RECOMBINATION-MEDIATED REGULATORY EVOLUTION OF HUMAN ENDOGENOUS RETROVIRUS HERVH

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
Thomas Carter
August 2021

RECOMBINATION-MEDIATED REGULATORY EVOLUTION OF THE HUMAN ENDOGENOUS RETROVIRUS HERVH

Thomas Carter, Ph. D.
Cornell University 2021

Human endogenous retrovirus type-H (HERVH) is specifically expressed in the pluripotent stem cells of the pre-implantation embryo. A subset of these elements appear to exert regulatory activities promoting pluripotency and self-renewal. How HERVH attained this specific expression pattern and incorporated itself into the human pluripotency network has not been revealed. In the first part of this dissertation, we aimed to elucidate the sequence features responsible for HERVH pluripotent transcription. We performed a “phyloregulatory” analysis of the long terminal repeat (LTR7) sequences, which harbors HERVH’s promoter, in which we layered genomic regulatory data onto a phylogenetic tree of LTR7 sequences. The results showed that LTR7 consists of at least 8 previously undefined subfamilies with unique evolutionary histories, transcription factor binding site (TFBS) profiles, and embryonic transcriptional niches. Only one of the youngest LTR7 subfamilies, we term 7up, exhibited promoter activity in pluripotent stem cells. We found that a complex series of deletions, duplications, and recombination events, led to the emergence of a cis-regulatory module unique to 7up, which is necessary for 7up’s pluripotent promoter activity. Together, these data highlight the unexpected role of inter-element recombination in driving the mosaic cis-regulatory evolution of an endogenous retrovirus. In the second part of this dissertation, we focused on the regulatory effects of intra-element recombination between the LTRs of HERVH. Intra-element recombination is a common mechanism by which a full-length endogenous retrovirus with two LTRs and coding genes collapses into one solitary (solo) LTR. We hypothesized that the full-length to solo transition may result in changes in an element's regulatory properties. To test this hypothesis, we compared the regulatory activity of solo and full-length LTR7up. We found that only full-length loci exhibit pluripotent promoter activity, while solo loci generally exhibited the hallmarks of active enhancers. Comparative genomics revealed the formation of solo LTR7 via recombination has occurred continuously throughout hominoid evolution, uncovering a potent mechanism of cis-regulatory variation in humans and great apes. This dissertation reveals the impact of both inter- and intra-element recombination in the evolution of regulatory DNA in the primate genome.

BIOGRAPHICAL SKETCH

Thomas Carter obtained his B.A. in Molecular, Cellular, and Developmental Biology from the University of Colorado, Boulder. During his undergraduate career, he worked under Dr. Bradley Olwin studying the effects of extra-cellular environments on satellite cell growth and differentiation. This work culminated in his honors thesis: *Developing a hydrogel culture system to promote muscle stem cell stemness* for which he was awarded *Summa cum laude* distinction upon graduation. Thomas then enrolled in the Molecular Biology program at the University of Utah, where he joined the lab of Cédric Feschotte in the department of Human Genetics. There, Thomas worked on a variety of transposon-centric projects including elucidating the role of transposable elements in lncRNA emergence in the cardiac stress response, the role of transposable elements in primate T-cell promoter and enhancer evolution, and the role of transposable elements in shaping the primate innate immune response. In the Fall of 2017, Thomas transferred to Cornell University with his lab mates and mentor. There he started two new projects entitled: LTR-LTR recombination as a cis-regulatory switch and Mosaic cis-regulatory evolution drives transcriptional partitioning of HERVH endogenous retrovirus in the human embryo. Following graduation, Thomas plans to pursue his interests in evolutionary genomics and stem cell biology.

To my partner, Nora Brown, and our benevolent overlord, Queen Jellybean

ACKNOWLEDGMENTS

No man is an island. Since I was born, I have had the intellectual, financial, and emotional support to imagine, pursue, and realize my scientific goals. A great deal of this support I owe to Mom and Dad. They gave me the early life experiences that allowed me to develop an appreciation for the natural things. They taught me to always be curious, always ask questions, and most importantly, to always *think for yourself*. But thinking is hard. It takes a lot of practice and a lot of teachers.

One of my first, was my 4th grade teacher, Becky Ramirez. Every portion of her class centered around science, particularly biology. That exposure gave me some idea of the breadth of wonder in the natural world. I carried this wonder with me throughout childhood, and it ultimately propelled me to pursuing a degree in Molecular, Cellular, and Developmental Biology at the University of Colorado. There, I joined the lab of Bradley Olwin. He gave me my first experience in doing real research. His help and that of my mentor Jenni Bernett, paved the way for my eventual graduate school applications and acceptance. My undergraduate period is also where I did my first solo project under Adam Cadwallader, a postdoc in Brad's lab. Adam mentored me in my first original research project, which was my first step to doing independent research. Adam's guidance enormously improved my scientific thinking. Adam also encouraged me to apply to his *alma mater*, the University of Utah, for graduate school. The transition to graduate school was difficult but made easier by the presence of my classmates, particularly Stephen Denham, Morgan Wambaugh, Sophia Praggastis, Ben 'Juicy' Jussila, Eric 'Bowling Shoes' Bogenschutz, Chase Bryan, and Jacob Cooper.

This is also where I met my current advisor, Cédric Feschotte. Due to tenuous funding, none of my formal rotations ended up securing me a position in a lab. Despite not rotating in his lab, Cédric took a risk and took me on. I knew I loved evolutionary biology, but my knowledge was shallow. Fortunately, Cédric is a naturalist. It's through this lens that he showed me the wonderful world of transposons. Cédric also had a zest for life that rubbed off on his more serious trainees. His passion and unending wonder and creativity have been a consistent source of inspiration. My entrance into Cédric's lab also marked a move from 'wet' to 'dry' lab work. This is a challenging and slow task, made better by the help of many lab mates including John McCormick, Rachel Cosby, Jainy Thomas, Ed Chuong, Alesia McKeown, John Frank, and most of all, Aurélie Kapusta. Her early trainings allowed me to eventually become independent in this new field and learn new tools for myself. After only three years of my time at Utah, the Feschotte lab moved to Cornell. This was made much easier by the support of my mentor and cross-country lab mates and friends Clément 'Clemmy' Goubert, Mitch Lokey, and Sylvia Chang who continue to be the greatest of companions. My time at Cornell was also bettered by the incredible new people here in Ithaca, especially Angela Rey, Sarah Kunz, Martik Chatterjee, Katie Gordon, Jonathan Wells, and Tram Ngyuen. I'd also like to thank the new members of the Feschotte lab, especially the expertise and banter of Manu and Julius. My work at Cornell was made possible by funding from the Cornell CVG and the NIH and the guidance and expertise of my committee members Charles Danko and Andy Clark. Most of all, I owe my continued mental stability and wellbeing to my most treasured companion and partner of 9 years, Nora Brown. She was my best friend at Colorado, Utah, and Cornell and has been my constant source of solace and support the entire time. I appreciate every day we have, and I love you dearly. And, of course, nothing I have done would have been possible without The Baby, Queen Jellybean the good. Your fluff has brought light to these dark times.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

DNA	Deoxyribonucleic acid
TF	Transcription Factor
LTR	Long Terminal Repeat
HIV	Human Immunodeficiency Virus
ERV	Endogenous Retrovirus
TFBS	Transcription Factor Binding Site
HERVH	Human Endogenous Retrovirus Type-H
iPSC	induced Pluripotent Stem Cell
ESC	Embryonic Stem Cell
TAD	Topologically Associated Domain
Int	Internal
HERVH-int	HERVH internal region
TE	Transposable Element
ESRG	Embryonic Stem Cell-Related Gene
UFbootstraps	Ultrafast bootstraps
KZFP	KRAB-Zinc Finger Protein
ChIP-seq	Chromatin Immunoprecipitation sequencing
ZNF	Zinc-finger protein
SNP	Single Nucleotide Polymorphism
GRO-seq	Global Run-On sequencing
RNA	Ribonucleic acid
RNA-seq	RNA-sequencing
scRNA-seq	Single-cell RNA sequencing

RVT	Reverse Transcriptase
LINE	Long interspersed nuclear element
CRE	Cis-regulatory element
TSS	Transcription Start Site
MYA	Million Years Ago
SGDP	Simons Genome Diversity Project
lncRNA	long non-coding RNA
CpG	Cytosine-phosphate-guanine
dREG	discriminatory Regulatory-Element from GRO-seq

CHAPTER 1

RETROVIRAL CIS-REGULATORY EVOLUTION

1.1 RETROVIRAL LONG TERMINAL REPEATS ARE POTENT CIS-REGULATORY UNITS

Retroviruses are unique in the viral realm. To propagate they must infiltrate a cell, access the host DNA, integrate itself into the host genome, and then use host machinery to transcribe its viral genes, eventually assembling its gene products into a replication-competent virion that exits the host cell to parasitize another host (Coffin et al., 1997a). The act of integration into a host genome provides the virus both advantages and challenges. One principal advantage is that the viral DNA becomes that of the host, increasing the difficulty of completely clearing an infection (Coffin et al., 1997b; Kane and Golovkina, 2010). One chief disadvantage is the reliance on host transcriptional machinery (Whitcomb and Hughes, 1992). Hosts tightly regulate DNA expression. In multicellular organisms, cell- and context-specific transcription factors (TF) (and nucleosomes) are largely responsible for a great deal of cell-type-specific regulatory environments (Dowell, 2010; Liu and Tjian, 2018). For a retrovirus to successfully transcribe itself in these environments, it must contain the right TF composition for host machinery to load POLII onto the viral promoter (Coffin, 1988). Retroviral promoters are within the 5' long terminal repeat (LTR). An identical 3' LTR flanks the internal coding genes, and provides other regulatory features, such as the poly-adenylation (poly-A) signal for immature transcripts (Eickbush and Malik, 2002).

Previous studies of retroviruses have associated point mutations and structural variation (indels) with changes in LTR promoter activity (Carvalho et al., 2021; Montano et al., 2020; Payne et al., 1990; Qu et al., 2016). Other studies examined how such cis-regulatory changes contributed to cell-type-specific adaptations (Ait-Khaled et al., 1995; Dampier et al., 2016; Leroux et al., 1997; Opijnen et al., 2020; Rohr et al., 2003) and disease progression (Duverger et al., 2013; Miller-Jensen et al., 2013; Nonnemacher et al., 2004). In HIV, for example, point mutations in AP1, SP1, and TATA core motifs have been found to alter LTR promoter activity that correlates with the switch between active and latent proviruses (Duverger et al., 2013; Miller-Jensen et al., 2013; Nonnemacher et al., 2004).

1.2 CIS-REGULATORY EVOLUTION OF ENDOGENOUS RETROVIRUSES

Like their exogenous counterparts from which they are derived (Eickbush and Malik, 2002; Johnson, 2019), endogenous retroviruses (ERVs) must integrate into host DNA and use host machinery to transcribe their genes and genomes (Coffin, 1988). ERVs have lost their ability to transmit horizontally within a population, and instead transpose within a genome. To amplify its genomic content between generations, an ERV must transpose in the germline or proto germline. Like their exogenous counterparts, an ERV’s LTRs must contain the right cocktail of TF binding sites (TFBS) for its genes to be transcribed by the host. Thus, the LTRs of prolific ERV families should have a TFBS repertoire that has been selected for strong transcription in germ or proto germ cells, such as the cells in early development. Many ERV subfamilies exhibit cell- and stage-specific expression within the human embryo (Chang et al., 2021; Göke et al., 2015; Hermant and Torres-Padilla, 2021; Peaston et al., 2004; Svoboda et al., 2004). In some cases, changes in TF binding profiles correlate with changes in expression profiles (Chuong et al., 2016; Göke et al., 2015).

1.3 HUMAN ENDOGENOUS RETROVIRUS TYPE-H HAS COLONIZED THE EARLY HUMAN EMBRYO

The human endogenous retrovirus type-H (HERVH) family is one of the most abundant ERV in humans, spanning more than 40 million years of evolution (Goodchild et al., 1993; Izsvák et al., 2016; Mager and Freeman, 1995). Thus far it is represented by four LTR subfamilies, all of which share the same HERVH-int: LTR7, LTR7b, LTR7c, and LTR7y (Bao et al., 2015; Kojima, 2018; Storer et al., 2021). These subfamilies have notable embryonic patterns of expression. LTR7 has been extensively studied for its expression in embryonic stem cells (ESCs) and induced pluripotent stem cells (iPSCs). LTR7 expression is presumably regulated by the binding of pluripotent TFs, including SP1, OCT4, NANOG, and SOX2 (Göke et al., 2015; Ito et al., 2017; Kelley and Rinn, 2012; Kunarso et al., 2010; Ohnuki et al., 2014; Pontis et al., 2019; Santoni et al., 2012). However, the mechanisms governing LTR7 TF binding, transcript initiation, and transcript elongation have yet to be elucidated. The combinatorial presence of these TF in ESC coincides with their binding at LTR7. Upon concomitant binding, these TF presumably load POLII onto the 5’ LTR7, where it transcribes the HERVH-int (Nelson et al., 1996). Transcription of some LTR7 loci produce transcripts that terminate in the 3’ LTR (Nelson et al., 1996; Wilkinson et al., 1990). Other loci produce so called ‘chimeric’ transcripts, where the RNA product consists of host and virus RNA (Römer et al., 2017). Many of these chimeric transcripts have been experimentally knocked out or knocked down in pluripotent stem cells. Perturbance of expression from singular or multiple HERVH or chimeric loci seem to disrupt pluripotent homeostasis leading to a loss of stemness and differentiation (Loewer et al., 2010; Lu et al., 2014; Ohnuki et al., 2014; Wang et al., 2014). Additionally, transcribed HERVH can create topologically associated domains (TADs) which can influence gene expression in differentiated daughter cells (Zhang et al., 2019). Strangely, perturbation of transcribed loci has led to inconsistent phenotypes with different degrees of cell death and what experimental cells differentiate into (Lu et al., 2014; Wang et al., 2014; Zhang et al., 2019).

