

Opinion

# The essential but enigmatic regulatory role of HERVH in pluripotency

Corinne E. Sexton,<sup>1</sup> Richard L. Tillett,<sup>2</sup> and Mira V. Han<sup>1,2,\*</sup>

**Human specific endogenous retrovirus H (HERVH) is highly expressed in both naive and primed stem cells and is essential for pluripotency. Despite the proven relationship between HERVH expression and pluripotency, there is no single definitive model for the function of HERVH. Instead, several hypotheses of a regulatory function have been put forward including HERVH acting as enhancers, long noncoding RNAs (lncRNAs), and most recently as markers of topologically associating domain (TAD) boundaries. Recently several enhancer-associated lncRNAs have been characterized, which bind to Mediator and are necessary for promoter–enhancer folding interactions. We propose a synergistic model of HERVH function combining relevant findings and discuss the current limitations for its role in regulation, including the lack of evidence for a pluripotency-associated target gene.**

## HERVH: an endogenous retrovirus essential for pluripotency

**Human endogenous retrovirus H (HERVH)** (see [Glossary](#)) is a primate-specific endogenous retrovirus (ERV) that is essential for maintenance of pluripotency in human stem cells [1,2]. Integrating about 35 million years ago [3], HERVH is one of the most recent and numerous ERVs with around 1000 copies present in the human genome [4]. Structurally, a full-length HERVH is flanked by two **long terminal repeats** (LTR7) and composed of the canonical viral genes *gag*, *pol*, and *env*. Because of extensive mutation accumulation, HERVH loci generally do not have protein-coding potential [5], and some loci are truncated (but, see [6] for specific HERVH protein expression in cancer). However, HERVH elements are better preserved in a full-length state or near-full-length state (many elements are missing the *env* protein) compared with other HERV families in the human genome, hinting at a selective force preferring full-length HERVH elements [7]. This, combined with the essentiality of HERVH transcription in the maintenance of human stem cell identity, implies exaptation of full-length HERVH in human host biology.

HERVH is used as a common marker for pluripotency, localizes almost exclusively in the nucleus and accounts for 2% of all poly-A RNA in **human embryonic stem cells (hESCs)** [1,8–12]. Although it is highly expressed in both **naïve stem cells** and **primed stem cells** (Box 1), the function of HERVH RNA in pluripotent cells is still ambiguously defined, due to the multifaceted role it seemingly plays in hESCs. Based on evidence observed so far, several roles have been assigned to HERVH including as functional **long noncoding RNAs (lncRNAs)** [1,2], enhancers and alternative promoters [13,14], and markers of **topologically associating domain (TAD)** boundaries [15]. In this opinion article, we synthesize the recent advances in the fields of enhancer regulation and genome organization in relation to the regulatory role of HERVH in stem cells. These new observations integrate the hypotheses of HERVH function and highlight gaps in our understanding of the role of HERVH transcripts in pluripotency.

## Highlights

Human specific endogenous retrovirus H (HERVH) is one of the most recently integrated and most numerous endogenous retroviruses in the human genome. It is highly expressed in stem cells, essential for pluripotency, and long been hypothesized to be an enhancer important in stem cells.

Mediator is a dynamic complex that changes conformation and partner proteins to bridge enhancers and promoters and initiate transcription.

Long noncoding RNAs (lncRNAs) transcribed from enhancers are shown to be functional participants of the enhancer regulation of target genes. One way that enhancer-associated lncRNAs function is through binding of MED12 and stabilizing the chromatin loops between enhancers and promoters.

Crosslinking immunoprecipitation experiments have shown specific binding of HERVH transcripts to MED12 and other Mediator kinase proteins.

Highly transcribed HERVH loci are capable of demarcating stem cell specific topologically associating domain (TAD) boundaries and defining 3D chromatin structures. Deleting HERVH elements eliminates TAD boundaries and reduces the transcription of nearby genes.

<sup>1</sup>School of Life Sciences, University of Nevada, Las Vegas, NV 89154, USA

<sup>2</sup>Nevada Institute of Personalized Medicine, Las Vegas, NV 89154, USA

\*Correspondence: [mira.han@unlv.edu](mailto:mira.han@unlv.edu) (M.V. Han).



### Box 1. HERVH expression across naïve and primed stem cells

The presence of HERVH expression in naïve versus primed pluripotent stem cells has been of some debate. Wang *et al.* found that naïve cells under 2i/L conditions (inhibition of ERK1 and GSK3b signalling with small molecules, and addition of leukemia inhibitory factor (LIF)) exhibited increased HERVH expression compared with primed hPSCs [2], and they later reported a protocol to pick naïve hESCs using a GFP–LTR7 reporter based on HERVH expression [50]. However, Theunissen *et al.* and Guo *et al.* found HERVH expression more prevalent in primed cells than in naïve cells, also showing that, in contrast to other transposable element families, HERVH/LTR7 are methylated in reset, naïve-like cells [8,51]. It has been shown that TRIM28 represses LTR7/HERVH in the naïve state [52], and ZNF534 which is highly expressed in naïve cells may be the sequence-specific partner to TRIM28 [8,26].

More recently, evidence for a middle ground has been presented. Yamauchi *et al.* showed that induction of an intermediate state between primed and naïve cells by overexpression of NR5A1 resulted in activation of a large number of HERVH elements with transient increased expression as compared to both primed and naïve states [53]. This confirms a previous characterization of the expression of HERVH in the reprogramming process of differentiated cells to iPSCs, wherein HERVH expression surges during reprogramming due to activation by transcription factors OCT4, SOX2, and KLF4, then falls to lower levels in the embryonic stem cell state [16].

