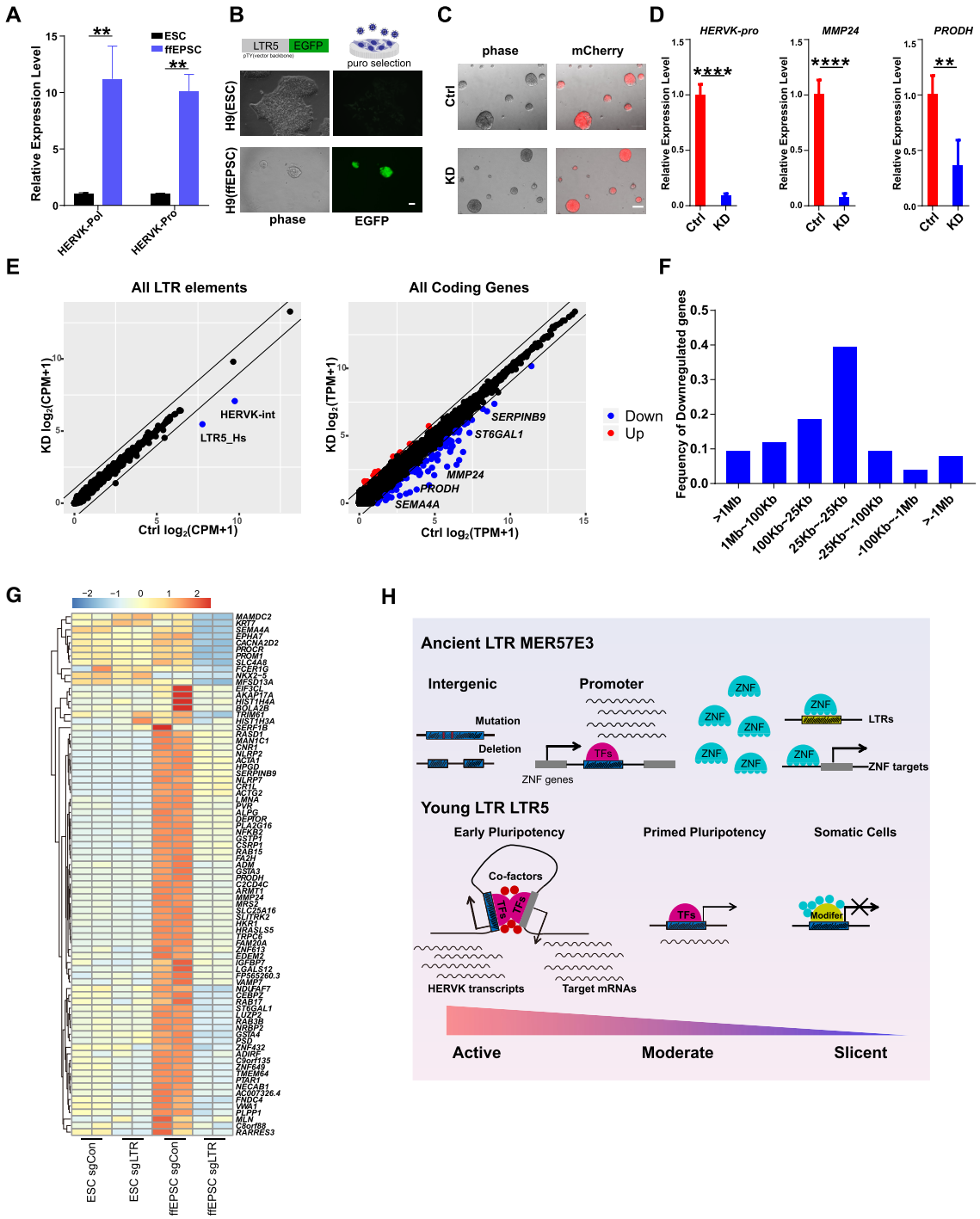

Figure 6. LTR5_Hs/LTR5 acts as a powerful *cis*-regulatory element in EPSCs. (**A**) RT-qPCR analysis of HERVK transcript relative level in ESC and ffEPSC. Error bars represent mean ± s.e.m. $^{**}P < 0.01$; unpaired Student *t*-test. At least three biological replicates were examined. (**B**) Top: schematic of generation of LTR5_Hs/LTR5-EGFP cell lines. Bottom: phase and EGFP contrast images of LTR5_Hs/LTR5-EGFP reporter human ESC and ffEPSC. Scale bars = 20 μm. (**C**) Phase and mCherry contrast images of LTR5_Hs/LTR5/HERVK inducible knockdown ffEPSC. Scale bars = 100 μm. (**D**) RT-qPCR analysis of HERVK target genes in HUES8-ffEPSC induced with dCas9-KRAB (CRISPRi). Error bars represent mean ± s.e.m. $^{**}P < 0.01$, $^{****}P < 0.0001$; unpaired Student *t*-test. At least three biological replicates were examined. (**E**) Left: RNA-seq analysis of all LTRs in CRISPRi experiment. Right: RNA-seq analysis of protein coding genes in CRISPRi experiment (data from this study) ($n = 2$ biological replicates for both conditions). (**F**) Histograms showing down-regulated genes were distinguished by different bins through LTR5_Hs/LTR5 distance from the TSS. (**G**) Heatmap of hierarchical clustering (Euclidean distance) using *Z*-score of scaled TPM of ESC and ffEPSC in LTR5_Hs/LTR5/HERVK CRISPRi experiment. (**H**) Top: Promoter MER57E3 is regulated by TEF and further affects the expression of ZNF gene and its downstream genes. Bottom: LTR5_Hs/LTR5 is more active in early pluripotency and acts as a *cis*-regulatory element to mediate the expression of early pluripotency genes. LTR5_Hs/LTR5 is bound by active/repressive epigenetic factors and maintains moderate expression level and limited regulatory ability in primed pluripotency, and then becomes completely silenced after exiting pluripotency.

regulate host genes (Figure 6H). In our study, MER57E3, as an ancient LTR, has undergone a long-term natural selection and accumulate mutations, and has lost the ability of autonomously amplification. Herein, we found the ZNF-MER57E3 has higher integrity compared with intergenic-MER57E3. Moreover, our data demonstrated the potential of MER57E3 to regulate ZNF genes (Figure 6H), which have been reported to directly bind other transposon elements, providing a new beneficial mechanism for the host genome stability.