All proposed functions of LTR7 in ESCs have focused on *transcribed* LTR7. This is despite transcribed LTR7 constituting a small minority of the entire subfamily. Understanding the mechanisms underlying its pluripotent transcription may yield insights into the functions of HERVH. Here, we explore two non-mutually exclusive possibilities. 1) Differential activity may be due to changes in LTR TF binding repertoires between insertions and might even be the result of several distinct subfamilies being lumped into LTR7. In Chapter 2, I address this possibility by layering regulatory genomic data onto an LTR7 phylogeny. 2) LTR7 promoter activity may require portions of the HERVH-int. To begin to probe the role of the internal region in transcription, I compare the regulatory signatures of full-length (LTR7-HERVH-LTR7) and solitary (solo) LTR7 who lack their internal region and one LTR.

REFERENCES

Ait-Khaled M, McLaughlin JE, Johnson MA, Emery VC. 1995. Distinct HIV-1 long terminal repeat quasispecies present in nervous tissues compared to that in lung, blood and lymphoid tissues of an AIDS patient. *AIDS* **9**:675–684.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**:11. doi:10.1186/s13100-015-0041-9

Carvalho M, Kirkland M, Derse D. 2021. Protein interactions with DNA elements in variant equine infectious anemia virus enhancers and their impact on transcriptional activity. *Journal of Virology* **67**.

Chang N-C, Rovira Q, Wells JN, Feschotte C, Vaquerizas JM. 2021. A genomic portrait of zebrafish transposable elements and their spatiotemporal embryonic expression. *bioRxiv* 2021.04.08.439009. doi:10.1101/2021.04.08.439009

Chuong EB, Elde NC, Feschotte C. 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**:1083–1087. doi:10.1126/science.aad5497

Coffin JM. 1988. Replication of Retrovirus GenomesRNA Genetics. CRC Press.

Coffin JM, Hughes SH, Varmus HE. 1997a. Immune Response to Retroviral Infection, Retroviruses. Cold Spring Harbor Laboratory Press.

Coffin JM, Hughes SH, Varmus HE, editors. 1997b. Retroviruses. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press.

Dampier W, Nonnemacher MR, Mell J, Earl J, Ehrlich GD, Pirrone V, Aiamkitsumrit B, Zhong W, Kercher K, Passic S, Williams JW, Jacobson JM, Wigdahl B. 2016. HIV-1 Genetic Variation Resulting in the Development of New Quasispecies Continues to Be Encountered in the Peripheral Blood of Well-Suppressed Patients. *PLOS ONE* **11**:e0155382. doi:10.1371/journal.pone.0155382

Dowell RD. 2010. Transcription factor binding variation in the evolution of gene regulation. *Trends in Genetics* **26**:468–475. doi:10.1016/j.tig.2010.08.005

Duverger A, Wolschendorf F, Zhang M, Wagner F, Hatcher B, Jones J, Cron RQ, van der Sluis RM, Jeeninga RE, Berkhout B, Kutsch O. 2013. An AP-1 Binding Site in the Enhancer/Core Element of the HIV-1 Promoter Controls the Ability of HIV-1 To Establish Latent Infection. *Journal of Virology* **87**.

Eickbush TH, Malik HS. 2002. Origins and Evolution of Retrotransposons. *Mobile DNA II* 1111–1144. doi:10.1128/9781555817954.ch49

Göke J, Lu X, Chan Y-S, Ng H-H, Ly L-H, Sachs F, Szczerbinska I. 2015. Dynamic Transcription of Distinct Classes of Endogenous Retroviral Elements Marks Specific Populations of Early Human Embryonic Cells. *Cell Stem Cell* **16**:135–141. doi:10.1016/j.stem.2015.01.005

Goodchild NL, Wilkinson DA, Mager DL. 1993. Recent Evolutionary Expansion of a Subfamily of RTVL-H Human Endogenous Retrovirus-like Elements. *Virology* **196**:778–788. doi:10.1006/viro.1993.1535

Hermant C, Torres-Padilla M-E. 2021. TFs for TEs: the transcription factor repertoire of mammalian transposable elements. *Genes Dev* **35**:22–39. doi:10.1101/gad.344473.120

Ito J, Sugimoto R, Nakaoka H, Yamada S, Kimura T, Hayano T, Inoue I. 2017. Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *PLOS Genetics* **13**:e1006883. doi:10.1371/journal.pgen.1006883

Izsvák Z, Wang J, Singh M, Mager DL, Hurst LD. 2016. Pluripotency and the endogenous retrovirus HERVH: Conflict or serendipity? *BioEssays* **38**:109–117. doi:10.1002/bies.201500096

Johnson WE. 2019. Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat Rev Microbiol* **17**:355–370. doi:10.1038/s41579-019-0189-2

Kane M, Golovkina T. 2010. Common Threads in Persistent Viral Infections. *Journal of Virology* **84**:4116–4123. doi:10.1128/JVI.01905-09

Kelley D, Rinn J. 2012. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biology* **13**:R107. doi:10.1186/gb-2012-13-11-r107

Kojima KK. 2018. Human transposable elements in Repbase: genomic footprints from fish to humans. *Mobile DNA* **9**:2. doi:10.1186/s13100-017-0107-y

Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, Chan Y-S, Ng H-H, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics* **42**:631–634. doi:10.1038/ng.600

Leroux C, Issel CJ, Montelaro RC. 1997. Novel and dynamic evolution of equine infectious anemia virus genomic quasispecies associated with sequential disease cycles in an experimentally infected pony. *Journal of Virology* **71**:9627–9639.

Liu Z, Tjian R. 2018. Visualizing transcription factor dynamics in living cells. *Journal of Cell Biology* **217**:1181–1191. doi:10.1083/jcb.201710038

Loewer S, Cabili MN, Guttman M, Loh Y-H, Thomas K, Park IH, Garber M, Curran M, Onder T, Agarwal S, Manos PD, Datta S, Lander ES, Schlaeger TM, Daley GQ, Rinn JL. 2010. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet* **42**:1113–1117. doi:10.1038/ng.710

Lu X, Sachs F, Ramsay L, Jacques P-É, Göke J, Bourque G, Ng H-H. 2014. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nature Structural & Molecular Biology* **21**:423–425. doi:10.1038/nsmb.2799

Mager DL, Freeman JD. 1995. HERV-H Endogenous Retroviruses: Presence in the New World Branch but Amplification in the Old World Primate Lineage. *Virology* **213**:395–404. doi:10.1006/viro.1995.0012

Miller-Jensen K, Skupsky R, Shah PS, Arkin AP, Schaffer DV. 2013. Genetic Selection for Context-Dependent Stochastic Phenotypes: Sp1 and TATA Mutations Increase Phenotypic Noise in HIV-1 Gene Expression. *PLOS Computational Biology* **9**. doi:10.1371/journal.pcbi.1003135

Montano MA, Nixon CP, Essex M. 2020. Dysregulation through the NF-κB Enhancer and TATA Box of the Human Immunodeficiency Virus Type 1 Subtype E Promoter. *Journal of Virology*.

Nelson DT, Goodchild NL, Mager DL. 1996. Gain of Sp1 Sites and Loss of Repressor Sequences Associated with a Young, Transcriptionally Active Subset of HERV-H Endogenous Long Terminal Repeats. *Virology* **220**:213–218. doi:10.1006/viro.1996.0303

Nonnemacher MR, Irish BP, Liu Y, Mauger D, Wigdahl B. 2004. Specific sequence configurations of HIV-1 LTR G/C box array result in altered recruitment of Sp isoforms and correlate with disease progression. *Journal of Neuroimmunology* **157**:39–47. doi:10.1016/j.jneuroim.2004.08.021

Ohnuki M, Tanabe K, Sutou K, Teramoto I, Sawamura Y, Narita M, Nakamura Michiko, Tokunaga Y, Nakamura Masahiro, Watanabe A, Yamanaka S, Takahashi K. 2014. Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *PNAS* **111**:12426–12431. doi:10.1073/pnas.1413299111

Opijnen T van, Jeeninga RE, Boerlijst MC, Pollakis GP, Zetterberg V, Salminen M, Berkhout B. 2020. Human Immunodeficiency Virus Type 1 Subtypes Have a Distinct Long Terminal Repeat That Determines the Replication Rate in a Host-Cell-Specific Manner. *Journal of Virology* **78**.

Payne SL, La Celle K, Pei XF, Qi XM, Shao H, Steagall WK, Perry S, Fuller F 1999. 1990. Long terminal repeat sequences of equine infectious anaemia virus are a major determinant of cell tropism. *Journal of General Virology* **80**:755–759. doi:10.1099/0022-1317-80-3-755

Peaston AE, Evsikov AV, Graber JH, Vries WN de, Holbrook AE, Solter D, Knowles BB. 2004. Retrotransposons Regulate Host Genes in Mouse Oocytes and Preimplantation Embryos. *Developmental Cell* **7**:597–606. doi:10.1016/j.devcel.2004.09.004

Pontis J, Planet E, Offner S, Turelli P, Duc J, Coudray A, Theunissen TW, Jaenisch R, Trono D. 2019. Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. *Cell Stem Cell* **24**:724-735.e5. doi:10.1016/j.stem.2019.03.012

Qu D, Li C, Sang F, Li Q, Jiang Z-Q, Xu L-R, Guo H-J, Zhang C, Wang J-H. 2016. The variances of Sp1 and NF-κB elements correlate with the greater capacity of Chinese HIV-1 B′-LTR for driving gene expression. *Scientific Reports* **6**:1–11. doi:10.1038/srep34532

Rohr O, Marban C, Aunis D, Schaeffer E. 2003. Regulation of HIV-1 gene transcription: from lymphocytes to microglial cells. *Journal of Leukocyte Biology* **74**:736–749. doi:https://doi.org/10.1189/jlb.0403180

Römer C, Singh M, Hurst LD, Izsvák Z. 2017. How to tame an endogenous retrovirus: HERVH and the evolution of human pluripotency. *Current Opinion in Virology*, Animal models for viral diseases • Paleovirology **25**:49–58. doi:10.1016/j.coviro.2017.07.001

Santoni FA, Guerra J, Luban J. 2012. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology* **9**:111. doi:10.1186/1742-4690-9-111

Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA* **12**:2. doi:10.1186/s13100-020-00230-y

Svoboda P, Stein P, Anger M, Bernstein E, Hannon GJ, Schultz RM. 2004. RNAi and expression of retrotransposons MuERV-L and IAP in preimplantation mouse embryos. *Developmental Biology* **269**:276–285. doi:10.1016/j.ydbio.2004.01.028

Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, Cai H, Besser D, Prigione A, Fuchs NV, Schumann GG, Chen W, Lorincz MC, Ivics Z, Hurst LD, Izsvák Z. 2014. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* **516**:405–409. doi:10.1038/nature13804

Wilkinson DA, Freeman JD, Goodchild NL, Kelleher CA, Mager DL. 1990. Autonomous expression of RTVL-H endogenous retroviruslike elements in human cells. *Journal of Virology* **64**:2157–2167.