It is important to note a subdivision in the HERVH family between type I, type Ia, and type II based on the repeat structure of the U3 region (LTR region) [54]. Based on RNA-seq in embryonic cells and hESCs/hPSCs, Gemmell *et al.* found that type I HERVH loci were almost twice as highly transcribed in H1 hESCs as other HERVH loci and type II HERVH loci were associated with higher transcription in early embryonic cells [55]. This could settle the dispute about expression of HERVH in naïve and primed cells as different individual loci could be contributing high expression in different cell states.

Similarly, Göke *et al.* found that specific LTR7 types were upregulated in stage-specific patterns, LTR7B in morula and eight-cells, LTR7Y in epiblasts and blastocysts, and LTR7 in hESCs and epiblasts, suggesting that specific loci can be used as both cell-type- and stage-specific markers [13]. This LTR-specific expression clustering lends more evidence to the conclusion of Gemmell *et al.* that the sequence of the LTR may be a driving factor in stage-specific locus expression.

### HERVH as lncRNA

Based on their length and their transcriptional activity, HERVHs themselves should be considered as lncRNAs, although they are not annotated as such in reference annotations. Several RNAi knockdown experiments provide strong evidence for a functional role of HERVH RNA transcripts in maintaining pluripotency. Depletion of HERVH RNA in H1 hESCs by short hairpin RNA (shRNA) results in cells morphing to a fibroblast-like appearance as well as downregulation of hallmark pluripotency transcription factors OCT4 (*POU5F1*), SOX2, and NANOG coupled with an upregulation of differentiation markers [1]. Similarly, in H9 hESCs, a shRNA depletion of HERVH led to loss of self-renewal and downregulation of OCT4 and NANOG [2].

Additionally, while the knockdown of HERVH transcripts pushes cells towards differentiation, overexpression of HERVH creates a differentiation-deficient phenotype in reprogrammed **induced pluripotent stem cells (iPSCs)** as shown in two separate perturbed systems: (i) KLF4 overexpression leads to LTR7 activation and HERVH overexpression [16]; and (ii) NAT1 knockdown decreases TUT7 expression, which then fails to degrade HERVH transcripts [17]. Both perturbations result in reduced differentiation and reversal rescues the differentiation-deficient phenotype [16,17].

Although a general knockdown of HERVH expression is sufficient to disrupt pluripotency, individual HERVH-derived lncRNAs with unique chimeric sequences have also been shown to have specific roles in pluripotent cells (Box 2). Strikingly, in H1 hESCs, HERVH-associated lncRNA expression is up to eightfold higher than non-HERVH-associated lncRNA elements [18]. Based on the hg38 reference genome and Gencode v36 annotations, HERVH/LTR7 elements specifically provide 58 transcription start sites (TSS) for Gencode annotated lncRNA genes. Of these, a few have been studied in detail, mainly through RNAi knockdown in hESCs (Box 2).

Although these specific examples provide some context for possible functions of HERVH transcripts in hESCs, distinct HERVH-derived lncRNA seems to have distinct mechanisms, for example, acting

### Glossary

**Gene deserts:** regions of the genome lacking protein-coding genes.

**Human embryonic stem cells**

**(hESCs):** pluripotent cells collected from blastocyst-stage embryos with the potential to differentiate into all other cell types. Commonly used hESC cell lines include H1, H7, and H9.

**Human endogenous retrovirus H**

**(HERVH):** a primate-specific retroviral element, the full length of which is 9 kb and contains the viral genes *gag*, *pol*, and *env*. Around 1000 copies exist in the human genome and most are not full length.

**Induced pluripotent stem cells**

**(iPSCs):** pluripotent cells that are generated from already differentiated cell types by reprogramming through expression of Oct-3/4, KLF4, SOX2, and c-Myc factors.

**Long noncoding RNA (lncRNA):** RNA transcripts of length >200 bp that are not translated into proteins.

**Long terminal repeats (LTRs):**

several hundred base pairs of repetitive sequences of DNA that often flank both ends of a retrotransposon or a retroviral provirus in eukaryotic genomes. Once integrated, the LTRs serve as the promoter responsible for the expression of the retrotransposon/provirus. Can also refer to the large class of LTR retrotransposons which are characterized by these flanking repeats.

**Mediator complex:** a multiprotein complex that is a coactivator of eukaryotic gene expression. Mediator recruits RNA Pol II to gene promoters and is generally required for all eukaryotic gene expression.

**miRNA sponges:** RNA molecules that competitively inhibit binding of miRNA to other molecules by providing alternative miRNA binding sites.

**Naïve stem cells:** cells in a ground state of development with full pluripotent capacity corresponding to the inner cell mass of preimplantation blastocysts in mammals. Several cultivation techniques exist to induce naïve stem cells *in vitro*.

**Primed stem cells:** cells in a more advanced state of development with restricted pluripotency corresponding to the postimplantation epiblast. hESCs are classified as primed stem cells.

**R-loop:** three-stranded nucleic acid structure, composed of a DNA:RNA hybrid and the associated nontemplate single-stranded DNA.

