Regulation and Biological Consequence of Retrotransposon Activation in Human Pluripotency

by

Joseph William Martin

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Lin He, Chair
Professor Don C. Rio
Professor Jacob E. Corn
Professor Maria M. Conboy

Summer 2018

Regulation and Biological Consequence of Retrotransposon Activation in Human Pluripotency

Abstract

Regulation and Biological Consequence of Retrotransposon Activation in Human Pluripotency

by

Joseph William Martin

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Lin He, Chair

Mammalian genomes contain millions of transposable element sequences, but the overall impact they have on host biological processes remains poorly characterized. Retrotransposon insertions are known to actively shape embryonic development of model organisms, mainly by influencing transcription of host genes or by contributing functional protein or RNA products. Retrotransposons are also transcribed in human preimplantation embryos, and characterization of exapted retrotransposon families will help explain how human embryogenesis differs from other organisms. The second chapter of this dissertation presents studies on Human Endogenous Retrovirus-H (HERVH), which is a primate specific retrotransposon family functional in human embryonic stem cells (hESCs). The human genome contains over 1000 HERVH insertions and is associated with four distinct promoters, complicating efforts to understand host transcriptional regulation. Also, while HERVH RNA is clearly essential to maintain stemness, the mechanism of its action is unknown. To investigate the regulation of HERVH, we utilize bioinformatic prediction and reporter assays to identify transcription factor binding sites unique to the LTR7Y subfamily. A major finding of this dissertation is that HERVH regulation is recapitulated in naive and primed hESC culture, and the LTR7Y can be used to mark the naive state. We also find evidence that the primary mechanism of the HERVH RNA is in *trans*. We use genome engineering to delete an individual HERVH insertion and find it does not phenocopy family-wide HERVH knockdown or effect adjacent genes. Furthermore, ectopic overexpression of HERVH RNA effects the dynamics of BMP4/LY294002 induced mesendoderm differentiation, establishing a *trans* role for HERVH RNA in the postimplantation primitive streak. HERVH is the most active endogenous retrovirus (ERV) in preimplantation embryos, but we find other retrotransposons classes are more likely to modify adjacent cellular genes. In the third chapter of this dissertation, we identify a short interspersed element (SINE) insertion within ZBTB16, a developmentally important zinc-finger transcription factor. This SINE element serves as an alternative promoter, generating a transcript isoform found in the human oocyte that is capable of producing a truncated protein product. Using in vitro cell culture assays, we find the truncated ZBTB16 protein retains its function as a cell cycle regulator but shows altered subcellular localization and resistance to post translational degradation, implying it may act to slow the cell cycle of early human embryos. The studies in this dissertation serve to clarify the complex regulation and mechanisms of function for retrotransposon families that are essential for human development. These findings advance the field of hESC biology by establishing clear markers of different pluripotent cell types, and ultimately add substantially to our understanding of how retrotransposons have shaped human pluripotency.

Dedication

To my parents and to Hanna, whose love and support have made this possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Transposable elements: a brief history

Transposable elements are genetic sequences capable of replicating throughout the genomes of nearly all organisms. They have been extraordinarily proficient, comprising around 10% of the C. elegans genome([1]), 35-45% of the mouse and human genomes([2],[3]), and up to 80% of the maize genome([4]). Their success may seem somewhat paradoxical, as early work on the P and I elements in Drosophila showed the activity of mobile DNA can be harmful, causing increased mutations rates and sterility([5],[6]). The prevailing wisdom became that transposons were a class of "junk DNA" that were selfish and conferred no selective advantage([7]). However, Barbara McClintock, who had initially discovered transposable elements in the 1940s, proposed an alternative hypothesis; mobile DNA can regulate genes, and thus shape evolution. Ultimately, McClintock's ideas were validated as transposable elements were shown to be operative in a number of biological processes, including, but not limited to, the protection of telomeres in Drosophila([8]), the expression of developmental genes in rodents([9]), and the fusion of cells that occurs during placentation in mammals([10],[11]). As Christian Biémont points out in his 2010 *Genetics* review, it is not clear that genomes with a large transposon load, such as the silkworm *Bombyx mori*, have more "evolvability" than those that contain less, like the western honey bee *Apis mellifera*([12]). What is clear, however, is that transposable elements have contributed sequence that functions in normal development; some of which is shared by many organisms, and some that is species-specific. Studying transposons deepens our understanding of how mammalian evolution occurs and ultimately contributes to our knowledge of development, health, and disease. This chapter introduces the pertinent information necessary to contextualize the main chapters by first reviewing transposon regulation in development and then by detailing the current knowledge of Human Endogenous Retrovirus H (HERVH) in pluripotency.

1.2 Retrotranspons: classification, structure, regulation, and exaptation in mammals

Classification and structure

Transposable elements can be categorized into DNA transposons and retrotransposons. DNA transposons "jump" through the genome using a single or double stranded DNA intermediate([13]). They comprise the majority of transposable element content in many protozoan and insect species, but are less numerous in mammals such as mouse and human([14]). Retrotransposons replicate using an RNA-intermediate, and can be classified into three major categories; Long Interspersed Nuclear Elements (LINES), Short Interspersed Nuclear Elements (SINES), and Endogenous Retroviruses (ERVs) (Figure 1.1). LINEs, SINEs, and ERVs replicate using distinct mechanisms. LINEs contain an internal promoter that is transcribed by Pol-II and have two major open reading frames (ORF1 and ORF2). These genes code for proteins that reverse transcribe retrotransposon RNA into DNA and integrate it into the genome([15],[16]). SINES lack any open reading frames, instead relying on the proteins encoded by LINEs to retrotranspose. They are transcribed by Pol-III, and likely evolved from 7SL, the RNA component of the signal recognition particle([17]). Finally, endogenous retroviruses (ERVs) are similar in structure to exogenous retroviruses and have an internal region flanked by long terminal repeats (LTRs). The provirus is transcribed by Pol-II and encodes the proteins necessary for retrotransposition, including capsid proteins, proteins involved in retroviral DNA synthesis, integrases, and envelope proteins([18]). ERVs are self-sufficient for retrotransposition, although there exist rare examples of ERV mobilization in-trans([19]).

Detailing the structure of LTRs is important because these sequences modify the host transcriptome by serving as developmentally regulated promoters for proviral DNA, alternative promoters for cellular genes, or by functioning as classical enhancers([20],[21],[22]). Intact proviral LTRs contain U3, R, and U5 regions. The U3 sequence is located upstream of the transcriptional start site and contains a high concentration of transcription factor binding sites. This sequence is highly variable between subfamilies, and allows for cell type specific expression([23],[9]). For example, the LTR of the murine endogenous retrovirus-L (MERVL) binds a number of transcription factors active at the two cell stage, including the master 2-cell regulator Dux([24]). Human endogenous retrovirus-H (HERVH) binds Nanog and Oct4 in human embryonic stem cells([25]), and the murine-specific ERV RLTR13 binds Cdx2, Eomes and Elf5, core factors essential for trophectoderm development([9]). The core promoter region usually contains a consensus TATA box and recruits Pol-II and core transcription factors([26]). The transcription start site (TSS) defines the start of the R region, which contains sequence that stabilizes the nascent RNA. The best characterized example is of the TAR region of human immunodeficiency virus (HIV), which binds the virally encoded TAT protein and increases polymerase processivity([27],[28]). A similiar phenomenon is also observed for the murine endogenous retrovirus MLV-SL3([29]).

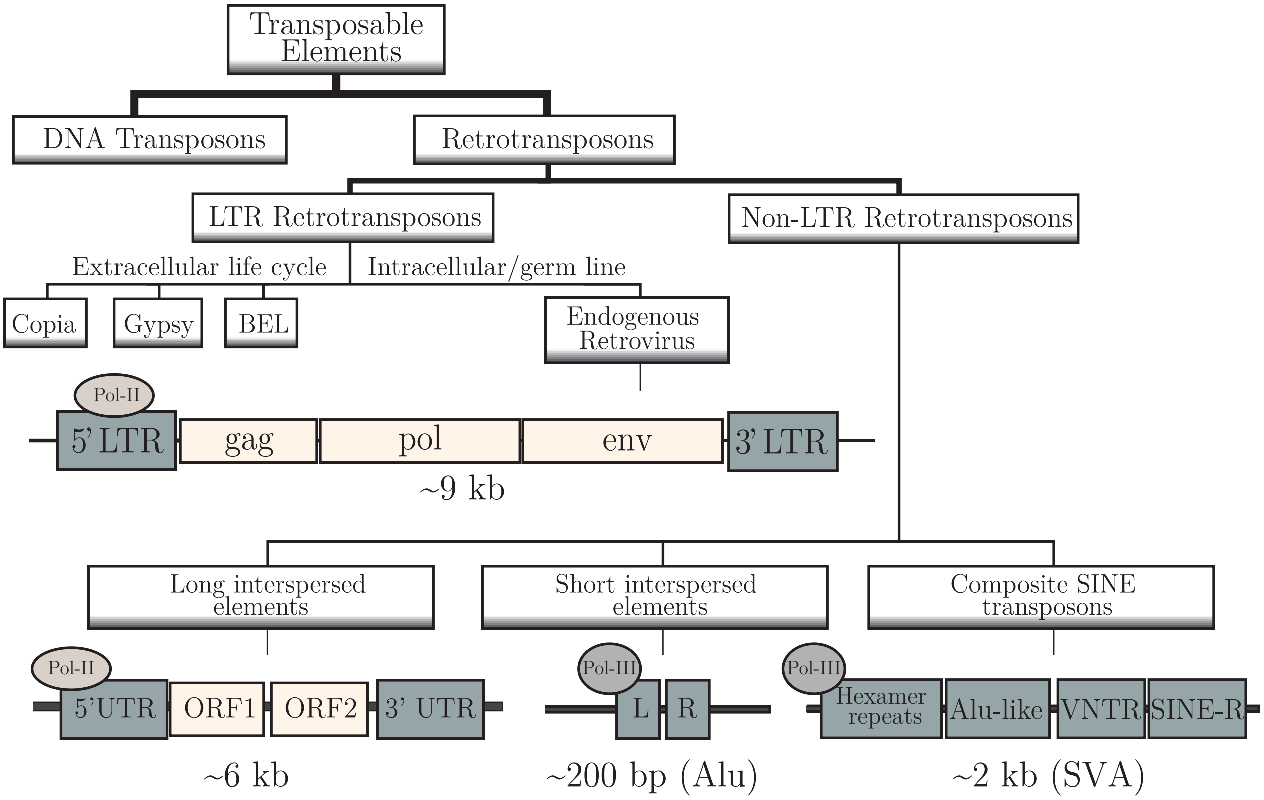


Figure 1.1: Classification and structure of retrotransposons

The organization of the major families of LTR and non-LTR retrotransposons found in Metazoa are depicted. LTR transposons are subdivided into Copia, Gypsy and BEL, which have extracellular mobility and are similar to retroviruses, and endogenous retroviruses, which are not infectious but propagate through the germ line. Non-LTR retrotransposons can be subdivided into long interspersed elements (LINEs), short interspersed elements (SINEs), and composite SINE retrotransposons.
The genomic structure of endogenous retrovirus, LINE, ALU, and SVA elements are shown. Endogenous retroviruses have *gag*, *pol*, and *env* genes flanked by LTRs and are transcribed by Pol-II. LINEs have two internal open reading frames encoding ORF1p and ORF2p polypeptides, are flanked by non repeating 5' and 3' UTRs, and are also transcribed by Pol-II. SINEs are composed of a left and right subunit derived from 7SL RNA, are transcribed by Pol-III, and are non-autonomous. Composite retrotransposons such as SVA contain more complex structures with hexamer repeats, Alu-like and SINE-like regions, and contain a Variable Number of Tandem Repeats (VNTR). Non-coding sequence is gray and coding regions are in yellow.

Host repression of retrotransposons

A defining characteristic of both DNA transposons and retrotransposons is their ability to replicate and then transmit vertically from parent to offspring. However, this activity can be harmful if it disrupts the function of an essential gene, so multiple mechanisms have evolved to prevent retrotransposition. This section will briefly mention the major post-transcriptional mechanisms that repress retrotransposons before focusing on transcriptional regulation because of its recognized role in modifying host gene expression. One major post-transcriptional protective pathway involves the Piwi-piRNA complex, which is well characterized in both Drosophila and mammals. In germ cells and early embryos, small RNAs are processed from inactivated transposable element clusters and complex with argonaute like Piwi proteins to target to complementary transposable element RNAs for cleavage(for review see[30]). In Drosophila, Piwi also interacts with linker histone H1 and Heterocromatin Protien 1 to induce heterochromatin formation around target transposable elements([31]). However, elimination of nuclear Piwi in Drosophlia ovaries does not impact host gene expression([32]). Another post-transcriptional repressive mechanism is the APOBEC family of cytidine deaminases, which target endogenous retrovirus RNA and catalyze extensive C-U mutagensis(for review see [33]). Finally, other less studied regulators include direct or indirect suppression by microRNAs (miRNAs)([34],[35]), endogenous small interfering RNAs (endo-siRNAs)([36]), and transfer RNA-derived fragments (tRFs)([37],[38]).

Retrotransposons are regulated at the transcriptional level via epigenetic DNA methylation and histone modification. In post-mitotic germ cells and somatic cells DNA methylation is critically important. The de novo methyl transferases DNMT1, DNMT3A, DNMT3B, and DNMT3L act cooperatively to methylate specific retrotransposon families. *Dnmt1* knockout mouse embryonic stem cell (mESC) lines show derepression of both ERV (IAP) and LINE (L1) families([39]), while *Dnmt3a* knockout mice show derepression of SineB1. *Dnmt3B* is also implicated in IAP and L1 repression([40]). DNMT3L lacks catalytic activity but recruits other methyltransferases to repress L1 activity([41]). However, in early embryogenesis the genomes of both mouse and human are relatively hypomethylated([42]), suggesting histone modifications are responsible for retrotransposon silencing at this stage. Deficiency of maternally deposited LSD1, a histone demethylase, leads to activation of Murine Endogenous Retrovirus-L (MERVL) in 2-cell embryos and in mESCs([43]). The H3K9me2 methyltransferase G9A is also required for MERVL repression([44]), while the H3K9me3 methyltransferase SETDB1 is necessary to repress IAP, ETn and MMERVK10C ERVs in mESC lines([45],[46]). How these repressive factors target specific retrotransposon classes is unclear, but in some cases they may complex with sequence specific binding proteins. For example, SETDB1 interacts with the transcriptional corepressor TRIM28 (also known as KAP1), and in mESCs, the SETDB1/KAP1 complex is recruited to murine leukemia virus (MLV) elements via the zinc finger protein ZFP809. ZFP809 targets the primer binding site of MLV, effectively silencing MLV transcription([47],[48]).

ZFP809 is one charactered example of a Kruppel-associated box-zinc finger protein (KRAB-ZFP), which is a major protein family responsible for repressing many retrotransposon classes. ZNF subfamilies have rapidly expanded through mammalian evolution([49]), and it is proposed this diversity results from their role in binding many retrotransposon families with unique promoter sequences. In primates, ZFP91 and ZFP93 have been shown to repress SVA and LINE1 elements([50]), providing evidence that ZNFs serve a defensive role in the evolutionary "arms race" against RT activation. Large scale efforts have attempted to identify the retrotransposon targets of each KRAB-ZFP. One effort used phylogenetic and genomic studies to identify 222 human KRAB-ZFPs, 159 of which are enriched for binding to at least one retrotransposon family([51]). It is likely many of these function as silencing factors, as binding and transcriptional repression were confirmed experimentally using a reporter system. Interestingly, this study found genes nearby retrotransposons bound by ZFP-KRABs are more active in cell types where those retrotransposons have active H3K4me1 and H3K27ac marks. Conversely, the same genes are less active in cell types where these nearby retrotransposons contain the repressive H3K9me3 modification. This correlation suggests that retrotransposon silencing can prevent aberrant activation of nearby host genes. Indeed, some cellular genes overlapping retrotransposons are upregulated in genetic backgrounds that disrupt KRAB-ZFP function([52],[53],[51]). One convincing example is in mouse, where an IAP element represses the *Zfp575* gene in a TRIM28 dependent manner([52]). While the mechanism is not entirely clear, it is likely that without ZFP575, H3K9me3 marks are lost, allowing for the deposition of active H3K4me3 and H3K27me3 marks by an unknown factor. However, this IAP directly overlaps with the 3' UTR of ZFP575, and it is still unclear if retrotransposons can actively repress gene transcription at greater distances. The advancement of genome engineering technology promises to expedite validation of candidate retrotransposon-gene pairs.

Retrotransposon exaptation

Retrotransposons have contributed to mammalian evolution in two primary ways. Firstly, the conflict between retrotransposon activation and host silencing mechanisms is a driving force that contributes to mammalian evolution, ultimately resulting in both new retrotransposon sequences and novel cellular proteins. Secondly, retrotransposon sequences have been directly co-opted by their hosts and are utilized as protein coding genes or gene regulatory elements in development.

Retrotransposons are constant conflict with their hosts. They are actively silenced through previously described mechanisms, but new elements can colonize the genome existing retrotransposons can mutate to escape repression. The risk of insertional mutagenesisis places the host under selective pressure to evolve a response, setting up a scenario for an evolutionary arms race. It has been proposed the KRAB-ZFP protein family is the result of such conditions, as it has rapidly expanded through primate evolution to become the largest transcription factor family in humans([54],[49]). While the arms race hypothesis is difficult to prove experimentally, it was shown that ∼15 MYA, ZFP93 evolved in the primate genome to repress an active LINE1 element. Then, ∼12.5 MYA, a L1PA3 insertion deleted the ZFP93 binding site responsible for its repression. This allowed rapid amplification through ape genomes before being silenced by an unknown factor, providing evidence of the mechanism that drives the co-evolution of retrotransposons and species specific zinc-finger proteins([50]).

Retrotransposon derived proteins can also be captured by the host to perform essential cellular functions. This is seen in placentation, a process that involves the cell-cell fusion of trophectoderm cells to form the syncytiotrophoblast layer that invades the uterine wall. In the viral life cycle, envelope (*env*) proteins mediate fusion between the viral and host cell membranes, and multiple env genes have been "captured" by mammals to serve this function in development. In humans, SYNCYTIN-1 and SYNCYTIN-2 are proteins derived from *env* genes of HERVW and HERV-FRD retrotransposons. Knockdown of SYNCYTIN-2 prevents cell-cell fusion in a trophoblast cell line([55]), implying this protein may be necessary for human development. Interestingly, retrotransposons have contributed to placental evolution independently in multiple organisms([56]). For example, the previously mentioned SYNCINTIN-2 is unique to primates and is produced from a relatively intact ERVFRD-1 element on chromosome 6. However, Bovidae have their own co-opted *env* gene, Fematrin-1, which is derived from the BERV-K1 ERV and mediates the formation of trinucleate cells in the placenta([57]). Trinucleate cells are unique to Bovidae and are important for supporting long gestation periods, providing evidence that exapted retrotransposons proteins can contribute to speciation.

Retrotransposon promoters contain numerous transcription factor binding sites and some have been repurposed to enhance nearby gene expression. Chuong et al([58]) investigated the role of ERV enhancers in the STAT1 mediated proinflammatory cytokine interferon-γ (IFNG) immune response. The authors identified MER41 elements adjacent to genes induced by IFNG, some of which contain numerous STAT1 binding sites. CRISPR deletion of candidate MER41 insertions decreased the induction of nearby genes, directly implicating MER41 in the human immune response. There is also evidence that retrotransposons can serve enhancers in development. In hESCs, many retrotransposons are unmethylated, marked by active histone marks, and bound by transcription factors such as ESR1, TP53, POU5F1, SOX2, and CTCF([59]). This signature is correlated with nearby gene expression and suggests productive enhancer-promoter loops, but experimental evidence is needed to identify retrotransposon derived enhancers important for mammalian development.

If a retrotransposon insertion is disrupted but the promoter remains intact, it can be repurposed to transcribe a nearby protein coding gene. Often the retrotransposon promoter alters the gene's expression pattern, changes the transcript isoform, or both. In mice, this is shown convincingly for the Dicer gene, where an intronic MTC element drives an oocyte specific Dicer isoform (Dicer-O) that lacks the n-terminal DExD helicase domain. This truncation fundamentally changes the enzyme's activity as Dicer-O, unlike the full length Dicer, can cleave double stranded RNAs into endo-siRNAs. Genetic deletion causes upregulation of siRNA targets and female sterility, showing the insertion of this MTA element in the Dicer locus has become essential for murine development([21]). Other examples include but are not limited to an ORR1A0 driven dominant negative isoform of PU.1 that is functional in erythroid differentiation([60]), and an ERV9-driven isoform of p63 that functions in the male germline to regulate DNA-damage induced apoptosis([61]). High-throughput sequencing of pluripotent stem cells and preimplantation embryos has revealed the presence of many more retrotransposon driven chimeric transcripts for protein coding genes([62],[63]), long non-coding RNAs (lincRNAs)([64]) and ERV derived antisense transcripts([65]). It is still unclear what percentage of putative chimeric transcripts may be functional. Additionally, the activation of a single retrotransposon family can coordinately activate many chimeric transcripts across the genome([66]), so it is possible that obvious phenotypes may not be observed unless multiple chimeric transcripts are disrupted. One well characterized example of coordinated gene expression comes from mouse, where MERVL is among the first sequences transcribed from the murine zygotic genome at the two cell stage. The MERVL associated LTR is found to drive not only the full length MERVL transcript but also a number of nearby cellular genes([67]). It appears as though MERVL has been wired into the transcriptional network of early development, and because it is no longer mobile, MERVL activation is not harmful but rather helps define the murine 2-cell transcriptome([68]). Another possibility is that MERVL gag protein, which is detectable at the two-cell stage, may preform some function important for the first few cell divisions of the fertilized egg([69]). Indeed, one report shows knockdown of MERVL transcripts induces arrest of 2-cell embryos([70]), but additional work is needed to establish MERVL as essential for murine development.

1.3 Stem cells as a model system to study retrotransposons

The developing embryo is the ideal system to study the effects of retrotransposon activation. However, embryos are a scarce resource, can be difficult to manipulate, and the use of human embryos in research is of ethical concern. Therefore, most research on retrotransposon function is done in-vitro using cell lines derived from the embryo, which will be reviewed here.

Much our knowledge about the molecular underpinnings of mammalian pluripotency comes from mESCs, which were first derived by culturing cells of the inner cell mass on a layer of mouse embryonic fibroblast cells([71],[72]). The defining characteristics of mESCs are their capacity to self renew and their ability to contribute to all tissues of a developing embryo. Early studies used the chimera assay and other techniques to establish the importance of cell signaling pathways stimulated by the cytokine leukemia inhibitory factor (LIF), which functions by signaling through the JAK-STAT, PI3K/AKT and MAPK pathways to maintain the expression of a complex, interconnected network of transcription factors([73],[74], for review see[75]). These proteins, which include OCT4, SOX2, and NANOG, serve as self-reinforcing core network that promote mESC self renewal and inhibit differentiation([76]). Other markers for the murine pluripotent state include widespread genomic hypomethylation([77]) and the presence of two active X-chromosomes([78]). Many retrotransposon families are regulated in mESCs. As reviewed in the previous section, the ERVs IAP, MLV, and MERVL, as well as the LINE family L1 and SINE family SineB1, are repressed ([39],[40],[41],[43],[44],[45],[46],[47],[48]). It is unclear if baseline expression of ERVs is important for pluripotency, but recently LINE1 RNA was directly implicated in mESC maintenance. Knockdown experiments revealed that in mESCs, LINE1 RNA mediates the binding of NUCLEOLIN and KAP1 to ribosomal DNA (rDNA), facilitating ribosome biosynthesis and mESC self-renewal([79]). Interestingly, LINE1 knockdown induces the master 2-cell transcription factor *Dux*, which in turn activates 2-cell specific MERVL transcripts. This mirrors many other studies that show perturbation of mESC factors can elevate levels of MERVL([67],[35], for review see([68]). Cells marked by MERVL expression are considered to have expanded cell fate because, unlike pluripotent mESCs, they can contribute to extraembryonic lineages in morula injection assays. However, a functional role for MERVL in this process remains to be established.

Like mESCs, hESCs were first derived from the inner cell mass and grown on a layer of mouse embryonic fibroblasts([80]). However, they lack a requirement for LIF and instead require the cytokine Fibroblast Growth Factor 2 (FGF2), which signals through the SMAD pathway to maintain the expression of the OCT4/SOX2/NANOG transcription factor network. Furthermore, hESCs differ from mESCs in their morphology, have one inactivated X-chromosome, and differ in their expression of pluripotency associated transcription factors such as TFCP2L1(reviewed in[81]). In fact, conventional hESCs are proposed to be more like the post implantation epiblast, and while still pluripotent, are considered “primed” for differentiation and are less capable of contributing to chimeric animals([82]). Understanding why conventional hESCs resemble the primed version of mESCs remains an active area of research, but recent work has attempted to generate ground state or “naive” hESCs that more resemble ground state mESCs and the human preimplatation epiblast. One promising effort is by Theunissen et al. 2014, who used a ground state fluorescent reporter for the OCT4 distal enhancer and screened a library of small molecule inhibitors that modulate cell signaling pathways. They found media supplemented with human LIF and inhibitors of MEK, GSK-3, B-Raf, and SRC was sufficient to activate naive state markers TFCP2L1, KLF4, and STELLA, and these preimplantation epiblast like cells were able to differentiate into all three germ layers in a teratoma assay([83]). Takashima et al. 2015 adopted a transgene approach, where ectopic expression of KLF2 and NANOG was combined with a media formulated with human LIF, protein kinase C (PKC) inhibitor Gö6983, and GSK3 inhibitor CH. These “reset”, or “naive” cells also showed expression of ground state markers, enhanced single cell recovery, and differentiated efficiently into all three germ layers([84]). Importantly, later work showed this media can be used to derive naive hESCs directly from the human embryo, without the need for transgenes([85]). In summary, conventional human embryonic stem cells are similar to postimplantation epiblast, while recently generated naive cells represent preimplantation epiblast and are more similar to mouse ESCs (Figure 1.2 A-B). While it is not known why hESCs require special conditions to reach the ground state, recent work has revealed some primate-specific retrotransposons are differentially expressed between primed and naive conditions([86]). It has been proposed that if these sequences have been exapted during evolution and are functional, they may partially explain how the human pluripotent state has diverged from other animals([87],[86]). Currently, the only retrotransposon family implicated in hESC maintenance is HERVH, which will be reviewed in detail in the next section.