Zhang Y, Li T, Preissl S, Amaral ML, Grinstein JD, Farah EN, Destici E, Qiu Y, Hu R, Lee AY, Chee S, Ma K, Ye Z, Zhu Q, Huang H, Fang R, Yu L, Izpisua Belmonte JC, Wu J, Evans SM, Chi NC, Ren B. 2019. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nature Genetics* **51**:1380–1388. doi:10.1038/s41588-019-0479-7

CHAPTER 2

MOSAIC CIS-REGULATORY EVOLUTION DRIVES TRANSCRIPTIONAL PARTITIONING OF HERVH ENDOGENOUS RETROVIRUS IN THE HUMAN EMBRYO[1]

2.1 ABSTRACT

The human endogenous retrovirus type-H (HERVH) family is expressed in the preimplantation embryo. A subset of these elements are specifically transcribed in pluripotent stem cells where they appear to exert regulatory activities promoting self-renewal and pluripotency. How HERVH elements achieve such transcriptional specificity remains poorly understood. To uncover the sequence features underlying HERVH transcriptional activity, we performed a phyloregulatory analysis of the long terminal repeats (LTR7) of the HERVH family, which harbor its promoter, using a wealth of regulatory genomics data. We found that the family includes at least 8 previously unrecognized subfamilies that have been active at different timepoints in primate evolution and display distinct expression patterns during human embryonic

[1] This chapter has been submitted for publication and is available on bioRxiv and as "Thomas A. Carter, Manvendra Singh, Gabrijela Dumbović, Jason D. Chobirko, John L. Rinn, and Cédric Feschotte (2021) Mosaic cis-regulatory evolution drives transcriptional partitioning of HERVH endogenous retrovirus in the human embryo" and is reprinted here with permission. The author contributions are as follows: Carter TA designed and performed all phylogenetic, comparative genomic, phyloregulatory, and recombinatory analyses and generated accompanying figures. Singh M performed all single-cell and SOX2/3 ChIP-seq pileup analyses. Dumbović G performed reporter assay experiments. Chobirko J performed all TFBS identification. Rinn J assisted with experimental planning. Feschotte C aided in project definition and manuscript preparation.

development. Notably, nearly all HERVH elements transcribed in ESCs belong to one of the youngest subfamilies we dubbed LTR7up. LTR7 sequence evolution was driven by complex mutational processes, including multiple recombination events between subfamilies, that led to transcription factor binding motif modules characteristic of each subfamily. Using a reporter assay, we show that one such motif, a predicted SOX2/3 binding site unique to LTR7up, is essential for robust promoter activity in induced pluripotent stem cells. Together these findings illuminate the mechanisms by which HERVH diversified its expression pattern during evolution to colonize distinct cellular niches within the human embryo.

2.2 MAIN

Transposable elements (TEs) are genomic parasites that use the host cell machinery for their own propagation. To propagate in the host genome, they must generate new insertions in germ cells or their embryonic precursors, as to be passed on to the next generation (Charlesworth and Langley, 1986; Cosby et al., 2019; Haig, 2016). To this end, many TEs have evolved stage-specific expression in germ cells or early embryonic development (Faulkner et al., 2009; Fort et al., 2014; Göke et al., 2015; Miao et al., 2020; Urusov et al., 2011). But how does this precise control of TE expression evolve?

Many endogenous retroviruses (ERVs) are known to exhibit highly stage-specific expression during early embryonic development (Chang et al., 2021; Göke et al., 2015; Hermant and Torres-Padilla, 2021; Peaston et al., 2004; Svoboda et al., 2004).

ERVs are derived from exogenous retroviruses with which they share the same prototypical structure with two long terminal repeats (LTRs) flanking an internal region encoding products promoting their replication (Eickbush and Malik, 2002). There are hundreds of ERV families and subfamilies in the human genome, each associated to unique LTR sequences (Kojima, 2018; Vargiu et al., 2016). Each family has infiltrated the germline at different evolutionary timepoints and have achieved various levels of genomic amplification (Bannert and Kurth, 2004; Vargiu et al., 2016). One of the most abundant families is HERVH, a family derived from a gamma retrovirus that first entered the genome of the common ancestor of apes, Old World monkeys, and New World monkeys more than 40 million years ago (mya) (Goodchild et al., 1993; Izsvák et al., 2016; Mager and Freeman, 1995).

There are four subfamilies of HERVH elements currently recognized in the Dfam (Storer et al., 2021) and Repbase (Bao et al., 2015; Kojima, 2018) databases and annotated in the reference human genome based on distinct LTR consensus sequences: LTR7 (formerly known as Type I), 7b (Type II), 7c, and 7y (Type Ia) (Bao et al., 2015; Goodchild et al., 1993; Jern et al., 2005, 2004). Additional subdivisions of HERVH elements were also proposed based on phylogenetic analysis and structural variation of their internal gene sequences (Gemmell et al., 2019; Jern et al., 2005, 2004). However, all HERVH elements are currently annotated in the human genome using a single consensus sequence for the internal region (HERVH_int) and the aforementioned four LTR subfamilies.

HERVH has been the focus of extensive genomic investigation for its high level of RNA expression in human embryonic stem cells (ESCs) and induced pluripotent stem cells (iPSCs) (Fort et al., 2014; Gemmell et al., 2015; Izsvák et al., 2016; Kelley and Rinn, 2012; Loewer et al., 2010; Römer et al., 2017; Santoni et al., 2012; Zhang et al., 2019). Several studies showed that family-wide HERVH knockdown results in the loss of pluripotency of human ESC and reduced reprogramming efficiency of somatic cells to iPSC (Lu et al., 2014; Ohnuki et al., 2014; Wang et al., 2014). Others reported similar phenotypes with the knockdown of individual HERVH-derived RNAs such as those produced from the *lincRNA-RoR* and *ESRG* loci (Loewer et al., 2010; Wang et al., 2014) or the deletion of individual HERVH loci acting as boundaries for topological associated domains (Zhang et al., 2019). These results converge on the notion that HERVH products (RNA or proteins) exert some modulatory effect on the cellular homeostasis of pluripotent stem cells. However, it is important to emphasize that different HERVH knockdown constructs produced variable results and inconsistent phenotypes (Lu et al., 2014; Wang et al., 2014; Zhang et al., 2019), and a recent knockout experiment of the most highly transcribed locus (*ESRG*) failed to recapitulate its previous knockdown phenotype (Takahashi et al., 2021). Despite intense study, which expressed HERVH loci, if any, are necessary for the maintenance of pluripotency remain unclear.

The mechanisms regulating the transcription of HERVH also remain poorly understood. RNA-seq analyses have established that HERVH expression in human ESCs, iPSCs, and the pluripotent epiblast can be attributed to a relatively small subset of loci (estimated between 83 and 209) driven by LTR7 (sensu stricto) sequences (Göke et al., 2015; Wang et al., 2014; Zhang et al., 2019). The related 7y sequences are known to be expressed in the pluripotent epiblast of human embryos (Göke et al., 2015) and a distinct subset of elements associated with 7b and 7y sequences are expressed even earlier in development at the onset of embryonic genome activation (Göke et al., 2015). These observations suggest that the HERVH family is composed of subsets of elements expressed at different timepoints during embryonic development and that these expression patterns reflect, at least in part, the unique cis-regulatory activities of their LTRs. While it has been reported that several transcription factors (TFs) bind and activate HERVH LTRs, including the pluripotency factors OCT4, NANOG, SP1, and SOX2 (Göke et al., 2015; Ito et al., 2017; Kelley and Rinn, 2012; Kunarso et al., 2010; Ohnuki et al., 2014; Pontis et al., 2019; Santoni et al., 2012), it remains unclear how TF binding contributes to the differential expression of HERVH subfamilies and why only a minority of HERVH are robustly transcribed in pluripotent stem cells and embryonic development.

To shed light on these questions, we focused this study on the cis-regulatory evolution of LTR7 elements. We use a "phyloregulatory" approach combining phylogenetic analyses and regulatory genomics to investigate the sequence determinants underlying the partitioning of expression of HERVH/LTR7 subfamilies during early embryonic development.

2.2.1 LTR7 CONSISTS OF 8 PREVIOUSLY UNDEFINED SUBFAMILIES

We began our investigation by examining the sequence relationships of the four LTR7 subfamilies currently recognized in the human genome: LTR7 sensu stricto (748 proviral copies; 711 solo LTRs), 7b (113; 524), 7c (24; 223), and 7y (77; 77). We built a maximum likelihood phylogenetic tree from a multiple sequence alignment of a total of 781 5' LTR and 1073 solo LTR sequences of near complete length (>350 bp) representing all intact LTR subfamilies extracted from the RepeatMasker output of the hg38 human reference assembly. While 7b and 7y sequences cluster, as expected, into clear monophyletic clades with relatively short internode distances and little subclade structure, sequences from the 7c and LTR7 subfamilies were much more heterogeneous and formed many subclades **(Fig. 2.2.1A)**. Notably, sequences annotated as LTR7 were split into distinct monophyletic clades indicative of previously unrecognized subfamilies within that group. The branch length separating some of these LTR7 subclades were longer from one another than they were from those falling within the 7b, 7c, and 7y clades, indicating that they represent subfamilies as different from each other as those previously recognized **(Fig. 2.2.1A)**.

We next sought to classify LTR7 elements more finely by performing a phylogenetic analysis using a multiple sequence alignment of all intact LTR7 sequences (>350 bp) along with the consensus sequences for the other LTR7 subfamilies for reference. We defined high-confidence subfamilies as those forming a clade supported by >95% ultrafast bootstrap (UFbootstrap) and internal branches >0.015 (1.5 nucleotide substitutions per 100 bp) separating subgroup nodes. Based on these criteria, LTR7 elements could be divided into 8 subfamilies **(Fig. 2.2.1B)**.

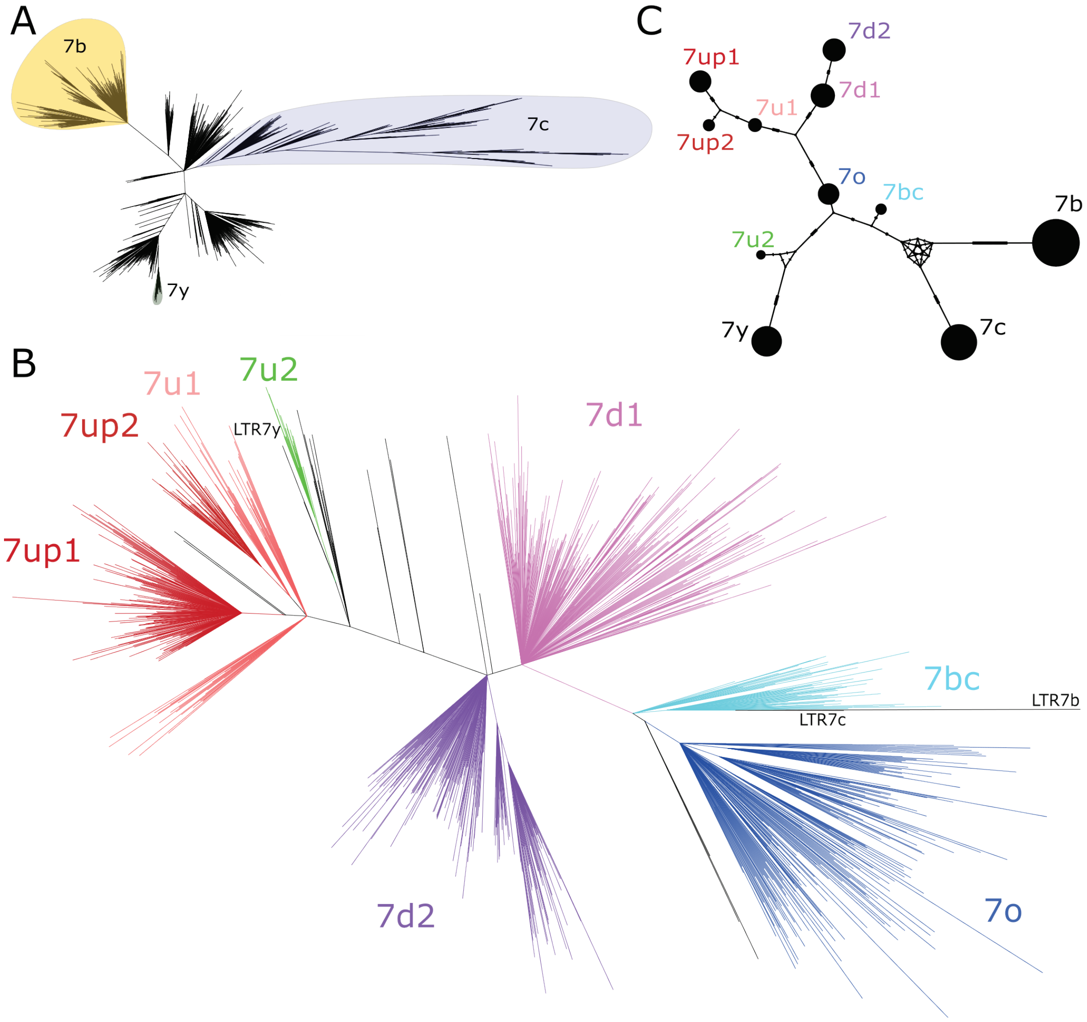


Fig. 2.2.1: Phylogenetic analysis of LTR7 sequences.
A) Unrooted phylogeny of all solo and 5' LTR7 sequences. All nodes with UFbootstraps >0.95, >10 member insertions, and >1.5 substitutions / 100 bp (~6 base pairs) are grouped and colored (see methods). Previously listed consensus sequences from 7b/c/y were included in the alignment and are shown in black. B) Unrooted phylogeny of all solo and 5' LTR7 subfamilies from 1a, 7b, 7c, and 7y. Colors denote clades consisting of previously annotated 7b, 7c, and 7y with >95% concordance. C) Median joining network analysis of all LTR7 and related majority rule consensus sequences. Ticks indicate the number of SNPs at non-gaps between consensus sequences. The size of circles is proportional to the number of members in each subfamily. Only LTR7 insertions that met filtering requirements (see methods) are included while 7b/c/y counts are from dfam.