### Box 2. Characterized HERVH-derived lncRNAs

As an example of HERVH-derived lncRNA, *linc-ROR* is a lncRNA necessary for reprogramming fibroblasts into iPSCs [56]. Kelley and Rinn found that the TSS for *linc-ROR* is within an LTR7 and the first two exons contain HERVH sequence [18]. Its proposed mechanism in hESCs is as an miRNA sponge. Specifically, *linc-ROR*, *OCT4*, *SOX2*, and *NANOG* share regulatory miRNAs and therefore transcription of *linc-ROR* prevents degradation of those transcription factors by miRNAs [19].

*LINC00458* also contains an LTR7-derived TSS and HERVH sequence in its first two exons. An siRNA knockdown of it in H1 ESCs resulted in both the loss of *OCT4* expression, as well as a transcriptome profile similar to *NANOG* knockdown. Although *LINC00458* has been shown to localize in the nucleus, it has not been shown to be significantly crosslinked with either *OCT4* or *SOX2* transcription factors [57]. Instead, *LINC00458* has been shown to bind to the SMAD2/SMAD3 transcription factor complex in response to soft substrate culture and is required for specification of endodermal lineage in H9 hESCs [20].

In the knockdown of the HERVH-derived lncRNA gene *ESRG*, expression of pluripotency factors is decreased and hESCs committed to differentiation [2]. The function of *ESRG* has also been investigated by a recent study with the opposite result; *ESRG* was dispensable for pluripotency maintenance through a deletion of the entire *ESRG* region [58]. The discrepancy between the studies could be attributed to the different approaches, with Wang *et al.* using an RNAi knockdown and Takahashi *et al.* a gene knockout [2,58].

An *in vivo* siRNA knockdown of *HPAT2* and *HPAT3*, lncRNA elements derived from HERVH sequence, found that they likely contribute to formation of the inner cell mass of embryos [59]. Because *HPAT2*-expressing blastocyst cells also express genetic markers of epiblast, endoderm, and trophoectoderm lineages, they could represent precursor cells capable of differentiation to any of the three lineages. *HPAT2* and *HPAT3* also are regulated by *NANOG*, and their transcripts bind specifically to *OCT4*, *CDK8*, and other mediator complex proteins (*MED6*, *MED12*, *MED21*, and *MED27*), confirming the results of Lu *et al.* and suggesting a regulatory role [1,40].

as **miRNA sponges** [19] or directly interacting with the transcription factor SMAD2/3 in the nucleus [20]. Unlike HERVH-derived chimeric lncRNAs, HERVH elements themselves share similar sequences. Whether they also share a common function across separate loci is still unclear.

### LTR7/HERVH as enhancers

LTR7/HERVH have long been hypothesized to exhibit enhancer activity, proposed originally based on the enrichment of *OCT4*, *NANOG*, and *SOX2* transcription factor binding sites specifically in the LTR7 regions of HERVH elements [1,2,12,16,21]. In an *OCT4* knockdown in hESCs, ERV1-associated *OCT4* binding sites were threefold enriched near downregulated genes [12], while depletion of *NANOG* after cloning a *NANOG*-bound LTR7 region into the promoter of a luciferase reduced reporter activity [13], both studies providing evidence of a regulatory role for the LTR7 sequence.

In addition to transcription factor binding, several genome-wide assays on epigenetic marks also support the potential for LTR7/HERVH to function as enhancers. LTR7 is significantly hypomethylated specifically in ESCs [22]. DNase I sensitivity data showed a 20-fold enrichment of LTR7 elements observed in open chromatin regions in H7 ESCs [14]. In addition, the histone modifications H3K4me3 and H3K27ac are significantly enriched in LTR7 elements [2,9,18,23,24]. While H3K4me3 is generally associated with promoters, and H3K27ac with enhancers, the contrasts between these modifications and their association with either element are not as distinct [25].

HERVH also overlaps with several annotated **superenhancers** [26]. For example, 11 HERVHs overlap with individual enhancer elements defined within superenhancers in H1 and H9 ESCs, and about 20 HERVH elements are found within the total region of each superenhancer [27].

Despite this evidence, functional effects of LTR7/HERVH candidate enhancers have not been assayed in a comprehensive manner, in contrast to HERVK's LTR5 [26,28]. Two early studies

**Superenhancers:** clusters of active enhancers that are generally associated with important cell-type-specific genes and activate higher transcriptional activity of target genes than regular enhancers.

**Topologically associating domain (TAD):** a structural compartment of the 3D genome where the sequences within the TAD region physically interact more often than with DNA outside of the TAD.

showed that two specific LTR7 loci could drive gene expression through luciferase reporter assays [1,12]. A CRISPRi experiment instating a repressive H3K9me3 mark on LTR7Y/B enhancers resulted in inconclusive global transcriptional changes regardless of gene proximity to LTR7Y/B loci [26], suggesting that a specific knockdown at a targeted locus could be more effective for investigating enhancer activity of LTR7 loci. However, a recent study utilized chromatin immunoprecipitation with a massively parallel reporter assay (ChIP-STARR-seq) to monitor enhancer activity in hESCs in an unbiased manner [29]. Although the study was not aimed specifically at LTR7/HERVH elements, the ChIP-Seq was targeted for NANOG, OCT4, H3K4me1, and H3K27ac and therefore included many LTR7 loci in its targets. Barakat *et al.* found that LTR7 was one of the most enriched transposable element families among active enhancers and showed enrichment in both naïve and primed ESCs [29]. Although LTR7 is, by definition, a repetitive sequence element, not all loci can be automatically treated as having the same function, as shown by the fact that not all of the LTR7 loci that are bound by OCT4 or NANOG are able to enhance transcription of a reporter gene; a sign of the importance of locus-specific functional assays and analysis. Despite these limitations, these data show the most definitive evidence so far that a subset of LTR7/HERVH genomic loci can indeed function as enhancers.