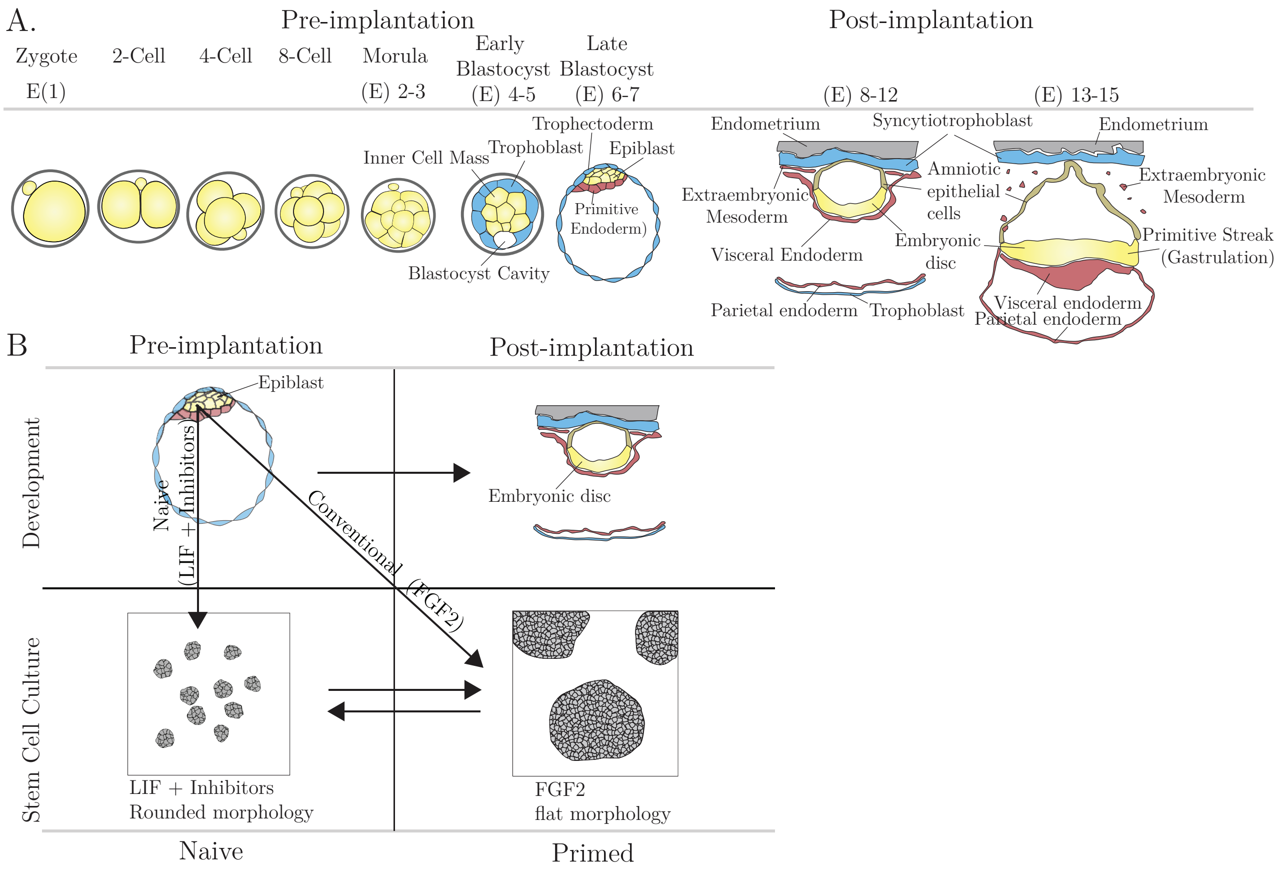


Figure 1.2: Summary of mammalian development and stem cell derivation

A. Development of human preimplantation (left) and postimplantation (right) embryos. Human embryogenesis is marked by the development of a mature blastocyst containing distinct trophectoderm (CDX2+), epiblast (NANOG+) and primitive endoderm (GATA6+) by embryonic day 6-7. After implantation, the distal epiblast cells mature into the embryonic disk while the proximal epiblast cells differentiate into amniotic epithelial cells, a cell type not found in murine embryos. The primitive endoderm differentiates into the parietal endoderm, visceral endoderm, and extraembryonic mesoderm. The trophectoderm generates the trophoblast and the syncytiotrophoblast that invades the endometrium (uterus). By embryonic day 13-15, the visceral endoderm expands and the embryonic disc further matures, forming the primitive streak which is the site of gastrulation. (E) is shorthand for embryonic day. Adapted from ([88]). For review see([88]). **B.** Diagram showing the source of human embryonic stem cells. hESCs are derived from the epiblast of human blastocysts. Under conventional culture conditions, outgrowths will generate flat monolayers with transcriptional profiles that resemble post implantation embryos and rely on TGFβ signaling. Outgrowth in media containing LIF and inhibitors results in naive hESCs with a rounded morphology, expression of naive transcription factors, and reliance on LIF with independence from TGFβ signaling. Figure adapted from Boroviak et al. 2017([88]).

1.4 Human Endogenous Retrovirus H

Background

Human endogenous retrovirus H (HERVH) is an endogenous retrovirus that is highly transcribed in hESCs cells and its non-coding RNA product has been shown to be necessary for the maintenance of the pluripotent state ([89],[87]). Although this primate specific retrovirus has clearly contributed to human evolution, elucidating the mechanism of its function has proved difficult. In order understand HERVH regulation and function, a more complete understanding of HERVH proviral structure and its evolution in the human genome is necessary.

Human endogenous retrovirus H (HERVH) bears similarity to gammaretroviruses and its possesses a classical ERV proviral structure with two LTRs flanking an internal region. The ancestral HERVH provirus likely became endogenized in higher primates approximately 30 to 35 million years ago, as HERVH-like elements can be found in both old and new world monkeys([90],[91],[92]). There is evidence of significant HERVH amplification from approximately 13 to 9 million years ago as homininae, the primate subfamily that contains gorilla, chimpanzee, and human, share a number of HERVH insertions that orangutans lack([93]). The sequence of the ancestral HERVH provirus has been approximated using an in silico approach that involves alignment of 926 relatively complete HERVH insertions from the human genome([94]). This work revealed a putative 9-kb ancestral provirus with *gag*, *pro*, *pol*, and *env* genes and a consensus histidine tRNA primer binding site. Ancestral HERVH contains a consensus TATA box, multiple SP1 transcription factor binding sites, and conserved splice donor (SD) and splice acceptor (SA) sites (Figure 1.3 A). Interestingly, the authors observed a skewed 15% G and 33%C nucleotide frequency not present in other ERVs, as well as an elongated 5' leader region between the primer binding site and the gag protein, but any functional significance of these unique features is unknown. It is likely that HERVH was silenced via extensive mutation before the split of gorilla from homininae, as there is no evidence of any unique HERVH insertions in humans or chimpanzee and no copies remain that encode for all of the full length ERV proteins([64]). Most HERVH insertions contain some DNA that originally coded for *gag* and *pol*, but almost all lack large stretches of *env* sequence([95]). The lack of predicted protein coding potential([87]), as well as the nuclear localization of HERVH in hESCs([89]), suggest the full length HERVH RNA product likely functions as a lincRNA.

HERVH can be classified into four separate subfamilies based on the primary sequence of its associated LTRs; LTR7, LTR7B, LTR7C, and LTR7Y. It is thought these subfamilies were mobile at different points during primate evolution, and LTR7Y is considered the most recent to evolve([93]). However, a comprehensive study describing HERVH subfamily radiation in primate evolution has not been published. The core promoter region surrounding the TATA box is well conserved among subfamilies, but the U3 region from 100 bp to 230 bp is highly variable (Figure 1.3 B). The U3 region of ERVs is known to be a hotbed for transcription factor binding sites([96]). Indeed, only the LTR7 family is expressed in hESCs, so it is possible a transcription factor binds a unique motif in this region. The solo-LTRs of some ERV subfamilies, such as the MERVL associated MT2C-Mm, remain active and can serve as an alternative promoter for cellular protein coding genes([67],[43]). But in contrast to complete HERVH insertions, solo LTR7, LTR7B, LTR7C, and LTR7Y loci show little evidence of transcription in either pluripotent or somatic cells([87]). The mechanistic reason for this is unknown, but it is possible solo LTR7/B/C/Y loci contain extensive mutations or the HERVH-internal region is somehow necessary to prevent cell mediated LTR silencing, possibly via an internal enhancer similar to that seen in HIV viruses([97]). In hESCs, LTR7-HERVH loci are highly active, contributing close to 2% of all poly-adenylated cellular RNA([25]). Most of this is full length, non spliced RNA representing the HERVH internal region and the transcribed portion of the 5' and 3' LTRs. However, a number of insertions show extensive splicing, either with adjacent non-coding genomic sequence or with nearby protein coding genes, creating a diverse array of unique transcripts (Figure 1.3 C).

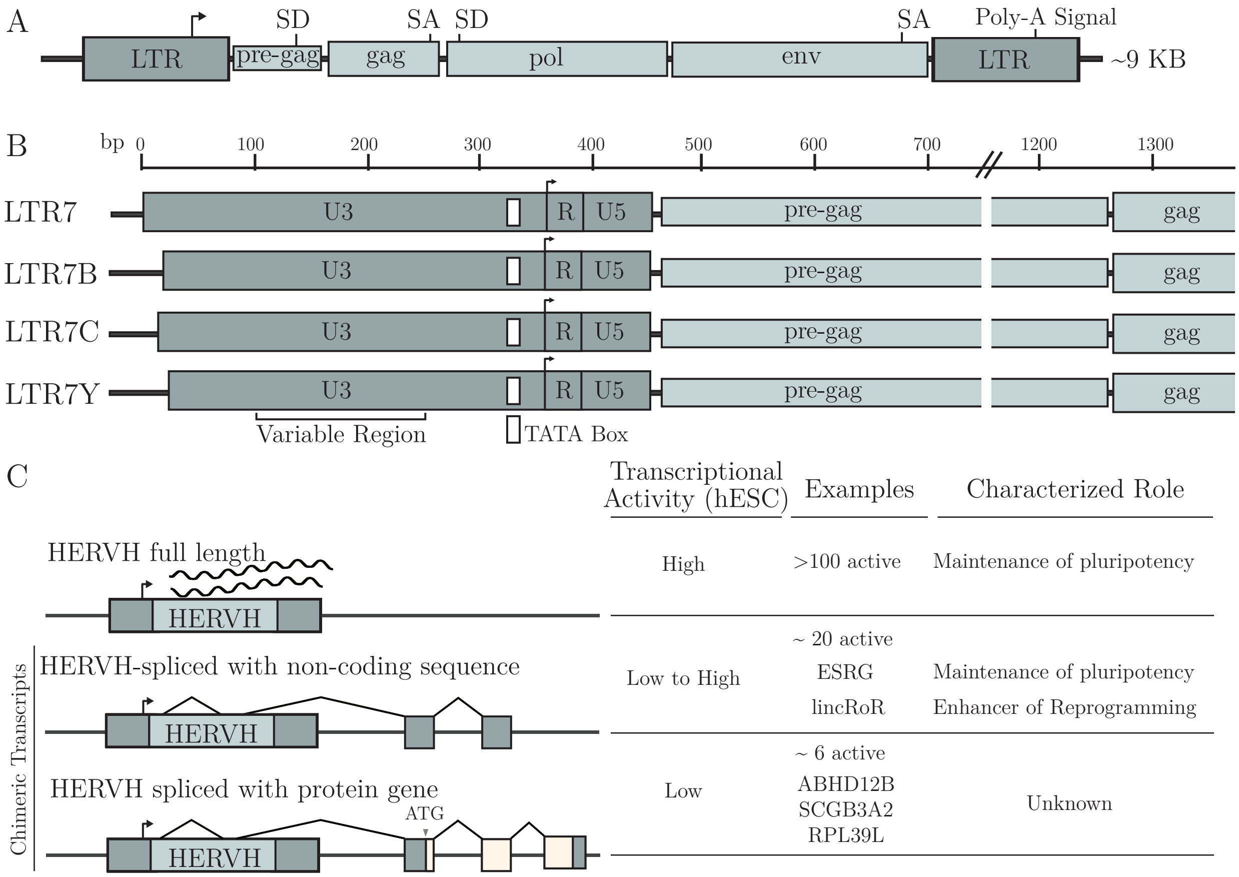


Figure 1.3: Summary of HERVH structure and HERVH derived transcripts

A. Ancestral HERVH proviral structure, containing flanking LTRs, extended pre-gag region, gag, pol, and env proteins. Identified poly-A signal site, splice donor (SD) and splice acceptor (SA) sites are shown. Drawing not to scale. **B.** Diagram of the four HERVH subfamilies, identifying the U3, R, and U5 regions, the TATA box, and the variable region within U3. Scale is in base pairs. **C.** Representative depictions of HERVH transcript classes with identified examples from the literature. HERVH insertions producing full-length RNA products are most common in hESCs, followed by HERVH spliced lincRNAs and then HERVH gene chimeras. Transcripts that splice with non-HERVH sequence are termed "chimeric transcripts". Transcriptional activity reported in primed hESCs from Wang et al. 2014([87]). ESRG characterized in Wang et al. 2014([87]). lincRoR characterized in Loewer et al. 2010([98]). The approximate number of loci in each class identified in Wang et al. 2014([87]).

Role of HERVH in pluripotency

The observation that many HERVH insertions are coordinately activated activated in pluripotent cells was first made by Jeremy Luban's group in 2012, almost 15 years after the derivation of the first hESCs line. The authors used publicly available RNA-Seq data to establish that HERVH is the most transcriptionally active ERV in hESCs (H1, H9 and I3 lines) and iPSCs (iPS-15b and iPS-11a lines), and makes up approximately 2% of all poly-adenylated RNA. HERVH activity falls rapidly upon differentiation into neural progenitor cells and is highly correlated with the pluripotency factors NANOG and OCT. Analysis of existing ChIP-Seq data showed that NANOG and OCT4 bind directly the HERVH LTR, further establishing that HERVH has been wired into the core pluripotency network([25]). The authors did not consider the individual LTR subfamilies associated with HERVH, but subsequent studies have shown the LTR7 family is most active in hESCs, suggesting the expression patterns and transcription factor binding profiles described in this study describe LTR7 insertions.

HERVH was shown to be essential for the maintenance of pluripotency by two separate groups in 2014([89],[87]). These studies induced family-wide HERVH knockdown using shRNAs targeting the internal region of multiple highly expressed copies. A 60% family-wide knockdown was sufficient to produce a loss of pluripotency phenotype in both the H1 and H9 lines, as measured by a reduction in pluripotency markers and upregulation of differentiation genes. While a detailed mechanism was not addressed, the authors suggested HERVH knockdown may effect the expression of HERVH adjacent genes, which will be discussed in more detail in the following section. These studies established HERVH as necessary for stem cells maintenance, but its importance in human development was still unproven. This was addressed by Durruthy-Durruthy et al. 2015, who characterized HERVH derived chimeric linc-RNAs in human embryos. The authors identified three lincRNAs with high expression in development; two derived from HERVH and one derived from HUERS-P1, an LTR8 retrotransposon. The authors injected siRNAs targeting all three lincRNAs into one cell of a two cell embyro and observed the injected cells did not contribute to the inner cell mass([99]). The interpretation of this result with regards to HERVH is complicated by two factors. One, the authors simultaneously targeted HERVH and HUERS-P1 lincRNAs, and thus it is possible the phenotype is caused by the HUERS-P1 knockdown. Two, the authors only analyzed three embryos for each condition, raising worries about low sample size. Despite these caveats, this work ultimately provided much needed evidence that retrotransposon lincRNAs, possibly HERVH derived ones, are important in human development.

A number of groups have found HERVH RNA enhances cellular reprogramming. Knockdown of HERVH greatly reduces the efficiency of iPSC generation([100],[87],[89]). Furthermore, overexpression of the HERVH internal region can enhance reprogramming efficiency, showing HERVH RNA can function in-trans([87]). This result may be in agreement with previous work done by John Rinn's lab at Harvard. His group identified a non-coding RNA transcript important for iPSC generation and termed it linc-Regulator of Reprogramming (linc-RoR). Linc-ROR is actually an HERVH derived lincRNA, containing sequence from the pre-gag and gag region of an HERHV insertion from chromosome 18. In agreement with the previously mentioned studies on HERVH reprogramming, linc-RoR knockdown impairs reprogramming efficiency and linc-RoR overexpression enhances it([101]). However, it is currently unknown if the function of linc-ROR is dependent its unique sequence (i.e. unique to this transcript) or on its HERVH derived sequence.

Proposed mechanisms of HERVH function

HERVH is transcribed from many insertions and produces a diverse array of full length, spliced, and chimeric transcripts. Thus, it is possible HERVH contributes to pluripotency through multiple, possibly distinct mechanisms acting in cis or in trans (Figure 1.4 A-B). The field has primarily addressed HERVH function as an enhancer or as a scaffold for cellular transcription factors.

A number of reports suggest HERVH functions by enhancing the expression of adjacent cellular genes. LTR7 can enhance the expression of luciferase when placed downstream in a plasmid([89]), suggesting it does have enhancer capability. Furthermore, multiple groups report that after shRNA knockdown, nearby cellular genes are preferentially downregulated([87],[89]), and there is at least one report that shows genes nearby HERVH loci are preferentially upregulated in reprogramming intermediates([102]). However, HERVH is known to serve as the promoter for a number of coding and non-coding chimeric transcripts. Because chimeric transcripts were not removed, these studies are describing both promoter activity and enhancer activity. Ultimately, because no example case of cis-regulation has been characterized, it is still not clear if HERVH functions as an enhancer in hESCs. Assuming that it does, it is also unknown if this effect has any impact on the maintenance of pluripotency.

To our knowledge the only experiments studying HERVH trans activity explore its potential as a scaffolding RNA. Lu et al. 2014([89]) performed RNA cross-linking and immunoprecipitation (RNA-CLIP) assays in H1 hESCs, and found that HERVH RNA is associated with activators CBP, P300, MED6, MED12, and OCT4, but not with repressors ESET, HDAC1 and PRC2. This is not surprising as HERVH is actively transcribed by these factors, but the authors confirmed this interaction via RNA-CLIP after transfection of HERVH and flag-tagged OCT4 into HEK293T cells. To our knowledge OCT4 does not have RNA binding domains, and no efforts were made to characterize the OCT4-HERVH interaction, so how the RNA may bind OCT4 is unclear. The authors speculated this interaction was important for directly regulating cellular genes, but it is still unknown if the interaction between HERVH RNA and cellular transcription factors contributes to HERVH's function in pluripotency. More unbiased approaches characterizing HERVH interaction partners could be performed using RNA-Antisense Purification (RAP), giving insight into a more comprehensive HERVH RNA interactome.

Determining the mechanism of HERVH function in pluripotency remains elusive. However, the phenotype described by multiple groups does have one commonality; published shRNAs that cause loss of stemmness all target conserved sequence in the HERVH internal region. Because the internal region is spliced out of the vast majority of chimeric lincRNAs and chimeric protein coding genes, it appears likely the full length HERVH RNA is indeed the primary effector of HERVH function. This model does have direct experimental evidence, as overexpression of a portion of the HERVH internal region can enhance reprogramming efficiency([87]). Chimeric transcripts or adjacent cellular genes may then contribute a secondary effect, becoming downregulated only after the pluripotency network has been disrupted by loss of the HERVH full length RNA.

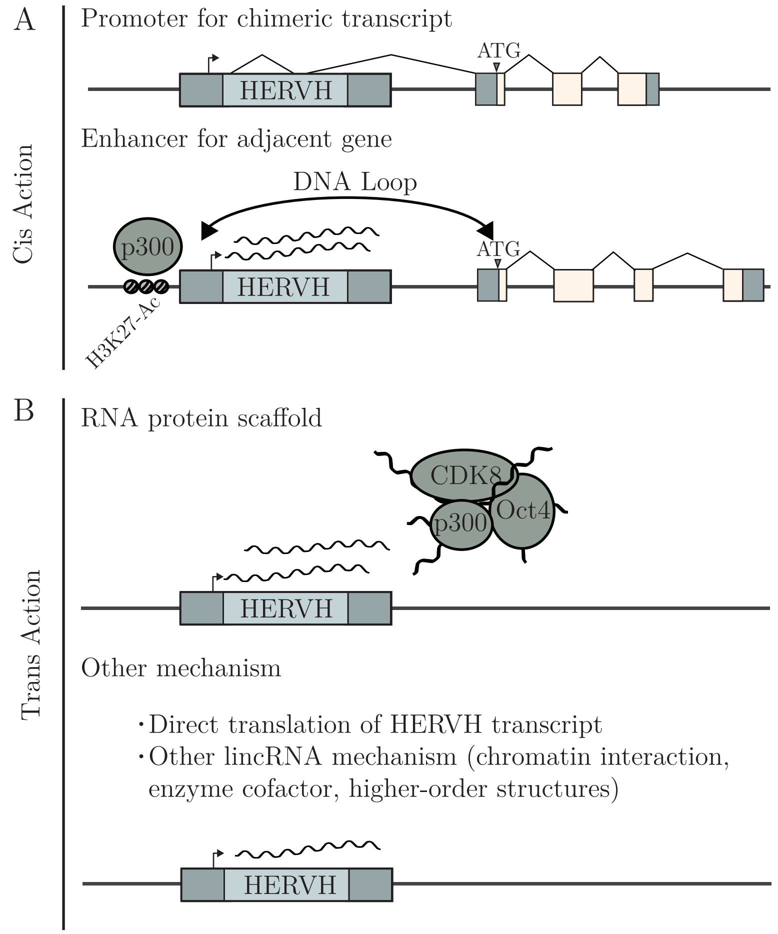

A
Cis Action
Promoter for chimeric transcript
HERVH
ATG
Enhancer for adjacent gene
DNA Loop
p300
HERVH
ATG
H3K27-Ac
B
Trans Action
RNA protein scaffold
CDK8
p300
Oct4
HERVH
Other mechanism
·Direct translation of HERVH transcript
·Other lincRNA mechanism (chromatin interaction, enzyme cofactor, higher-order structures)
HERVH

Figure 1.4: Proposed mechanisms of HERVH function

A. Cis action for HERVH function. *HERVH as an promoter for cellular genes.* LTR7 drives the expression of a number of protein coding genes uncharacterized in pluripotency, including RPL39L, ABHD12B, and SCGB3A2([**Wang2015**]), suggesting chimeric transcripts could be functional. *HERVH as an enhancer for cellular genes.* Evidence for an enhancer effect includes the observation that genes nearby HERVH loci are downregulated after HERVH knockdown([89]) as well as ChIP seq data showing binding of the enhancer related factor P300 to HERHV loci([103]). **B.** Trans action for HERVH function. *HERVH as a protein scaffold.* HERVH may bind cellular transcription factors such as OCT4, CDK8, and P300 to modify their function. *Other mechanism.* HERVH may function through another mechanism described for lincRNAs, such associating with chromatin, acting as an enzyme cofactor, or participating in higher order structures (for review see([104])).

HERVH: a marker for the naive state?

Wang et al. 2014([87]) reported that HERVH transcription, while active in primed hESCs, actually marks a rare cell population that possesses transcriptional signatures of the human naive state. To show this, the authors used the piggyBac transposon system to generate a stable HERVH-reporter line in H9 hESCs. This reporter uses the LTR7 promoter to drive GFP expression, and after generating clones containing a single copy insertion, they observed heterogeneity in GFP signal. After sorting for cells with highest GFP intensity, the authors report these cells have high levels of TBX3, NANOG, OCT4 and PRDM14, all transcription factors that are enriched in the the naive state. Furthermore, the LTR7-high cells propagate in 3i-LIF (naive) media and differentiate with a delay relative to normal primed hESCs. Finally, they report that microarray data of these cells cluster with human ICM using principle component analysis. The interpretation that LTR7-HERVH marks the naive state was drawn into question by Rudolf Jaenisch's group, who published a paper detailing transcripts differentially expressed after primed to naive reversion. They found LTR7-HERVH was not induced in the naive state and suggested other retrotransposons such as SINE-VNTR-Alu (SVA) may serve as better markers([86]). Additionally, Goke et al. reported that in human embryos, the LTR7 family is active in human epiblast (ground state) but is upregulated after outgrowth into hESCs (primed conditions). They suggested the LTR7Y subfamily may be be a better marker, and generated a LTR7Y reporter using a 3'LTR from an inactive LTR7Y insertion. While it did show basal activity in the H1 hESC line, its activity was increased in some cells after naive reversion in 3i-LIF, implying LTR7Y may be a better marker than LTR7 for the naive state. Currently there is a need to better understand the regulation of HERVH subfamilies as this promises not only to clarify how HERVH is regulated but also will assist in generating high quality naive state hESCs.

Chapter 2

HERVH

2.1 Background

Retrotransposons are mobile genetic elements that replicate using an RNA intermediate. While generally silenced via DNA methylation in somatic cells and the germ line([39], [105], [106], [107]), some endogenous retrovirus (ERVs) escape this repression in early embryonic development. Most of the transcriptionally active ERVs are no longer mobile (for review see[108]), and some families show evidence for purifying selection([43],[109],[110],[59],[111]), suggesting they may have been co-opted by the host to to play an active role in developmental processes. For example, work in mouse has revealed the murine-specific endogenous retrovirus MERVL is highly expressed at the two-cell stage and drives the expression of a number of protein coding genes that help define the two cell transcriptome([67], [68], [22]). Humans have their own unique suite of ERV insertions, and analysis of single cell RNA-sequencing of human embryos([112],[113]) reveals the many ERVs are active, including HERV9, THE1A, ERV1, HERVL, HERVK and HERVH([95]). Determining the functional importance of ERV activation in human embryogenesis is difficult because of ethical concerns surrounding the use of human embryos for research. However, the HERVH family is highly transcribed in human embryonic stem cells and is regulated by OCT4 and NANOG([25]), giving researchers a model system to study the biological consequences of ERV activation. Knockdown of HERVH RNA causes a loss of stemness phenotype([87],[89]), establishing this primate-specific ERV is essential to maintain human pluripotency.

Two main outstanding questions remain to address the role HERVH plays in human development. Firstly, the regulation of HERVH subfamilies is not well understood. HERVH-internal region is associated with four LTR subfamilies that display distinct primary sequences; LTR7, LTR7B, LTR7C, and LTR7Y (reviewed in detail in chapter 1, Figure 1.3). Only the LTR7 class is transcribed in hESCs, but one paper suggests LTR7 transcription actually marks naive stem cell state, which is more representative of the human preimplantation epiblast([87]). Other reports suggest LTR7 is transcribed in both naive and conventional (primed) hESCs, and that LTR7Y may serve a better marker for naive pluripotency([95],[86]). Elucidation of the transcription factors that regulate these subfamilies promises to resolve this discrepancy, may identify key regulators of the primed to naive transition, and ultimately will allow for the derivation of more high quality naive hESCs. Secondly, while HERVH is clearly essential to maintain human pluripotency, the main mechanism of its action is unknown. HERVH is proposed to function either in cis, through regulation of adjacent genes, or in trans, via an RNA-scaffold (reviewed in detail in chapter 1, Figure 1.4). HERVH does act as an alternative promoter for a few protein coding genes([114],[115],[20]), and it is possible that LTR7 possesses enhancer activity([89]), but whether these phenomenon explain the HERVH knockdown phenotype is unknown. There is also evidence to support a trans model of HERVH function. HERVH RNA is reported to act as a protein scaffold by interacting with OCT4, P300, and members of the mediator complex([89]). But again, it remains to be shown if this interaction is necessary for HERVH function. This chapter presents studies elucidating how HERVH transcription is regulated in development and in hESC culture, and clarifies the likely mechanisms of HERVH function.

2.2 Results - Studies of HERVH Regulation

HERVH transcription in the human embryo and in pluripotent stem cells

Recent advancements in single cell RNA-seq technology has revealed the transcriptome of individual blastomeres from human pre-implantation embryos([113],[112]). Multiple groups have described RT expression patterns by developmental stage, and they have recognized the HERVH family as the most highly transcribed ERV at the blastocyst stage([95],[87]), however, the degree to which individual HERVH loci expression is recapitulated in pluripotent stem cell culture has not been systematically addressed.

The RepeatMasker annotation identifying human retrotransposon sequences separates the HERVH internal region from adjacent LTRs, complicating RNA-seq quantification. In order to assign RNA-seq reads that fall inside the internal region to the proper subfamily, we wrote a python script that rationally annotates full length, incomplete, and solo-LTRs (see methods for details). Our analysis reveals 2,615 independent HERVH insertions in the human genome. When considering all subfamilies together, we find similar numbers of solo-LTR and complete structures, while the incomplete and internal only structures are relatively rare. LTR7 is the largest subfamily, but the LTR7B, LTR7C, and LTR7Y subfamilies have a higher percentage of complete insertions (Table 2.1).