While long internal branches with high UFbootstrap support separate LTR7 subfamilies, intra-subfamily internal branches with >95% UFbootstrap support were shorter (<0.015), suggesting that each subfamily was the product of a rapid burst of amplification of a common ancestor. To approximate the sequence of these ancestral elements we generated majority-rule consensus sequence for each of the 8 newly defined LTR7 subfamilies (7o, 7bc, 7up, etc.). The consensus sequences were deposited at www.dfam.org.

To investigate the evolutionary relationships among the newly defined and previously known LTR7 subfamilies, we conducted a median-joining network analysis (Leigh and Bryant, 2015) of their consensus sequences **(Fig. 2.2.1C)**. The network analysis provides additional information on the relationships between subfamilies and approximates the shortest and most parsimonious paths between them (Bandelt et al., 1999; Cordaux et al., 2004; Posada and Crandall, 2001). The results place 7o in a central position from which two major lineages are derived. One lineage led to two sub-lineages, formed by 7up1, 7up2, and 7u1 (with 7up1 and 7up2 being most closely related) and by 7d1 and 7d2. The other lineage emanating from 7o rapidly split into two sub-lineages; one gave rise to 7u2 and then to 7y and the other gave rise to 7bc which is connected to the two more diverged subfamilies 7b and 7c **(Fig. 2.2.1C)**. Together these results indicate that the LTRs of HERVH elements can be divided into additional subfamilies than those previously recognized.

2.2.2 THE AGE OF LTR7 SUBFAMILIES SUGGESTS THREE MAJOR WAVES OF HERVH PROPAGATION

The genetic differences between LTR7 subfamilies suggest that they may have been active at different evolutionary timepoints. To examine this, we used reciprocal *liftover* analysis to infer the presence/absence of each human LTR7 locus across five other primate genomes. Insertions shared at orthologous genomic position across a set of species are deemed to be ancestral to these species and thus can be inferred to be at least as old as the divergence time of these species (Johnson, 2019).

The results of this cross-species analysis indicate that LTR7 subfamilies have been transpositionally active at different timepoints in the primate lineage **(Fig. 2.2.3A)**. The subfamilies 7o, 7bc, and 7c are the oldest since the majority of their insertions are found at orthologous position in rhesus macaque, an Old World Monkey (OWM). These three subfamilies share similar evolutionary trajectories, with most of their proliferation occurring prior to the split of OWM and hominoids, ~25 mya **(Fig. 2.2.2a)**. Members of the 7b subfamily (the most numerous, 637 solo and full-length insertions) appear to be overall younger, since only 22% of the human 7b elements could be lifted over to rhesus macaque and the vast majority appeared to have inserted between 10 and 20 mya (**Fig. 2.2.2A,** Figure supplement 1). Only 5 of the 550 elements in the 7d1 and 7d2 subfamilies could be retrieved in rhesus macaque, but ~30% were shared with gibbon and ~75% were shared with orangutan. Thus, these two subfamilies are largely hominoid-specific and achieved most of their proliferation prior to the split of African and Asian great apes ~14 mya **(Fig. 2.2.3A)**. Members of the 7u1 subfamily also emerged in the hominoid ancestor, but the majority (55%) of 7u1 elements present in the human genome inserted after the split of gibbons in the great ape ancestor, between 14 and 20 mya. Thus, the 7b, 7d1/2, and 7u1 subfamilies primarily amplified during the same evolutionary window, 14 to 20 mya.

The 7up1/2, 7y, and 7u2 subfamilies represent the youngest in the human genome, with most of their proliferation occurring between ~10 and ~14 mya, in the ancestor of African great apes **(Fig. 2.2.3A)**. Based on these results, these subfamilies seem to have experienced a burst of transposition after the divergence of African and Asian great apes but before the split of the pan/homo and gorilla lineages. For example, only 14 of the 208 (6.7%) human 7up1 elements can be retrieved in orangutan, but 178

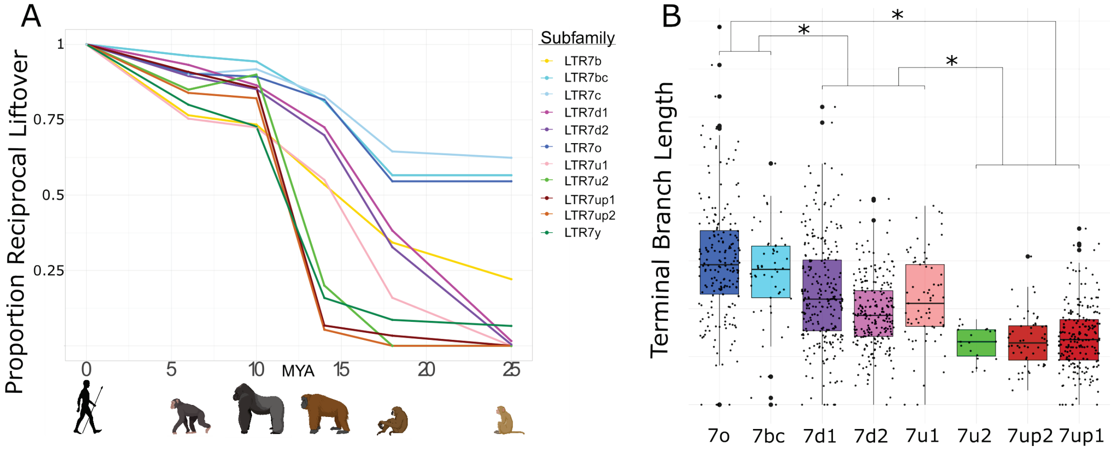


Fig. 2.2.2: Age analysis of LTR7 subfamilies.
A) Proportion of a given subfamily that have 1:1 orthologous insertions between human and other primate species. LTR7 subfamilies are from trees in Figs. 1a and 2a; 7b/c/y subfamilies are from RepeatMasker annotations. Non-human primates are spaced out on the X axis in accordance with their approximate divergence times to the human lineage. B) Terminal branch lengths of all LTR7 insertions from Fig. 2.2.1a. Groups with similar liftover profiles were merged for statistical testing (see methods). Differences with padj<1e-15 are denoted with * (Wilcox rank-sum test with Bonferroni correction).

(85.6%) can be found in gorilla. These data indicate that the three youngest LTR7 subfamilies mostly expanded in the ancestor of African great apes **(Fig. 2.2.2C)**.

As an independent dating method, we used the terminal branch length separating each insertion from its nearest node in Fig. 2.2.1B **(Fig. 2.2.2B)**. Here, the terminal branch lengths are proportional to nucleotide divergence accumulated after insertion and can thus approximate each insertion's relative age. This method largely corroborated the results of the *liftover* analysis and revealed three age groups among LTR7 subfamilies characterized by statistically different mean branch lengths (p(adj)< 1e-15; Wilcox rank-sum test). By contrast, we found no statistical difference between the mean branch length of the subfamilies within these three age groups, suggesting that they were concomitantly active. Taken together, our dating analyses distinguish 3 major waves of HERV propagation: an older wave 25-40 mya involving 7c, 7o, and 7bc elements, an intermediate wave 9-20 mya involving 7b, 7d1/2 and 7u1, and a most recent wave 4-10 mya implicating primarily 7up1/2, 7u2 and 7y elements.

2.2.3 ONLY LTR7UP SHOWS ROBUST TRANSCRIPTION IN HUMAN ESC AND IPSC

Our data thus far indicate that LTR7 is composed of genetically and evolutionarily distinct subfamilies. Because a subset of HERVH elements linked to LTR7 were previously reported to be transcribed in pluripotent stem cells (human ESCs and iPSCs), we wondered whether this activity was restricted to one or several of the LTR7 subfamilies newly defined herein. To investigate this, we performed a "phyloregulatory" analysis, where we layered locus-specific regulatory data obtained from publicly available genome-wide assays in ESCs (mostly from the H1 cell line, see methods) for each LTR insertion on top of a phylogenetic tree depicting their evolutionary relationship. We called an individual LTR7 insertion as positive for a given feature if there is overlap between the coordinates of the LTR and that of a peak called for this mark (see methods). We predicted that if transcriptional activity was an ancestral property of a given subfamily, evidence of transcription and "activation" marks should be clustered within the cognate clade. Alternatively, if transcription and activation marks were to be distributed throughout the tree, it would indicate that LTR7 transcriptional activity in pluripotent cells was primarily driven by post-insertional changes or context-specific effects. Differences in the proportion of positive insertions for a given mark between LTR7 subfamilies were tested using a chi-square test with Bonferroni correction. Unless otherwise noted, all proportions compared thereafter were significantly different (padj< 0.05).

The results **(Fig. 2.2.3A)** show that HERVH elements inferred to be "highly expressed" (fpkm > 2) based on RNA-seq analysis (Wang et al., 2014) were largely confined to two closely related subfamilies, 7up1 and 7up2, together referred to as 7up hereafter. Indeed, we estimated that 33% of 7up elements (88 loci) are highly expressed according to RNA-seq compared with only 2% of highly expressed elements from all other subfamilies combined (17 loci). Nascent RNA mapping using GRO-seq data (Estarás et al., 2015) recapitulated this trend with 22% of 7up loci with visible signal (Figure supplement 2), compared with only 4% of other LTR7 loci (**Fig. 2.2.3D**, Figure supplement 2). Half of the loci displaying GRO-seq signal (53/96) also showed evidence of mature RNA product (supp. file 1). Thus, HERVH transcriptional activity in H1 ESCs is largely limited to loci driven by 7up sequences.

As previously noted from ChIP-seq data (Ohnuki et al., 2014), we found that KLF4 binding is a strong predictor of transcriptional activity: KLF4 ChIP-seq peaks overlap 91% of 7up loci and KLF4 binding is strongly enriched for the 7up subfamilies relative to other subfamilies **(Fig. 2.2.3A,B,D)**. NANOG binding is also enriched for 7up (97.7% of loci overlap ChIP-seq peaks) but is observed to varying degrees at other LTR7 loci that do not show evidence of active transcription based on GRO-seq and/or RNA-seq (85% of 7u1 loci, 32% 7d1, 45% 7d2, 13% 7o, 8.7% 7bc, and 0% of 7u2). Other TFs with known roles in pluripotency are also enriched at 7up loci, such as SOX2 (32% LTR7up, 1-3% all other LTR7), FOXP1(49%, 0-4.3%), and FOXA1(28%, 0-1.4%). In fact, FOXA1 binds only a single non-7up insertion in our dataset, making it the most exclusive feature of 7up loci among the TFs examined in this analysis. In contrast, OCT4 binds merely 12% of 7up loci (see supp. file 8 for full statistical analysis of all marks).