### HERVH demarcates TAD boundaries

Most recently, a few HERVH loci have been shown to demarcate TAD boundaries in hESCs [15]. TADs are structural compartments of the 3D genome in which DNA regions within the TAD physically interact together more often than with DNA outside of the TAD [30]. TAD boundaries are commonly marked by elevated presence of cohesin and CCCTC-binding factor (CTCF), which contribute to DNA loop formation [31]. Through Hi-C techniques, Zhang *et al.* showed that deletion of HERVH loci led to the elimination of TAD boundaries and conversely the *de novo* insertion of HERVH could create new TAD boundaries in hESCs [15]. TAD boundary formation by HERVH is dependent on high expression of HERVH, as repression of HERVH transcription reduces TAD boundary formation. Zhang *et al.* posited that through HERVH transcription, TAD boundaries are formed when the cohesin complex is positioned by RNA polymerase II (RNA Pol II) at the 3' end of HERVH. Deletion of TAD boundaries reduced gene expression upstream of HERVH, supporting the hypothesis that HERVH/LTR7 function as enhancers [15].

Canonically, CTCF is the most common marker of TAD boundaries; most boundaries are created as cohesin is loaded at TSS of active genes and travels along with RNA Pol II until it accumulates at CTCF sites [32]. However, insertion of 8-kb HERVH sequences at random locations creates TAD boundaries irrespective/independent of the existence of a CTCF binding motif [15], thus begging the question as to how HERVH contributes to boundary formation. (i) Do HERVH elements contain a noncanonical CTCF motif at their 3' ends, or (ii) is transcription of HERVH alone sufficient for TAD creation without CTCF?

Despite a lack of canonical CTCF motifs, TAD boundaries associated with the top 50 highly expressed HERVH loci show CTCF binding peaks at the 3' end of HERVH, as well as RNA Pol II and cohesin (SMC3) peaks [15], consistent with the canonical model.

Although CTCF defines 75–95% of TAD boundaries in mammals, there is a minor proportion of TAD boundaries that are not associated with CTCF, which are instead associated with active transcription or chromatin type [30,33]. RNA Pol II and transcription are key factors in cohesin positioning and boundary formation [32,34–36]. In fact, a recent study showed that with a high resolution 3C-based technology known as Micro-C, which captures nucleosome level chromatin interactions, we can see numerous promoter–enhancer interaction loops within TADs that are independent of CTCF and based on active transcription [37].

There could also be an alternative limiting factor in place of CTCF that stalls the RNA Pol II and cohesin, such as an **R-loop**. R-loops have recently been proposed to induce stalling of RNA Pol II to the 3' end of transcription [38]; therefore, the presence of R-loops in hESCs at LTR7/HERVH loci could be part of TAD boundary creation [39].

Although it is clear that HERVH sequence demarcates TAD boundaries, the role of HERVH expression is not understood. The association of many TAD boundaries with highly expressed genes indicates that certain 3D genome structures are important for the regulated expression of such genes, but it does not mean that those genes are expressed for the purpose of creating TAD boundaries. Similarly, TAD boundaries are likely the effect of active HERVH transcription, rather than the function. Understanding the relationship between high HERVH expression, pluripotency maintenance, and TAD formation is essential to discovering the true function of HERVH in stem cells.

### Mediator binding: the key to HERVH function?

In the seminal work by Lu *et al.*, they hypothesized that the HERVH RNA transcript plays a role in the enhancer function of LTR7, and tested to see which proteins are bound to the HERVH transcript [1]. RNA crosslinking and immunoprecipitation (RNA-CLIP) assays on a set of candidate proteins showed that while the HERVH RNA did not bind to the chromatin modifiers that they examined, it bound to OCT4, p300, and several **Mediator complex** proteins, specifically CDK8, MED12, and MED6 [1], which was confirmed by Durruthy-Durruthy *et al.* [40]. The Mediator complex generally recruits RNA Pol II to promoters [41]. Mediator also associates with enhancer regions through transcription factor binding interactions. The specific proteins bound by HERVH RNA, CDK8, and MED12, are part of the mediator kinase module, which is transiently bound to the full Mediator complex through MED12. Unlike the head and middle units of the Mediator complex that are essential for transcription, the tail unit and CDK8 kinase seem to have a more regulatory function [42,43].

Recent developments in Mediator research makes the observations by Lu *et al.* particularly interesting. In addition to its role in transcription initiation, studies have shown that Mediator acts as a dynamic bridge between the enhancer and the core promoter via a chromatin loop [41,44]. lncRNAs can interact with the Mediator complex in this process to regulate promoter–enhancer looping and transcription activation [45], and several studies have characterized a class of lncRNAs called enhancer-associated lncRNAs that are both transcribed from enhancers and regulate enhancer activity [46].

We hypothesize that HERVH is another case of an enhancer-associated lncRNA that interacts with Mediator kinase proteins to regulate pluripotency-associated gene transcription through regulation of promoter–enhancer looping. There are alternative ways that enhancer RNA can regulate enhancer function, for example, chromatin remodeling, recruitment of RNA Pol II, or promoting the release of negative elongation factor, NELF [46]. We believe that the above mechanism is the most likely, as this synthesizes all four lines of evidence as follows: (i) LTR7 is an enhancer [13,23], (ii) which transcribes the functional enhancer-associated lncRNA HERVH [1,47], (iii) which then interacts with Mediator proteins [1,40], (iv) to influence TAD boundaries and regulate pluripotent gene expression [15].