Table 2.1: HERVH insertions in the human genome

	All	LTR7	LTR7B	LTR7C	LTR7Y	Complex
Complete	946	639	95	42	51	20
Solo LTR	1286	664	393	183	71	91
Incomplete	383	199	82	25	14	10
Internal Only	36					
Total	2615	1502	570	250	136	121

Complete insertions are defined as 2 LTRs flanking an internal region. Incomplete LTRs are defined as insertions with an LTR present on only one side an internal region. Solo-LTRs have no associated internal region. Complex insertions are defined as those having LTRs of at least two different subfamilies.

We then asked to what extent HERVH subfamilies may contribute to overall HERVH expression in human development and in ESCs. Single-cell RNA-seq data is published for every stage of human preimplantation development([112]). We used our custom HERVH annotation to quantify the expression of HERVH subfamilies from the zygote through blastocyst stages, as well as in multiple hESC lines. We find distinct patters of expression for each subfamily. LTR7 is activated from the 2-4 cell stages, and then again only in the epiblast cells of the blastocyst, and further increases after the outgrowth of epiblast into hESCs. LTR7B activity peaks at the 8-cell stage, while LTR7C is restricted to the pre-blastocyst stages. LTR7Y is expressed in all cells of the blastocyst, including trophoblast cells, and is silenced after the outgrowth epiblast into hESCs (Figure 2.1 A). In summary, it appears HERVH subfamilies show cell type specific expression, but together HERVH assocated LTRs express HERVH throughout preimplantation development and into hESC culture.

We next explored the expression of individual loci. Visualizing differentially expressed HERVH loci in heatmap form reveals the LTR7 insertions active in early development are largely different than those expressed at blastocyst and in hESCS (Figure 2.1 C). However, most active LTR7 loci in the blastocyst are also upregulated in hESCs.

To validate these findings, we turned to available cell culture systems. Conventional, or “primed” hESCs are thought to model post implantation epiblast, while ground state, or “naive” hESCs are proposed to resemble preimplantation epiblast([116]). We reverted wild type hESCs to the naive state using 5i/Lif/ActA media and tested for LTR7 and LTR7Y transcription using qPCR primers targeted at the family level and at individual loci. In agreement with our predictions from the RNA-seq of the human embryo, overall HERVH levels are slightly reduced in naive cells. We find that while LTR7 lincRNAs are downregulated, multiple LTR7Y loci are significantly increased, suggesting the overall reduction of HERVH transcription is due to LTR7 (Figure 2.1 B). We detected no amplification of spliced LTR7Y lincRNAs in primed cells, but did detect amplification when using primers targeted at multiple LTR7Y family members. This suggests that some LTR7Y loci are extremely specific to the naive stage, while other LTR7Y are expressed in both cell types. An alternate explanation is that LTR7Y family primers have some off target amplification.

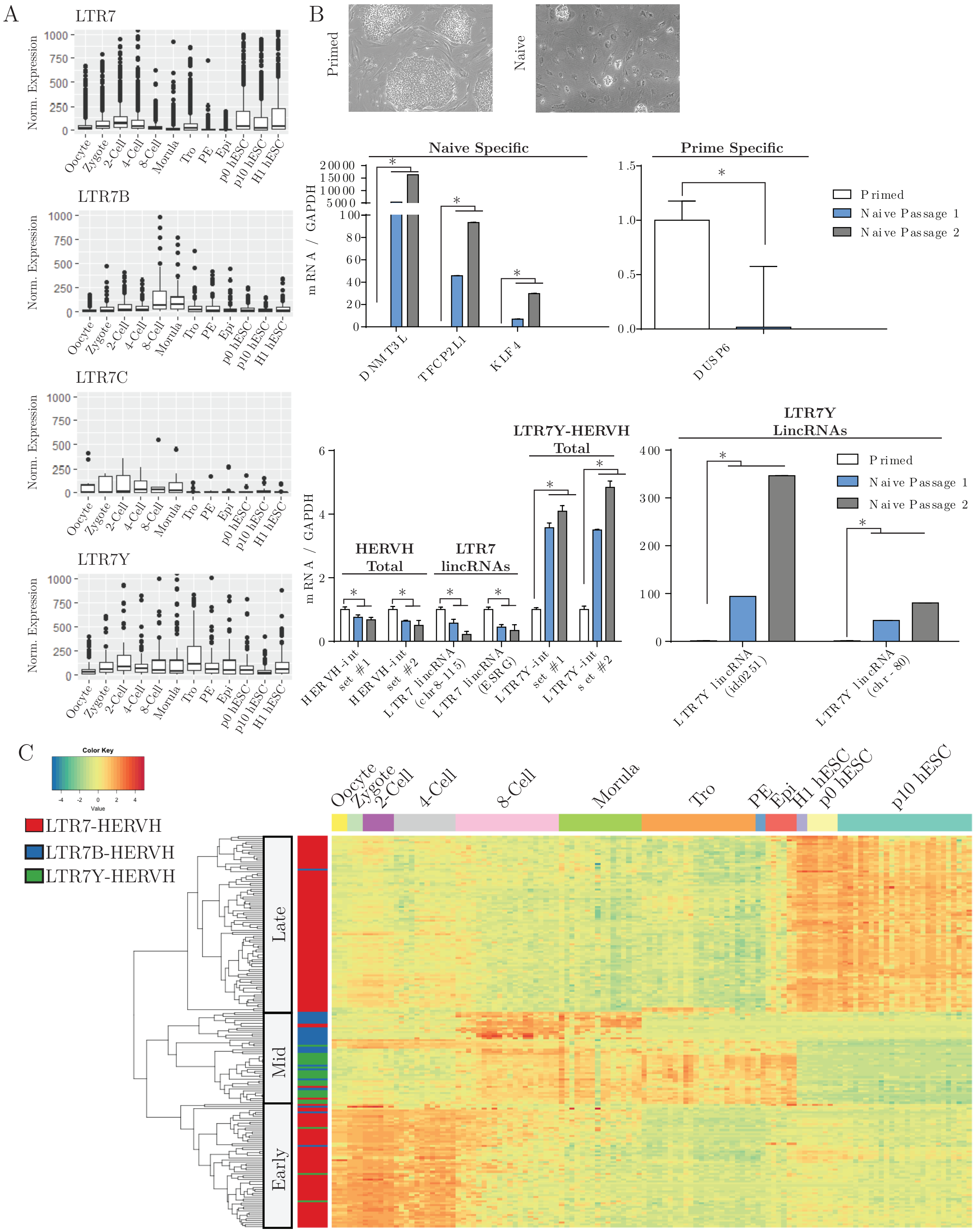

A
LTR7
LTR7B
LTR7C
LTR7Y
Norm. Expression
1000
750
500
250
0
Oocyte
Zygote
2-Cell
4-Cell
8-Cell
Morula
Tro
PE
Epi
p0 hESC
p10 hESC
H1 hESC
B
Primed
Naive
Naive Specific
Prime Specific
mRNA / GAPDH
20000
15000
10000
5000
100
80
60
40
20
0
*
DNMT3L
TFCP2L1
KLF4
1.5
1.0
0.5
0.0
DUSP6
Primed
Naive Passage 1
Naive Passage 2
HERVH Total
LTR7 lincRNAs
LTR7Y-HERVH Total
LTR7Y LincRNAs
6
4
2
400
300
200
100
HERVH-int set #1
HERVH-int set #2
LTR7 lincRNA (chr8-115)
LTR7 lincRNA (ESRG)
LTR7Y-int set #1
LTR7Y-int set #2
LTR7Y lincRNA (id:0251)
LTR7Y lincRNA (chr-80)
C
Color Key
-4
-2
0
2
4
Value
LTR7-HERVH
LTR7B-HERVH
LTR7Y-HERVH
Late
Mid
Early
Oocyte
Zygote
2-Cell
4-Cell
8-Cell
Morula
Tro
PE
Epi
H1 hESC
p0 hESC
p10 hESC

Figure 2.1: HERVH transcription dynamics in human embryos and hESCs

A. Expression of HERVH by subfamily in preimplantation development and in hESCs. Each data point is the average expression of an HERVH insertion among all single cell blastomers from that stage, displayed as box plot with outliers shown as points (data from([112]). Tro - trophectoderm, PE - primitive endoderm, Epi - Epiblast. **B.** Experimental validation of RNA seq predictions using primed and naive hESC culture. After reversion from primed (FGF2 media) to naive (5i/LIF/ActA media), naive hESC colonies were recovered and expression of naive and primed genes were assayed by qPCR for two passages. Data presented is representative data from three independent primed to naive reversions. Error bars are SEM from three technical measurements. **C.** Heat map of differentially expressed HERVH insertions from same dataset as in A. Differentially expressed HERVH loci are displayed in rows and sorted with heiarchical clustering. Developmental time is fixed on the x-axis. LTR7C was excluded due to its relatively low expression. Tro - trophectoderm, PE - primitive endoderm, Epi - Epiblast.

Sequence Alignment of expressed HERVH insertions

We next asked if the active subfamilies contained obvious differences in their primary sequences that may account for their differential expression or suggest divergent function. We aligned the LTR7, LTR7Y, and LTR7B loci expressed in naive hESCs and compared them to the ancestral HERVH sequence([94]). Interestingly, the sequence encoding for *gag* and the first half of *pol* is relatively intact, with large stretches of over 95% sequence identity between all subfamilies. The region directly preceding *gag* is marked by a number of small insertions, all of which appear in at least two HERVH loci. We believe this region was likely intact in the ancestral provirus, possibly encoding for a larger *gag* protein, and has since been extensively mutated. We find large deletions in the second half of the pro gene, and very few HERVH copies have *env* sequence, confirming these insertions are likely non-coding (Figure 2.2 A). While the majority of internal sequence is conserved between subfamilies, a subset of LTR7Y loci are divergent in their pre-*gag* and pro regions (Figure 2.2 B). It is possible these insertions represent an independent radiation event. Interestingly, a subset of LTR7Y loci appear similar to LTR7B in their 5' LTR promoters, which we term “LTR7Y-2” or “B-Like” (Figure 2.2 C). To determine if LTR7Y-2/B-like is regulated differently, we compared LTR7 and LTR7Y-2/B expression from primed and naive cells([86]). We find that when compared to LTR7, both LTR7Y subgroups are more specific to the naive state (Wilcoxon test, $p< 0.001$). When comparing between LTR7Y subgroups, we find LTR7Y-2/B-like loci are on average higher upregulated than LTR7Y-1, but not significantly so (Figure 2.2 D). However, 5 out of 6 of the most differentially expressed loci belong to the LTR7Y-2/B-like group, suggesting it may be useful to subdivide LTR7Y when analyzing other datasets.

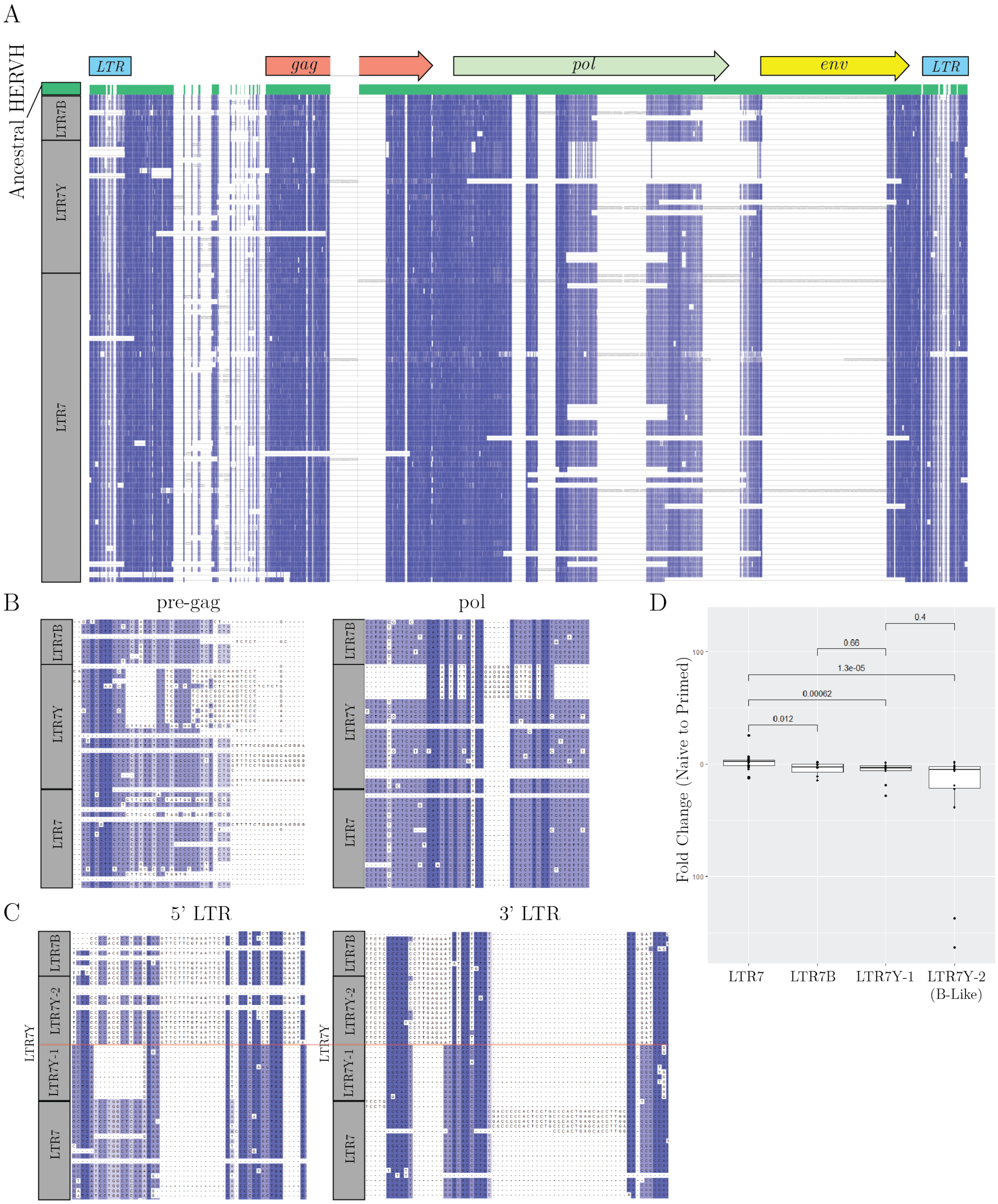

A
Ancestral HERVH
LTR
gag
pol
env
LTR
LTR7B
LTR7Y
LTR7
B
pre-gag
pol
LTR7B
LTR7Y
LTR7
C
5' LTR
3' LTR
LTR7B
LTR7Y-2
LTR7Y-1
LTR7
LTR7Y
D
Fold Change (Naive to Primed)
100
0
100
0.4
0.66
1.3e-05
0.00062
0.012
LTR7
LTR7B
LTR7Y-1
LTR7Y-2
(B-Like)

Figure 2.2: Primary sequence alignment of HERVH subfamilies

A. Primary sequence alignment of LTR7, LTR7B, and LTR7Y HERVH insertions expressed in naive hESCs compared to the ancestral HERVH consensus sequence (green, on top). Gaps in alignment caused by a single sequence were removed. **B.** Example of divergent sequences within the pre-*gag* and *pro* internal regions, with subset of LTR7 loci shown. **C.** Examples of divergent sequences within the 5' and 3' LTR regions, with subset of LTR7 loci shown. **D.** Box and whisker plot showing the fold change of HERVH loci after reversion of primed to naive cells. Negative fold change value signifies upregulation in naive cells and positive fold change value indicates upregulation in primed cells. Pairwise comparisons are performed using a Wilcoxon rank sum test. RNA-Seq data from Theunissen et al. 2016([86]).

Regulation of LTR7Y transcription in the naive state

We find LTR7Y activity is specific to naive state hESCs. To confirm this is the result of biological regulation, we aimed to determine protein factors important for its transcription. Presumably, LTR promoters that obtain mutations in important transcription factor binding sites will be lose activity or become silenced. Using this logic, the DNA sequence of highly transcribed LTR7Y promoters was compared to inactive loci using the HOMER motif analysis tool. A number of known transcription factor sequence motifs were significantly overrepresented (supplementary Table 2.2). One promising candidate is the CCCCACCC motif which is found one time in LTR7Y loci and in three tandem copies in LTR7Y-2/B-like and LTR7B loci (Figure 2.3 A, B). This is the consensus binding sequence for KLF4 and KLF17, transcription factors that are markers of the naive state([86]). Furthermore, this motif is absent in LTR7 promoters, suggesting it may contribute to differences observed in the expression pattern of LTR7 and LTR7Y (Figure 2.3 B).

To determine the importance of the CCCCACCC motif to LTR7Y transcription, we cloned LTR7Y promoters containing all three motifs into a reporter vector driving tdTomato, as well as a mutant reporter lacking all three motifs (LTR7Δ). We also cloned an LTR7 reporter (similar to those previous published([87])), for comparison. In agreement with our previous results, the LTR7 reporter was highly active in almost every cell in primed hESCs, while the LTR7Y reporter showed no signal. Reversion to the naive state activated both LTR7Y and LTR7 promoters, while LTR7Δ showed no activity, suggesting these motifs are essential for LTR7Y transcription (Figure 2.3 D). Of note, the LTR7 reporter shows a higher signal in the naive state compared to the LTR7Y reporters, which is surprising because RNA-seq data predicts these two subfamilies are transcribed at roughly equal levels.

To address if KLF4 is sufficient to drive the expression of LTR7Y, we cloned KLF4 into the piggyBac system and overexpressed it in hESCs. We observed widespread cell death and were not able to recover cells with stable KLF4 expression, suggesting the expression of this naive specific transcription factor is not compatible with primed culture. However, overexpression of KLF4 in HEK293Ts was sufficient to activate silenced LTR7Y loci, but not LTR7 loci (Figure 2.3 C). In all, these data suggest KLF proteins, possibly KLF4 or KLF17, specifically activate LTR7Y transcription via a conserved CCCCACCC motif.

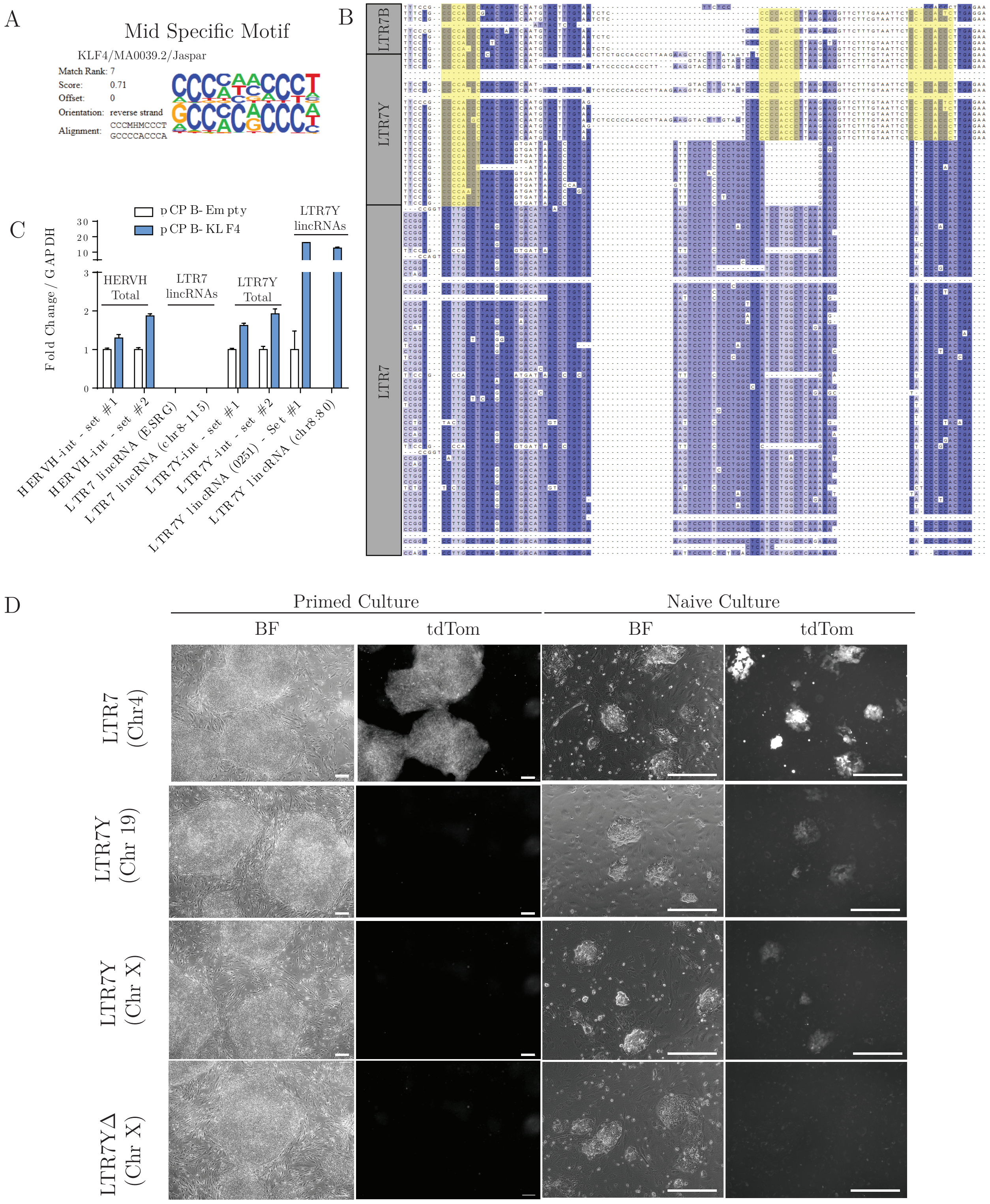

A
Mid Specific Motif
KLF4/MA0039.2/Jaspar
Match Rank: 7
Score: 0.71
Offset: 0
Orientation: reverse strand
Alignment: CCCMHMCCCT
GCCCCACCCA
B
LTR7B
LTR7Y
LTR7
C
pCPB-Empty
pCPB-KLF4
Fold Change / GAPDH
0
1
2
3
10
20
30
HERVH Total
LTR7 lincRNAs
LTR7Y Total
LTR7Y lincRNAs
HERVH-int - set #1
HERVH-int - set #2
LTR7 lincRNA (ESRG)
LTR7 lincRNA (chr8-115)
LTR7Y-int - set #1
LTR7Y-int - set #2
LTR7Y lincRNA (0251) - Set #1
LTR7Y lincRNA (chr8:80)
D
Primed Culture
Naive Culture
BF
tdTom
BF
tdTom
LTR7 (Chr4)
LTR7Y (Chr 19)
LTR7Y (Chr X)
LTR7YΔ (Chr X)

Figure 2.3: Regulation of LTR7Y transcription in naive and primed hESCs

A. KLF4 motif enriched in active LTR7Y loci, generated from the HOMER motif analysis tool. The primary sequence of active LTR7Y LTRs were used as querry and inactive LTR7Y loci were used as background. **B.** Sequence alignment showing expressed LTR7B, LTR7Y, and LTR7 loci from naive hESCs. The identified CCCCACCC motif is highlighted in yellow.
C. Overexpression of KLF4 in HEK293Ts. pCPB-KLF4 or control vector was transfected into HEK293Ts, and after 48 hours cells were assayed for HERVH transcripts via qPCR. Data is representative of two independent experiments. Error bars are standard error of three technical measurements. **D.** Reporter vectors in primed and naive hESCs. Reporters lines with LTR7, LTR7Y, or LTR7YΔ driving tdTomato were generated in primed culture and then revered to Naive culture conditions. Naive hESC images taken 10 days after beginning naive reversion using brightfield and fluorescence microscopy. Scale bar is 100 μM. D. Timecourse of mesendoderm differentiation. Mesendoderm differentation was induced with a BMP4/Ly based protocol and pluripotency, mesendoderm, and HERVH transcripts were determined at various time points with qPCR. Data is representative from 3 independent experiments. Error bars are standard error of three technical measurements.

Regulation of LTR7 in Primed Pluripotency

An existing report suggests primed hESCs with high LTR7 transcription are linked to the naive state([87]). However, our data and other groups have observed LTR7 is highly active in the primed state, suggesting LTR7 may not be a good naive marker([86]). We reanalyzed the published microarray data from HERVH high cells, and found they do not show upregulation of naive transcription factors such as KLF17 and DUSP1([86]), but rather genes involved in the specification of the primitive streak, including T, EOMES, and MESP1 (Figure 2.4 A, Table 2.2). To further investigate this we generated clonal cell lines for our LTR7 promoter, and observed clones that showed reporter activity in every cell, but with heterogeneity in signal intensity within a colony (Figure 2.4 B). After sorting into high and low populations, we find enrichment for primitive streak markers but not naive markers (Figure 2.4 C). HERVH derived transcripts have been shown to be essential for the maintenance of the pluripotent state, but its potential role in the specification of early cell fate decisions has not been explored. Because there is enrichment for primitive streak markers after sorting for LTR7-high cells, we hypothesized HERVH may remain active after exit of pluripotency into mesendoderm, which is a precursor to endoderm and mesoderm and representative of the embryonic primitive streak([117]). We subjected hESCs to mesendoderm differentiation mediated by BMP4 and PI3 kinase inhibitor LY294002. In agreement with previously published reports([118]), we observed an early peak in mRNA level for the pluripotency factor NANOG, as well as induction of mesendoderm transcripts EOMES, GSC, MESP1, T (BRACHYURY) TBX6, and MIXL1. LTR7 was induced after only 24 hours and remained elevated throughout the treatment (Figure 2.4 D). This induction appears specific, as multiple LTR7Y lincRNAs were not detected.