MER57E3 is mostly located in the promoter region of the ZNF genes. Previous reports showed that there was a feedback loop between LTRs and ZNF genes (74,75), a reflection of LTRs adaption to the host. Herein, we found the chromatin accessibility near the full-length MER57E3 was much higher than which of the truncated MER57E3, and the more accumulated mutations it has, the lower the chromatin accessibility of MER57E3 is. Those data represent MER57E3 might be co-evolved with the host ZNF genes. In addition, we proved that TEF of the bZIP gene family could directly bind to MER57E3 and regulate transcription activity evidenced by luciferase assay in HEK293T and overexpression assay in human ESCs. Many ZNFs related to MER57E3 are activated in human ESCs (Figure 3A), and these ZNFs can bind at least one LTR. There might be a possibility that MER57E3 controls the expression of adjacent ZNF proteins to regulate the expression of host genes and the activity of other LTRs. Our inducible CRISPRi data (Figure 3E and I) clearly showed that MER57E3 can act as a proximal regulatory element to control the expression of ZNF genes. Then we found ZNF730, a ZNF gene related to MER57E3 can directly bind to TSS to regulate host gene expression. Our study provides an insight to understand how activation of transposon elements could contribute to certain related disorders by regulating the ZNF genes.

As the young LTRs, LTR5_Hs/LTR5 seems to regulate host gene expression directly. Our data indicate that most of the *HERVK* transcripts are located in the nucleus, and combined with histone modification, YY1 and POLII enrichment (Supplementary Table S2 and S3), indicating a possibility that *HERVK* may act as eRNA (76). Interestingly, a couple of previous studies reported that the viral protein encoded by HERVK could be detected in pluripotent cells (61,77), suggesting a potential diverse role of HERVK. Through motif enrichment analysis, we found that LTR5_Hs/LTR5 enriched transcription factors, such as KLF4/5 and TFAP2C, highly related to early embryo status, so we further surveyed the early embryo cells and found that LTR5_Hs/LTR5 is more active in naïve ESCs and ffEPSCs (Figure 5D). By converting LTR5_Hs/LTR5 CRISPRi ESCs into ffEPSCs, we found that LTR5_Hs/LTR5 becomes more active with more powerful *cis*-regulation ability. Very interestingly, we observed the reactivation of certain neural-related genes upon higher activity of LTR5_Hs/LTR5. This implies that the human-specific LTR5_Hs/LTR5 fine-tunes the host gene expression during early embryogenesis and might affect the human-specific brain development process as well.