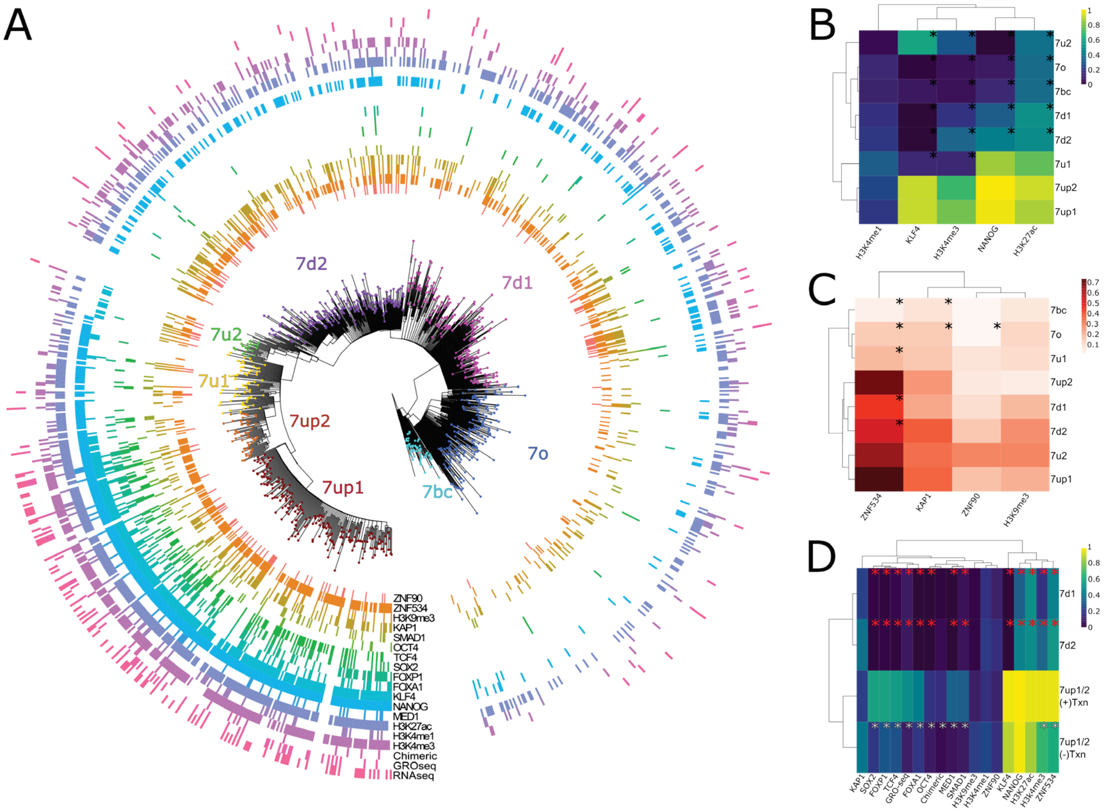


Fig. 2.2.3: Phyloregulatory analysis of LTR7.
A) "Phyloregulatory" map of LTR7. The phylogenetic analysis to derive the circular tree is the same as for the tree in Fig. 2.2.1A but rooted on the 7b consensus. Subfamilies defined in Fig. 2.2.1 are denoted with dotted colored tips. Positive regulatory calls for each insertion are shown as tick marks of different colors and no tick mark indicates a negative call. All marks are derived from ESC except for ZNF90 and ZNF534, which are derived from ChIP-exo data after overexpression of these factors in HEK293 cells (see methods) B) Heatmap of major activation and repression profiles. Proportions indicate the proportion of each group positive for a given characteristic. Trees group LTR7 subfamilies on regulatory signature, not sequence similarity. Asterisks denote statistical differences between given group and 7up1 (padj> 0.05 Wilcox rank-sum with Bonferroni correction). C) Heatmap done in similar fashion to Fig. 2.2.3B but for repression marks. D) Heatmap of transcribed (>2 fpkm) and untranscribed 7up1/2 (<2 fpkm) and all 7d1/2. Red asterisks denote statistical differences between 7d1/2 and 7up1 (padj< 0.05 chi-square Bonferroni correction). White asterisks denote differences between transcribed and untranscribed LTR7up.

Congruent with having generally more TF binding and transcriptional activity, 7up loci also have a propensity to be decorated by H3K4me3, a mark of active promoters (76% LTR7up vs 19% all others) and the broader activity mark H3K27ac (89% vs 48%) **(Fig. 2.2.3A,B)**. By contrast, H3K4me1, a mark typically associated with low POLII loading as seen at enhancers as opposed to promoters, is spread rather evenly throughout the tree of LTR7 sequences (26% vs. 18%) **(Fig. 2.2.2A,B)**. Thus, promoter marks are primarily restricted to 7up loci, but a broader range of 7up loci display putative enhancer marks.

Taken together, our phyloregulatory analysis suggests that strong promoter activity in ESCs is restricted to 7up elements.

2.2.4 DIFFERENTIAL ACTIVATION, RATHER THAN REPRESSION, EXPLAIN THE DIFFERENTIAL TRANSCRIPTIONAL ACTIVITY OF LTR7 SUBFAMILIES IN ESCS

The pattern described above could be explained by two non-mutually exclusive hypotheses: (i) 7up elements (most likely their progenitor) have acquired unique sequences (TF binding sites, TFBS) that promote Pol II recruitment and active transcription, and/or (ii) they somehow escape repressive mechanisms that actively target the other subfamilies, preventing their transcription. For instance, 7up elements may lack sequences targeted by transcriptional repressors such as KRAB-Zinc Finger proteins (KZFP) that silence the other subfamilies in ESCs. KZFP are well-known for binding TEs in a subfamily-specific manner where they nucleate inheritable epigenetic silencing (Ecco et al., 2017; Jacobs et al., 2014; Wolf et al., 2020; Yang et al., 2017) and several KZFPs are known to be capable of binding LTR7 loci (Imbeault et al., 2017). To examine whether KZFPs may differentially bind to LTR7 subfamilies, we analyzed the loading of the corepressor KAP1 and the repressive histone mark H3K9me3 typically deposited through the KZFP/KAP1 complex, across the LTR7 phylogeny using ChIP-seq data previously generated for ESCs (Imbeault et al., 2017; Theunissen et al., 2016). We found that KAP1 and H3K9me3 loading were neither enriched nor depleted for 7up elements relative to other subfamilies **(Fig. 2.2.3A,C)**. Overall, there were no significant differences in the level of H3K9me3 marking across subfamilies and the only difference in KAP1 binding was a slight but significant depletion for 7bc and 7o compared to all other subfamilies including 7up (14% vs. 35% - padj< 0.05 chi-square Bonferroni correction). Furthermore, KAP1 and H3K9me3 loading were found in similar proportions in expressed and unexpressed 7up elements (padj> 0.05) **(Fig. 2.2.2C)**. This was also the case for CpG methylation, whose presence was not differential between subfamilies (padj> 0.05 Wilcox rank-sum with Bonferroni correction) (Figure supplement 2). Thus, KAP1 binding and repressive marks at LTR7 in ESCs poorly correlate with their transcriptional activity and differential repression is unlikely to explain the differential promoter activity of LTR7 subfamilies in ESCs.

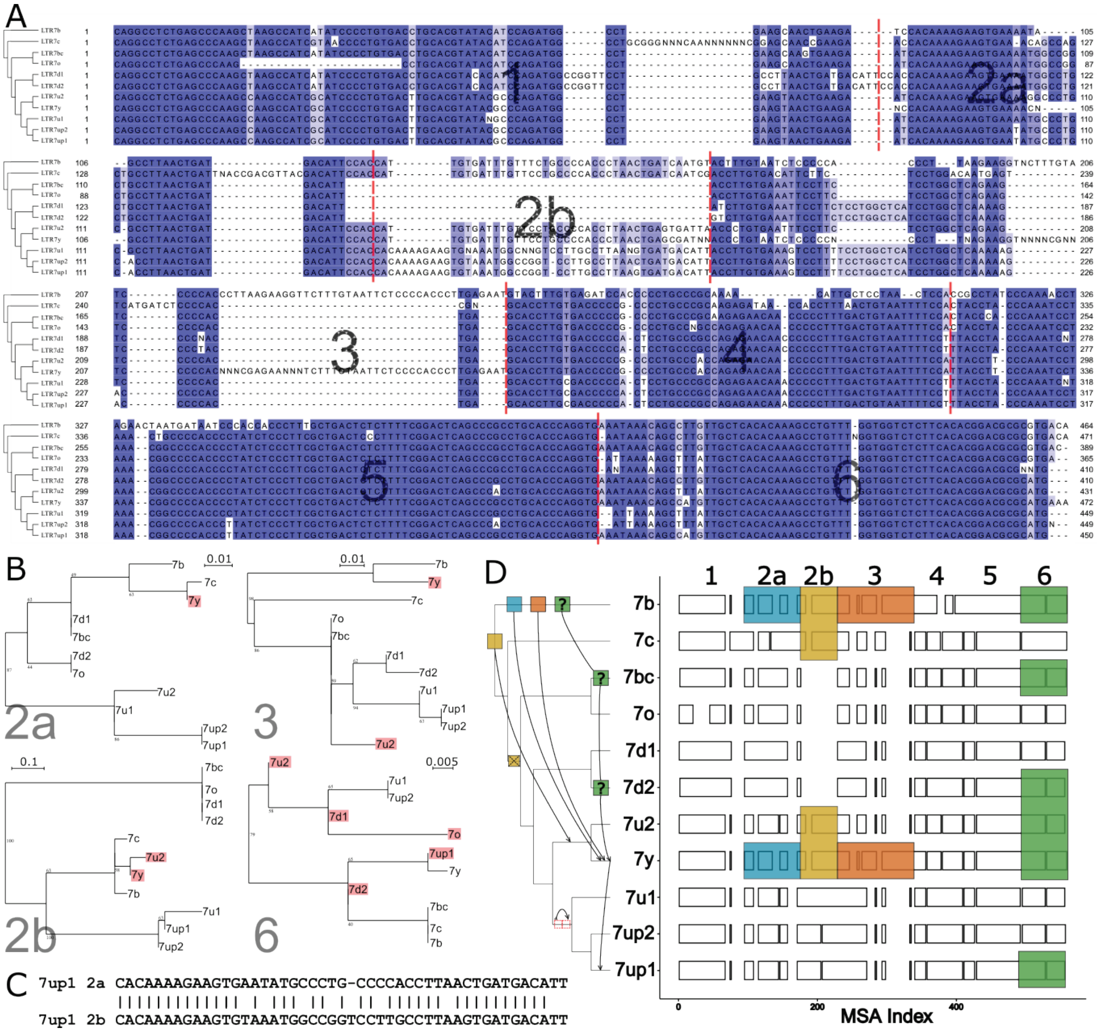
We also examined the binding profile of ZNF534 and ZNF90, two KZFPs previously reported to be enriched for binding LTR7 elements using ChIP-exo data in human embryonic kidney 293 cells (Imbeault et al., 2017), in order to examine whether they bind a particular subset of elements in our LTR7 phylogeny. We found that while ZNF90 bound all LTR7 subfamilies to a similar extent, ZNF534 preferentially bound members of the 7up subfamily (72% of LTR7up vs. 34-53% of non-LTR7up). However, ZNF534 binding in 293 cells did not correlate with transcriptional activity of 7up elements in ESCs nor with KAP1 binding or H3K9me3 deposition in these cells **(Fig. 2.2.3A,D)**. In other words, there was no significant enrichment for ZNF534 binding within untranscribed 7up elements nor depletion within the 7up elements we inferred to be highly transcribed in ESCs. These observations could simply reflect the fact that ZNF534 itself is not highly expressed in ESCs (Figure supplement 3) and do not preclude that ZNF534 represses 7up in other cellular contexts or cell types. Collectively these data suggest that differential LTR binding of KZFP/KAP1 across subfamilies cannot readily explain their differential regulatory activities in ESCs. Thus, differential activation is the most likely driver for the promoter activity of 7up elements in ESCs.

To determine which factors are associated and potentially determinant for 7up promoter activity, we compared the set of "highly expressed" 7up loci to 7up loci which are apparently poorly expressed, using 7d1/d2 as outgroups **(Fig. 2.2.3D)**. While known regulators of LTR7 transcription, KLF4 and NANOG, are enriched for binding to 7up elements, their occupancy alone cannot distinguish transcribed from untranscribed 7up loci **(Fig. 2.2.3D)**. Thus, other factors must contribute to the transcriptional activation of 7up elements. Our analysis of pluripotent transcriptional activators SOX2, FOXA1, FOXP1, OCT4, TCF4, and SMAD1 (Boyer et al., 2005;

Chambers and Smith, 2004; Niwa, 2007) binding profiles show that all of these TFs are enriched in robustly transcribed 7up loci compared to non-transcribed loci **(Fig. 2.2.3D)**. Intriguingly, when overexpressed in HEK293 cells, the potential KZFP repressor ZNF534 preferably binds ESC-transcribed 7up over untranscribed 7up, suggesting that ZNF534 may suppress transcription-competent 7up in cellular contexts where this factor is expressed. Together these data suggest that differential repression cannot explain the differential promoter activity of LTR7 subfamilies in ESCs but rather that highly expressed LTR7up loci are preferentially bound by a cocktail of transcriptional activators that are less prevalent on poorly-expressed loci.

2.2.5 INTER-ELEMENT RECOMBINATION AND INTRA-ELEMENT DUPLICATION DROVE LTR7 SEQUENCE EVOLUTION

The data presented above suggest that the transcriptional activity of 7up in ESCs emerged from the gain of a unique combination of TFBS. To identify sequences unique to 7up relative to its closely related subfamilies, we aligned the consensus sequences of the newly defined LTR7 subfamilies and those of 7b/c/y consensus sequences. This multiple sequence alignment revealed blocks of sequences that tend to be highly conserved across subfamilies, only diverging by a few SNPs, while other regions showed insertion/deletion (indel) segments specific to one or a few subfamilies **(Fig. 2.2.4A)**. These indels resulted in substantial gain and loss of DNA between closely related subfamilies, with the longest consensus (7y) having a length of 472-bp and the shortest (7o) a mere 365-bp. These observations suggest that segmental rearrangements have played an important role in the evolution of LTR7 sequences.



```
7up1  2a  CACAAAAGAAGTGAATATGCCCTG-CCCCACCTTAACTGATGACATT
          ||||||||||||| | ||| || | ||   |||||| ||||||||||
7up1  2b  CACAAAAGAAGTGTAAATGGCCGGTCCTTGCCTTAAGTGATGACATT
```

Fig. 2.2.4: Modular block evolution of LTR7 subfamilies.
A) A multiple sequence alignment of LTR7 subfamily consensus sequences. The phylogenetic topology from Figure 1 is shown on the left. The MSA is broken down into sequence blocks (red lines) with differential patterns of relationships. B) Parsimony trees from Fig. 4a sequence blocks. Subfamilies whose blocks do not match the overall phylogeny are highlighted in red. Bootstrap values >0 are shown. C) Blastn alignment of LTR7up1 block 2a and 2b. D) A multiple sequence alignment of majority-rule consensus sequences from each LTR7 subfamily detailing shared structure. Blocks show aligned sequence; gaps represent absent sequence. Colored sections identify putative phylogeny-breaking events. Recombination events whose directionality can be inferred (via aging) are shown with blocks and arrows on the cladogram. Recombination events with multiple possible routes are denoted with “?”. The deletion of 2b is denoted on the cladogram with a red “X”; the duplication of 2a is denoted with 2 red rectangles.