Table 2.2: Gene ontology terms enriched in upregulated genes in LTR7 high cells

GO biological process complete	expected	Fold Enrichment	raw P value	FDR
Endodermal cell fate specification	0.04	84.17	1.65E-05	1.51E-02
Primitive streak formation	0.07	45.91	6.98E-05	3.63E-02
SMAD protein signal transduction	0.36	16.83	2.56E-06	9.98E-03
Somatic stem cell population maintenance	0.31	15.88	2.36E-05	1.85E-02
Mesoderm formation	0.4	12.56	6.76E-05	3.64E-02
Regulation of anatomical structure morphogenesis	5.93	2.87	9.91E-05	4.69E-02
Positive regulation of developmental process	7.75	2.71	3.11E-05	2.31E-02
Positive regulation of multicellular organismal process	9.37	2.46	5.70E-05	3.56E-02
Regulation of multicellular organismal development	11.23	2.32	5.25E-05	3.42E-02
Animal organ development	18.28	1.97	6.03E-05	3.49E-02
Negative regulation of biological process	30.22	1.72	2.15E-05	1.76E-02

False Discovery Rate < 0.05
Data from Wang et al. 2014([87])

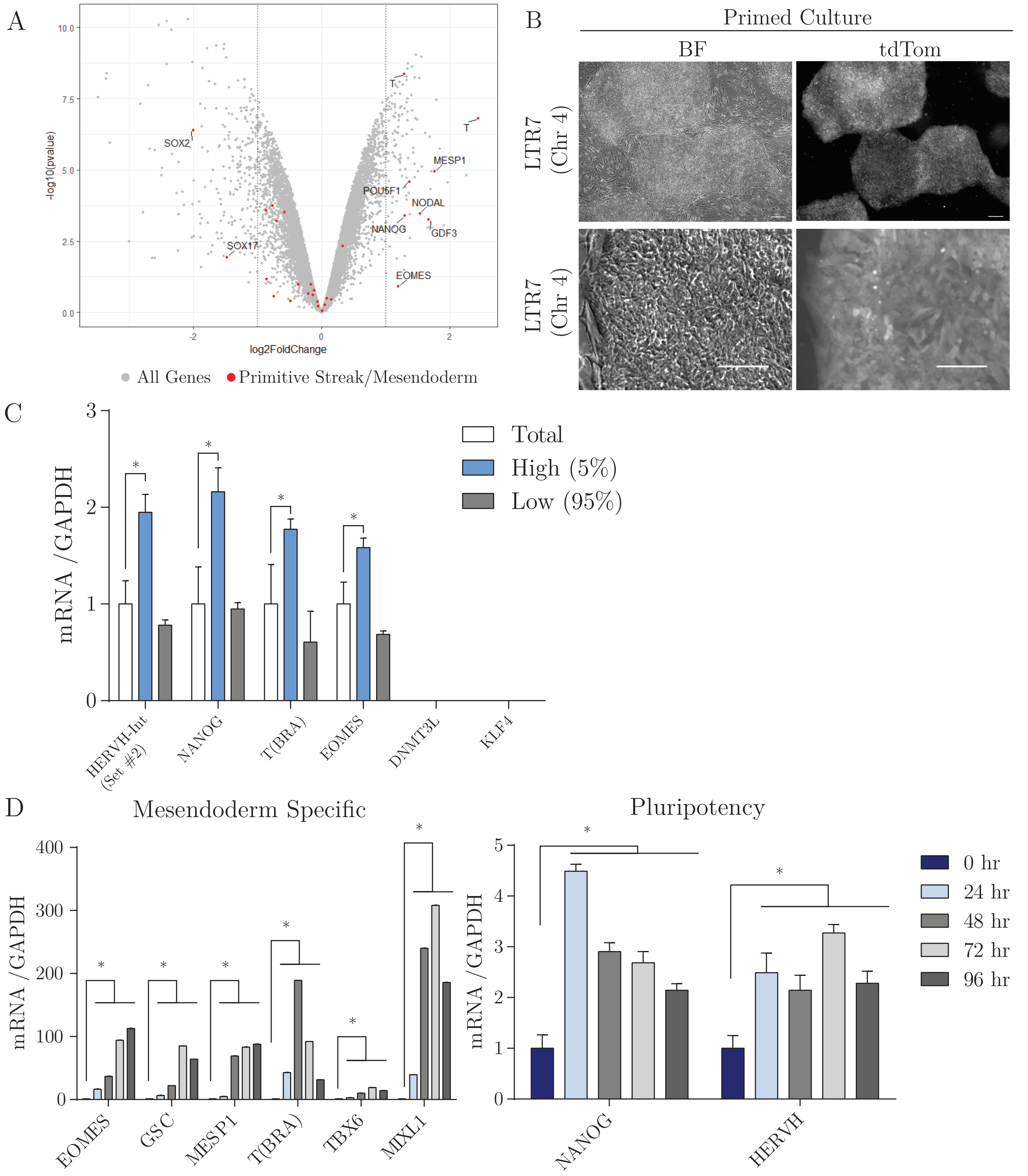

A
10.0
7.5
5.0
2.5
0.0
-log10(pvalue)
-2
0
2
log2FoldChange
T
T
SOX2
MESP1
POU5F1
NODAL
NANOG
GDF3
SOX17
EOMES
● All Genes ● Primitive Streak/Mesendoderm
B
Primed Culture
BF
tdTom
LTR7 (Chr 4)
LTR7 (Chr 4)
C
mRNA /GAPDH
3
2
1
0
*
*
*
*
Total
High (5%)
Low (95%)
HERVH-Int (Set #2)
NANOG
T(BRA)
EOMES
DNMT3L
KLF4
D
Mesendoderm Specific
mRNA /GAPDH
400
300
200
100
0
*
*
*
*
*
*
EOMES
GSC
MESP1
T(BRA)
TBX6
MIXL1
Pluripotency
mRNA /GAPDH
5
4
3
2
1
0
*
*
NANOG
HERVH
0 hr
24 hr
48 hr
72 hr
96 hr

Figure 2.4: LTR7 activity in primed hESCs and mesendoderm

A. Volcano plot of gene expression after sorting for LTR7 high cells. Genes involved in primitive streak formation are highlighted in red and labeled. Data from Wang et al. 2014([87]). **B.** LTR7 Reporter vector in primed hESCs. Clonal LTR7 reporter lines driving tdTomato were generated and imaged with brightfield (BF) and fluorescence microscopy. LTR7 sequence is from a highly expressed insertion on chromosome 4. Bottom panels show closeup of hESC colony to reveal heterogeneity in signal. Scale bar is 100 μM. **C.** qPCR analysis of LTR7 high vs low cells (sorted top 5% vs lower 95% of cells). Data is representative from three independently generated clones and error bars are standard deviation from three technical measurements. **D.** Timecourse of gene expression during BMP4/LY294002 induced mesendoderm differentiation. Data presented is representative from 4 independent experiments. Error bars are standard error from 3 technical measurements.

Regulation of LTR7 in mesendoderm specification

NANOG is proposed to regulate LTR7 transcription via directing binding to the LTR7 promoter, and LTR7 regulation is highly correlated with NANOG expression([25]). However, during mesendoderm differentiation, LTR7 remains elevated at 96 hours while NANOG trends down. Therefore, we hypothesized that additional transcription factors may regulate LTR7 during this process. To address this we analyzed existing ChIP-seq data generated by Faial et al. 2015([119], accession GSE60606). The authors differentiated the H9 hESC line to mesendoderm using similar Bmp4 / LY294002 based protocol and performed ChIP-seq at 36 hours using an antibody against endogenous Brachyury protein. Because the original analysis did not consider retrotransposon sequences, we reannotated the called peaks using the Homer Annotate Peaks software. We found significant enrichment of peaks nearby to or overlapping with LTR7 loci (p=2.35e-5, hypergeometric probability test), suggesting direct binding of Brachyury to LTR7 during Bmp4/Ly induced differentiation. Brachyury binding appears specific to LTR7, as ChIP-seq peaks do not significantly overlap with HERVK, an ERV of similar size to HERVH([66]) elements (p=0.35). We visually inspected the Brachyury ChIP-Seq peaks and identified they are centered around both 5' and 3' LTR7s (Figure 2.5).

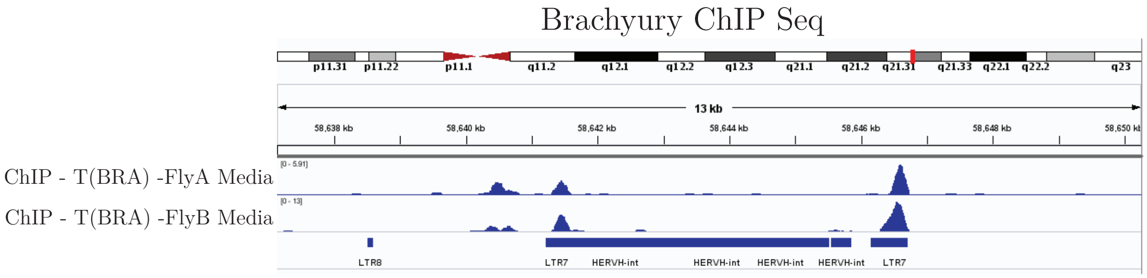


Figure 2.5: Brachyury binds LTR7 in mesendoderm differentiation

Visualization of Brachyury binding to LTR7 in mesendoderm. ChIP seq data from Faial et al. 2015([119]) was analyzed for repetitive element content. Fly-A media contains FGF2, LY294002, and Activin A. Fly-B media contains FGF2, LY294002, and BMP4. Visualization done using IGV browser and shows a representative HERVH insertion on chromosome 5.

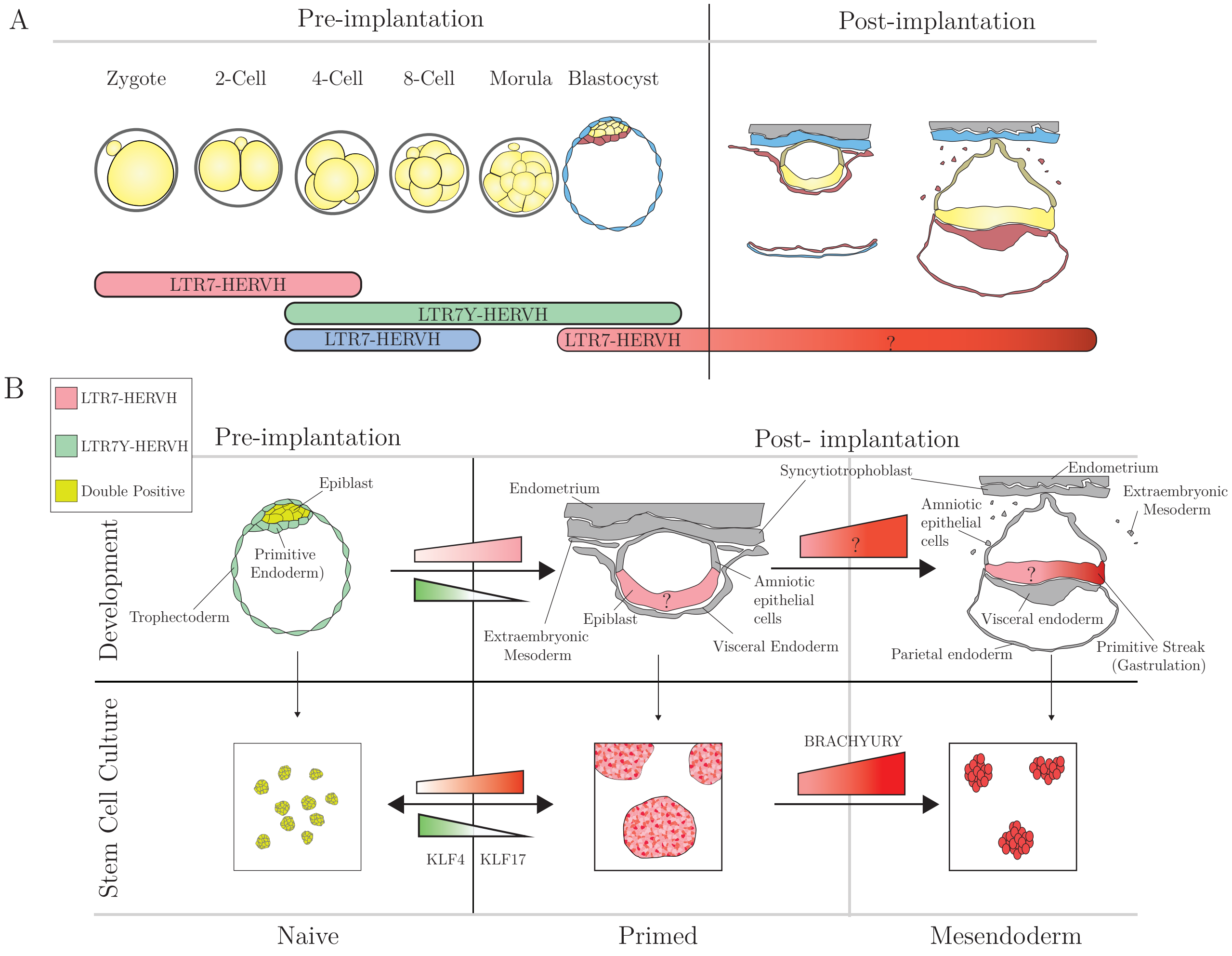


Figure 2.6: Model of HERVH regulation

A. HERVH regulation in preimplantation development. LTR7, LTR7B, and LTR7Y combine to express HERVH RNA throught every cell of preimplantation development. Human post implantation RNA seq data does not exist, but we propose LTR7 is expressed post implantation because LTR7-HERVH transcripts increase during hESCs derivation and again during mesendoderm differentiation. **B.** LTR7 and LTR7Y regulation in development and in hESC culture. LTR7 and LTR7Y are both active in the epiblast and in naive stem cell culture. The transition from primed to naive cells upregulated silenced LTR7Y loci, and this is likely dependent on KLF4 or KLF17. Primed hESCs show heterogeneous LTR7 expression and high cells are marked by a primitive streak gene signature. Upon mesendoderm differentiation, LTR7Y remains silenced but LTR7 is further upregulated, a process possibly regulated by direct binding of Brachyury to the LTR7 promoter. While RNA-seq data does not exist for human post implantation, we speculate LTR7-HERVH may be active at the primitive streak at the onset of gastrulation.

2.3 Results - Studies of HERVH Function

Functional validation of LTR7Y transcripts in naive pluripotency

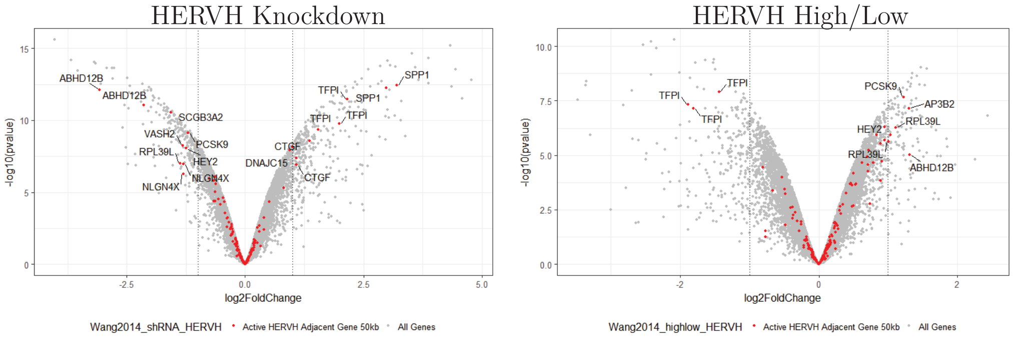
LTR7 is necessary to maintain pluripotency, but it is currently unknown if LTR7Y RNA is also functional. There are a multiple of reasons to believe LTR7Y may be necessary for the maintenance of naive pluripotency even while LTR7 is expressed. One, LTR7 is downregulated in naive cells, so LTR7Y transcription may be important to maintain HERVH levels over a critical threshold. Two, there are differences in transcribed sequences within the the pre-*gag* and *pol* internal regions that may bind distinct protein effectors. Finally, LTR7Y is obviously located next to different cellular genes and LTR7Y cis-regulation could be important for the naive state. We attempted to address a potential functional difference between LTR7 and LTR7Y loci by specifically targeting LTR7Y with shRNAs before reverting primed cells to naive. Due to limited sequence differences between LTR7 and LTR7Y internal regions, only 10 shRNAs were designed, none of which successfully knocked down LTR7Y during naive reversion (data not shown). Investigating LTR7Y function in naive cells using a knockdown strategy seems non-feasible with our shRNA systems. CRISRPi, which silences genes by recruiting inactive Cas9-KRAB to promoters and enhancers([120],[121]), could theoretically target the CCCCACCC motif and provide an alternative method for future work.

Analysis of adjacent gene expression in primed hESCs

Previous studies of HERVH suggest it functions to maintain the pluripotent state in cis by acting as a classical enhancer or an alternative promoter to regulate adjacent protein coding genes and lincRNAs([89],[87],[102]). To investigate a potential enhancer role, we reanalyzed published microarray data of hESCs after HERVH knockdown as well as hESCs sorted into HERVH-high and HERVH-low populations. We hypothesized that if HERVH functions as an enhancer, nearby genes will decrease after HERVH knockdown and increase in cells with high HERVH transcription.

Using the GREAT gene association tool (http://great.stanford.edu/), we identified the two closest genes within 50 kb of active HERVH loci. In agreement with previous results([89]), we find genes nearby HERVH loci are significantly downregulated after HERVH knockdown (Chi-square with Yates' correction, p=1.2e-06). However, inspection of these genes reveals that ABHD12B, SCGB3A2, RPL39L, and PCSK9 are HERVH-gene chimeras, where the transcript is under the control of the LTR promoter. Removing these genes makes the enrichment no longer statistically significant(p=0.20, Figure 2.7 A, left panel). This analysis was repeated for HERVH high vs low cells, and again it showed significant upregulation of adjacent genes, but not when direct chimeric transcripts are removed (p=4.13e-06 and p=0.2963, Figure 2.7 A, right panel). In summary, we do not find widespread evidence that active HERVH loci can regulate adjacent cellular genes through a classical enhancer type mechanism. We then repeated this analysis using inactive HERVH loci. Surprisingly, we do find significant enrichment of downregulated adjacent genes adjacent to inactive HERVH loci after shRNA knockdown (p=3.67e-10, Figure 2.7 B, left panel), as well as upregulation of adjacent genes after sorting (p=0.001, Figure 2.7 B, right panel). This analysis suggests it is not active but inactive loci that may function as enhancers.

A **Active HERVH, 50 kb window (two closest genes)**

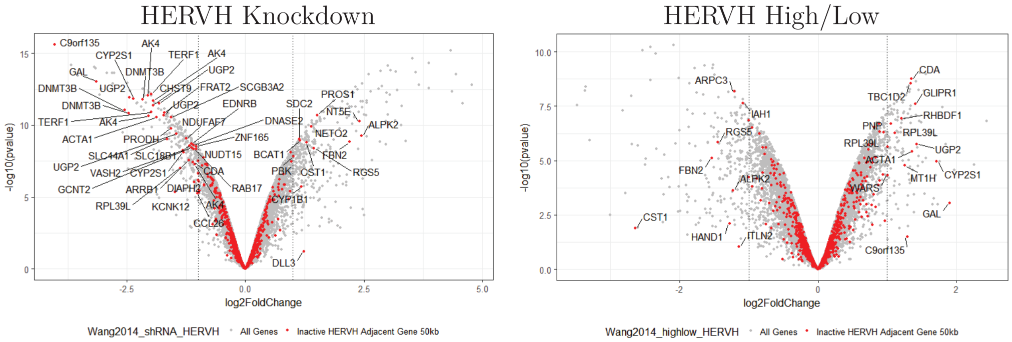


p-value = 1.212e-06 p-value = 0.184 p-value = 1 p-value = 4.131e-06

chimeric removed
p-value = 0.2017

chimeric removed
p-value = 0.2963

B **Inactive HERVH, 50 kb window (two closest genes)**



p-value = 3.656e-10 p-value = 0.8373 p-value = 1 p-value = 0.001061

Figure 2.7: Effect of HERVH modulation on HERVH adjacent genes

A. Volcano plot showing gene expression after HERVH knockdown (left) and after sorting for LTR7 high cells (right). The two closest adjacent genes nearby active HERVH are labeled in red (50 kb window). Log2 fold change cutoff of -1 and 1 is indicated with a dashed line. HERVH adjacent genes that are differentially expressed are labeled. Data from Wang et al. 2014([87]). P values are calculated using Chi-square with Yates' correction
B. Identical to A, except adjacent genes nearby inactive HERVH loci are for selected for analysis.

Investigation of HERVH chimeric transcripts

Previous papers have shown HERVH maintains the pluripotent state by knockdown with shRNAs targeting the HERVH-internal region([89],[87]). However, the majority of LTR7 chimeric transcripts splice out this sequence, suggesting they are not primary effectors of HERVH function. To determine if any chimeric transcripts are targeted by published shRNAs, we systematically identified all putative HERVH driven chimeric transcripts using the AceGene database, which is a comprehensive database of available mRNA and EST data([122]). We find 145 total putative HERVH-chimeric transcripts (supplementary Table 2.4). All previously identified HERVH chimeric transcripts are found([87],[114],[103]), indicating this is a comprehensive method that likely overestimates the true number of HERVH-driven chimeric transcripts in hESCs. We then we made a custom BLAST database of all HERVH putative HERVH chimeric transcripts and blasted the most effective HERVH shRNA sequences against it (shRNAS #3 and #4 from Wang et al. 2014([87])). We found no protein coding genes expressed in hESCs that are targeted by either shRNA. The only expressed chimeric lincRNA that was targeted was a lincRNA annotated as *Embryonic Stem Cell Related Gene* (ESRG). ESRG is actually the highest expressed HERVH derived transcript in hESCs and is targeted by all published shRNAs we are aware of([89],([87], [100]). This led us to the hypothesize that if chimeric transcripts are primary effectors of HERVH function, disrupting ESRG locus should phenocopy family wide shRNA knockdown.

We deleted ESRG using CRISPR/Cas9 by targeting unique sequence flanking the HERVH insertion that drives ESRG expression (Figure 2.8 A). Screening 56 clones revealed 5 heterozygous and 1 knockout clone, which was sequenced validated as a homozygous deletion of ESRG. Initially, we observed normal morphology and normal expression of pluripotency factors in the ESRG knockout cell line. We continued culturing for 8 passages (48 days), and observed no upregulation of differentiation factors or abnormal morphology. Furthermore, the genes nearby ESRG, which are silenced in hESCs, were not activated (Figure 2.8 B). After comparing these data to the phenotype induced by shRNA mediated knockdown (Figure 2.8 C), we conclude loss of a single HERVH chimeric loci does not phenocopy family wide knockdown. Because ESRG knockout does not effect stemness, and other spliced chimeric HERVH transcripts are not predicted targets of published shRNAs, chimeric transcripts are likely not primary effectors of HERVH function.

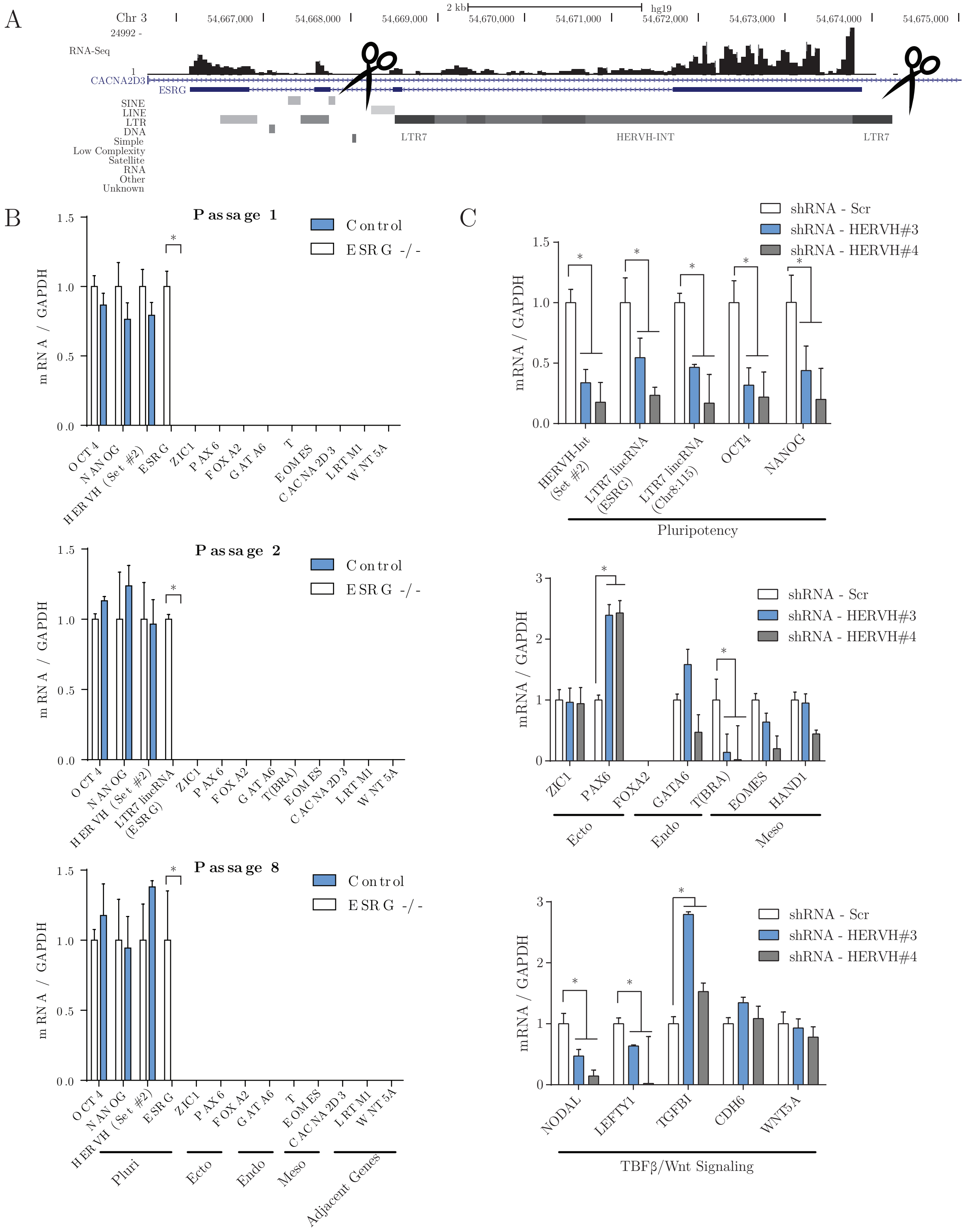

A
Chr 3
2 kb
hg19
54,667,000
54,668,000
54,669,000
54,670,000
54,671,000
54,672,000
54,673,000
54,674,000
54,675,000
24992 -
RNA-Seq
1
CACNA2D3
ESRG
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
LTR7
HERVH-INT
LTR7
B
Passage 1
Control
ESRG -/-
mRNA / GAPDH
1.5
1.0
0.5
0.0
*
OCT4
NANOG
HERVH (Set #2)
ESRG
ZIC1
PAX6
FOXA2
GATA6
T
EOMES
CACNA2D3
LRTM1
WNT5A
Passage 2
Control
ESRG -/-
mRNA / GAPDH
*
OCT4
NANOG
HERVH (Set #2)
LTR7 lincRNA (ESRG)
ZIC1
PAX6
FOXA2
GATA6
T(BRA)
EOMES
CACNA2D3
LRTM1
WNT5A
Passage 8
Control
ESRG -/-
mRNA / GAPDH
*
OCT4
NANOG
HERVH (Set #2)
ESRG
ZIC1
PAX6
FOXA2
GATA6
T
EOMES
CACNA2D3
LRTM1
WNT5A
Pluri
Ecto
Endo
Meso
Adjacent Genes
C
shRNA - Scr
shRNA - HERVH#3
shRNA - HERVH#4
mRNA / GAPDH
1.5
1.0
0.5
0.0
*
HERVH-Int (Set #2)
LTR7 lincRNA (ESRG)
LTR7 lincRNA (Chr8:115)
OCT4
NANOG
Pluripotency
shRNA - Scr
shRNA - HERVH#3
shRNA - HERVH#4
mRNA / GAPDH
3
2
1
0
*
ZIC1
PAX6
FOXA2
GATA6
T(BRA)
EOMES
HAND1
Ecto
Endo
Meso
shRNA - Scr
shRNA - HERVH#3
shRNA - HERVH#4
mRNA / GAPDH
3
2
1
0
*
NODAL
LEFTY1
TGFBI
CDH6
WNT5A
TBFβ/Wnt Signaling

Figure 2.8: Knockout of a highly expressed LTR7 chimeric gene

A. Schematic showing the strategy for the CRISPR deletion of ESRG. Visualization uses the UCSC genome browser. RNA-seq data is from the ENCODE RNA-seq track of the UCSC genome browser for H1 hESCs. **B.** ESRG knockout clone was passaged with wildtype clones for 8 pasages and differentiation markers were tracked with qPCR. Data is from the single ESRG knockout clone. Error bars are standard error from three technical measurements. Empty data signifies these genes were not detected with qPCR (ct >32). **C.** shRNA knockdown of HERVH. HERVH knockdown was induced by two separate shRNAs targeting the HERVH-internal region (sequence from Wang et al. 2014 [87]). Data is representative from three independent experiments. Error bars are standard error from three technical measurements.