The catalog of functional enhancer-associated lncRNAs that contribute to loop formation is short but increasing. Originally discovered by Lai *et al.* [45] as activating noncoding (nc)RNAs, there are now several well-studied cases (Box 3). The common observations across these studies are that the enhancer is transcribed into a lncRNA, which interacts with MED12 of the Mediator subunit, and that knockdown of the lncRNA impairs enhancer–promoter loops.



### Box 3. Enhancer-associated lncRNAs interacting with Mediator

Lai *et al.* found a class of lncRNAs with enhancer-like activity that they termed ncRNA-activating (ncRNA-a) [45]. ncRNA-as coeluted with affinity-purified Mediator and were enriched in the RNAs crosslinked to MED12 measured by UV-RNA immunoprecipitation. Both knockdown of MED1/MED12 and ncRNA-as similarly decreased expression of their targets *SNAI1*, *AURKA*, and *TAL1* *in vivo*. Depletion of the ncRNA-as or MED1/MED12 reduces the chromosomal looping between ncRNA-a and its target loci measured by chromosome conformation capture (3C) [45].

Hsieh *et al.* demonstrated that the *KLK3* enhancer RNA *KLK3e* facilitates the chromatin looping between the *KLK3* enhancer and its target *KLK2* promoter. Silencing of the *KLK3e* RNA impaired both *KLK2* gene expression and the interaction of the *KLK2* promoter with the *KLK3* enhancer, while suppression of MED1 also reduced the interaction of the *KLK3/2* loci as measured by 3C-qPCR [60].

Since then, further examples have been discovered that implicate Mediator–ncRNA interactions in the 3D genome organization [61]. Trimarchi *et al.* found that lncRNA *LUNAR1* has high frequency contact with *IGF1R* promoter in Hi-C, and recruits the Mediator complex on the promoter and regulates its transcription [62]. Papadopoulou *et al.* found that ncRNA *Dlx1as* binds MED12 and controls the expression of *Hoxd11* of the HoxD gene cluster [63]. In this case, MED12 association with the ncRNA *Dlx1as* is dependent on *RING1B* of the PRC1 complex, and only after *RING1B* dissociates from *Dlx1as* upon differentiation, is the *Dlx1as*–MED12 complex able to activate the target *Hoxd11* [63]. Tan *et al.* found that enhancer RNA *ARIEL* recruits mediator proteins to the *ARID5B* enhancer, promotes enhancer–promoter interactions, and activates expression of *ARID5B*. Loss of *ARIEL* reduced enhancer–promoter interactions and enhancer occupancy of TAL1 complex members, partner proteins, mediator proteins, RNA Pol II, and *ARID5B* itself [64]. If enhancer RNAs bound by other structural proteins such as cohesin are included, the list expands [65].

Although examples in Box 3 support the role of enhancer RNA in the formation of loops, there are numerous other examples where enhancer (e)RNA transcription is inconsequential or unrelated to loop formation [46]. In order to test the mechanistic hypothesis, one needs to show that HERVH is necessary for the loop formation between HERVH and the target loci. This would require a chromosome conformation capture (3C) assay to detect interaction between HERVH and the targets with knockdown of HERVH transcripts and/or knockdown of MED12 to establish that both HERVH transcripts and MED12 are required for the specific DNA contact. The first part of the experiment was completed by Zhang *et al.*, who showed the necessity of HERVH transcription in TAD formation [15]. The additional link to MED12 would provide the mechanistic basis for the loop formation. Ultimately, additional assays would need to show that the loop is necessary to activate target transcription. This leads to the central challenge of our problem, the missing targets of HERVH.

### Challenges uncovering HERVH targets

To truly understand the function of HERVH in pluripotent cells, there must be a confidently linked target for its regulatory role [48]. So far, the connection between its essentiality for pluripotency and its actual target of regulation remains a mystery (see Outstanding questions). As a regulatory element essential for pluripotency, the basest explanation of its function would be that HERVH is regulating genes which are essential for pluripotency. However, HERVH loci are mostly located in **gene deserts**, far from characterized pluripotency genes and other protein-coding genes [1]. Although HERVH knockdown analysis causes downregulation of pluripotency markers and upregulation of differentiation markers, it has been difficult to distinguish direct *cis* versus *trans* effects [1].

Zhang *et al.* performed the most complete study that aimed to link HERVH to a target by perturbing two HERVH loci using CRISPR-Cas9 and seeing the effects in the transcriptome [15]. They showed that there were more genes in the vicinity of the ablated HERVH that were downregulated compared with upregulated, supporting the enhancer function. Among the downregulated genes was the lncRNA *HBL1* that supports pluripotency by suppressing differentiation [15]. Even as the only positive pluripotency associated target that is regulated by HERVH,

the picture seems to be more complicated than the model proposed. This is because *HBL1* resides on the 3' side of HERVH, outside the TAD that was defined by the knocked out HERVH, and that *HBL1* is also regulated by SOX2 [49].