Overall, here we characterize a functional class of LTRs (i.e. MER57E3 elements), which is located in downstream of the TSS and regulates the adjacent ZNF gene expression to affect host gene expression. Furthermore, LTR5_Hs/LTR5, as young LTRs, could regulate gene expression through long-range effects and play a different role in pluripotent stem cells with different pluripotency states.

DATA AVAILABILITY

All information is displayed in materials and methods. All raw sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE186430. The source of the published sequencing data can be viewed in (Supplementary Table S1).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We sincerely thank the core facility of the Medical Research Institute at Wuhan University for the technical support.
Author contributions: W.J. and T.Z. conceived the project and designed the experiment. T.Z. performed most of the bench experiments. R.Z. performed conversion of ffEPSC and contributed to the figure preparation. M.L. performed Dual-Luciferase Reporter Assays. B.T. constructed consecutive sgRNAs into lenti-guide plasmid. T.Z. analyzed the next-generation sequencing data with help from C.Y. X.L. provided PCGF6 ChIP-seq data. T.Z. drafted the manuscript, and W.J. and T.Z. finalized the manuscript together with P.L. All authors contributed to and approved the final manuscript.

FUNDING

National Key Research and Development Program of China [2016YFA0503100]; National Natural Science Foundation of China [31970608, 91740102]; Science and Technology Department of Hubei Province [2020CFA017, 2021CFA049]; Fundamental Research Funds for the Central Universities in China [2042021kf0207]. Funding for open access charge: Fundamental Research Funds for the Central Universities in China.
Conflict of interest statement. None declared.

REFERENCES

1. Doolittle,W.F. and Sapienza,C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature*, **284**, 601–603.
2. Sharif,J., Endo,T.A., Nakayama,M., Karimi,M.M., Shimada,M., Katsuyama,K., Goyal,P., Brind'Amour,J., Sun,M.A., Sun,Z. *et al.* (2016) Activation of endogenous retroviruses in dnmt1(-/-) ESCs involves disruption of SETDB1-Mediated repression by NP95 binding to hemimethylated DNA. *Cell Stem Cell*, **19**, 81–94.
3. Fasching,L., Kapopoulou,A., Sachdeva,R., Petri,R., Jonsson,M.E., Manne,C., Turelli,P., Jern,P., Cammas,F., Trono,D. *et al.* (2015) TRIM28 represses transcription of endogenous retroviruses in neural progenitor cells. *Cell Rep.*, **10**, 20–28.
4. Brattas,P.L., Jonsson,M.E., Fasching,L., Nelander Wahlestedt,J., Shahsavani,M., Falk,R., Falk,A., Jern,P., Parmar,M. and Jakobsson,J. (2017) TRIM28 controls a gene regulatory network based on endogenous retroviruses in human neural progenitor cells. *Cell Rep.*, **18**, 1–11.