HERVH effects differentiation dynamics in trans

Our previous results suggest full length HERVH RNA may be the primary effector of HERVH function. To test this, we asked if full length HERVH RNA can impact the pluripotent state in trans. We cloned what represents the HERVH RNA product, selected from the same insertion as our LTR7 reporter construct, into the *piggBac* transposon system (Figure 2.9 A). To obtain levels of overexpression significantly higher than background, we replaced the LTR7 promoter with the Human Elongation Factor-1 alpha (EF-1-α) promoter, and also selected for successfully transected cells using puromycin. Ectopic overexpression of LTR7-HERVH RNA did not effect pluripotency transcripts or induce detectable levels of mesendoderm associated genes in primed culture conditions (Figure 2.9 B). However, upon mesendoderm differentiation, we observed a significant increase in the induction of T, EOMES, and MIXL, but not NANOG or POU5F1 (OCT4) (Figure 2.9 C). In all, these experiments show LTR7 is active in mesoderm and the LTR7-HERVH RNA product can act in trans to effect the dynamics of early cell fate decisions.

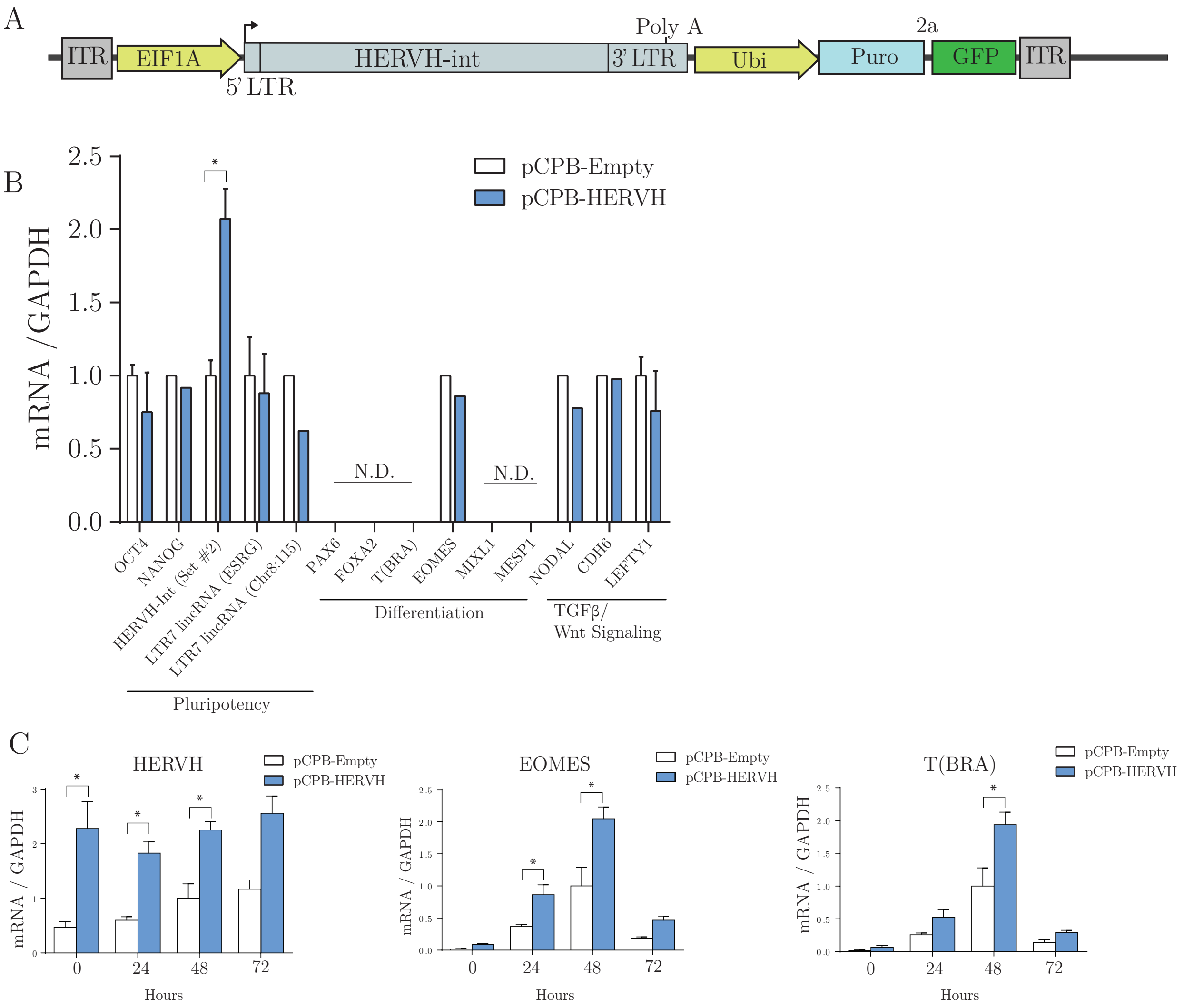

A
ITR
EIF1A
5' LTR
HERVH-int
Poly A
3' LTR
Ubi
Puro
2a
GFP
ITR
B
mRNA /GAPDH
2.5
2.0
1.5
1.0
0.5
0.0
*
pCPB-Empty
pCPB-HERVH
N.D.
N.D.
OCT4
NANOG
HERVH-Int (Set #2)
LTR7 lincRNA (ESRG)
LTR7 lincRNA (Chr8:115)
PAX6
FOXA2
T(BRA)
EOMES
MIXL1
MESP1
NODAL
CDH6
LEFTY1
Pluripotency
Differentiation
TGFβ/
Wnt Signaling
C
HERVH
pCPB-Empty
pCPB-HERVH
mRNA / GAPDH
3
2
1
0
*
*
*
0
24
48
72
Hours
EOMES
pCPB-Empty
pCPB-HERVH
mRNA / GAPDH
2.5
2.0
1.5
1.0
0.5
0.0
*
*
0
24
48
72
Hours
T(BRA)
pCPB-Empty
pCPB-HERVH
mRNA / GAPDH
2.5
2.0
1.5
1.0
0.5
0.0
*
0
24
48
72
Hours

Figure 2.9: HERVH effects mesoderm differentiation in trans

A. Diagram of the pCPB-HERVH overexpression construct. **B.** Overexpression of HERVH does not effect pluripotency. HERVH overexpression in hESCs was obtained by transfection with either pCPB-HERVH or control vector and subsequent selection of positive clones with puromycin. Data is representative from 5 independent experiments. Error bars are standard error from 3 technical measurements. **C.** HERVH overexpression effects mesendoderm differentiation. pCPB-HERVH was overexpressed in hESCs and mesendoderm differentation was induced. Transcripts are quantified with qPCR for various timepoints. Data is representative from 3 independent experiments. Error bars are standard error of three technical measurements.

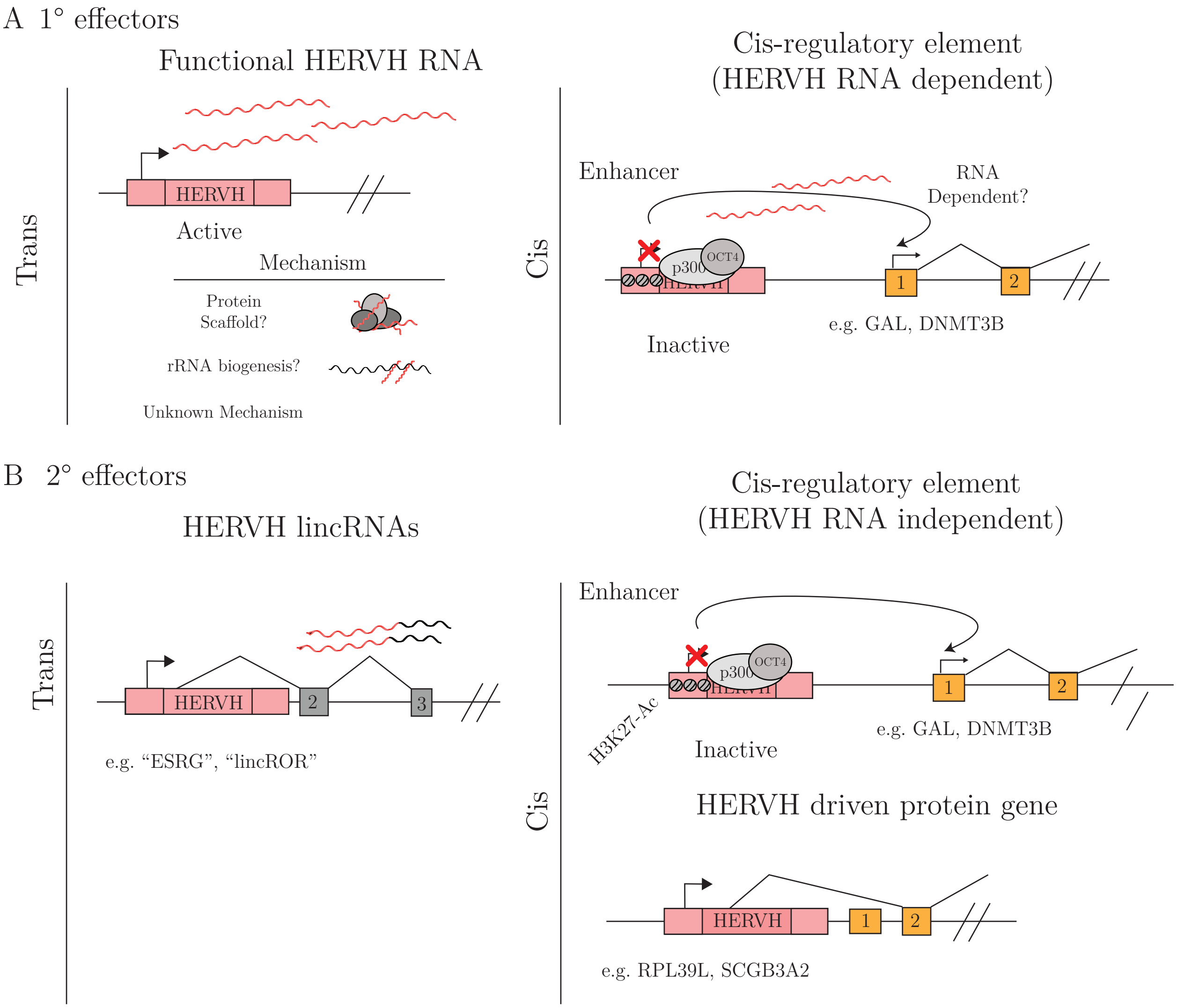


Figure 2.10: Model of HERVH function

A. Primary effectors of HERVH function. We propose that HERVH full length RNA is the primary effector of HERVH function because it is preferentially targeted by published shRNAs. The exact mechanism by which it functions is still unknown. One possibility is that the potential enhancer activity of inactive loci may be somehow dependent on the full length RNA. **B.** Secondary effectors of HERVH function. We provide evidence that HERVH lincRNAs and protein coding genes are not primary effectors of HERVH function as they are downstream of shRNA knockdown. However, they may constitute a secondary effect that contributes to loss of pluripotency after full length HERVH RNA knockdown. The observed enhancer effect may also be a secondary effector if it functions independently of the HERVH full length RNA.

2.4 Discussion

In recent years there has been a surge of interest surrounding the phenomenon of retrotransposon activation in early development. In mouse, the MERVL and LINE1 families have been directly implicated in zygotic genome activation([70],[79]), and an individual insertion from the MTA family has clearly been co-opted to drive a Dicer isoform essential for germ line maturation([21]). In humans, many retrotransposon families are also activated, showing distinct expression patters from zygote to late blastocyst([95]). However, the HERVH family is by far the highest expressed family in pluripotent stem cell culture([25]), and so far is the only one demonstrated to be essential for pluripotency([89],[87]). Here we propose that HERVH is expressed in multiple pluripotent cell types and that its primary mechanism of action is in trans.

We find that the LTRs for HERVH combine to promote the expression of HERVH internal region in every cell of the human preimplantation embryo, and in cell culture, through the formation of the primitive streak. Furthermore, we find that the LTR7 may not be a good marker of the naive state as previously reported([87]), as we show it is expressed in naive and primed hESCs, and is further activated during differentiation into mesendoderm. Of note, we observe upregulation of LTR7 reporter activity in naive cells but see downregulation of endogenous LTR7 trancripts. It is possible full length LTR7 insertions have internal regulatory sequence that is not included in the LTR7-tdTomato reporter. An interesting candidate factor is the repressive KRAB-ZFP ZNF534, which binds HERVH in vitro and is differentially regulated between naive and primed states([86]). We also find that Brachyury binds directly to LTR7 loci in differentiating cells. This implies that upregulation of LTR7 in mesendoderm is biolgically relevant, and also suggests there is positive feedback loop between LTR7 and Brachyury, as HERVH overexpression increases T levels during differentiation.

We find certain members of the LTR7Y subfamily are excellent markers of naive hESCs, and this transcriptional specificity is likely conferred by a consensus KLF binding motif. KLF4 and KLF17 are promising candidate factors to bind this sequence because they are both differentially expressed between primed and naive states. Because multiple KLF proteins have highly similar binding motifs, it also possible other KLF family members could bind the identified CCCCACCC motif and contribute to LTR7Y transcription. Ultimately, ChIP-seq experiments should address which factors bind LTR7Y in naive hESCs. Our method used to identify the CCCCACCC motif could be applied to other active ERVs. For example, repeating this analysis for LTR7B could identify transcription factors important for human 8-cell or morula.

While the primary mechanism of HERVH function is still unclear, we do narrow potential possibilities. First, we show HERVH chimeric transcripts are likely not essential to maintain pluripotency. The only chimeric transcript predicted to be targeted by published shRNAs is ESRG, and after genetic knockout we find no loss of stemness. This is in contrast to work from Wang et al. 2014, who report ESRG knockdown causes upregulation of differentiation markers. This discrepancy could be explained by the use of different hESCs lines (H9 in Wang et al. 2014, WIBR3 in this study). Alternatively, it is possible that during clonal selection, surviving colonies adapted in some way to the loss of ESRG. However, we feel our genetic knockout does show ESRG is not fundamentally essential for pluripotency. We also show that active HERVH loci are unlikely to enhance the expression of adjacent genes, as genes within 50 kb of active HERVH loci are not dynamic with HERVH knockdown or high HERVH transcription. However, we find differentially expressed genes are more likely to be near inactive HERVH copies. Enhancer sequences are normally not transcribed but rather serve to bind transcription factors and engage in DNA loops with nearby promoters(for review see[123]). We speculate that if inactive loci are enhancers, they likely retain the binding motifs for pluripotency related transcription factors that bind LTR7, such as NANOG, OCT4, and P300([25]), and are free to engage in productive DNA loops with adjacent genes. It is important to acknowledge that if LTR7 does serve as an enhancer, it may do this independently from the RNA produced by active HERVH insertions. Alternatively, this observation could be purely circumstantial. It is possible HERVH preferentially integrated nearby active genes in pluripotency, as many viruses show a bias for integration near active genes([124]). If so, these genes may simply be pluripotency genes, and are downregulated upon differentiation independent of nearby HERVH loci. Ultimately, experimental validation by deletion of candidate HERVH enhancers is needed to show if inactive HERVH loci can function as enhancers and determine if this effect is important for pluripotency.

We find that overexpression of HERVH internal region is able to effect the dynamics of primitive steak formation. This result supports the trans model of function, as ectopic expression of HERVH is being driven by the EIF1α promoter from randomly integrated insertions. This is in agreement with studies done in reprogramming, where overexpression of HERVH transcripts enhance reprogramming efficiency([87],[101]). Future work could use this differentiation assay to identify the minimal region necessary to enhance mesendoderm differentiation and identify the protein factors that bind it.

Finally, we offer an explanation for the claim that LTR7-HERVH high cells mark the naive state. We find that every cell in hESCs culture is positive for a LTR7 reporter, but we do observe heterogeneity in signal intensity. We find that sorting for LTR7-HERVH does not reveal markers of naive pluripotency (TFCP2L1, DPPA3, KLF17), but rather markers of the primitive streak (T, MIXL1, EOMES). However, a number of genes show elevated transcription in both naive hESCs and in the primitive streak, including NANOG, NODAL, LEFT2, GDF3, WNT5B, WNT3, and WNT3A. During primed to naive reversion, primed cells with naturally elevated NANOG levels and high Wnt signaling are reported to be more likely to survive([125]). The LTR7 high cells display such a signature, and isolating them before reversion to naive would explain the observation of increased reversion efficiency. In summary, LTR7 high cells in primed culture are poised for differentiation into mesendoderm. However, they also possess the capacity to revert efficiently to the naive state, presumably due to high levels of NANOG and Wnt signaling necessary for mesendoderm differentiation.

The major outstanding challenge for future studies of HERVH is to demonstrate the mechanism LTR7 employs to maintain human pluripotency. Additionally, it remains to be determined if other HERVH subfamilies function similarly, or have unique functions earlier in development. Our data suggest that the unspliced, non-chimeric HERVH RNA is likely to be the primary effector of HERVH function in hESCs, and additional experiments identifying proteins that bind this RNA are likely to be informative. The HERVH overexpression experiments provide an assay that should allow the identification of a minimal functional region, and mutation of putative RNA motifs could be used to interrogate a hypothetical RNA binding protein-HERVH interaction. In all, this work significantly contributes to the understanding of HERVH regulation and function in pluripotency. It will likely assist with future efforts to derive new pluripotent cell cultures and also helps explain how pluripotency and human development differs from other mammals.

2.5 Methods

Custom HERVH annotation

The repeat masker database was downloaded from the UCSC genome browser for the Feb. 2009 assembly of the human genome (hg19, GRCh37, (GCA000001405.1)). HERVH elements within 50 base pairs were merged into a single annotation suing a custom python script. The new merged annotations were then assigned a class (LTR7, LTR7B, LTR7C, or LTR7Y), a strand (+/-), and a structure(complete, incomplete, or solo-LTR). Complete insertions are defined as an annotation with 2 LTRs flanking an internal region. Incomplete LTRs are annotations with an LTR present on only one side an internal region. Solo-LTRs have no associated internal region. Complex insertions are defined as those having LTRs of at least two different subfamilies.

Bioinformatic analysis of single cell RNA seq data from human embryos

Single cell RNA-seq data from the human embryo was retrieved from Gene Expression Omnibus (GSE36552,([112])). RNA-Seq data was mapped using TopHat with the -G option, which first mapped reads to virtual transcriptome containing genes in ensebml release 74 and our custom HERVH annotation. The reads that did not fully map to custom transcriptome were then mapped onto the genome. Up to two mismatches were allowed. For reads that aligned multiple times, alignments with the best score were reported (up to 20, default parameters). Read counts per gene were quantified using FeatureCounts([126]) using fragments that have both ends successfully aligned (options -B -p). Gene expression was normalized by gene length and differential expression between stages was performed using DESeq (negative binomial distribution). For the box plot, the expression of each HERVH loci was determined by taking the average expression value in all blastomeres for each developmental stage. The cell types of the blastocyst (primitive endoderm, trophectoderm, and epiblast) were defined in Guo et al. 2015([127]). The heat map was generated using all differentially expressed HERVH loci (log2 fold change >1 or <1) using the Superheat package in R.

Statistical Analysis

All numerical results are presented as the mean with standard deviation from three technical replicates or the number of biological replicates (independent experiments) stated in the figure legend. Comparison between control and experimental or between timepoints was performed using a two-sided upaired Student's T test using with significance level of 0.05.

Quantitative real-time PCR

RNA was prepared from harvested cells using TRIZOL according to manufacturers instructions (Life Technologies, Cat. 15596). RNA was treated with DNAse for 15 minutes (Invitrogen, cat. 18068015) and reverse transcribed using iScript Advanced Reverse-Transcriptase (Bio-Rad, Cat. 1725037). Quantitative real time PCR was performed with SYBR FAST qPCR Master Mix (Kapa Biosystems, cat. KK4604) and Applied Biosystems StepOnePlus Real Time System. The primer sequences are included in the supplemental data.

Cell culture

For primed hESC culute, hESCs lines WIBR3 (NIH stem cell registry 0079) and H9 (NIH stem cell registry 0062) were cultured on irradiated mouse embryonic fibroblast (MEFs) feeder layers in hESC media (DMEM/F12 supplemented with 20% knockout serum replacement (Gibco Cat. 10828028), 1 mM glutamine, 1% non-essential amino acids (Invitrogen Cat. M7145), 0.1 mM β-mercaptoethanol (Sigma-Aldrich Cat. M6250), and 5 ng/ml FGF2 (Stem Cell Technologies, Cat. 78003.1). hESCs were passaged every 5 to 7 days. To passage, cells were washed in PBS and then incubated in DMEM/F12 containing Collagenase IV (2 mg/mL, Cat. 17104019). After 20 minutes, cells were dislodged with hESC wash media (DMEM/F12 supplemented with 5% FBS) and feeder cells were removed using gravity separation in a 15 mL conical tube and broken into small clumps and plated. hESCs used in experiments were maintained under passage 40.
For naive hESC culture, naive hESCs were generated using the WIBR3 cell line cultured on irradiated MEFs feeder layers. For conversion of primed to naive, 2 x 10^5 primed hESCs were first incubated in physiological oxygen conditions (5%O2, 3%CO2) for 10 days. Before conversion, cells were incubated with primed hESC media supplemented with 10 μM Y27632 (Stemcell Technologies) for 24 hours. Cells were then passaged using Accutase (Gibco) onto a MEF feeder layer and incubated with hESC media supplemented with 10 μM Y27632 for two days. Media was switched to 5i/L/A (naive media) and widespread cell death was observed. After 10 days cells recovered cells were passaged polyclonally using Accutase. Naive culture media is defined as follows: a base media of 50% DMEM/F12 (Invitrogen Cat. 11320) and 50% Neurobasal media (Invitrogen Cat. 21103) supplemented with 1x N2 (Invitrogen Cat. 17502048), 1x B27 (Invitrogen Cat. 17504044), 1 mM glutamine (Invitrogen), 1% non-essential amino acids (Gibco Cat. 11140076), penicillin-streptomycin (Invitrogen), 50 μg BSA (Sigma Cat. 9048-46-8), 0.5% knockout serum replacement (Gibco Cat. 10828028), 8 ng/uL FGF2 (Stemcell technologies Cat. 78003), 20 ng/μ Human LIF (Stemgent Cat. 03-0016), 20 ng/μL ACTIVIN-A (Peprotech Cat. 120-14E) and the following small molecule inhibitors: 10 μM Y27632 (Stemcell Technologies) 1 μM PD0325901 (Stemgent Cat. 04-0006), 1 μM WH-4-0230 (Stemgent Cat. 4-0079), 0.5 μM SB5908850 (Stemgent Cat. 4-0080), 1 μM IM-12 (Stemgent Cat. 04-0081).
HEK293T cells were cultured in DMEM/F12 with 10% FBS and 5% Penn/Strep

(ThermoFisher Cat. 15070063) and passaged at 90% confluence using trypsin (Gibco).

Plasmid Transfections

Transfection of HEK293T cells: HEK293T cells were seeded onto a 6 well plate (0.5x10^6 cells). The next day 3 ug pCPB control or pCBP-KLF4 vector was incubated with 1 mg/ml PEI (Sigma Cat. 408727) in reduced serum OptiMEM (ThermoFisher Cat. 31985062) for 10 minutes and then added to cells. Cells were harvested at 48 hours. **Transfection of WIBR3 hESCs**: WIBR3 cells were passaged according to normal protocol onto MEF feeders. When cells reached 40$ confluence, they were transfected with Lipofectamine Stem transfection reagent (ThermoFisher Cat. STEM00001). For a six well plate, 6 μg total DNA was incubated with 12 uL Lipofectamine Stem transfection reagent in OptiMEM reduced serum media for 10 minutes before adding to hESCs. For experiments involving stable cell generation, cells were allowed to recover until confluencey before splitting onto puromycin resistant feeders. The next day, media was changed to hESC media supplemented with 1 ug/mL puromycin (Gibco Cat. A1113802) and selection continued for 3 days before chagning back to normal hESC media.

Definition of expressed and non-expressed HERVH loci

The alignment of HERVH used a subset of highly expressed HERVH loci. HERVH loci were defined as highly expressed if their RPKM value exceeded 20 in either epiblast, trophectoderm, naive hESCs or primed hESCs. RPKM values were determined for epiblast, trophectoderm, and primed hESCs from our analysis of single cell human embryos (see Bioinformatic analysis of single cell RNA seq data from human embryos), resulting in a list of 95 expressed HERVH insertions.

HERVH primary sequence alignment

The primary sequence of insertions expressed in naive hESCs were aligned to the HERVH ancestral sequence using MUSCLE with the -stable option (do not rearrange sequences). Gaps in the sequence alignment caused by only 1 alignment were removed. Visualization was done using UGENE software (ugene.net). Sequences (rows) were manually rearranged within families in order to emphasize sequence similarities between loci. Hover, this ordering was only done once and all alignments show the same ordering of loci.

Motif analysis

The HOMER (Hypergeometric Optimization of Motif EnRichment) suite was downloaded and the findMotifs.pl script was used to find motifs enriched in expressed HERVH sufamilies. To do this the primary sequences of loci defined as either highly or lowly expressed loci were downloaded from UCSC genome browser in fasta format. These sequences were passed to findMotifs.pl and highly expressed loci were used as input and lowly expressed loci were selected as user defined background genes (-bg option).

Analysis of LTR subfamilies fold changes between naive and primed cells

Fold changes of retrotransposon insertions from primed to naive were downloaded from Theunissen et al. 2016([86]). Fold change values for LTR7, LTR7Y, LTR7Y-2(B-Like) were extracted for HERVH insertions previously defined as highly expressed and were plotted using ggplot2 and ggpubr in R. Pairwise statistical analysis was performed using a Wilcoxon rank sum test.

LTR7 and LTR7Y reporter lines

hESCs with stable of expression of LTR7 and LTR7Y reporters were generated in WIBR3s using the PiggyBac transposon system. 4.5 μg pCPB transfer vector containing the LTR reporter was cotransfected with 1.5 μg pBASE transposase vector according to the protcol detailed in the **Plasmid Transfections** section of the methods. Cells were selected using puromycin (1 μg/mL) to insure 100% transfection efficiency. Lines were then reverted to the naive state using the protocol described in the **Cell Culture** section and images were taken at 10 days using a Zeiss Z1 fluorescence microscope. RNA samples were harvested at 10 and 16 days (passage 1 and 2).
To generate clonal lines for the LTR7 reporter, stable LTR7 reporter lines were incubated with primed hESC media supplemented with 10 μM Y27632 (Stemcell Technologies) for 24 hours. Cells were then passaged using Accutase (Gibco) onto a MEF feeder layer and incubated with hESC media supplemented with 10 μM Y27632 for two days. After 10 days, clones were picked and expanded. To ensure clonal selection, this process was repeated before clonal lines were used for experiments. To sort LTR7 reporter lines into high and low populations, clonal LTR7 lines were expanded to confluence in 6 well plates. hESCs were trypsinized to single cell and passed through a 35μm nylon mesh (Corning Cat 352235) to remove clumps. 5 x 10^6 cells were sorted into total, high (top 5% of signal), and low (lower 95% of signal)) populations using the BD Influx Cell Sorter (UC Berkeley, Li Ka Shing Builidng).

Genome editing in hESCs

Guide RNAs targeting unique region flanking the ESRG locus were designed using the MIT crispr design tool (http://crispr.mit.edu/). The top hits were synthesized as oligos and cloned into the px458 vector (Addgene 48138). This vector was cotransfected into hESCs with pCPB-GFP vector, which contains EF1a promoter driving GFP. This transfection protocol is described in **Plasmid Transfections**. After 48 hours, GFP positive cells were isolated using the BD Influx Cell Sorter (UC Berkeley, Li Ka Shing

Builidng). Cells were plated in media supplemented with 10 μM Y27632 (Stemcell Technologies) for 3 days. The media was then switched to normal hESC media and surviving colonies were picked approximately 12 days later. These clones were genotype using PCR and WT and KO cells were passaged for 8 passages.