Linking enhancers to targets based on functional evidence is difficult even for regular enhancers, and there are hundreds of thousands of candidate enhancers based on biochemical marks, but only a small list of validated and target-linked human enhancers [48]. However, there are a few HERVH specific challenges. In addition to the complication rising from the auto-feedback loop between HERVH/LTR7 and OCT4, NANOG, and other pluripotency factors, there is also the extra complication arising from the plurality and potential redundancy or interaction among HERVH loci. Although the HERVH knockouts (KOs) altered TAD boundaries and expression patterns for hundreds of genes, both KOs maintained pluripotency [15]. This indicates that we may need to combine methods that target multiple loci, while pursuing detailed investigation on each loci, before we can understand the full functionality of HERVH with regards to pluripotency.

Although the functional connection to pluripotency remains enigmatic even with data as comprehensive as in [15], and may seem discouraging, reviewing the research reveals that we are at an exciting time in regulatory genomics and in understanding elements such as HERVH. Especially with single-cell CRISPR-based enhancer screens, soon we may see CRISPR guide (g)RNAs designed to target several HERVH loci across cells, whose excision can then be examined for their transcriptome-wide effects using single-cell RNA-seq. Combined with higher resolution Hi-C assays that capture detailed enhancer-promoter interactions, we may be able to narrow down targets in the near future.

An alternative possibility is that we are overlooking uncharacterized targets, such as the lncRNA genes in the vicinity, that could be potentially important for pluripotency. The list of differentially regulated genes in HERVH KOs, for example, the 43 genes identified in [15], could be the first candidate genes to examine for potentially undiscovered functions in pluripotency. It is also possible that there are additional roles of the HERVH transcript unrelated to enhancer activity, which could be uncovered through an unbiased proteomic study of proteins binding to HERVH RNA.

### Concluding remarks

The knowledge gained in the past few years about HERVH [15,26], enhancers [23], and lncRNAs [45] convincingly argues that HERVH is an enhancer that regulates pluripotency. More specifically, we propose that HERVH is an enhancer-associated lncRNA that binds both Mediator and pluripotency-associated transcription factors and contributes to the loop formation between the LTR7 enhancer and the target promoter (Figure 1, Key figure). However, to truly establish HERVH/LTR7 as enhancers and understand their role in pluripotency, targets that contribute to the pluripotency gene network must be identified. Future perturbation studies on additional HERVH loci, examining transcriptome and conformational changes should be continued, and with the scalability of new technologies, we should eventually be able to identify important targets.

In addition, as we study each HERVH loci, one at a time, or in combination, it will be important to disentangle the confusion that arises due to the intrinsic plural and repetitive nature of HERVs. We discuss HERVs as a whole set, but each study has examined a few select loci in detail, and those loci do not overlap for all studies. Are the HERVH/LTR7 loci that are bound by OCT4 and NANOG identical to the loci that are positive in reporter assays or to the loci that demarcate TAD boundaries? A meta-analysis that integrates separate observations on HERVH at the locus level would be a worthwhile effort.

### Outstanding questions

What are the targets of HERVH/LTR7 enhancers and their role in pluripotency? Do HERVH/LTR7 enhancers regulate genes that maintain pluripotency?

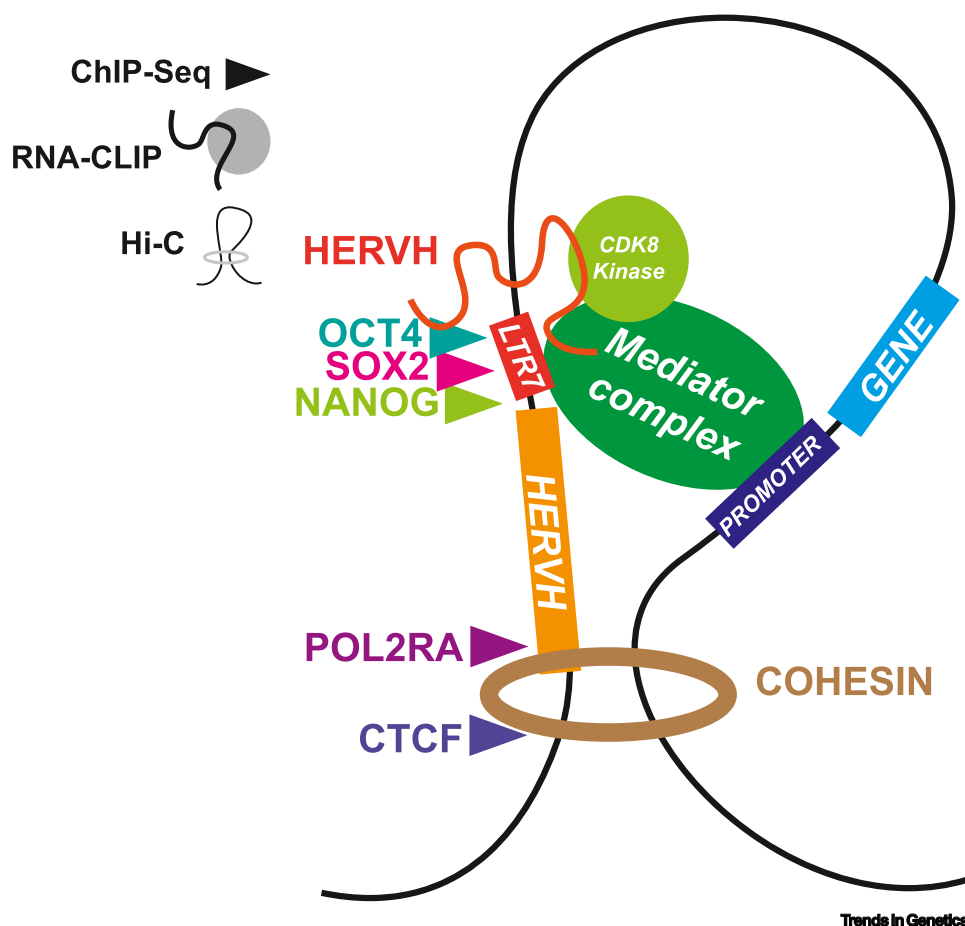
Is transcription of HERVH RNA necessary for enhancer function? If yes, are HERVH RNA transcripts required for all functional LTR7 enhancers?