Upon closer scrutiny, we noticed that the indels characterizing some of the subfamilies were at odds with the evolutionary relationship of the subfamilies defined by overall phylogenetic and network analyses. This was particularly obvious in segments we termed block 2b (where 7y and 7u2 share a large insertion with 7b and 7c) and block 3 (where 7y and 7b share a large insertion). This led us to carefully examine the multiple sequence alignment of the LTR7 consensus sequences to identify indels with different patterns of inter-subfamily relationships. Based on this analysis, we defined seven sequence blocks shared by a different subset of subfamilies, pointing at relationships that were at odds with the overall phylogeny of the LTR7 subfamilies **(Fig. 2.2.4A-B)**. These observations suggested that some of the blocks have been exchanged between LTR7 subfamilies through recombination events.

To systematically test if recombination events between elements drove the evolution of LTR7 subfamilies, we generated parsimony trees for each block of consensus sequences and looked for incongruences with the overall consensus phylogeny. We found a minimum of 6 instances of clades supported in the block parsimony trees that were incongruent with those supported by the overall phylogeny **(Fig. 2.2.4B,D)**. We also found some blocks evolved via tandem duplication. Notably, block 2b was absent from 7d1/2 and 7bc/o but present in all other subfamilies. However, block 2b from 7b, 7c, 7u2, and 7y aligned poorly with block 2b from 7up and 7u1. Instead, block 2b from 7up/u1 2b was closely related (~81% nucleotide similarity) to block 2a from the same subfamilies **(Fig. 2.2.4D)**, suggestive that it arose via tandem duplication in the common ancestor of these subfamilies. To further clarify the evolutionary history of the 2a-2b duplication, we aligned all 2a and 2b blocks from all subfamilies and generated a parsimony tree (Figure supplement 4). This analysis indicated that the 2b block from 7up/u1 most closely resembles the 2a block from 7d. The results above suggest that the evolution of HERVH was characterized by extensive diversification of LTR sequences through a mixture of point mutations, indels, and recombination events.

2.2.6 HERVH SUBFAMILIES SHOW DISTINCT EXPRESSION PROFILES IN THE PREIMPLANTATION EMBRYO

We hypothesized that the mosaic pattern of LTR sequence evolution described above gave rise to TFBS combinations unique to each family that drove shifts in HERVH expression during early embryogenesis. To test this, we aimed to reanalyze the expression profiles of newly defined LTR7 subfamilies during early human embryogenesis and correlate these patterns with the acquisition of embryonic TF binding motifs within each of the subfamilies.

To perform this analysis, we first reannotated the hg38 reference genome assembly using Repeatmasker with a custom library consisting of the consensus sequences for the 8 newly defined LTR7 subfamilies plus newly generated consensus sequences for 7b, 7c, and 7y subfamilies redefined from the phylogenetic analysis presented in Fig. 2.2.1B (Figure supplement 5) (see methods). Our newly generated Repeatmasker annotations (supp. file 2) did not drastically differ from previous annotations of LTR7 and 7c, where 90% and 86% of insertions, respectively, were concordant with the old Repeatmasker annotations (though LTR7 insertions were now assigned to one of the 8

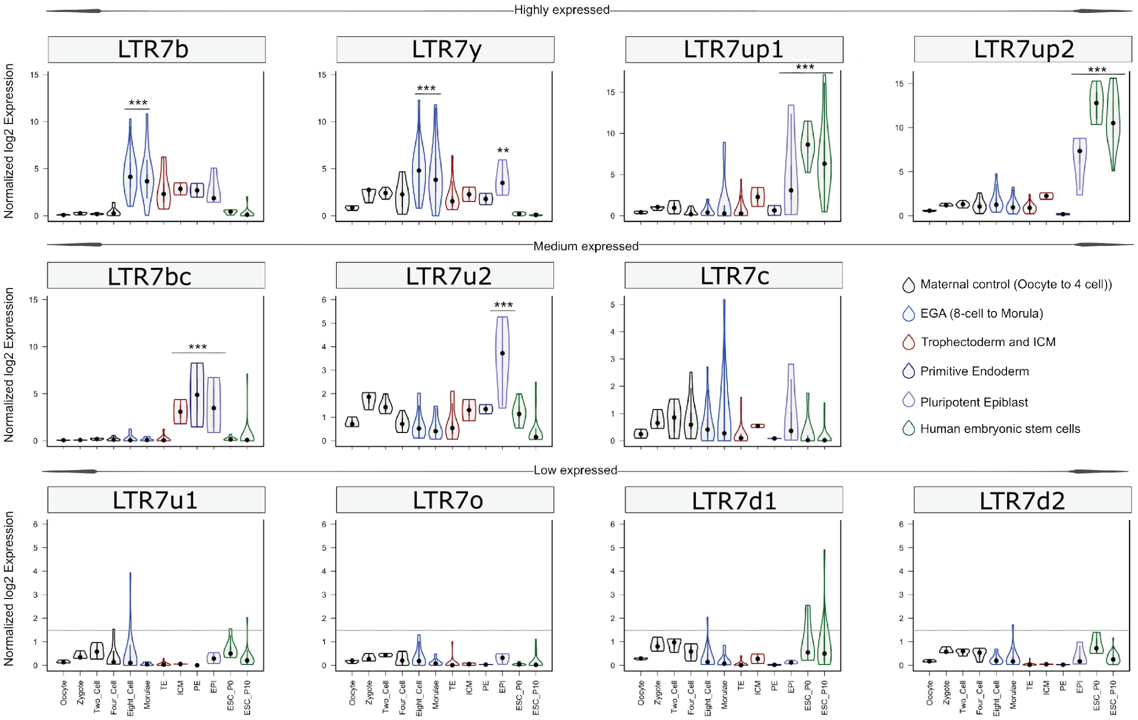


Fig. 2.2.5: Expression profile of LTR7 subfamilies in human preimplantation embryonic lineages and ESCs.
The solid dots and lines encompassing the violins represent the median and quartiles of single cellular RNA expression. The color scheme is based on embryonic stages, defined as maternal control of early embryos (Oocytes, Zygote, 2-cell and 4-cell stage), EGA (8-cell and Morula), inner cell mass (ICM), trophectoderm (TE), epiblast (EPI) and primitive endoderm (PE) from the blastocyst, and ESCs at passages 0 and 10.

newly defined subfamilies). 7y and 7b annotations, however, shifted significantly. Only 33% of previously annotated 7y reannotated concordantly with 53% now being annotated as 7u2 and only 52% of 7b reannotated concordantly, with 22% now annotated as 7y. These shifts can be largely explained by the fact that 7u2 and 7y are closely related **(Fig. 2.2.1A-C)** and 7y and 7b share a great deal of sequence through recombination events **(Fig. 2.2.4B-C)**.

Next we used the newly generated Repeatmasker annotations to examine the RNA expression profiles of the different LTR7 subfamilies using scRNA-seq data from human pre-implantation embryos and RNA-seq data from human ESCs (Blakeley et al., 2015; Tang et al., 2010) (see methods).

As expected, we found that the 7up subfamilies were highly expressed in the pluripotent epiblast and in ESCs **(Fig. 2.2.5)**. 7up expression was highly specific to these pluripotent cell types, with little to no transcription at earlier developmental time points. As previously observed (Göke et al., 2015), the 7b subfamily exhibited expression at the 8-cell and morula stages, coinciding with EGA **(Fig. 2.2.5)**. Another remarkable expression pattern was that of 7u2 which was restricted to the pluripotent epiblast **(Fig. 2.2.5)**. Interestingly, the 7y subfamily combined the expression of 7b and 7u2 (8-cell and morula plus epiblast), perhaps reflecting the acquisition of sequence blocks from both subfamilies **(Fig. 2.2.4B-C)**. Despite very similar sequence and age (**Fig. 2.2.1, Fig. 2.2.2, Fig. 2.2.4A)**, 7bc and 7o elements show stark contrast in their expression profiles. 7o elements show no significant transcription at any time point in early development, while 7bc elements display RNA expression throughout the blastocyst, including trophectoderm and inner cell mass, primitive endoderm, and pluripotent epiblast **(Fig. 2.2.5)**. Previous expression analysis of the oldest LTR7 subfamily, 7c, did not find robust stage-specific expression (Göke et al., 2015). Our analysis revealed that some 7c elements display moderate RNA expression at various developmental stages **(Fig. 2.2.5)**. This pattern may reflect the relatively high level of sequence heterogeneity within this subfamily **(Fig. 2.2.1)**.

In summary, our analysis indicates that LTR7 subfamilies have distinct but partially overlapping expression profiles during human early embryonic development that appear to mirror their complex history of sequence diversification.

2.2.7 A PREDICTED SOX2/3 MOTIF UNIQUE TO 7UP IS REQUIRED FOR TRANSCRIPTIONAL ACTIVITY IN PLURIPOTENT STEM CELLS

We hypothesized that differences in embryonic transcription among LTR7 subfamilies were driven by the gain and loss of TF binding motifs, and that one or more of these mutations led to 7up's pluripotent-specific transcription. To find TF motifs enriched within each LTR7 subfamily relative to the others, we performed an unbiased motif enrichment analysis using the program HOMER to calculate enrichment scores of known TF motifs within each segmental block defined in Fig. 2.2.4A in a pairwise comparison of each subfamily against each of the other subfamilies (see methods). The results yielded a slew of TF motifs enriched for each subfamily relative to the others (see **Fig. 2.2.6A** for 7up1 and enrichment for all HERVH subfamilies in supp. files 3,4). These results suggested that each LTR7 subfamily possesses a unique repertoire of TF binding motifs, which could explain their differential expression during embryonic development.

Next, we sought to pinpoint mutational events responsible for the gain of TF motifs responsible for the unique expression of 7up in ESC. The single most striking motif distinguishing the 7up clade from the others was a SOX2/3 motif which coincided with an 8-bp insertion in block 2b **(Fig. 2.2.6A,B)**. Note this motif (and insertion) was also present in 7u1, the closest relative to 7up **(Fig. 2.2.4C)**, but absent in all other subfamilies **(Fig. 2.2.6B)**.