Differentiation of hESCs into mesendoderm

Before mesendoderm induction, conditioned media was prepared by incubating hESC media without FGF2 overnight on MEF feeders and the media was collected the next day. hESCs were passaged onto plates coated with 80 μg/mL Matrigel (Corning Cat. 356231) using standard passaging procedure (see **Cell culture**) and incubated overnight in conditioned media supplemented with 10 ng/mL (Stemcell technologies Cat. 78003). Mesendoderm induction was initiated by changing the media to conditioned media without supplemented FGF but with 10 ng/mL BMP4 (Gibco Cat. PHC9534) and 10uM LY294002 (Invitrogen Cat. PHZ1144) for 96 hours.

Analysis of published Brachyury ChIP-seq data

To analyze Brachyury binding to HERVH loci in mesendodederm, we downloaded published called ChIP seq peaks from Faial et al. 2015([119]). To determine if ChIP seq peaks were enriched for LTR7 sequences, we used the HOMER Annotate Peaks software to label peaks within 100 basepairs of ERV sequences. To determine enrichment for HERVH LTRs, we calculated the number of members of each ERV family that contain a ChIP seq peak. After normalizing for the size and number of insertions for each family, statistical enrichment was determined using a hypergeometric probability test.

shRNA knockdown of LTR7 and LTR7Y transcripts

Viral transfer vectors containing shRNAs targeting HERVH were generated in either the SGEP or PLKO.1 vectors (see **DNA plasmids** section. To generate infectious lentivirus containing the desired shRNAs, HEK293T cells in 10 cm dish were transfected at 30% confluence using 24 uL PEI in OptiMEM with 2.5 μg g shRNA construct, 1.875 μg psPAX2 (Addgene 12260), and 0.625 μg pMD2.g (Addgene 12259). Virus was harvested at 48 h and 72 h after transfection. hESCs were transduced using 1 mL of virus containing supernatant supplemented with 5 μ/mL polybrene (SC Biotechnology, Cat. #sc-134220) for 12 h before switching to back regular hESC media. Positive cells were selected using 2 μ/ml pyromycin for 3 days. For shRNAs targeting LTR7Y insertions, cells were then reverted to the naive state using the protocol described in the **Cell culture** section.

Analysis of published LTR7 high vs low microarray data

LTR7 high vs low microarray data was downloaded from Gene Expression Omnibus (accession GSE54726). Differential expression of gene probes was determined using an R script provided by GEO2R (https://www.ncbi.nlm.nih.gov/geo/geo2r) which calculates differential expression using the limma package. Differentially expressed genes were defined as those with a log2fold change >1 or <1. Gene ontology was performed using differentially expressed genes and the online PANTHER gene ontology tool (http://pantherdb.org/). Gene ontology analysis was done using PANTHER Overrepresentation Test (Released 20171205) and statistics were calculated using Fisher's Exact with FDR multiple test correction.

Analysis of published HERVH knockdown microarray data

HERVH knockdown microarray data was downloaded from Gene Expression Omnibus (accession GSE54726) and differential gene expression was done identically to the LTR7 high vs low microarray analsysis. To determine HERVH adjacent genes, we used the GREAT online gene association tool (http://great.stanford.edu/), and genes adjacent to transcribed HERVH loci were retrieved using the “Two Nearest Genes” option (50 kb). This generated a list of 177 genes. Duplicate entries (i.e. adjacent to more than one HERVH loci) were removed, yielding a list of 114 genes. To determine if the genes nearby HERVH are enriched in the list of differentially expressed genes, statistical tests were performed (Chi-square with Yates' correction). Differentially expressed genes were defined as genes with a log2Fold change of >1 or <-1.

Determination of HERVH-chimeric genes

To identify HERVH driven chimeric transcripts, we utilized the AceGene database, which is a comprehensive database of available mRNA and EST data([122]). To generate a list of putative chimeric genes driven by HERVH, we intersected the transcriptional start site of each Acegene transcript with all HERVH loci. This list contains 431 total transcripts, 45 of which are listed in the UniProt database. All previously identified HERVH chimeric transcripts are found([87],[114],[103]), indicating this is a comprehensive method that likely overestimates the true number of HERVH-driven chimeric transcripts.

Prediction of HERVH chimeric trancripts targeted by published shRNAs

To determine what putative chimeric transcripts are targeted by published shRNAs, we made a custom BLAST database of all HERVH putative HERVH chimeric transcripts. We then used blastn (NCBI) to compare the shRNA targeting sequence from published shRNA sequences([87]) with the following parameters; -task blastn-short -penalty -1. Alignments with three or more mismatches in core positions 3-19 were removed, generating separate lists of putative chimeric transcripts targeted by each shRNA.

DNA constructs

shRNA constructs To generate LTR7 knockdown constructs, the sequence of shRNAs #2 and #3 from Wang et al. 2014([87]) were synthesized as oligos and ligated into the pLKO.1 vector (Addgene 10878) using AgeI and EcoRI sites. For LTR7Y knockdown, sequences unique to the LTR7Y internal region were analyzed with the SplashRNA algorithm. http://splashrna.mskcc.org/([128]). The top 5 hits were synthesized as oligos and cloned into the SGEP vector (Addgene 111170). These unique sequences were also analyzed with the Broad Institute pLKO shRNA design tool (https://portals.broadinstitute.org) and the top 5 hits were synthesized as oligos and cloned into the pLK0.1 vector.
KLF4 Overexpression Vector: kpCPB-EF1-α Empty Vector: The piggyback destination vector was modified to contain a BSMBI cloning site directly after the EF-1α promoter, making it suitable for Golden Gate Cloning. KLF4 open reading frame was amplified from the pCXLE-hSK vector (Addgene 27078) and cloned into pCPB-Empty vector using a BsmbI restriction enzyme based Golden Gate Cloning.
HERVH Overexpression Vector: An HERVH insertion at location chr4:180087630-180091722 (hg19 genome) was amplified from the human genome using a nested PCR strategy. The first PCR used primers flanking the HERVH insertion, allowing for unique amplication, and the second PCR amplified HERVH from the predicted TSS to the end of the 3' LTR. This transcript was cloned into pCPB-EF1-α Empty Vector using a BsmbI restriction enzyme based Golden Gate Cloning reaction.
LTR reporter constructs: pCPB-Empty Vector: The piggyback destination vector was modified to contain a BSMBI cloning site proceeded by no promoter. LTR sequences were amplified directly from the human genome. The forward primer for all constructs lies directly before the 5' LTR in unique genomic sequence and the reverse lies approximately 110 basepairs inside the HERVH-internal region. The tdTomato protein was amplified from the 2c:tdTomato reporter (Addgene 40281) and constituted a second piece in a BsmbI restriction enzyme based Golden Gate Cloning reaction.

2.6 Supplementary Information

Table 2.3: Putative HERVH-chimeric transcripts

Protein Coding	lincRNA annotated	lincRNA	(Acegene Annotation)	
ABHD12B	ESRG	basmer	nunayu	waglaw
ACTR3C	FLJ26245	beewer	peeter	warubo
ATXN3	LOC100126447	blydawbu	pervawbu	watumi
CALB1	LOC100133317	cherchee	plajorbo	wyku
CCDC141	LOC100287242	chodybo	ploydaw	yayayu
CLEC12A	LOC146880	chyfaw	pober	yutora
CYP11A1	LOC348926	dybo	porrobo	zarswoy
DNAJC15	LOC349408	fergar	poydybo	zyskawby
FUCA1	LOC729739	flachabu	poytabo	
FUT3	LRRC2-AS1	flajabu	ranare	
GABRP	MGC32805	florstuby	rarkey	
GSDMB	NCRNA00263	gardybo	rarspubu	
GZMA	UCA1	gawdybo	rusimo	
HHLA1	LOC79015	geedybo	sardybo	
HRG	psiTPTE22	geyskeybu	sawame	
HTR7		glospubu	shawlabu	
IL34		gojey	sheespar	
KIF1B		hamuyo	sheyfeyby	
KLKB1		hanaya	shorblabu	
LRRC61		jerdaw	siyamu	
MACC1		jufloy	skarporbu	
MOK		jyree	skerpu	
PCSK9		jyshorbu	skoyvarbo	
PMM2		kawame	slardybo	
RPL39L		kerpu	sleegar	
SCGB3A2		klawnerbu	sleeger	
SEC23B		kleyspeybu	slorswoybu	
SELP		kloswoybu	smarrarby	
SEMA3E		klygorbo	smublerby	
SLCO1B1		kogerbo	snarkerbo	
SLCO1B3		kohumu	snoyvy	
SPG20		leegar	snudybo	
SPTLC1		leyspeybu	sorseyby	
TFPI		lywawby	spurar	
TUBB2A		meysorby	sutime	
VRK2		moyboy	swarubo	
VWA3B		moyneybu	sygabo	
WDR41		muhumi	syzo	
ZNF107		muklyby	tacho	
ZNF114		naya	tygoybo	
ZNF177		noyvy	vardy	

Table 2.4: Alignment of shRNA sequences against HERVH chimeric transcripts

Targeting sequence GCAACTCGTCCCAAATCTTCCT

```
>HESRG.dAug10 range=chr3:54666151-54673884strand=- Length=3138
Score = 44.1 bits (22),  Expect = 4e-07
Identities = 22/22 (100%), Gaps = 0/22 (0%)
Strand=Plus/Plus

Query  1     GCAACTCGTCCCAAATCTTCCT  22
             ||||||||||||||||||||||
Sbjct  1197  GCAACTCGTCCCAAATCTTCCT  1218

>MOK.wpAug10 range=chr14:102700027-102707527strand=- Length=335
Score = 30.2 bits (15),  Expect = 0.005
Identities = 18/19 (95%), Gaps = 0/19 (0%)
Strand=Plus/Minus

Query  1   GCAACTCGTCCCAAATCTT  19
           ||||||| |||||||||||
Sbjct  38  GCAACTCATCCCAAATCTT  20
```

Targeting sequence GCCGAGCTAGGTCCCAATTCTT

```
>FUT3.fAug10 range=chr19:5844837-5848812strand=- Length=629
Score = 44.1 bits (22),  Expect = 4e-07
Identities = 22/22 (100%), Gaps = 0/22 (0%)
Strand=Plus/Minus

Query  1   GCCGAGCTAGGTCCCAATTCTT  22
           ||||||||||||||||||||||
Sbjct  59  GCCGAGCTAGGTCCCAATTCTT  38

>HESRG.eAug10 range=chr3:54667622-54673900strand=- Length=1412
Score = 44.1 bits (22),  Expect = 4e-07
Identities = 22/22 (100%), Gaps = 0/22 (0%)
Strand=Plus/Plus

Query  1    GCCGAGCTAGGTCCCAATTCTT  22
            ||||||||||||||||||||||
Sbjct  314  GCCGAGCTAGGTCCCAATTCTT  335

>zarswoy.aAug10 range=chr14:38660249-38662219strand=+ Length=735
Score = 36.2 bits (18),  Expect = 9e-05
Identities = 21/22 (95%), Gaps = 0/22 (0%)
Strand=Plus/Plus

Query  1    GCCGAGCTAGGTCCCAATTCTT  22
            |||||||||| |||||||||||
Sbjct  204  GCCGAGCTAGTTCCCAATTCTT  225

>VRK2.rAug10 range=chr2:58344552-58373563strand=+ Length=630
Score = 36.2 bits (18),  Expect = 9e-05
Identities = 18/18 (100%), Gaps = 0/18 (0%)
Strand=Plus/Minus

Query  5   AGCTAGGTCCCAATTCTT  22
           ||||||||||||||||||
Sbjct  30  AGCTAGGTCCCAATTCTT  13
```

Chapter 3

Chimeric Isoforms in Human Preimplantation Development

3.1 Background

Retrotransposons are genetic elements that replicate throughout the genome using an RNA intermediate. Most sequences that remain in mouse and human are no longer mobile, and a subset have been exapted by their hosts for use as enhancers([52],[129]), promoters for cellular genes([21],[9],[130]), or as non-coding RNA products([64],[79] for review see[131]). In humans, retrotransposons are most active in germ cells and in preimplantation embryos, leading to the hypothesis that some have been exapted and are functional in development. The endogenous retrovirus HERVH is involved in the maintenance of human pluripotency([89],[87]), but the potential benefit of of LINES and SINES is in early human development is less clear. There is evidence to suggest some SINE families may be functional. Mammalian-wide interspersed repeats (MIRs), which are SINEs that actively transposed around 130 million years ago, are enriched in enhancer sequences, implying they have been selected to function as regulators of gene expression([132]). Alu elements, another SINE family present in over 1 million insertions in the human genome, are found in the protein coding regions of around 4% human genes([133]). Many of these present in alternatively spliced transcripts that create premature stop codons or frameshift mutations and are deleterious, but some may generate functional protein products([134]). Recent advancements in singe-cell RNA sequencing promises to expedite the identification of promising candidates for characterization. This chapter presents studies on one such candidate, ZBTB16, a developmentally important zinc finger protein whose expression and protein product is altered by a MIRb/Alu retrotransposon insertion.

3.2 Results

Identification of retrotransposon-gene chimeric transcripts in human development

To investigate the potential role retrotransposon-gene chimeric genes have in human development, a collaborator, Anne Bitton Ph.D. (Institut Pasteur), analyzed publicly available single cell RNA-seq data from blastomeres of human preimplantation embryos and human embryonic stem cells (unpublished, data from accession GSE36552, [112]). She identified 48,501 active retrotransposon loci (at least 1 count per million reads), and found over 20% are correlated with the expression of an adjacent gene. Around 20% of these gene-retrotransposon pairs contain junction reads directly joining the gene and retrotransposon, suggesting activation of retrotransposons is directly modifying the transcriptome and possibly the proteome of human preimplanation embryos. In all, Dr. Bitton's work has identified hundreds of protein coding genes that splice with nearby retrotransposon sequences. These junctions are highly expressed (> 30 junction read counts in at least one stage, see Methods), and can be found in multiple data sets and with multiple mapping methods. It is likely these transcripts exist in the human embryo, as a postdoc in our lab, Andrew Modzelewski, was able to validate many blastocyst specific chimeric transcrips using hESCs (data not shown).

We investigated if the genes found to make chimeras are statistically enriched for any biological terms. Using Gene Ontology enrichment analysis, we identified overrepresentation of genes with WD repeat, RNA binding, or Pleckstrin homology domains, AAA ATPAses, developmentally relevant transcription factors and Zinc Finger proteins, among others (Table 3.1). We chose to further explore the enrichment of zinc-finger containing proteins due to their recognized role in retrotransposon regulation.

Structure of ZNF-retrotransposon chimeric transcripts

To understand how the ZNF-chimeric transcripts are regulated, we selected four candidates where the chimeric isoform comprises at least 50% of the total transcription of the gene in at least one stage; ZNF226, ZNF605, ZNF544, and ZBTB16. Visualization of these transcripts reveals they all disrupt the ZNF protein product in some way, although through different mechanisms. The retrotransposons that splice with ZNF226 and ZNF544 are both downstream of the transcript and replace the last canonical exon (Figure 3.1 A, B). ZNF605 contains an antisense, intronic MER75-int element that appears to terminate the transcript after the first coding exon (Figure 3.1 C). Finally, MIRb is intronic to ZBTB16, but appears to act as an alternate promoter, potentially creating a truncated protein lacking the n-terminal domains (Figure 3.1 D). We also quantified the expression of the canonical and chimeric isoforms, and observe their expression patterns can be independent.

Table 3.1: Gene ontology terms enriched for genes that make chimeric transcripts in human embryos

Term	p.value	p.adj	significance
Protein phosphatase 1 regulatory subunits	0.0000006	0.00011	100.00
WD repeat domain containing	0.0000005	0.00011	100.00
AAA ATPases	0.0000209	0.00265	64.98
Pleckstrin homology domain containing	0.0002442	0.01885	43.49
RNA binding motif containing	0.0002474	0.01885	43.49
Zinc fingers	0.0005023	0.03190	37.73
Anaphase promoting complex	0.0010413	0.04408	34.19
C2 and WW domain containing	0.0009820	0.04408	34.19
MutS homologs	0.0009820	0.04408	34.19

Adjusted p-value significance cutoff of 0.05.

For example, the canonical ZNF226 is detected from 4-cell through morula, but the chimeric ZNF226:ERVL-E-int is specific to 4-cell (Figure 3.1 A-D).

We chose to further characterize the ZBTB16:MIRb chimera because it is (1) the highest expressed retrotransposon:ZNF transcript in human preimplantation embryos, (2) generates a putative protein product that may differ in function from the canonical protein, and (3) is a previously studied gene that is known to be important for development([135],[136]). Characterizing the properties of the ZBTB16:MIRb protein could give insight into how retrotransposons shape the development of the human embryo.

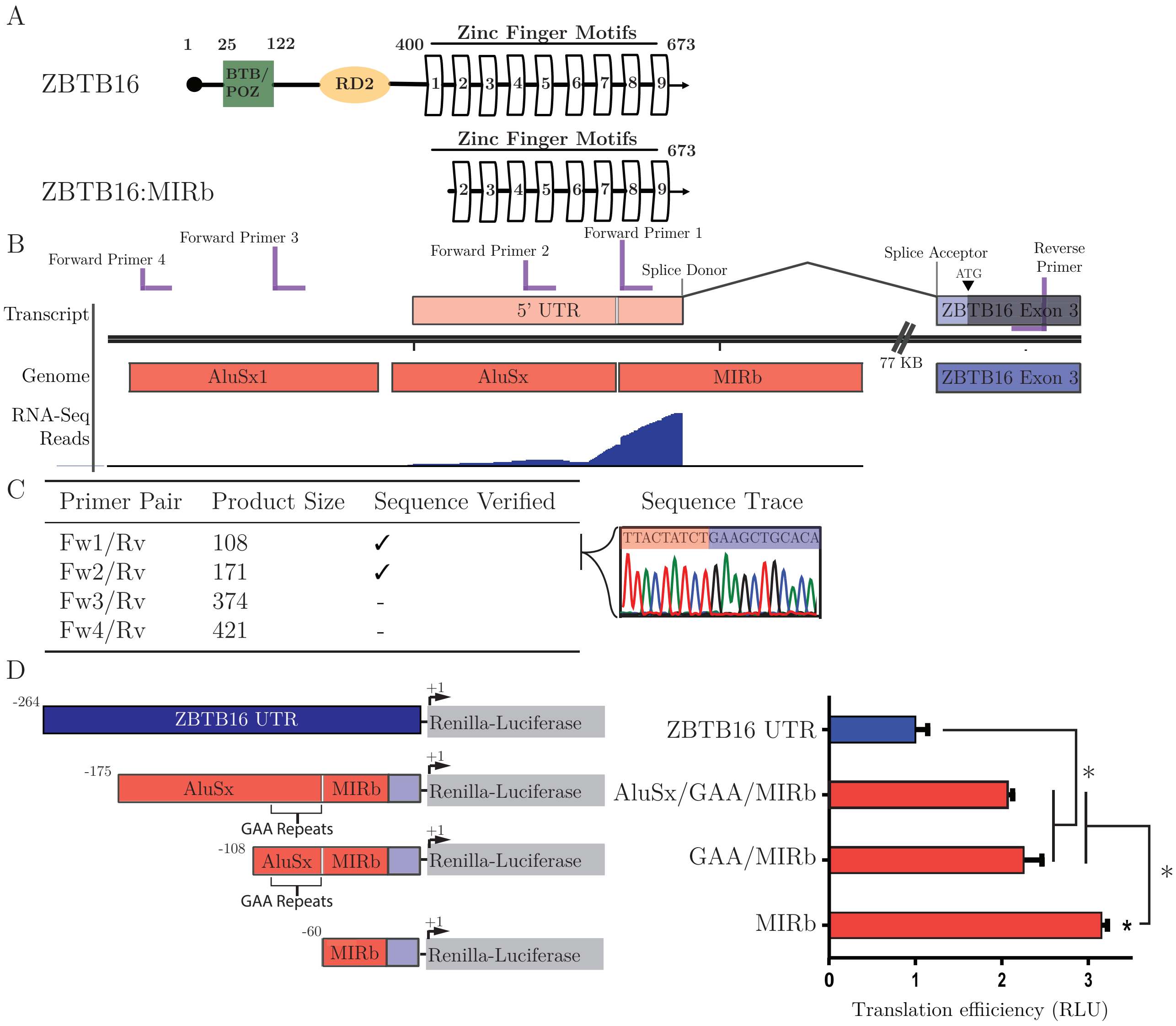
The primary objectives of this study are to determine the likelihood that ZBTB16:MIRb chimeric protein is actually transcribed and translated in the human embryo, and to determine if its expression may have functional consequence in human embryogenesis. The major limitation in studying a human-specific protein isoform in development is that human embryos are extremely difficult to obtain and manipulate. While previous studies have knocked down retrotransposon sequences in developing human embryos([99]), we do not have access to the material or ethical approval to manipulate ZBTB16 levels during human preimplantation development. Furthermore, this is a human specific isoform; we cannot disrupt it using a model system like mouse. Therefore, we aimed to validate that the ZBTB16:MIRb chimeric transcript is expressed in human oocytes, and we address the functional importance of the ZBTB16:MIRb chimera in vitro.

Figure 3.1: Candidate ZNF:retrotransposon chimeric genes show distinct expression patterns in human preimplantation embryos

A. Schematic showing ZNF226:ERVL-E-int chimeric product **B.** ZNF606:MER76-int chimera **C.** ZNF544:MLT1C chimera **D**. ZBTB16:MIRB chimeric product. Canonical transcripts are in blue and retrotransposon containing chimeric transcripts are in red. Junction dept is the number of uniquely mapped reads that span the junction between the gene and retrotransposon (for chimeric) or been two exons not part of the chimeric transcript (for canonical). Error bar is standard deviation from all single cell RNA seq replicates from that stage. Drawings not to scale.

Validation of the ZBTB16:MIRb chimera

We first aimed to validate that the ZBTB16:MIRb chimeric transcript is expressed in the human oocyte. Our ideal experimental strategy was to use cDNA from human oocytes and perform 5' Rapid Amplification of cDNA Ends (5' RACE), capturing the junction between MIRb and ZBTB16, as well as identifying the transcriptional start site. However, due to extremely limited quantities of human oocyte derived cDNA, we instead devised a conventional PCR based strategy. Briefly, we designed four forward primers within the AluSx/MIRb element and one reverse primer that anneals inside the third exon of ZBTB16. Collaborators at the Third Military Medical University in Chongqing, China ran three PCR reactions per primer pair using cDNA generated from human ooctyes. They successfully amplified bands of the expected size for Fw1/Rv and Fw2/Rv primer pairs, but not for Fw3/RV or Fw4/RV reverse (Figure 3.2 B-C). Sequencing the products validates the junction is identical to the one predicted by RNA-seq, verifying the ZBTB16:MIRb transcript is present in human oocytes (Figure 3.2 C). The actual splice site is within MIRb and is found inside a splicing motif (TCTgtatgt) recognized by Human Splicing Finder software([137]), suggesting the splicing may occur via canonical pathways. It is interesting that the Fw3 and Fw4 primers failed to amplify. We believe this is because they lie before the predicted transcriptional start site (Figure 3.2 B). However, it is possible they failed simply due to technical reasons.



Primer Pair	Product Size	Sequence Verified
Fw1/Rv	108	✓
Fw2/Rv	171	✓
Fw3/Rv	374	-
Fw4/Rv	421	-

Figure 3.2: ZBTB16:MIRb in present in human oocytes and the ZBTB16:MIRb 5'UTR enhances translation efficiency

A. Schematic of ZBTB16 and ZBTB16:MIRb protein structure. Canonical ZBTB16 has n-terminal BTB/POZ and RD2 domains, as well as 9 zinc fingers. The BTB/POZ and RD2 domains are truncated in ZBTB16:MIRb. **B.** Schematic of the Alu/MIRb promoter active in human oocytes. Primers used for validatation experiment are shown in purple. RNA seq is from a representative 2-cell blastomere. **C.** Table summarizing the validation of ZBTB16 in the human embryo, with the sequenced junction shown on right. **D**. ZBTB16:MIRb 5' UTR luciferase assay. Depicted 5' UTRs were in vitro transcribed in front of renilla-luciferase and then transfected into HEK293T cells. Data is representative from three independent experiments and and error is standard deviation from three technical replicates. Signficance tested with a two sided Student's t-test.

Stability and localization of ZBTB16:MIRb chimeric protein

While validation of the ZBTB16:MIRb chimeric transcript in oocytes is crucial, it does not address the likelihood the chimeric mRNA may be capable of translation or that the truncated protein is stable. To determine if the chimeric 5' UTR is capable of translation, we tested its translation efficiency using a luciferase reporter in HEK293Ts. We find chimeric 5' UTR produces significantly more luciferase than the canonical ZBTB16 5' UTR (Figure 3.2 D). We are not certain of the precise transcriptional start site, so we also created two truncations which remove sections of the Alu element. Removing the left arm of ALU while retaining a set of GAA repeats did not effect translation, but when the GAA repeats are removed we observe an increase in translation efficiency. GAA repeats are known to form an ordered single-stranded RNA structure due to stacking interactions([138]), so it is possible the chimeric ZBTB16:MIRb 5'UTR forms secondary structure that regulates translation. In summary, it appears likely the chimeric ZBTB16:MIRb transcript is a good substrate for translation initiation.

We next determined if the truncated protein product is stable. We tagged both isofroms with a 3xFLAG peptide and and overexpressed them in HEK293Ts, finding both proteins are produced at the expected sizes, with little evidence of degradation. Interestingly, we observed two bands for ZBTB16:MIRb, suggesting it may be subject to post-translational modification (Figure 3.3 A). The truncated isoform of ZBTB16:MIRb lacks the n-terminal BTB/POZ and RD2 domains (Figure 3.2 A), which are important for nuclear-cytoplasmic shuttling([139]). To determine if ZBTB16:MIRb protein has altered subcellular localization, we overexpressed both isoforms in U2OS cells. In agreement with previous studies([140]) canonical ZBTB16 showed both nuclear and cytoplasmic signal. However, ZBTB16:MIRb expression was entirely nuclear, suggesting the Alu/MIRb insertion in human embryos creates a ZBTB16 protein isoform with altered subcellular localization (Figure 3.3 B).