Do HERVH RNA transcripts act *in cis* or *in trans* on LTR7 enhancers?

What is the functional significance of HERVH RNA binding by Mediator kinase? Does mediator binding affect the 3D conformation between HERVH and target promoter?

**Key figure**

Proposed mechanism of action for human endogenous retrovirus H (HERVH) that synthesizes all lines of evidence



**Figure 1.** HERVH long noncoding RNA (lncRNA; represented by the orange line) is an activating enhancer-associated lncRNA that binds to Mediator kinase and OCT4, contributing to the loop between long terminal repeat (LTR7), which acts as an enhancer, and the target promoter. Additionally, RNA polymerase II and CCCTC-binding factor (CTCF) accumulate at the 3' end of the HERVH sequence, corresponding with topologically associating domain (TAD) boundaries. Chromatin immunoprecipitation sequencing (ChIP-Seq) analysis is represented by triangles, RNA crosslinking immunoprecipitation (RNA-CLIP) evidence is represented by lines and circles, and Hi-C contact is represented by the circular cohesin loop and black DNA line.

**Acknowledgments**

We would like to thank Edwin Oh for his insights. This work was supported by National Institute of General Medical Sciences of the National Institutes of Health under award number P20GM121325, and the National Science Foundation under Grant No. 1750532 and No. 1946082.

**Declaration of interests**

The authors declare no competing interests.



## References

- Lu, X. *et al.* (2014) The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat. Struct. Mol. Biol.* 21, 423–425
- Wang, J. *et al.* (2014) Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* 516, 405–409
- Mager, D.L. and Freeman, J.D. (1995) HERV-H endogenous retroviruses: presence in the new world branch but amplification in the old world primate lineage. *Virology* 213, 395–404
- Jern, P. *et al.* (2004) Definition and variation of human endogenous retrovirus H. *Virology* 327, 93–110
- de Parseval, N. *et al.* (2001) Characterization of the three HERV-H proviruses with an open envelope reading frame encompassing the immunosuppressive domain and evolutionary history in primates. *Virology* 279, 558–569
- Mullins, C.S. and Linnebacher, M. (2012) Endogenous retrovirus sequences as a novel class of tumor-specific antigens: an example of HERV-H env encoding strong CTL epitopes. *Cancer Immunol. Immunother.* 61, 1093–1100
- Gemmell, P. *et al.* (2016) Phylogenetic analysis reveals that ERVs “die young” but HERV-H is unusually conserved. *PLoS Comput. Biol.* 12, e1004964
- Theunissen, T.W. *et al.* (2016) Molecular criteria for defining the naive human pluripotent state. *Cell Stem Cell* 19, 502–515
- Santoni, F.A. *et al.* (2012) HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology* 9, 111
- Battle, S.L. *et al.* (2019) Enhancer chromatin and 3D genome architecture changes from naive to primed human embryonic stem cell states. *Stem Cell Rep.* 12, 1129–1144
- Mareschi, K. *et al.* (2019) Human endogenous retrovirus-H and K expression in human mesenchymal stem cells as potential markers of stemness. *Intervirology* 62, 9–14
- Kunarto, G. *et al.* (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* 42, 631–634
- Göke, J. *et al.* (2015) Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell* 16, 135–141
- Jacques, P.-É. *et al.* (2013) The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet.* 9, e1003504
- Zhang, Y. *et al.* (2019) Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat. Genet.* 51, 1380–1388
- Ohnuki, M. *et al.* (2014) Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proc. Natl. Acad. Sci. U. S. A.* 111, 12426–12431
- Takahashi, K. *et al.* (2020) Critical roles of translation initiation and RNA uridylation in endogenous retroviral expression and neural differentiation in pluripotent stem cells. *Cell Rep.* 31, 107715
- Kelley, D. and Rinn, J. (2012) Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.* 13, R107
- Wang, Y. *et al.* (2013) Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. *Dev. Cell* 25, 69–80
- Chen, Y.-F. *et al.* (2020) Control of matrix stiffness promotes endodermal lineage specification by regulating SMAD2/3 via lincRNA LINC00458. *Sci. Adv.* 6, eaay0264
- Ito, J. *et al.* (2017) Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *PLoS Genet.* 13, e1006883
- Xie, M. *et al.* (2013) DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nat. Genet.* 45, 836–841
- Cao, Y. *et al.* (2019) Widespread roles of enhancer-like transposable elements in cell identity and long-range genomic interactions. *Genome Res.* 29, 40–52
- Friedli, M. *et al.* (2014) Loss of transcriptional control over endogenous retroelements during reprogramming to pluripotency. *Genome Res.* 24, 1251–1259
- Kim, T.-K. and Shiekhattar, R. (2015) Architectural and functional commonalities between enhancers and promoters. *Cell* 162, 948–959
- Pontis, J. *et al.* (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.e5
- Jiang, Y. *et al.* (2019) SEdb: a comprehensive human super-enhancer database. *Nucleic Acids Res.* 47, D235–D243
- Fuentes, D.R. *et al.* (2018) Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. *eLife* 7, e35989
- Barakat, T.S. *et al.* (2018) Functional dissection of the enhancer repertoire in human embryonic stem cells. *Cell Stem Cell* 23, 276–288.e8
- Dixon, J.R. *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380
- Fudenberg, G. *et al.* (2016) Formation of chromosomal domains by loop extrusion. *Cell Rep.* 15, 2038–2049
- Busslinger, G.A. *et al.* (2017) Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. *Nature* 544, 503–507
- Bonev, B. *et al.* (2017) Multiscale 3D genome rewiring during mouse neural development. *Cell* 171, 557–572.e24
- Nora, E.P. *et al.* (2017) Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* 169, 930–944.e22
- Heinz, S. *et al.* (2018) Transcription elongation can affect genome 3D structure. *Cell* 174, 1522–1536.e22
- van Steensel, B. and Furlong, E.E.M. (2019) The role of transcription in shaping the spatial organization of the genome. *Nat. Rev. Mol. Cell Biol.* 20, 327–337
- Hsieh, T.-H.S. *et al.* (2020) Resolving the 3D landscape of transcription-linked mammalian chromatin folding. *Mol. Cell* 78, 539–553.e8
- Niehrs, C. and Luke, B. (2020) Regulatory R-loops as facilitators of gene expression and genome stability. *Nat. Rev. Mol. Cell Biol.* 21, 167–178
- Yan, P. *et al.* (2020) Genome-wide R-loop landscapes during cell differentiation and reprogramming. *Cell Rep.* 32, 107870
- Durruthy-Durruthy, J. *et al.* (2016) The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. *Nat. Genet.* 48, 44–52
- Soutourina, J. (2018) Transcription regulation by the Mediator complex. *Nat. Rev. Mol. Cell Biol.* 19, 262–274
- Cevher, M.A. *et al.* (2014) Reconstitution of active human core Mediator complex reveals a critical role of the MED14 subunit. *Nat. Struct. Mol. Biol.* 21, 1028–1034
- Plaschka, C. *et al.* (2015) Architecture of the RNA polymerase II–Mediator core initiation complex. *Nature* 518, 376–380
- Jeronimo, C. *et al.* (2016) Tail and kinase modules differently regulate core mediator recruitment and function *in vivo*. *Mol. Cell* 64, 455–466
- Lai, F. *et al.* (2013) Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* 494, 497–501
- Li, W. *et al.* (2016) Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat. Rev. Genet.* 17, 207–223
- Wang, J. *et al.* (2014) Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* 516, 405–409
- Gasperini, M. *et al.* (2020) Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.* 21, 292–310
- Liu, J. *et al.* (2017) HBL1 is a human long noncoding RNA that modulates cardiomyocyte development from pluripotent stem cells by counteracting MIR1. *Dev. Cell* 42, 333–348.e5
- Wang, J. *et al.* (2016) Isolation and cultivation of naive-like human pluripotent stem cells based on HERVH expression. *Nat. Protoc.* 11, 327–346