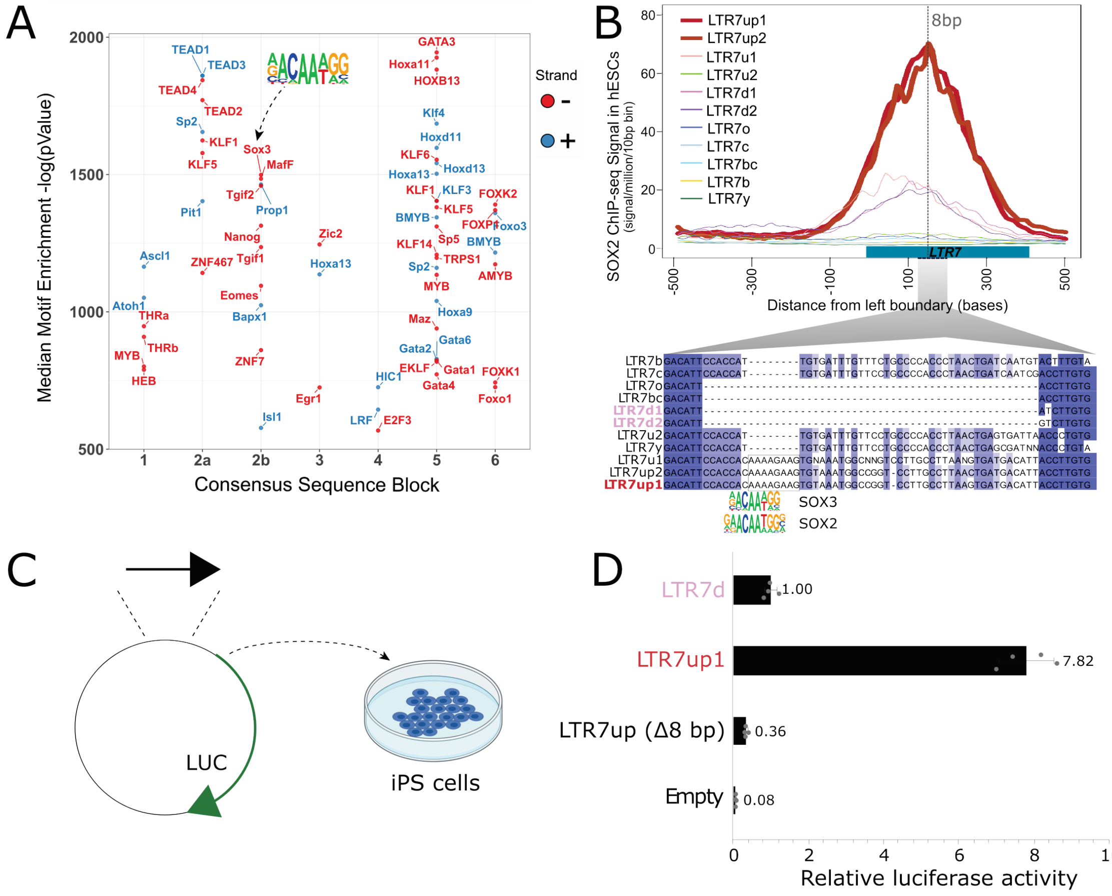


Fig. 2.2.6: An 8-bp insertion, SOX2/3 binding site necessary for LTR7up transcription.
A) (log) p-values >500 for HOMER motifs enriched in 7up1 insertion's sequence blocks vs the same blocks from other insertions from other HERVH subfamilies are shown. B) Line plots show SOX2 ChIP-seq signal at LTR7 subfamily loci in human ESCs. Signal from genomic loci was compiled relative to position 0. The 7up/u1 8bp insertion position is shown with a dotted line. Region 2b harboring SOX2/3 TFBS is detailed below. C) Scheme of DNA fragments cloned into pGL3-basic vector driving luciferase gene expression (LUC) with identified SOX2/3 motifs. 3 constructs were analyzed: Entire LTR7up (7up1), 7d1/2 consensus sequence (approximate ancestral sequence for all LTR7d) and LTR7up with 8 nucleotides deleted (LTR7up (Δ8bp - AAAAGAAG)) (see panel B). D) Normalized relative luciferase activity of tested fragments compared to LTR7 down; n = 4 measurements; bars, means across replicates; error bars, standard deviation of the mean, dots, individual replicates.

We hypothesized that the 8-bp insertion provided a binding motif for SOX2 and/or SOX3 contributing to 7up promoter activity in ESCs. Indeed, SOX2 and SOX3 bind a highly similar motif (Bergsland et al., 2011; Heinz et al., 2010), activate an overlapping set of genes and play a redundant function in pluripotency (Corsinotti et al., 2017; Niwa et al., 2016; Wang et al., 2012). In addition, we observed that both SOX2 and SOX3 are expressed in human ESCs but SOX3 was more highly and more specifically expressed in ESCs (Figure supplement 6A,C). While SOX3 binding has not been profiled in human ESCs, ChIP-seq data available for SOX2 indicated that it binds preferentially 7up in a region coinciding with the 8-bp motif **(Fig. 2.2.6B)**. Together these observations suggest that 7up promoter activity in ESCs might be conferred in part by the gain of a SOX2/3 motif in block 2b.

To experimentally test this prediction, we used a luciferase reporter to assay promoter activity of three different LTR7 sequences in iPSCs (see methods). The first consisted of the full-length 7d consensus sequence (predicted to be inactive in iPSCs), the second contained the full-length 7up1 consensus (predicted to be active) and the third used the same 7up1 consensus sequence but lacking the 8-bp motif unique to 7up1/2 and 7u1 elements overlapping the SOX2/3 motif **(Fig. 2.2.6B,C)**. The results of the assays revealed that the 7d construct exhibited, as predicted, only weak promoter activity in iPSC compared to the empty vector **(Fig. 2.2.6D)**, while the 7up1 construct had much stronger promoter activity, driving on average 7.8-fold more luciferase expression than 7d and 100-fold more than the empty vector **(Fig. 2.2.6D)**. Strikingly, the promoter activity was essentially abolished in the 7up1 construct lacking the 8-bp motif, which drove minimal luciferase expression (on average, 3-fold less than LTR7d and 20-fold less than the intact LTR7up sequence). These results demonstrate that the 8-bp motif in 7up1 is necessary for robust promoter activity in iPSCs, likely by providing a SOX2/3 binding site essential for this activity.

2.3 DISCUSSION

The HERVH family has been the subject of intense investigation for its transcriptional and regulatory activities in human pluripotent stem cells. These studies often have treated the entire family as one homogenous, monophyletic entity and it has remained generally unclear which loci are transcribed and potentially important for pluripotency. This is in part because HERVH/LTR7 is an abundant and young family which poses technical challenges to interrogate the activity of individual loci and design experiments targeting specific members of the family (Chuong et al., 2017; Lanciano and Cristofari, 2020). Here, we applied a 'phyloregulatory' approach that integrates regulatory genomics data to a phylogenetic analysis of LTR7 sequences to reveal several new insights into the origin, evolution, and transcription of HERVH elements. In brief, our results show that: (i) LTR7 is a polyphyletic group composed of at least eight monophyletic subfamilies; (ii) these subfamilies have distinct evolutionary histories and transcriptional profiles in human embryos and a single and relatively small subgroup (~264 loci), LTR7up, exhibits robust promoter activity in ESC; (iii) LTR7 evolution is characterized by the gain, loss, and exchange of cis-regulatory modules likely underlying their transcriptional partitioning during early embryonic development.

2.3.1 PHYLOREGULATORY ANALYSIS OF LTR7 DISENTANGLES THE CIS-REGULATORY EVOLUTION OF HERVH

Previous studies have treated LTR7 *sensu stricto* insertions as equivalent representatives of their subfamilies (Bao et al., 2015; Gemmell et al., 2019; Göke et al., 2015; Izsvák et al., 2016; Storer et al., 2021; Wang et al., 2014; Zhang et al., 2019). While some of these studies were able to detect differential transcriptional partitioning between LTR7, LTR7y, and LTR7b (Göke et al., 2015), the amalgamating of LTR7 loci limited the ability to detect transcriptional variations among LTR7 and to identify key sequence differences responsible for divergent transcription patterns. Our granular parsing of LTR7 elements and their phyloregulatory profiling has revealed striking genetic, regulatory, and evolutionary differences amongst these sequences. Importantly, a phylogeny based on the coding sequence (RVT domain) of HERVH provided less granularity to separate the subfamilies than the LTR sequences (Figure supplement 7). The classification of new subfamilies within LTR7 enabled us to discover that they have distinct expression profiles during early embryonic development **(Fig. 2.2.5)** that were previously obscured by their aggregation into a single group of elements. For example, the 7u2 subfamily is, to our knowledge, the first subfamily of human TEs reported to have preimplantation expression exclusively in the epiblast.

It has been observed for some time that only a small subset of HERVH elements are expressed in ESCs (Gemmell et al., 2019; Göke et al., 2015; Ohnuki et al., 2014; Santoni et al., 2012; Schön et al., 2001; Wang et al., 2014; Zhang et al., 2019). Some have attributed this property to variation in the internal region of HERVH, context-dependent effects (local chromatin or cis-regulatory environment) and/or age (Gemmell et al., 2019; Zhang et al., 2019). Our results provide an additional, perhaps simpler explanation: we found that HERVH elements expressed in ESCs are almost exclusively driven by two closely related subfamilies of LTR7 (7up) that emerged most recently in hominoid evolution. We identified one 8-bp sequence motif overlaps a predicted SOX2/3 binding site unique to 7up that is required for promoter activity in pluripotent stem cells. These results highlight that the primary sequence of the LTR plays an important role in differentiating and diversifying HERVH expression during human embryonic development.

The phyloregulatory approach outlined in this study could be applied to illuminate the regulatory activities of LTR elements in other cellular contexts. In addition to embryogenesis, subsets of LTR7 and LTR7y elements are known to be upregulated in oncogenic states (Babaian and Mager, 2016; Glinsky, 2015; Kong et al., 2019; Yu et al., 2013). It would be interesting to explore whether these activities can be linked to the gain of specific TFBS using the new LTR7 annotations and regulatory information presented herein. Other human LTR families, such as MER41, LTR12C, or LTR13 have been previously identified as enriched for particular TF binding and cis-regulatory activities in specific cellular contexts (Chuong et al., 2016; Deniz et al., 2020; Ito et al., 2017; Krönung et al., 2016; Sundaram et al., 2014). In each case, TF binding enrichment was driven by a relatively small subset of loci within each family. We suspect that some of the intrafamilial differences in TF binding and cis-regulatory activity may be caused by unrecognized subfamily structure and subfamily-specific combinations of TFBS, much like we observe for LTR7.

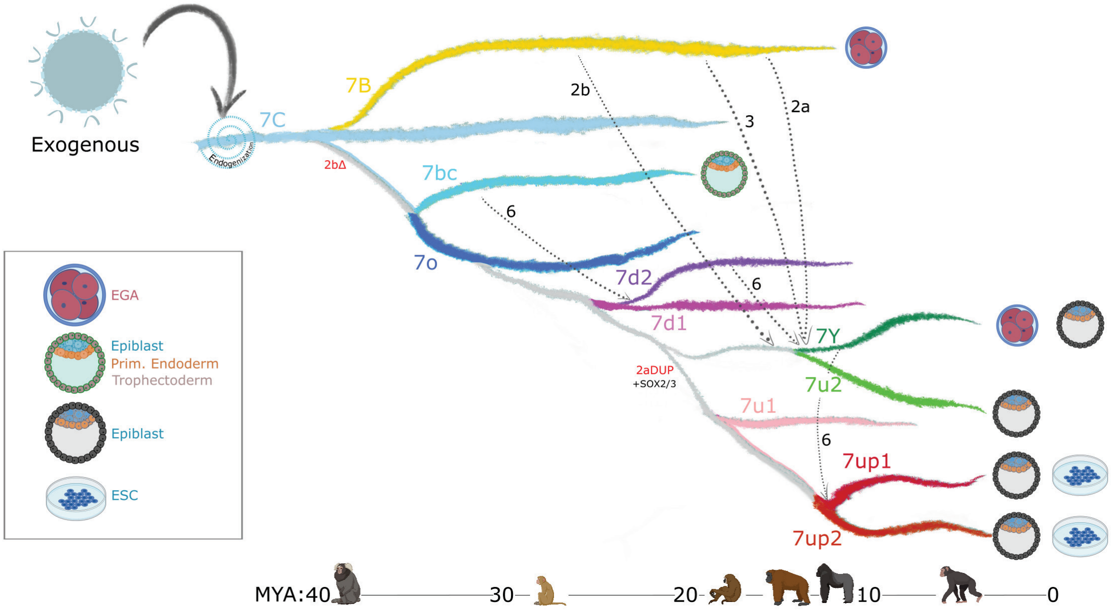


Fig. 2.3.1: Model of LTR7 subfamily evolution.
Estimated LTR7 subfamily transpositional activity in mya are listed with corresponding approximate primate divergence times (bottom). The positioning and duration of transpositional activity are based on analysis from Fig. 3b. The grey connections between subfamilies indicate average tree topology which is driven by overall pairwise sequence similarity. Dashed lines indicate likely recombination events which led to the founding of new subfamilies. Stage-specific expression profiles from Fig. 5a are detailed to the right of each corresponding branch.

2.3.2 RECOMBINATION AS A DRIVER OF LTR CIS-REGULATORY EVOLUTION

Recombination is a common and important force in the evolution of exogenous RNA viruses (Jetzt et al., 2000; Pérez-Losada et al., 2015; Simon-Loriere and Holmes, 2011) and endogenous retroviruses (Vargiu et al., 2016). Traditional models of recombination describe recombination occurring due to template switching during reverse transcription, a process that requires the co-packaging of RNA genomes, a feature of retroviruses and some retrotransposons (Lai, 1992; Matsuda and Garfinkel, 2009). Previous studies proposed that the HERVH family had undergone inter-element recombination events of both its coding genes (Mager and Freeman, 1987; Vargiu et al., 2016) and LTR (Goodchild et al., 1993). Specifically, it was inferred that recombination event between Type I LTR (i.e., LTR7) and Type II LTR (LTR7b) led to the emergence of Type Ia (LTR7y).