5. Hisada,K., Sanchez,C., Endo,T.A., Endoh,M., Roman-Trufero,M., Sharif,J., Koseki,H. and Vidal,M. (2012) RYBP represses endogenous retroviruses and preimplantation- and germ line-specific genes in mouse embryonic stem cells. *Mol. Cell. Biol.*, **32**, 1139–1149.
6. Zhang,H., Zhang,F., Chen,Q., Li,M., Lv,X., Xiao,Y., Zhang,Z., Hou,L., Lai,Y., Zhang,Y. *et al.* (2021) The piRNA pathway is essential for generating functional oocytes in golden hamsters. *Nat. Cell Biol.*, **23**, 1013–1022.
7. Perron,H. and Lang,A. (2010) The human endogenous retrovirus link between genes and environment in multiple sclerosis and in multifactorial diseases associating neuroinflammation. *Clin. Rev. Allergy Immunol.*, **39**, 51–61.
8. Gruchot,J., Kremer,D. and Kury,P. (2019) Neural cell responses upon exposure to human endogenous retroviruses. *Front. Genet.*, **10**, 655.
9. Manghera,M., Ferguson-Parry,J. and Douville,R.N. (2016) TDP-43 regulates endogenous retrovirus-K viral protein accumulation. *Neurobiol. Dis.*, **94**, 226–236.
10. Wu,Z., Zhou,J., Zhang,X., Zhang,Z., Xie,Y., Liu,J.B., Ho,Z.V., Panda,A., Qiu,X., Cejas,P. *et al.* (2021) Reprogramming of the esophageal squamous carcinoma epigenome by SOX2 promotes ADAR1 dependence. *Nat. Genet.*, **53**, 881–894.
11. Singh,M., Cai,H., Bunse,M., Feschotte,C. and Izsvak,Z. (2020) Human endogenous retrovirus k rec forms a regulatory loop with MITF that opposes the progression of melanoma to an invasive stage. *Viruses*, **12**, 1303.
12. Wang,R., Li,H., Wu,J., Cai,Z.Y., Li,B., Ni,H., Qiu,X., Chen,H., Liu,W., Yang,Z.H. *et al.* (2020) Gut stem cell necroptosis by genome instability triggers bowel inflammation. *Nature*, **580**, 386–390.
13. Modzelewski,A.J., Shao,W., Chen,J., Lee,A., Qi,X., Noon,M., Tjokro,K., Sales,G., Biton,A., Anand,A. *et al.* (2021) A mouse-specific retrotransposon drives a conserved cdk2ap1 isoform essential for development. *Cell*, **184**, 5541–5558.
14. Schmidt,D., Schwalie,P.C., Wilson,M.D., Ballester,B., Goncalves,A., Kutter,C., Brown,G.D., Marshall,A., Flicek,P. and Odom,D.T. (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, **148**, 335–348.
15. Kunarso,G., Chia,N.-Y., Jeyakani,J., Hwang,C., Lu,X., Chan,Y.-S., Ng,H.-H. and Bourque,G. (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.*, **42**, 631–634.
16. Sundaram,V., Cheng,Y., Ma,Z., Li,D., Xing,X., Edge,P., Snyder,M.P. and Wang,T. (2014) Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.*, **24**, 1963–1976.
17. Haring,N.L., van Bree,E.J., Jordaan,W.S., Roels,J.R.E., Sotomayor,G.C., Hey,T.M., White,F.T.G., Galland,M.D., Smidt,M.P. and Jacobs,F.M.J. (2021) ZNF91 deletion in human embryonic stem cells leads to ectopic activation of SVA retrotransposons and up-regulation of KRAB zinc finger gene clusters. *Genome Res.*, **31**, 551–563.
18. Xiong,F., Wang,R., Lee,J.H., Li,S., Chen,S.F., Liao,Z., Hasani,L.A., Nguyen,P.T., Zhu,X., Krakowiak,J. *et al.* (2021) RNA m(6)A modification orchestrates a LINE-1-host interaction that facilitates retrotransposition and contributes to long gene vulnerability. *Cell Res.*, **31**, 861–885.
19. Liu,L., Leng,L., Liu,C., Lu,C., Yuan,Y., Wu,L., Gong,F., Zhang,S., Wei,X., Wang,M. *et al.* (2019) An integrated chromatin accessibility and transcriptome landscape of human pre-implantation embryos. *Nat. Commun.*, **10**, 364.
20. Grow,E.J., Flynn,R.A., Chavez,S.L., Bayless,N.L., Wossidlo,M., Wesche,D.J., Martin,L., Ware,C.B., Blish,C.A., Chang,H.Y. *et al.* (2015) Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature*, **522**, 221–225.
21. Goke,J., Lu,X., Chan,Y.S., Ng,H.H., Ly,L.H., Sachs,F. and Szczerbinska,I. (2015) Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell*, **16**, 135–141.