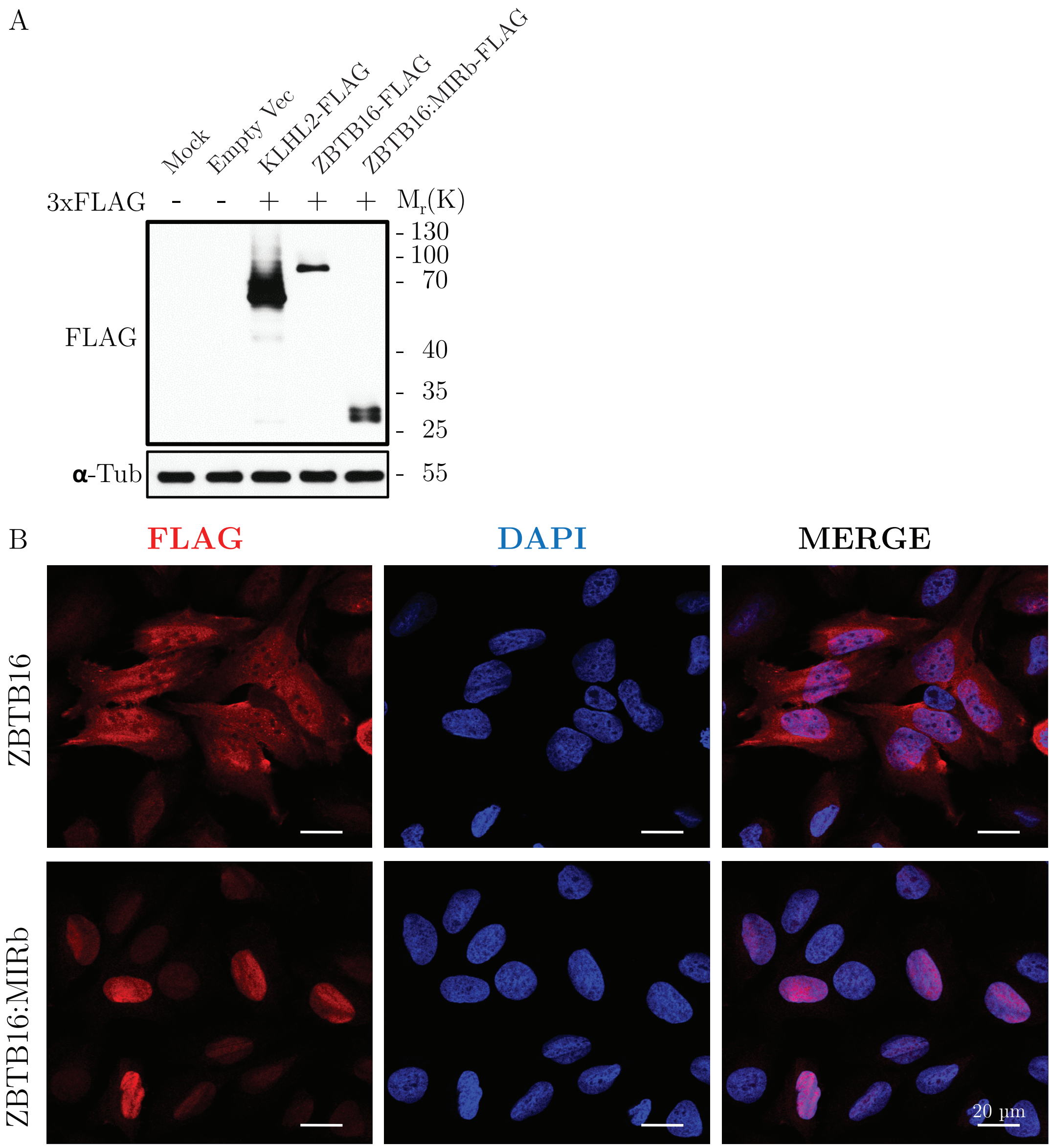


Figure 3.3: ZBTB16:MIRb generates a stable protein with altered subcellular localization

A. To address stability of ZBTB16:MIRb, flag-tagged ZBTB16 and ZBTB16:MIRb were overexpressed in HEK293T cells and analyzed by anti-flag western blot. KLHL2-FLAG is included as a positive control and anti-tubulin was used to show equal loading. Data is representative from 4 independent experiments. **B.** To address protein subcellular localization of ZBTB16:MIRb, stable U2OS cell lines expressing flag-tagged ZBTB16 and ZBTB16:MIRb were generated and immunofluorescence was performed. Data is representative from 3 independent experiments.

Effect of ZBTB16 on the cell cycle

CDK2 phosphorylates ZBTB16 at serine 197 and tyrosine 282 and targets it for ubiquitin mediated degradation, antagonizing the ZBTB16 repression of Cyclin-A2 and allowing entry into the cell cycle([140],[139]). However, the truncated ZBTB16:MIRb isoform lacks these phosphorylation sites (Figure 3.2 A) and shows only nuclear localization, suggesting it may retain its function as a cell cycle regulator while resisting CDK2 mediated degradation. To address if ZBTB16:MIRb can regulate the cell cycle, we overexpressed both isoforms in synchronized HEK293T cells. Both the full length and truncated versions of ZBTB16 caused an increase in the percentage of cells in G1 phase, before gradually declining by 12 hours post release (Figure 3.4 A). The expression of Cylin A2 transcripts was also repressed at 12 hours in both ZBTB16 and ZBTB16:MIRb overexperssion conditions, suggesting both isoforms slow the cell cycle through repression of Cyclin-A2 (Figure 3.4 B). Data generated in our lab by Sebastian Henkel shows that in HEK293T cells, canonical ZBTB16 is subject to proteosome mediated degradation while ZBTB16:MIRb is not (data not shown). In summary, our data suggests ZBTB16:MIRb retains its function as a cell cycle regulator but is insensitive to ubiquitin mediated degradation.

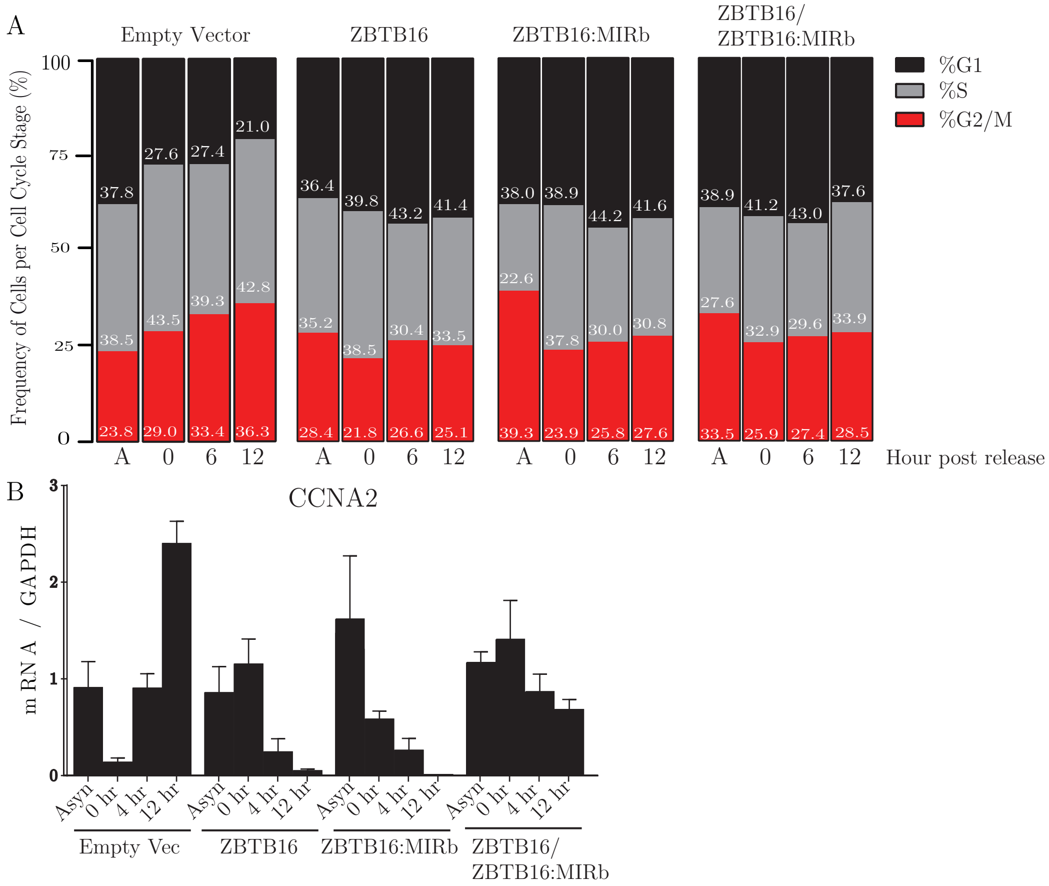


Figure 3.4: ZBTB16:MIRb effects the cell cycle and represses CCNA2

A. To access ZBTB16:MirB activity in the cell cycle, both ZBTB16 and ZBTB16:MIRb isoforms were overexpressed in HEK293T cells. Asynchonous (A) cells were synchronized using serum starvation and then released. DNA content per cell was quantified with cytofluorometric analysis at various timepoints using propidium iodide staining. Stacked bar plot showing percentage of cells in G2,S, and G2/M. Data is representative from two independent experiments. **B.** qPCR analysis of CCNA2(Cyclin-A2) from the cells analyzed in A. Error bars are standard error of three technical measurements.

3.3 Discussion

We present evidence that an isoform of ZBTB16, an important cell cycle regulator, is transcribed from a Alu/MIRb insertion early human preimplantation development. This isoform produces a truncated ZBTB16 protein that shows altered subcellular localization and retains is function as a cell cycle regulator. ZBTB16 is acetylated by P300 within the 9th zinc finger, a process that enhances DNA binding and is required for ZBTB16 cell cycle regulation([141]). This domain is retained in ZBTB16:MIRb, suggesting a possible mechanistic explanation for how ZBTB16:MIRb maintains cell cycle regulation. Work in the lab by Sebastian Henkel shows that ZBTB16:MIRb is resistant to proteosome degradation. Ubiquitin mediated degradation of CDK2 phosphorylates n-terminal residues that are truncated in the ZBTB16:MIRb isoform, so future studies may determine if ZBTB16:MIRb is a more potent cell cycle inhibitor than canonical ZBTB16.

We would also like to note that ZBTB16 was first identified in an individual with acute promyelotic leukemia, where a chromosomal translocation joined ZBTB16 with the RARA gene coding for Retinoic Acid Receptor Alpha([142]). This translocation event occurred between exons 2 and 3 of the ZBTB16 gene, generating a RARA-ZBTB16 fusion protein that contains the same 7 zinc fingers that are retained in the truncated ZBTB16:MIRb protein. RARA-ZBTB16 was shown to confer retinoic acid resistance in an acute myeloid leukemia cell line through recruitment of p300 to the promoter of CRABPI([143]). We believe this observation is circumstantial evidence supporting our model that the terminal 7 zinc fingers constitute a functional protein domain.

Finally, we would like to speculate about the function of ZBTB16 in the human embryogenesis. Mouse embryos reach the late blastocyst stage after 84-96 hours, while human embryos take an additional 24-30 hours([144]). Cyclin-A2 is expressed during this time([145]), so ZBTB16:MIRb expression could effect its dynamics and explain, at least in part, slower embryonic maturation. Alternatively, ZBTB16:MIRb could act during oocyte maturation, where Cyclin-A2 is known to regulate the cell cycle in mouse([146]). While direct manipulation of developing human embryos is unlikely due to ethical concerns, future experiments could introduce the ZBTB16:MIRb chimera into mouse embryos to test for effects on the embryonic cell cycle.

3.4 Methods

Identification of chimeric transcripts in human embryogenesis

RNA seq data from human preimplantation embryos (accession number GSE36552) was mapped with TopHat (v. 2.0.11) to the hg19 reference genome. Retrotransposon-gene junctions (chimeric reads) were defined as reads that overlap on one end with an annotated exon of an Ensembl gene and on the other end with an annotated retrotransposon. We retained reads that contain at least 10 counts in two samples, and normalized single cell RNA-seq samples using generalized linear model using edgeR. Gene ontology analysis was performed for chimeric genes that using MSigDB gene sets (http://software.broadinstitute.org/gsea/msigdb/collections.jsp) and significance was calculated using Fisher's hypergeometric test. This bioinformatic analysis was performed by our collaborator Anne Biton (Institut Pasteur).

ZNF:retrotransposon candidate evaluation

Candidate chimeric ZNFs were inspected manually using The Integrative Genomics Viewer (IGV). The top 30 out of 60 expressed zinc finger genes were individually evaluated for the presence of the RT-exon junction. Junction read depth was determined manually for every developmental stage. Chimeric junctions with less than 30 reads only appeared in a some datasets and often showed unrealistic splicing patterns, so the candidate list was refined to chimeric transcripts showing over 30 reads in at least 50% of the data sets of one developmental stage. This list was further pruned to candidates that showed a higher percentage (>50%) of chimeric transcript compared to canonical isoform.

Validation of ZBTB16 in human oocytes

To determine the presence of ZBTB16:MIRb in human oocytes, primers were designed spanning the junction between MIRb and ZBTB16 exon 3. PCR amplification from using cDNA from human oocytes was performed by collaborators at the Third Military Medical University in Chongqing, China. Primer sequences are
ZBTB16 FW1: 5' GCACCCTGGATGAAGACTCA 3'.
ZBTB16 FW2: 5' GTTGCACTCCAGCCCAAGAC 3'.
ZBTB16 FW3: 5' GGCAGGAAAATTGCTTGAAGG 3'.
ZBTB16 FW4: 5'AATTAGCTGGGTGGCAGGTG 3'.
ZBTB16 RV: 5' GCAAACTATCCAGGAACCGC 3'.

Luciferase assay

Translation efficiency of ZBTB16 was determined using luciferase assays. mRNA was in vitro transcribed from psiCHECK-2 plasmids containing ZBTB16 and ZBTB16:MIRb 5'

UTR-renilla luciferase using HiScribe™ T7 ARCA mRNA kit (New England BioLabs, Cat. E2065S). It was then 5' capped, and poly-A tailed also using HiScribe™ T7 ARCA mRNA kit. mRNA was purified using MEGAclear transcription clean up kit (Ambion, Cat: AM1908) then cotransfected with mRNA encoding firefly luciferase into into HEK293t cells using lipofectamine 2000. Dual-Luciferase® Reporter Assay (Promega, Cat. E1910) was performed in triplicate TECAN Infinite® F200 microplate reader. Luciferase signal between samples was normalized to firefly luciferase.

Western blot

Western blot was used to determine the stability of ZBTB16 protein. pLV-GFP plasmids encoding ZBTB16-3x FLAG or ZBTB16:MIRb-3x FLAG were transfected into HEK293T cells using PEI. Cells were collected in RIPA lysis buffer and protein concentration was measured with DC Protein Assay (Bio-Rad, Cat. 5000111). 20 μg total protein was subjected to SDS-PAGE and transfered to 0.2 μm nitrocellulose membranes. After blocking with TBST and 5% non-fat milk, primary antibodies (anti-FLAG M2, Sigma-Aldrich Cat. F3165, 1:1000 dillution and anti-tubulin, Abcam Cat. ab7291, 1:1000 dillution) were incubated for 1 h at room temperature followed by incubation for 1 h with the appropriate horseradish peroxidase-conjugated secondary antibodies. Proteins were visualized by chemiluminescence (Pierce ECL Western Blotting Substrate, Thermo Scientific Cat. 32109).

Generation of U2OS ZBTB16 expressing stable cell lines

To generate infectious lentivirus containing the desired constructs, HEK293T cells were transfected with three plasmids: transfer plasmid pLV-eGFP (Addgene plasmid 36083) containing either ZBTB16-3xFLAG or ZBTB16:MIRb-3xFLAG, packaging plasmid VSVG and envelope plasmid ΔVPR. Cells were transfected using PEI and virus was harvested at 24 h and 48 h after transfection. U2OS cells were transduced with 1 mL of virus containing supernatant supplemented with 5 μ/mL polybrene (SC Biotechnology, Cat. sc-134220) for 24 h. Positive cells were selected using 2 μ/ml pyromycin.

Immunofluorescence

Subcellular localization of ZBTB16 protein was determined using immunofluorescence. U20S cell lines with stable ZBTB16 or ZTB16:MIRb expression were seeded onto poly-lysine coated glass cover slides (100 μ/ml-poly lysine, Sigma-Aldrich Cat. 25988-63-0). Cells were fixed for 10 minutes with 3.7% formaldehyde and permeabilized for 10 minutes with 0.5% Triton X-100 in PBS. Blocking was performed for 1 h at RT using PBS containing 3% BSA. Primary antibody staining (rabbit anti-FLAG (SIGMA, Cat. F7425, 1:300 dilution) was performed overnight, followed by incubation with secondary antibody (goat anti-rabbit-IgG, Alexa Flour 594, Life technology, Cat. A11005, 1:200

dillution). (Life technology, Cat. A11037). DAPI staining was performed for 5 min at RT using 1 μ/ml DAPI in PBS. The slides were mounted with ProLong® Gold Antifade Reagent (Invitrogen, Cat. P36930) and images were captured with a Zeiss LSM700 confocal laser scanning microscope.

Cell cycle analysis

The effect of ZBTB16 expression on the cell cycle was determined by propidium iodine staining cell cyle analysis. HEK293T cells were transfected with plasmids encoding ZBTB16 or ZTB16:MIRb using polyethylene glycol (PEI) and 24 hours later were serum starved for 18 hours to synchronize the cell cycle. Cells were harvested at six hour timepoints and fixed in 70% ethanol and then stained with propidium iodide (50 μg/ml). The cell cycle profiles were determined by flow cytometry using the BD LSRFortessa cell analyzer. Compensation and final analysis was performed using FlowJo.

Quantitative real-time PCR

RNA was prepared from transfected cells using TRIZOL according to manufacturers instructions (Life Technologies, Cat. 15596). RNA was treated with DNAse for 15 minutes (Invitrogen, Cat. 18068015) and reverse transcribed using iScript Advanced Reverse-Transcriptase (Bio-Rad, Cat. 1725037). Quantitative real time PCR was performed with SYBR FAST qPCR Master Mix (Kapa Biosystems, Cat. KK4604) and Applied Biosystems StepOnePlus Real Time System. The primer sequences for CCNA2: Fw- 5' CCAGGAGAATATCAACCCGGA 3', RV- 5' GGTGCAACCCGTCTCGT 3' and GAPDH: Fw- 5' AGCCACATCGCTCAGACAC 3', Rv- 5' GCCCAATACGACCAAATCC 3'.

Bibliography

1. C. Elegans Consortium, S. Genome sequence of the nematode C. elegans: a platform for investigating biology. **282,** 1–8 (2005).
2. Mouse Genome Consortium, S. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
3. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Hong, M. L., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., De La

Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Patrinos, A. & Morgan, M. J. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

4. San-Miguel, P., Tikhonov, A., Jin, Y. K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P. S., Edwards, K. J., Lee, M., Avramova, Z. & Bennetzen, J. L. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274,** 765–768 (1996).

5. Picard, G., Bregliano, J., Bucheton, A., Lavige, J., Pelisson, A. & Kidwell, M. Non-mendelian female sterility and hybrid dysgenesis in Drosophila melanogaster. *Genetical Research* **32,** 275–287 (1978).

6. Rubin, G. M., Kidwell, M. G. & Bingham, P. M. The molecular basis of P-M hybrid dysgenesis: The nature of induced mutations. *Cell* **29,** 987–994 (1982).

7. Doolittle, F. & Sapienza, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284,** 601–603 (1980).

8. Pardue, M.-L. & DeBaryshe, P. Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. *Annual Review of Genetics* **37,** 485–511 (2003).

9. Chuong, E. B., Karim Rumi, M. A., Soares, M. J. & Baker, J. C. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nature Genetics* **45,** 325–9 (Mar. 2013).

10. Sha, M., Lee, X., ping Li, X., Veldman, G. M., Finnerty, H., Racie, L., LaVallie, E., Tang, X. Y., Edouard, P., Howes, S., Keith, J. C. & McCoy, J. M. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403,** 785–789 (2000).

11. Dupressoir, A., Vernochet, C., Harper, F., Guegan, J., Dessen, P., Pierron, G. & Heidmann, T. A pair of co-opted retroviral envelope syncytin genes is required for formation of the two-layered murine placental syncytiotrophoblast. *Proceedings of the National Academy of Sciences* **108,** E1164–E1173 (2011).

12. Biémont, C. *A brief history of the status of transposable elements: From junk DNA to major players in evolution* 2010.

13. Herron, P. R. *Mobile DNA II* (2002).

14. Feschotte, C. & Pritham, E. J. DNA Transposons and the evolution of eukaryotic genomes. *Annual Review of Genetics* **41,** 331–368 (2007).

15. Swergold, G. D. Identification, characterization, and cell specificity of a human LINE-1 promoter. *Molecular and cellular biology* (1990).

16. Furano, A. V. The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. *Progress in nucleic acid research and molecular biology* (2000).

17. Dewannieux, M., Dupressoir, A., Harper, F., Pierron, G. & Heidmann, T. Identification of autonomous IAP LTR retrotransposons mobile in mammalian cells. *Nature Genetics* (2004).

18. Bannert, N. & Kurth, R. The evolutionary dynamics of human endogenous retroviral families. *Annual Review of Genomics and Human Genetics* (2006).

19. Ribet, D., Dewannieux, M. & Heidmann, T. An active murine transposon family pair: Retrotransposition of ”master” MusD copies and ETn trans-mobilization. *Genome Research* (2004).

20. Ohnuki, M., Tanabe, K., Sutou, K., Teramoto, I., Sawamura, Y., Narita, M., Nakamura, M., Tokunaga, Y., Nakamura, M., Watanabe, a., Yamanaka, S. & Takahashi, K. Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proceedings of the National Academy of Sciences* **111,** 12426–12431 (2014).

21. Flemr, M., Malik, R., Franke, V., Nejepinska, J., Sedlacek, R., Vlahovicek, K. & Svoboda, P. A retrotransposon-driven dicer isoform directs endogenous small interfering rna production in mouse oocytes. *Cell* (2013).

22. Eckersley-Maslin, M. A., Svensson, V., Krueger, C., Stubbs, T. M., Giehr, P., Krueger, F., Miragaia, R. J., Kyriakopoulos, C., Berrens, R. V., Milagre, I., Walter, J., Teichmann, S. A. & Reik, W. MERVL/Zscan4 Network Activation Results in Transient Genome-wide DNA Demethylation of mESCs. *Cell Reports* **17,** 179–192 (2016).

23. Sjottem, E., Anderssen, S. & Johansen, T. The promoter activity of long terminal repeats of the HERV-H family of human retrovirus-like elements is critically dependent on Sp1 family proteins interacting with a GC / GT box located immediately 3 ' to the TATA box. *Journal of Virology* **70,** 188–198 (1996).

24. Hendrickson, P. G., Doráis, 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., Nix, D. A., Peterson, C. M., Tapscott, S. J., Carrell, D. T. & Cairns, B. R. Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nature Genetics* **49,** 925–934 (2017).

25. Santoni, F. a., Guerra, J. & Luban, J. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology* **9,** 111 (Jan. 2012).

26. Kim, Y. K., Bourgeois, C. F., Pearson, R., Tyagi, M., West, M. J., Wong, J., Wu, S. Y., Chiang, C. M. & Karn, J. Recruitment of TFIIH to the HIV LTR is a rate-limiting step in the emergence of HIV from latency. *EMBO Journal* (2006).

27. Berkhout, B., Silverman, R. H. & Jeang, K. T. Tat trans-activates the human immunodeficiency virus through a nascent RNA target. *Cell* (1989).

28. Soto-Rifo, R., Limousin, T., Rubilar, P. S., Ricci, E. P., Décimo, D., Moncorgé, O., Trabaud, M. A., André, P., Cimarelli, A. & Ohlmann, T. Different effects of the TAR structure on HIV-1 and HIV-2 genomic RNA translation. *Nucleic Acids Research* (2012).

29. Trubetskoy, A. M., Okenquist, S. A. & Lenz, J. R region sequences in the long terminal repeat of a murine retrovirus specifically increase expression of unspliced RNAs. *Journal of virology* (1999).

30. Iwasaki, Y. W., Siomi, M. C. & Siomi, H. PIWI-Interacting RNA: Its Biogenesis and Functions. *Annual Review of Biochemistry* **84,** 405–433 (2015).

31. Iwasaki, Y. W., Murano, K., Ishizu, H., Shibuya, A., Iyoda, Y., Siomi, M. C., Siomi, H. & Saito, K. Piwi Modulates Chromatin Accessibility by Regulating Multiple Factors Including Histone H1 to Repress Transposons. *Molecular Cell* **63,** 408–419 (2016).

32. Klenov, M. S., Lavrov, S. A., Korbut, A. P., Stolyarenko, A. D., Yakushev, E. Y., Reuter, M., Pillai, R. S. & Gvozdev, V. A. Impact of nuclear Piwi elimination on chromatin state in Drosophila melanogaster ovaries. *Nucleic Acids Research* **42,** 6208–6218 (2014).

33. Koito, A. & Ikeda, T. *Intrinsic immunity against retrotransposons by APOBEC cytidine deaminases* 2013.

34. Hu, X., Zhu, W., Chen, S., Liu, Y., Sun, Z., Geng, T., Wang, X., Gao, B., Song, C., Qin, A. & Cui, H. Expression of the env gene from the avian endogenous retrovirus ALVE and regulation by miR-155. *Archives of Virology* **161,** 1623–1632 (2016).

35. Choi, Y. J., Lin, C.-P., Risso, D., Chen, S., Kim, T. A., Tan, M. H., Li, J. B., Wu, Y., Chen, C., Xuan, Z., Macfarlan, T., Peng, W., Lloyd, K. C. K., Kim, S. Y., Speed, T. P. & He, L. Deficiency of microRNA miR-34 expands cell fate potential in pluripotent stem cells. *Science* **355** (2017).

36. Berrens, R. V., Andrews, S., Spensberger, D., Santos, F., Dean, W., Gould, P., Sharif, J., Olova, N., Chandra, T., Koseki, H., von Meyenn, F. & Reik, W. An endosiRNA-Based Repression Mechanism Counteracts Transposon Activation during Global DNA Demethylation in Embryonic Stem Cells. *Cell Stem Cell* **21,** 694–703.e7 (2017).

37. Sharma, U., Conine, C. C., Shea, J. M., Boskovic, A., Derr, A. G., Bing, X. Y., Belleannee, C., Kucukural, A., Serra, R. W., Sun, F., Song, L., Carone, B. R., Ricci, E. P., Li, X. Z., Fauquier, L., Moore, M. J., Sullivan, R., Mello, C. C., Garber, M. & Rando, O. J. Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals. *Science* **351,** 391–396 (2016).

38. Schorn, A. J., Gutbrod, M. J., LeBlanc, C. & Martienssen, R. LTR-Retrotransposon Control by tRNA-Derived Small RNAs. *Cell* **170,** 61–71.e11 (2017).

39. Walsh, C. P., Chaillet, R. & Bestor, T. H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature Genetics* **20,** 116–117 (Oct. 1998).

40. Kato, Y., Kaneda, M., Hata, K., Kumaki, K., Hisano, M., Kohara, Y., Okano, M., Li, E., Nozaki, M. & Sasaki, H. Role of the Dnmt3 family in de novo methylation of imprinted and repetitive sequences during male germ cell development in the mouse. *Human Molecular Genetics* (2007).

41. Bourc'his, D. & Bestor, T. H. *Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L* 2004.

42. Dean, W., Santos, F., Stojkovic, M., Zakhartchenko, V., Walter, J., Wolf, E. & Reik, W. Conservation of methylation reprogramming in mammalian development: Aberrant reprogramming in cloned embryos. *Proceedings of the National Academy of Sciences* **98,** 13734–13738 (2001).

43. Macfarlan, T. S., Gifford, W. D., Agarwal, S., Driscoll, S., Lettieri, K., Wang, J., Andrews, S. E., Franco, L., Rosenfeld, M. G., Ren, B. & Pfaff, S. L. Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. *Genes and Development* **25,** 594–607 (2011).

44. Maksakova, I. a., Thompson, P. J., Goyal, P., Jones, S. J., Singh, P. B., Karimi, M. M. & Lorincz, M. C. Distinct roles of KAP1, HP1 and G9a/GLP in silencing of the two-cell-specific retrotransposon MERVL in mouse ES cells. *Epigenetics & chromatin* **6,** 15 (Jan. 2013).

45. Matsui, T., Leung, D., Miyashita, H., Maksakova, I. A., Miyachi, H., Kimura, H., Tachibana, M., Lorincz, M. C. & Shinkai, Y. Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nature* **464,** 927–931 (2010).

46. Karimi, M. M., Goyal, P., Maksakova, I. a., Bilenky, M., Leung, D., Tang, J. X., Shinkai, Y., Mager, D. L., Jones, S., Hirst, M. & Lorincz, M. C. DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell stem cell* **8,** 676–87 (June 2011).