51. Guo, G. *et al.* (2017) Epigenetic resetting of human pluripotency. *Development* 144, 2748–2763
52. Tao, Y. *et al.* (2018) TRIM28-regulated transposon repression is required for human germline competency and not primed or naïve human pluripotency. *Stem Cell Rep.* 10, 243–256
53. Yamauchi, K. *et al.* (2020) Overexpression of nuclear receptor 5A1 induces and maintains an intermediate state of conversion between primed and naïve pluripotency. *Stem Cell Rep.* 14, 506–519
54. Goodchild, N.L. *et al.* (1993) Recent evolutionary expansion of a subfamily of RTVL-H human endogenous retrovirus-like elements. *Virology* 196, 778–788
55. Gemmell, P. *et al.* (2019) The exaptation of HERV-H: evolutionary analyses reveal the genomic features of highly transcribed elements. *Front. Immunol.* 10, 1339
56. Loewer, S. *et al.* (2010) Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat. Genet.* 42, 1113–1117
57. Ng, S.-Y. *et al.* (2012) Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *EMBO J.* 31, 522–533
58. Takahashi, K. *et al.* (2021) The pluripotent stem cell-specific transcript ESRG is dispensable for human pluripotency. *PLoS Genet.* 17, e1009587
59. Glinsky, G. *et al.* (2018) Single cell expression analysis of primate-specific retroviruses-derived HPAT lincRNAs in viable human blastocysts identifies embryonic cells co-expressing genetic markers of multiple lineages. *Heliyon* 4, e00667
60. Hsieh, C.-L. *et al.* (2014) Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. *Proc. Natl. Acad. Sci.* 111, 7319–7324
61. Allen, B.L. and Taatjes, D.J. (2015) The Mediator complex: a central integrator of transcription. *Nat. Rev. Mol. Cell Biol.* 16, 155–166
62. Trimarchi, T. *et al.* (2014) Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute leukemia. *Cell* 158, 593–606
63. Papadopoulou, T. *et al.* (2016) Dual role of Med12 in PRC1-dependent gene repression and ncRNA-mediated transcriptional activation. *Cell Cycle* 15, 1479–1493
64. Tan, S.H. *et al.* (2019) The enhancer RNA ARIEL activates the oncogenic transcriptional program in T-cell acute lymphoblastic leukemia. *Blood* 134, 239–251
65. Li, W. *et al.* (2013) Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* 498, 516–520