22. Hendrickson,P.G., Dorais,J.A., Grow,E.J., Whiddon,J.L., Lim,J.W., Wike,C.L., Weaver,B.D., Pflueger,C., Emery,B.R., Wilcox,A.L. *et al.* (2017) Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nat. Genet.*, **49**, 925–934.
23. Yang,F., Huang,X., Zang,R., Chen,J., Fidalgo,M., Sanchez-Priego,C., Yang,J., Caichen,A., Ma,F., Macfarlan,T. *et al.* (2020) DUX-miR-344-ZMYM2-Mediated activation of MERVL LTRs induces a totipotent 2C-like state. *Cell Stem Cell*, **26**, 234–250.
24. Macfarlan,T.S., Gifford,W.D., Driscoll,S., Lettieri,K., Rowe,H.M., Bonanomi,D., Firth,A., Singer,O., Trono,D. and Pfaff,S.L. (2012) Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*, **487**, 57–63.
25. Grow,E.J., Weaver,B.D., Smith,C.M., Guo,J., Stein,P., Shadle,S.C., Hendrickson,P.G., Johnson,N.E., Butterfield,R.J., Menafra,R. *et al.* (2021) p53 convergently activates dux/dux4 in embryonic stem cells and in facioscapulohumeral muscular dystrophy cell models. *Nat. Genet.*, **53**, 1207–1220.
26. Liu,J., Gao,M., He,J., Wu,K., Lin,S., Jin,L., Chen,Y., Liu,H., Shi,J., Wang,X. *et al.* (2021) The RNA m(6)A reader YTHDC1 silences retrotransposons and guards ES cell identity. *Nature*, **591**, 322–326.
27. Lu,X., Sachs,F., Ramsay,L., Jacques,P.E., Goke,J., Bourque,G. and Ng,H.H. (2014) The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat. Struct. Mol. Biol.*, **21**, 423–425.
28. Zhang,Y., Li,T., Preissl,S., Amaral,M.L., Grinstein,J.D., Farah,E.N., Destici,E., Qiu,Y., Hu,R., Lee,A.Y. *et al.* (2019) Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat. Genet.*, **51**, 1380–1388.
29. Pontis,J., Planet,E., Offner,S., Turelli,P., Duc,J., Coudray,A., Theunissen,T.W., Jaenisch,R. and Trono,D. (2019) Hominoid-Specific transposable elements and KZFPs facilitate human embryonic genome activation and control transcription in naive human ESCs. *Cell Stem Cell*, **24**, 724–735.
30. Encode Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
31. Zheng,R., Geng,T., Wu,D.Y., Zhang,T., He,H.N., Du,H.N., Zhang,D., Miao,Y.L. and Jiang,W. (2021) Derivation of feeder-free human extended pluripotent stem cells. *Stem Cell Rep.*, **16**, 1686–1696.
32. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with bowtie 2. *Nat. Methods*, **9**, 357–359.
33. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
34. Ramirez,F., Dundar,F., Diehl,S., Gruning,B.A. and Manke,T. (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.*, **42**, W187–W191.
35. Imbeault,M., Helleboid,P.Y. and Trono,D. (2017) KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*, **543**, 550–554.
36. Feng,J., Liu,T., Qin,B., Zhang,Y. and Liu,X.S. (2012) Identifying chip-seq enrichment using MACS. *Nat. Protoc.*, **7**, 1728–1740.
37. Pertea,M., Kim,D., Pertea,G.M., Leek,J.T. and Salzberg,S.L. (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, stringtie and ballgown. *Nat. Protoc.*, **11**, 1650–1667.
38. Quinlan,A.R. (2014) BEDTools: the swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics*, **47**, 11.12.1–11.12.34.
39. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
40. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Mol. Cell*, **38**, 576–589.
41. He,J., Fu,X., Zhang,M., He,F., Li,W., Abdul,M.M., Zhou,J., Sun,L., Chang,C., Li,Y. *et al.* (2019) Transposable elements are regulated by context-specific patterns of chromatin marks in mouse embryonic stem cells. *Nat. Commun.*, **10**, 34.
42. Kelley,D. and Rinn,J. (2012) Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.*, **13**, R107.
43. Wang,J., Singh,M., Sun,C., Besser,D., Prigione,A., Ivics,Z., Hurst,L.D. and Izsvak,Z. (2016) Isolation and cultivation of naive-like human pluripotent stem cells based on HERVH expression. *Nat. Protoc.*, **11**, 327–346.
44. Wang,J., Xie,G., Singh,M., Ghanbarian,A.T., Rasko,T., Szvetnik,A., Cai,H., Besser,D., Prigione,A., Fuchs,N.V. *et al.* (2014)

Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature*, **516**, 405–409.
45. Deniz,O., Ahmed,M., Todd,C.D., Rio-Machin,A., Dawson,M.A. and Branco,M.R. (2020) Endogenous retroviruses are a source of enhancers with oncogenic potential in acute myeloid leukaemia. *Nat. Commun.*, **11**, 3506.
46. Sakashita,A., Maezawa,S., Takahashi,K., Alavattam,K.G., Yukawa,M., Hu,Y.C., Kojima,S., Parrish,N.F., Barski,A., Pavlicev,M. *et al.* (2020) Endogenous retroviruses drive species-specific germline transcriptomes in mammals. *Nat. Struct. Mol. Biol.*, **27**, 967–977.
47. Pehrsson,E.C., Choudhary,M.N.K., Sundaram,V. and Wang,T. (2019) The epigenomic landscape of transposable elements across normal human development and anatomy. *Nat. Commun.*, **10**, 5640.
48. Benayoun,B.A., Pollina,E.A., Ucar,D., Mahmoudi,S., Karra,K., Wong,E.D., Devarajan,K., Daugherty,A.C., Kundaje,A.B., Mancini,E. *et al.* (2014) H3K4me3 breadth is linked to cell identity and transcriptional consistency. *Cell*, **158**, 673–688.
49. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
50. Storer,J., Hubley,R., Rosen,J., Wheeler,T.J. and Smit,A.F. (2021) The dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA*, **12**, 2.
51. Simonti,C.N., Pavlicev,M. and Capra,J.A. (2017) Transposable element exaptation into regulatory regions is rare, influenced by evolutionary age, and subject to pleiotropic constraints. *Mol. Biol. Evol.*, **34**, 2856–2869.
52. Oleksiewicz,U., Gladych,M., Raman,A.T., Heyn,H., Mereu,E., Chlebanowska,P., Andrzejewska,A., Sozanska,B., Samant,N., Fak,K. *et al.* (2017) TRIM28 and interacting KRAB-ZNFs control self-renewal of human pluripotent stem cells through epigenetic repression of Pro-differentiation genes. *Stem Cell Rep.*, **9**, 2065–2080.
53. Fang,F., Xia,N., Angulo,B., Carey,J., Cady,Z., Durruthy-Durruthy,J., Bennett,T., Sebastiano,V. and Reijo Pera,R.A. (2018) A distinct isoform of ZNF207 controls self-renewal and pluripotency of human embryonic stem cells. *Nat. Commun.*, **9**, 4384.
54. Zorzan,I., Pellegrini,M., Arboit,M., Incarnato,D., Maldotti,M., Forcato,M., Tagliazucchi,G.M., Carbognin,E., Montagner,M., Oliviero,S. *et al.* (2020) The transcriptional regulator ZNF398 mediates pluripotency and epithelial character downstream of TGF-beta in human PSCs. *Nat. Commun.*, **11**, 2364.
55. Senft,A.D. and Macfarlan,T.S. (2021) Transposable elements shape the evolution of mammalian development. *Nat. Rev. Genet.*, **22**, 691–711.
56. Liu,H., Chang,L.H., Sun,Y., Lu,X. and Stubbs,L. (2014) Deep vertebrate roots for mammalian zinc finger transcription factor subfamilies. *Genome Biol Evol*, **6**, 510–525.
57. Hunger,S.P., Li,S., Fall,M.Z., Naumovski,L. and Cleary,M.L. (1996) The proto-oncogene HLF and the related basic leucine zipper protein TEF display highly similar DNA-binding and transcriptional regulatory properties. *Blood*, **87**, 4607–4617.
58. Inukai,T., Inaba,T., Dang,J., Kuribara,R., Ozawa,K., Miyajima,A., Wu,W., Look,A.T., Arinobu,Y., Iwasaki,H. *et al.* (2005) TEF, an antiapoptotic bZIP transcription factor related to the oncogenic E2A-HLF chimera, inhibits cell growth by down-regulating expression of the common beta chain of cytokine receptors. *Blood*, **105**, 4437–4444.
59. Mandegar,M.A., Huebsch,N., Frolov,E.B., Shin,E., Truong,A., Olvera,M.P., Chan,A.H., Miyaoka,Y., Holmes,K., Spencer,C.I. *et al.* (2016) CRISPR interference efficiently induces specific and reversible gene silencing in human iPSCs. *Cell Stem Cell*, **18**, 541–553.
60. Reus,K., Mayer,J., Sauter,M., Scherer,D., Muller-Lantzsch,N. and Meese,E. (2001) Genomic organization of the human endogenous retrovirus HERV-K(HML-2.HOM) (ERVK6) on chromosome 7. *Genomics*, **72**, 314–320.