47. Wolf, D. & Goff, S. P. TRIM28 Mediates Primer Binding Site-Targeted Silencing of Murine Leukemia Virus in Embryonic Cells. *Cell* **131,** 46–57 (2007).

48. Wolf, D. & Goff, S. P. Embryonic stem cells use ZFP809 to silence retroviral DNAs. *Nature* **458,** 1201–1204 (2009).

49. Hamilton, a. T., Huntley, S., Gordon, L. & Stubbs, L. Evolutionary expansion and divergence in a large family of primate-specific zinc finger transcription factor genes. *Genome research* **16,** 584–594 (2005).

50. Jacobs, F. M. J., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A. D., Katzman, S., Paten, B., Salama, S. R. & Haussler, D. An evolutionary arms race between KRAB zinc finger genes 91/93 and SVA/L1 retrotransposons. *Nature* **516,** 242–245 (2014).

51. Imbeault, M., Helleboid, P. Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543,** 550–554 (2017).

52. Rowe, H. M., Kapopoulou, A., Corsinotti, A., Fasching, L., Macfarlan, T. S., Tarabay, Y., Viville, S., Jakobsson, J., Pfaff, S. L. & Trono, D. TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. *Genome research* **23,** 452–61 (Mar. 2013).

53. Wolf, G., Yang, P., Füchtbauer, A. C., Füchtbauer, E. M., Silva, A. M., Park, C., Wu, W., Nielsen, A. L., Pedersen, F. S. & Macfarlan, T. S. The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses. *Genes and Development* **29,** 538–554 (2015).

54. Tupler, R., Perini, G. & Green, M. R. Expressing the human genome. *Nature* **409,** 832–833 (2001).

55. Vargas, A., Moreau, J., Landry, S., LeBellego, F., Toufaily, C., Rassart, E., Lafond, J. & Barbeau, B. Syncytin-2 plays an important role in the fusion of human trophoblast cells. *J Mol Biol* (2009).

56. Roberts, M., Green, J. A. & Schulz, L. C. The evolution of the placenta. *Reproduction* **152,** 179–189 (2016).

57. Nakaya, Y., Koshi, K., Nakagawa, S., Hashizume, K. & Miyazawa, T. Fematrin-1 Is Involved in Fetomaternal Cell-to-Cell Fusion in Bovinae Placenta and Has Contributed to Diversity of Ruminant Placentation. *Journal of Virology* (2013).

58. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351,** 1083–1087 (2016).

59. Bourque, G., Leong, B., Vega, V. B., Chen, X., Yen, L. L., Srinivasan, K. G., Chew, J. L., Ruan, Y., Wei, C. L., Huck, H. N. & Liu, E. T. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Research* (2008).

60. Mak, K. S., Burdach, J., Norton, L. J., Pearson, R. C., Crossley, M. & Funnell, A. P. Repression of chimeric transcripts emanating from endogenous retrotransposons by a sequence-specific transcription factor. *Genome Biology* (2014).

61. Wang, T., Zeng, J., Lowe, C. B., Sellers, R. G., Salama, S. R., Yang, M., Burgess, S. M., Brachmann, R. K. & Haussler, D. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 18613–8 (Nov. 2007).

62. Peaston, A. E., Evsikov, A. V., Graber, J. H., de Vries, W. N., Holbrook, A. E., Solter, D. & Knowles, B. B. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev Cell.* arXiv: NIHMS150003 (2004).

63. Franke, V., Ganesh, S., Karlic, R., Malik, R., Pasulka, J., Horvat, F., Kuzman, M., Fulka, H., Cernohorska, M., Urbanova, J., Svobodova, E., Ma, J., Suzuki, Y., Aoki, F., Schultz, R. M., Vlahovicek, K. & Svoboda, P. Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes. *Genome Research* (2017).

64. Kapusta, A., Kronenberg, Z., Lynch, V., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M. & Feschotte, C. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genetics* **9,** 1–20 (Apr. 2013).

65. Conley, A. B., Miller, W. J. & Jordan, I. K. *Human cis natural antisense transcripts initiated by transposable elements* 2008.

66. 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., Pera, R. a. R. & Wysocka, J. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522,** 221–5 (2015).

67. Macfarlan, T. S., Gifford, W. D., Driscoll, S., Lettieri, K., Rowe, H. M., Bonanomi, D., Firth, A., Singer, O., Trono, D. & Pfaff, S. L. ES cell potency fluctuates with endogenous retrovirus activity. *Nature* **487,** 57–63 (2012).

68. Schoorlemmer, J., Pérez-Palacios, R., Climent, M., Guallar, D. & Muniesa, P. Regulation of mouse retroelement MuERV-L/MERVL expression by REX1 and epigenetic control of stem cell potency. *Frontiers in Oncology* **4,** 1–18 (Jan. 2014).

69. Blanco-Melo, D., Gifford, R. J. & Bieniasz, P. D. Reconstruction of a replication-competent ancestral murine endogenous retrovirus-L. *Retrovirology* **15,** 1–17 (2018).

70. Huang, Y., Kim, J. K., Do, D. V., Lee, C., Penfold, C. A., Zylicz, J. J., Marioni, J. C., Hackett, J. A. & Surani, M. A. Stella modulates transcriptional and endogenous retrovirus programs during maternal-to-zygotic transition. *eLife* (2017).

71. Evans, M. J. & Kaufman, M. H. Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292,** 154–156 (1981).

72. Martin, G. R. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proceedings of the National Academy of Sciences of the United States of America* **78,** 7634–7638 (1981).

73. Williams, R. L., Hilton, D. J., Pease, S., Willson, T. A., Stewart, C. L., Gearing, D. P., Wagner, E. F., Metcalf, D., Nicola, N. a. & Gough, N. M. Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells. *Nature* **336,** 684–687 (1988).

74. Smith, A. G., Heath, J. K., Donaldson, D. D., Wong, G. G., Moreau, J., Stahl, M. & Rogers, D. Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. *Nature* **336,** 688–690 (1988).

75. Burdon, T., Smith, A. & Savatier, P. Signalling, cell cycle and pluripotency in embryonic stem cells. *Trends in Cell Biology* **12,** 432–438 (2002).

76. Rizzino, A. The Sox2-Oct4 connection: critical players in a much larger interdependent network integrated at multiple levels. *Stem Cells* **21,** 1033–1039 (2013).

77. Chen, T., Ueda, Y., Dodge, J. E., Wang, Z. & Li, E. Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. *Molecular and Cellular Biology* **23,** 5594–5605 (2003).

78. Leahy, A. M. Y., Xiong, J.-w., Kuhnert, F. & Stuhlmann, H. Use of developmental marker genes to define temporal and spatial patterns of differentiation during embryoid body formation. *Journal of Experimental Zoology* **284,** 68–81 (1999).

79. Percharde, M., Lin, C.-j., Yin, Y., Huang, B., Shen, X., Ramalho-santos, M., Percharde, M., Lin, C.-j., Yin, Y., Guan, J., Peixoto, G. A. & Bulut-karslioglu, A. A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. *Cell* **174,** 1–15 (2018).

80. Thomson, J. A., Itskovitz-eldor, J., Shapiro, S. S., Waknitz, M. A., Swiergiel, J. J., Marshall, V. S. & Jones, J. M. Embryonic stem cell lines derived from human blastocysts. *Science* **282,** 1145–1147 (1998).

81. Weinberger, L., Ayyash, M., Novershtern, N. & Hanna, J. H. Understanding stem cell states : naïve to primed pluripotency in rodents and humans. *bioRxiv* **16,** 1–34 (2015).

82. Levine, S. & Grabel, L. The contribution of human/non-human animal chimeras to stem cell research. *Stem Cell Research* **24,** 128–134 (2017).

83. Theunissen, T. W., Powell, B. E., Wang, H., Mitalipova, M., Faddah, D. A., Reddy, J., Fan, Z. P., Maetzel, D., Ganz, K., Shi, L., Lungjangwa, T., Imsoonthornruksa, S., Stelzer, Y., Rangarajan, S., D'Alessio, A., Zhang, J., Gao, Q., Dawlaty, M. M., Young, R. A., Gray, N. S. & Jaenisch, R. Systematic identification of culture conditions for induction and maintenance of naive human pluripotency. *Cell Stem Cell* **15,** 471–487 (2014).

84. Takashima, Y., Guo, G., Loos, R., Nichols, J., Ficz, G., Krueger, F., Oxley, D., Santos, F., Clarke, J., Mansfield, W., Reik, W., Bertone, P. & Smith, A. *Erratum: Resetting transcription factor control circuitry toward ground-state pluripotency in human* 2015.

85. Guo, G., Von Meyenn, F., Santos, F., Chen, Y., Reik, W., Bertone, P., Smith, A. & Nichols, J. Naive Pluripotent Stem Cells Derived Directly from Isolated Cells of the Human Inner Cell Mass. *Stem Cell Reports.* arXiv: NIHMS150003 (2016).

86. Theunissen, T. W., Friedli, M., He, Y., Planet, E., O'Neil, R. C., Markoulaki, S., Pontis, J., Wang, H., Iouranova, A., Imbeault, M., Duc, J., Cohen, M. A., Wert, K. J., Castanon, R., Zhang, Z., Huang, Y., Nery, J. R., Drotar, J., Lungjangwa, T., Trono, D., Ecker, J. R. & Jaenisch, R. Molecular criteria for defining the naive human pluripotent state. *Cell Stem Cell* **19,** 502–515 (2016).

87. Wang, J., Xie, G., Singh, M., Ghanbarian, A. T., Rasko, T., Szvetnik, A., Cai, H., Besser, D., Prigione, A., Fuchs, N. V., Schumann, G. G., Chen, W., Lorincz, M. C., Ivics, Z., Hurst, L. D. & Izsvak, Z. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* **516,** 405–9 (2014).

88. Boroviak, T. & Nichols, J. Primate embryogenesis predicts the hallmarks of human naïve pluripotency, 175–186 (2017).

89. Lu, X., Sachs, F., Ramsay, L., Jacques, P.-É., Göke, J., Bourque, G. & Ng, H.-H. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nature Structural & Molecular Biology* **21,** 423–425 (Apr. 2014).

90. Subramanian, R. P., Wildschutte, J. H., Russo, C. & Coffin, J. M. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology* (2011).

91. Mager, D. & Freeman, D. HERV-H endogenous retroviruses: presence in the new world branch but amplification in the old world primate lineage. *Virology* **213,** 395–404 (1995).

92. Anderssen, S., Sjøttem, E., Svineng, G. & Johansen, T. Comparative analyses of LTRs of the ERV-H family of primate-specific retrovirus-like elements isolated from marmoset, African green monkey, and man. *Virology* **234,** 14–30 (1997).

93. Izsvák, Z., Wang, J., Singh, M., Mager, D. L. & Hurst, L. D. Pluripotency and the endogenous retrovirus HERVH: conflict or serendipity? *BioEssays* **38,** 109–117 (2016).

94. Jern, P., Sperber, G. O., Ahlsén, G. & Sperber, O. Sequence variability , gene structure , and expression of full-length Human Endogenous Retrovirus H. *Journal of Virology* **79,** 6325–6337 (2005).

95. Göke, J., Lu, X., Chan, Y. S., Ng, H. H., Ly, L. H., Sachs, F. & Szczerbinska, I. Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell* **16,** 135–141 (2015).

96. Jacques, P.-É., Jeyakani, J. & Bourque, G. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS genetics* **9,** e1003504 (May 2013).

97. Verdin, E., Becker, N., Bex, F., Droogmans, L. & Burny, A. Identification and characterization of an enhancer in the coding region of the genome of human immunodeficiency virus type 1. *Proceedings of the National Academy of Sciences of the United States of America* **87,** 4874–4878 (1990).

98. Loewer, S., Cabili, M. N., Guttman, M., Loh, Y. H., Thomas, K., Park, I. H., Garber, M., Curran, M., Onder, T., Agarwal, S., Manos, P. D., Datta, S., Lander, E. S., Schlaeger, T. M., Daley, G. Q. & Rinn, J. L. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nature Genetics* (2010).

99. Durruthy-Durruthy, J., Sebastiano, V., Wossidlo, M., Cepeda, D., Cui, J., Grow, E. J., Davila, J., Mall, M., Wong, W. H., Wysocka, J., Au, K. F. & Reijo Pera, R. A. The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. *Nature Genetics* (2015).

100. Ohnuki, M., Tanabe, K., Sutou, K., Teramoto, I., Sawamura, Y., Narita, M., Nakamura, M., Tokunaga, Y., Nakamura, M., Watanabe, A., Yamanaka, S. & Takahashi, K. Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proceedings of the National Academy of Sciences of the United States of America* (Aug. 2014).

101. Loewer, S., Cabili, M. N., Guttman, M., Loh, Y. H., Thomas, K., Park, I. H., Garber, M., Curran, M., Onder, T., Agarwal, S., Manos, P. D., Datta, S., Lander, E. S., Schlaeger, T. M., Daley, G. Q. & Rinn, J. L. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nature Genetics* **42,** 1113–1117 (2010).

102. Friedli, M., Turelli, P., Kapopoulou, A., Rauwel, B., Castro-Diáz, N., Rowe, H. M., Ecco, G., Unzu, C., Planet, E., Lombardo, A., Mangeat, B., Wildhaber, B. E., Naldini, L. & Trono, D. Loss of transcriptional control over endogenous retroelements during reprogramming to pluripotency. *Genome Research* **24,** 1251–1259 (2014).

103. Koyanagi-Aoi, M., Ohnuki, M., Takahashi, K., Okita, K., Noma, H., Sawamura, Y., Teramoto, I., Narita, M., Sato, Y., Ichisaka, T., Amano, N., Watanabe, A., Morizane, A., Yamada, Y., Sato, T., Takahashi, J. & Yamanaka, S. Differentiation-defective phenotypes revealed by large-scale analyses of human pluripotent stem cells. *Proceedings of the National Academy of Sciences.* arXiv: 0706.1062v1 (2013).

104. Marchese, F. P., Raimondi, I. & Huarte, M. *The multidimensional mechanisms of long noncoding RNA function* 2017.

105. Bakshi, A., Herke, S. W., Batzer, M. A. & Kim, J. DNA methylation variation of human-specific Alu repeats. *Epigenetics* (2016).

106. Phalke, S., Nickel, O., Walluscheck, D., Hortig, F., Onorati, M. C. & Reuter, G. Retrotransposon silencing and telomere integrity in somatic cells of Drosophila depends on the cytosine-5 methyltransferase DNMT2. *Nature Genetics* (2009).

107. Molaro, A., Falciatori, I., Hodges, E., Aravin, A. A., Marran, K., Rafii, S., Richard McCombie, W., Smith, A. D. & Hannon, G. J. Two waves of de novo methylation during mouse germ cell development. *Genes and Development* (2014).

108. Stoye, J. P. *Endogenous retroviruses: Still active after all these years?* 2001.

109. Lowe, C. B., Bejerano, G. & Haussler, D. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 8005–10 (May 2007).

110. Brady, T., Lee, Y. N., Ronen, K., Malani, N., Berry, C. C., Bieniasz, P. D. & Bushman, F. D. Integration target site selection by a resurrected human endogenous retrovirus. *Genes & development* (2009).

111. Gemmell, P., Hein, J. & Katzourakis, A. Phylogenetic Analysis Reveals That ERVs ”Die Young” but HERV-H Is Unusually Conserved. *PLoS Computational Biology* **12** (2016).

112. Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., Huang, J., Li, M., Wu, X., Wen, L., Lao, K., Li, R., Qiao, J. & Tang, F. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology* **20,** 1131–9 (Sept. 2013).

113. Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C. Y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y. E., Liu, J. Y., Horvath, S. & Fan, G. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500,** 593–597 (2013).

114. Gebefugi, E., Brunmeir, R., Weierich, C., Wolff, H., Brack-werner, R. & Leib-mösch, C. Activation of a HERV-H LTR induces expression of an aberrant calbindin protein in human prostate carcinoma cells. *Retrovirology* **1,** 3–15 (2009).

115. Kowalski, P. E., Freeman, J. D. & Mager, D. L. Intergenic splicing between a HERV-H endogenous retrovirus and two adjacent human genes. *Genomics* **57,** 371–9 (May 1999).

116. Ware, C. B. Concise Review: Lessons from Naive Human Pluripotent Cells. *Stem Cells* **35,** 35–41 (2017).

117. Funa, N. S., Schachter, K. A., Lerdrup, M., Ekberg, J., Hess, K., Dietrich, N., Honoré, C., Hansen, K. & Semb, H. β-Catenin Regulates Primitive Streak Induction through Collaborative Interactions with SMAD2/SMAD3 and OCT4. *Cell Stem Cell* **16,** 639–652 (2015).

118. Vallier, L., Touboul, T., Chng, Z., Brimpari, M., Hannan, N., Millan, E., Smithers, L. E., Trotter, M., Rugg-Gunn, P., Weber, A. & Pedersen, R. A. Early cell fate decisions of human embryonic stem cells and mouse epiblast stem cells are controlled by the same signalling pathways. *PLoS ONE* **4.** arXiv: NIHMS150003 (2009).

119. Faial, T., Bernardo, A. S., Mendjan, S., Diamanti, E., Ortmann, D., Gentsch, G. E., Mascetti, V. L., Trotter, M. W. B., Smith, J. C. & Pedersen, R. A. Brachyury and SMAD signalling collaboratively orchestrate distinct mesoderm and endoderm gene regulatory networks in differentiating human embryonic stem cells. *Development* **142,** 2121–2135 (2015).

120. Larson, M. H., Gilbert, L. a., Wang, X., Lim, W. a., Weissman, J. S. & Qi, L. S. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nature Protocols* **8,** 2180–2196 (2013).

121. Simeonov, D. R., Gowen, B. G., Boontanrart, M., Roth, T. L., Gagnon, J. D., Mumbach, M. R., Satpathy, A. T., Lee, Y., Bray, N. L., Chan, A. Y., Lituiev, D. S., Nguyen, M. L., Gate, R. E., Subramaniam, M., Li, Z., Woo, J. M., Mitros, T., Ray, G. J., Curie, G. L., Naddaf, N., Chu, J. S., Ma, H., Boyer, E., Van Gool, F., Huang, H., Liu, R., Tobin, V. R., Schumann, K., Daly, M. J., Farh, K. K., Ansel, K. M., Ye, C. J., Greenleaf, W. J., Anderson, M. S., Bluestone, J. A.,

Chang, H. Y., Corn, J. E. & Marson, A. Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature* **549,** 111–115 (2017).

122. Thierry-Mieg, D. & Thierry-Mieg, J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome biology* (2006).

123. Shlyueva, D., Stampfel, G. & Stark, A. *Transcriptional enhancers: From properties to genome-wide predictions* 2014.

124. Schröder, A. R. W., Shinn, P., Chen, H., Berry, C., Ecker, J. R., Bushman, F., Boeke, J., Devine, S., Brown, P., Bowerman, B., Varmus, H., Bishop, J., Bukrinsky, M., Sharova, N., McDonald, T., Pushkarskaya, T., Tarpley, G., Stevenson, M., Burke, W., Eickbush, D., Xiong, Y., Jakubczak, J., Eickbush, T., Bushman, F., Craigie, R., Butler, S., Hansen, M., Bushman, F., Carteau, S., Hoffmann, C., Bushman, F., Coffin, J., Hughes, S., Varmus, H., Corbeil, J., Sheeter, D., Genini, D., Rought, S., Leoni, L., Du, P., Ferguson, M., Masys, D., Welsh, J., Fink, J., Al., E., Davis, C., Dikic, I., Unutmaz, D., Hill, C., Arthos, J., Siani, M., Thompson, D., Schlessinger, J., Littman, D., Ellison, V., Abrams, H., Roe, T., Lifson, J., Brown, P., Farnet, C., Haseltine, W., Farnet, C., Bushman, F., Follenzi, A., Ailes, L., Bakovic, S., Gueuna, M., Naldini, L., Gaiano, N., Amsterdam, A., Kawakami, K., Allende, M., Becker, T., Hopkins, N., Gallay, P., Swingler, S., Song, J., Bushman, F., Trono, D., Geiss, G., Bumgarner, R., An, M., Agy, M., van't Wout, A., Hammersmark, E., Carter, V., Upchurch, D., Mullins, J., Katze, M., Hansen, M., Smith, G., Kafri, T., Molteni, V., Siegel, J., Bushman, F., Hartung, S., Jaenisch, R., Breindl, M., Ji, H., Moore, D., Blomberg, M., Braiterman, L., Voytas, D., Natsoulis, G., Boeke, J., Jordan, A., Defechereux, P., Verdin, E., Kafri, T., van Praag, H., Ouyang, L., Gage, F., Verma, I., Katz, R., Gravuer, K., Skalka, A., Katz, R., DiCandeloro, P., Kukolj, G., Skalka, A., Kirchner, J., Connolly, C., Sandmeyer, S., Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Al., E., Leclercq, I., Mortreux, F., Cavrois, M., Leroy, A., Gessain, A., Wain-Hobson, S., Wattel, E., Li, L., Olvera, J., Yoder, K., Mitchell, R., Butler, S., Lieber, M., Martin, S., Bushman, F., Miller, M., Farnet, C., Bushman, F., Mooslehner, K., Karls, U., Harbers, K., Panet, A., Cedar, H., Popik, W., Pitha, P., Pruss, D., Bushman, F., Wolffe, A., Pruss, D., Reeves, R., Bushman, F., Wolffe, A., Pryciak, P., Varmus, H., Pryciak, P., Muller, H.-P., Varmus, H., Rohdewohld, H., Weiher, H., Reik, W., Jaenisch, R., Breindl, M., Scherdin, U., Rhodes, K., Breindl, M., Scottoline, B., Chow, S., Ellison, V., Brown, P., Shih, C.-C., Stoye, J., Coffin, J., Simmons, A., Aluvihare, V., McMichael, A., Smit, A., Stevens, S., Griffith, J., Stevens, S., Griffith, J., Swingler, S., Gallay, P., Camaur, D., Song, J., Abo, A., Trono, D., Temin, H., Keshet, E., Weller, S., Venter, J., Adams, M., Myers, E., Li, P., Mural, R., Sutton, G., Smith, H., Yandell, M., Evans, C., Holt, R., Al., E., Vijaya, S., Steffan, D., Robinson, H., Weidhaas, J., Angelichio, E., Fenner, S., Coffin, J., Withers-Ward, E., Kitamura, Y., Barnes, J., Coffin, J.,

Zijlstra, M., Li, E., Sajjadi, F., Subramani, S. & Jaenisch, R. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* (2002).

125. Mendjan, S., Mascetti, V. L., Ortmann, D., Ortiz, M., Karjosukarso, D. W., Ng, Y., Moreau, T. & Pedersen, R. A. NANOG and CDX2 pattern distinct subtypes of human mesoderm during exit from pluripotency. *Cell Stem Cell* (2014).

126. Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research* **41** (2013).

127. Guo, G., Huss, M., Tong, G. Q., Wang, C., Li Sun, L., Clarke, N. D. & Robson, P. Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst. *Developmental Cell* **18,** 675–685 (2015).

128. Pelossof, R., Fairchild, L., Huang, C. H., Widmer, C., Sreedharan, V. T., Sinha, N., Lai, D. Y., Guan, Y., Premsrirut, P. K., Tschaharganeh, D. F., Hoffmann, T., Thapar, V., Xiang, Q., Garippa, R. J., Rätsch, G., Zuber, J., Lowe, S. W., Leslie, C. S. & Fellmann, C. Prediction of potent shRNAs with a sequential classification algorithm. *Nature Biotechnology* **35,** 350–353 (2017).

129. Gerdes, P., Richardson, S. R., Mager, D. L. & Faulkner, G. J. *Transposable elements in the mammalian embryo: Pioneers surviving through stealth and service* 2016.

130. Lynch, V. J., Leclerc, R. D., May, G. & Wagner, G. P. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nature Genetics* (2011).

131. Göke, J. & Ng, H. H. CTRL+INSERT: retrotransposons and their contribution to regulation and innovation of the transcriptome. *EMBO reports.* arXiv: arXiv:1011.1669v3 (2016).

132. Krull, M., Petrusma, M., Makalowski, W., Brosius, J. & Schmitz, J. Functional persistence of exonized mammalian-wide interspersed repeat elements (MIRs). *Genome Research* **17,** 1139–1145 (2007).

133. Nekrutenko, A. & Li, W. H. *Transposable elements are found in a large number of human protein-coding genes* 2001.

134. Sorek, R., Ast, G. & Graur, D. Alu-containing exons are alternatively spliced. *Genome Research* (2002).

135. Barna, M., Pandolfi, P. P. & Niswander, L. Gli3 and Plzf cooperate in proximal limb patterning at early stages of limb development. *Nature* **436,** 277–281 (2005).

136. Labbaye, C., Quaranta, M. T., Pagliuca, A., Militi, S., Lich, J. D., Testa, U. & Peschle, C. Plzf induces megakaryocytic development, activates tpo receptor expression and interacts with gata1 protein. *Oncogene* **21,** 6669–6679 (2002).

137. Desmet, F. O., Hamroun, D., Lalande, M., Collod-Bëroud, G., Claustres, M. & Béroud, C. Human Splicing Finder: An online bioinformatics tool to predict splicing signals. *Nucleic Acids Research* (2009).

138. Sobczak, K., de Mezer, M., Michlewski, G., Krol, J. & Krzyzosiak, W. J. RNA structure of trinucleotide repeats associated with human neurological diseases. *Nucleic Acids Research* (2003).

139. Yeyati, P. L., Shaknovich, R., Boterashvili, S., Li, J., Ball, H. J., Waxman, S., Nason-Burchenal, K., Dmitrovsky, E., Zelent, A. & Licht, J. D. Leukemia translocation protein PLZF inhibits cell growth and expression of cyclin A. *Oncogene* (1999).

140. Costoya, J. A., Hobbs, R. M. & Pandolfi, P. P. Cyclin-dependent kinase antagonizes promyelocytic leukemia zinc-finger through phosphorylation. *Oncogene* (2008).

141. Guidez, F., Howell, L., Isalan, M., Cebrat, M., Alani, R. M., Ivins, S., Hormaeche, I., Mcconnell, M. J., Pierce, S., Cole, P. A., Licht, J. & Zelent, A. Histone Acetyltransferase Activity of p300 Is Required for Transcriptional Repression by the Promyelocytic Leukemia Zinc Finger Protein. *Molecular and cellular biology* (2005).

142. Sitterlin, D., Tiollais, P. & Transy, C. The RAR alpha-PLZF chimera associated with Acute Promyelocytic Leukemia has retained a sequence-specific DNA-binding domain. *Oncogene* **14,** 1067–1074 (1997).

143. Guidez, F., Parks, S., Wong, H., Jovanovic, J. V., Mays, A., Gilkes, A. F., Mills, K. I., Guillemin, M.-C., Hobbs, R. M., Pandolfi, P. P., de The, H., Solomon, E. & Grimwade, D. RAR -PLZF overcomes PLZF-mediated repression of CRABPI, contributing to retinoid resistance in t(11;17) acute promyelocytic leukemia. *Proceedings of the National Academy of Sciences* **104,** 18694–18699 (2007).

144. Herrero, J., Tejera, A., Albert, C., Vidal, C., De Los Santos, M. J. & Meseguer, M. A time to look back: Analysis of morphokinetic characteristics of human embryo development. *Fertility and Sterility* (2013).

145. Ohashi, A., Imai, H. & Minami, N. Cyclin A2 Is Phosphorylated During the G2/M Transition in Mouse Two-Cell Embryos. *Molecular Reproduction and Development* **66,** 343–348 (2003).

146. Fuchimoto, D.-i. Posttranscriptional Regulation of Cyclin A1 and Cyclin A2 During Mouse Oocyte Meiotic Maturation and Preimplantation Development. *Biology of Reproduction* (2001).