61. Fuchs,N.V., Loewer,S., Daley,G.Q., Izsvak,Z., Lower,J. and Lower,R. (2013) Human endogenous retrovirus k (HML-2) RNA and protein expression is a marker for human embryonic and induced pluripotent stem cells. *Retrovirology*, **10**, 115.
62. Theunissen,T.W., Friedli,M., He,Y., Planet,E., O'Neil,R.C., Markoulaki,S., Pontis,J., Wang,H., Iouranova,A., Imbeault,M. *et al.* (2016) Molecular criteria for defining the naive human pluripotent state. *Cell Stem Cell*, **19**, 502–515.
63. Fuentes,D.R., Swigut,T. and Wysocka,J. (2018) Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. *Elife*, **7**, e35989.
64. Gu,B., Swigut,T., Spencley,A., Bauer,M.R., Chung,M., Meyer,T. and Wysocka,J. (2018) Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements. *Science*, **359**, 1050–1055.
65. Suntsova,M., Gogvadze,E.V., Salozhin,S., Gaifullin,N., Eroshkin,F., Dmitriev,S.E., Martynova,N., Kulikov,K., Malakhova,G., Tukhbatova,G. *et al.* (2013) Human-specific endogenous retroviral insert serves as an enhancer for the schizophrenia-linked gene PRODH. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 19472–19477.
66. Qin,J., Whyte,W.A., Anderssen,E., Apostolou,E., Chen,H.H., Akbarian,S., Bronson,R.T., Hochedlinger,K., Ramaswamy,S., Young,R.A. *et al.* (2012) The polycomb group protein L3mbtl2 assembles an atypical PRC1-family complex that is essential in pluripotent stem cells and early development. *Cell Stem Cell*, **11**, 319–332.
67. Geng,T., Zhang,D. and Jiang,W. (2019) Epigenetic regulation of transition among different pluripotent states: concise review. *Stem Cells*, **37**, 1372–1380.
68. Yang,Y., Liu,B., Xu,J., Wang,J., Wu,J., Shi,C., Xu,Y., Dong,J., Wang,C., Lai,W. *et al.* (2017) Derivation of pluripotent stem cells with in vivo embryonic and extraembryonic potency. *Cell*, **169**, 243–257.
69. Li,W., Lee,M.H., Henderson,L., Tyagi,R., Bachani,M., Steiner,J., Campanac,E., Hoffman,D.A., von Geldern,G., Johnson,K. *et al.* (2015) Human endogenous retrovirus-K contributes to motor neuron disease. *Sci. Transl. Med.*, **7**, 307ra153.
70. Chuong,E.B., Rumi,M.A., Soares,M.J. and Baker,J.C. (2013) Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat. Genet.*, **45**, 325–329.
71. Jansz,N. and Faulkner,G.J. (2021) Endogenous retroviruses in the origins and treatment of cancer. *Genome Biol.*, **22**, 147.
72. Dunican,D.S., Cruickshanks,H.A., Suzuki,M., Semple,C.A., Davey,T., Arceci,R.J., Greally,J., Adams,I.R. and Meehan,R.R. (2013) Lsh regulates LTR retrotransposon repression independently of dnmt3b function. *Genome Biol.*, **14**, R146.
73. Sun,L., Fu,X., Ma,G. and Hutchins,A.P. (2021) Chromatin and epigenetic rearrangements in embryonic stem cell fate transitions. *Front. Cell Dev. Biol.*, **9**, 637309.
74. Ito,J., Kimura,I., Soper,A., Coudray,A., Koyanagi,Y., Nakaoka,H., Inoue,I., Turelli,P., Trono,D. and Sato,K. (2020) Endogenous retroviruses drive KRAB zinc-finger protein family expression for tumor suppression. *Sci. Adv.*, **6**, eabc3020.
75. Thomas,J.H. and Schneider,S. (2011) Coevolution of retroelements and tandem zinc finger genes. *Genome Res.*, **21**, 1800–1812.
76. Sartorelli,V. and Lauberth,S.M. (2020) Enhancer RNAs are an important regulatory layer of the epigenome. *Nat. Struct. Mol. Biol.*, **27**, 521–528.
77. Babarinde,I.A., Ma,G., Li,Y., Deng,B., Luo,Z., Liu,H., Abdul,M.M., Ward,C., Chen,M., Fu,X. *et al.* (2021) Transposable element sequence fragments incorporated into coding and noncoding transcripts modulate the transcriptome of human pluripotent stem cells. *Nucleic Acids Res.*, **49**, 9132–9153.