Our findings of extensive sequence block exchange between 7y and 7b **(Fig. 2.2.4D)** are consistent with these inferences. Furthermore, our division of HERVH into at least 11 subfamilies, rather than the original trio (Type I, II, Ia), and systematic analysis of recombination events **(Fig. 2.2.4)** suggest that recombination has occurred between multiple lineages of elements and has been a pervasive force underlying LTR diversification. We identified a minimum of six recombination events spanning 20 million years of primate evolution (see **Fig. 2.2.4D** and summary model in **Fig. 2.2.7**). The coincidence of recombination events with changes in expression profiles **(Fig. 2.2.7)** suggests that these events were instrumental to the diversification of HERVH embryonic expression. The hybrid origin and subsequent burst of amplification of LTR7 subfamilies **(Fig. 2.2.1,2)** suggest they expanded rapidly after shifting their transcriptional profiles. The coincidence of niche colonization with a burst in transposition leads us to speculate that these shifts in expression were foundational to the formation and successful expansion of new HERVH subfamilies. It would be interesting to explore whether inter-element recombination has also contributed to the evolution of other LTR subfamilies and the diversification of their expression patterns. Previous work has highlighted the role of TEs, and LTRs in particular, in donating built-in cis-regulatory sequences promoting the evolutionary rewiring of mammalian transcriptional networks (Chuong et al., 2017; Feschotte, 2008; Hermant and Torres-Padilla, 2021; Jacques et al., 2013; Rebollo et al., 2012; Sundaram and Wysocka, 2020; Thompson et al., 2016). We show that recombination provides another layer to this idea, where combinations of TFBS can be mixed-and-matched, then mobilized and propagated, further accelerating the diversification of these regulatory DNA elements. As HERVH expanded and diversified, its newly evolved cis-regulatory modules became confined to specific host lineages **(Fig. 2.2.2)**. Thus, it is possible that the formation of new LTR via recombination and their subsequent amplification catalyzed cis-regulatory divergence across primate species.

2.3.3 LTR EVOLUTION ENABLED HERVH'S COLONIZATION OF DIFFERENT NICHES IN THE HUMAN EMBRYO

Our evolutionary analysis reveals that multiple HERVH subfamilies were transpositionally active in parallel during the past ~25 my of primate evolution **(Fig. 2.2.2,7)**. This is in stark contrast to the pattern of LINE1 evolution in primates, which is characterized by a single subfamily being predominantly active at any given time (Khan et al., 2006). We hypothesize that the ability of HERVH to colonize multiple cellular niches underlie this difference. Indeed, we observe that concurrently active HERVH subfamilies are transcribed at different developmental stages, such as 7up and 7u2 being transcribed in the pluripotent epiblast at the same time that 7y and the youngest 7b were transcribed at the 8 cell and morula stages **(Fig. 2.2.7)**. We posit that this partitioning allowed multiple HERVH subfamilies to amplify in parallel without causing overt genome instability and cell death during embryonic development.

Niche diversification may have also enabled HERVH to evade cell-type-specific repression by host-encoded factors such as KZFPs. KZFPs are thought to emerge and adapt during evolution to silence specific TE subfamilies in a cell-type specific manner (Bruno et al., 2019; Cosby et al., 2019; Ecco et al., 2017; Imbeault et al., 2017). For example, there is evidence that the progenitors of the currently active L1HS subfamily became silenced in human ESCs via KZFP targeting, but evaded that repression and persisted in that niche through the deletion of the KZFP binding site (Jacobs et al., 2014). HERVH may have persisted through another evasive strategy: changing their TFBS repertoire to colonize niches lacking their repressors. To silence all LTR7, any potential HERVH-targeting KZFP would need to gain expression in multiple cellular contexts. For example, one potential repressor, ZNF534, binds a wide range of LTR7 sequences, but is particularly enriched at 7up in HEK293 cells **(Fig. 2.2.3A,D)**. Our analysis shows that ZNF534 is most highly expressed in the morula, but dips in human ESC (Figure supplement 3). Thus, ZNF534 may repress 7up at earlier stages of development but is apparently unable to suppress 7up transcription in pluripotent stem cells. If true, this scenario would illustrate how LTR diversification facilitated HERVH persistence in the face of KZFP coevolution. Further investigation is needed to explore the interplay between KZFPs and HERVH subfamilies during primate evolution.

2.3.4 IMPLICATIONS FOR STEM CELL AND REGENERATIVE BIOLOGY

Lastly, our findings may provide new opportunities for stem cell research and regenerative medicine. Our data on 7up reinforces previous findings (Corsinotti et al., 2017; Wang et al., 2012) that place SOX2/3 as central players in pluripotency. Furthermore, our analysis identified a set of TFs whose motifs are uniquely enriched in different LTR7 subfamilies with distinct expression patterns in early embryonic cells, which may enable a functional discriminatory analysis of the role of these TFs in each cell type. HERVH/LTR7 has been used as a marker for human pluripotency (Ohnuki et al., 2014; Santoni et al., 2012; Wang et al., 2014), and recent work has revealed that HERVH/LTR7-positive cells may be more amenable to differentiation, and are therefore referred to as “primed” cells (Göke et al., 2015; Theunissen et al., 2016). However, primed cells are not as promising for regenerative medicine as so-called “naïve” cells (Nichols and Smith, 2009), which are less differentiated and resemble cells from late morula to epiblast, or so-called “formative” cells, which most closely resemble cells from the early post-implantation epiblast (Kalkan and Smith, 2014; Kinoshita et al., 2021; Rossant and Tam, 2017). Of relevance to this issue is our finding that elements of the 7u2 subfamily are highly and exclusively expressed in the pluripotent epiblast in vivo **(Fig. 2.2.5)**, but weakly so in H1 ESC, which consists of a majority of primed cells and a minority of naïve or formative cells (Gafni et al., 2013). Thus, it might be possible to develop a LTR7u2-driven reporter system to mark and purify naïve or formative cells from an heterogenous ESC population. Similarly, a MERVL LTR-GFP transgene has been used in mouse to purify rare 2-cell-like totipotent cells where this LTR is specifically expressed amidst mouse ESCs in culture (Hermant and Torres-Padilla, 2021; Macfarlan et al., 2012).

In conclusion, our study highlights the modular cis-regulatory evolution of an endogenous retrovirus which has facilitated its transcriptional partitioning in early embryogenesis. We believe that phyloregulatory dissection of endogenous retroviral LTRs has the potential to further our understanding of the evolution, impact, and applications of these elements in a broad range of biomedical areas.

2.4 ACKNOWLEDGEMENTS

We would like to thank Arian Smit and Robert Hubley for their assistance in remasking the human genome with the custom LTR7 library in accordance with RepeatMasker/Dfam specifications. We would also like to thank the Cornell statistical consulting unit (CSCU), and Stephen Parry in particular, for their help with statistical analyses. Finally, we would like to credit BioRender.com for its primate **(Fig. 2.2.2,7)** and embryonic cell **(Fig. 2.2.6,7)** graphics. This work was supported by funds from the National Institutes of Health to CF (GM112972; HG009391; GM122550); Cornell Center for Vertebrate Genomics to T.A.C.; and the Cornell Presidential Fellow Program to M.S. J.L.R. is supported by the Howard Hughes Medical Institute Faculty Scholar program and NIH (GM099117).

2.5 METHODS

2.5.1 HERVH LTR sequence identification

All HERVH-int and accompanying LTRs (LTR7, 7b, 7c, and 7y) were extracted from masked (RepeatMasker version 4.0.5 repeat Library 20140131 - (Smit et al., 2013)) GRCh38/hg38 (alt chromosomes removed). All annotated HERVH-int and HERVH LTR were run through OneCodeToFindThemAll.pl (Bailly-Bechet et al., 2014) followed by rename_mergedLTRelements.pl (Thomas et al., 2018) to identify solo and full-length HERVH insertions. 5' LTRs from full-length insertions >4kb were combined with and solo LTRs. LTRs >350bp were considered for future analysis.

2.5.2 Multiple sequence alignment, phylogenetic tree generation, and LTR7 subdivision

All HERVH LTRs **(Fig. 2.2.1A – supp. file 5)** or only LTR7s **(Fig. 2.2.1B – supp. file 6)** were aligned with mafft –auto (Nakamura et al., 2018) strategy: FFT-NS-2/Progressive method followed by PRANK (Löytynoja and Goldman, 2010) with options -showanc -support -njtree -uselogs -prunetree -prunedata -F -showevents. Uninformative structural variations were removed with Trimal (Capella-Gutierrez et al., 2009) with option -gt 0.01.

To visualize inter-insertion relationships, the MSA was input into IQtree with options -nt AUTO -m MFP -bb 6000 -asr -minsup .95 (Chernomor et al., 2016). This only displays nodes with ultrafast (UF) bootstrap support >0.95.

Clusters of >10 insertions sharing a node with UFbootstrap support that were separated from other insertions by internal branch lengths >0.015 (1.5subs / 100 bp) were defined as belonging to a new bona fide LTR7 subfamily **(Fig. 2.2.1B)**.

2.5.3 LTR7 consensus generation and network analysis

Majority rule (51%) was used to generate each LTR7 subfamily at nodes described in **Fig. 2.2.1**. Positions without majority consensus are listed as "N". Majority rule consensus sequences were aligned with MUSCLE in SEAVIEW (supp. file 7) (Edgar, 2004; Gouy et al., 2010). Alignment was visualized with Jalview2 (Waterhouse et al., 2009) **(Fig. 2.2.4A)** and ggplot2 **(Fig. 2.2.4)**.

Non-gap SNPs from the muscle alignment were used to construct a median-joining network (Bandelt et al., 1999) with POPART (Leigh and Bryant, 2015).

2.5.4 Reverse Transcriptase Domain extraction, alignment, and tree generation

The reverse transcriptase (RT) domain was extracted from HERVH-int consensus via repbrowser (Fernandes et al., 2020):

CACCCTTACCCCGCTCAATGCCAATATCCCATCCCACAGCATGCTTTAAAAGGATTAAAGCCTGTTATCACTCGCCTGCTACAGCATGGCCTTTTAAAGCCTATAAACTCTCCTTACAATTCCCCCATTTTACCTGTCCTAAAACCAGACAAGCCTTACAAGTTAGTTCAGGATCTGTGCCTTATCAACCAAATTGTTTTGCCTATCCACCCCATGGTGCCAAACCCATATACTCTCCTATCCTCAATACCTCCCTCCACAACCCATTATTCTGTTCTGGATCTCAAACATGCTTTCTTTACTATTCCTTTGCACCCTTCATCCCAGCCTCTCTTCGCTTTCACTTGGA

This sequence was blated (best hit) against all annotated HERVH-int in the human genome and matches were extracted. Corresponding LTR7 subdivision annotations from figure 1 were matched with these HERVH-int RT domains. Mafft alignment and IQTree generation were done identically to the Mafft and IQTree run for the LTRs (see corresponding methods section).

2.5.5 Peak calling

ChIP-seq datasets representing transcription factors (TFs), histone modifications, and regulatory complexes in human embryonic stem cells and differentiated cells were retrieved from GSE61475 (38 distinct TFs and histone modifications), GSE69647 (H3K27Ac, POU5F1, MED1 and CTCF), GSE117395 (H3K27Ac, H3K9Me3, KLF4, and KLF17), and GSE78099 (An array of KRAB-ZNFs and TRIM28) (Imbeault et al., 2017). ZNFs enriched in LTR7 binding (ZNF90, ZNF534, ZNF75, ZNF69B, ZNF257, ZNF57, and ZNF101) from HEK293 peaks were all evaluated, but only ZNF90 and ZNF534 bound >100 LTR7 insertions (data not shown). The others were dropped from the analysis.

ChIP-seq reads were aligned to the hg19 human reference genome using the Bowtie2. All reads with phred score less than 33 and PCR duplicates were removed using bowtie2 and Picard tools respectively. ChIP-seq peaks were called by MACS2 with the parameters in “narrow” mode for TFs and “broad” mode for histone modifications, keeping FDR < 1%. ENCODE-defined blacklisted regions were excluded from called peaks. For phyloregulatory analysis **(Fig. 2.2.2)**, we then converted hg19 to hg38 (no alt) coordinates via UCSC *liftover* (100% of coordinates lifted) and intersected these peak with the loci from LTR7 subfamilies using bedtools with any overlap. For ChIP-seq binding enrichment on a subset of marks following motif analysis (Fig 5), 70% overlap of peak and LTR was required. Enrichment of a given TF within LTR7 subfamilies was calculated using enrichR package in R, using the customized in-house codes (see the codes on GitHub for the detailed analysis pipelines and calculation of enrichment score).

2.5.6 Phyloregulatory analysis

Peaks from external ChIP-seq datasets were intersected with LTR7 insertions (Quinlan and Hall, 2010). LTR7 insertions that intersected with >1bp of peaks were counted as positive for the respective mark. We repeated this analysis with a range of overlap requirements from extending the LTR 500bp into unique DNA to 70% overlap and found few differential calls (data not shown). The phylogenetic tree rooted on 7b (ggtree) was combined with these binary data (ggheat).

"Highly transcribed" (fpkm >2) and "chimeric" HERVH from H1 cells (GSE54726) (Wang et al., 2014) were intersected with LTR7 similarly to ChIP-seq data. Those which intersected LTR7 were marked as "RNA-seq" or "chimeric" respectively.

GRO-seq profiles from H1 cells (Estaras et al.) (GSE64758) were created for windows 10bp upstream and 8kb downstream of 5' and solo LTR7 (Ramírez et al., 2016). The most visible signal was confined to the top 7th of insertions (Figure supplement 2). All LTR7 were subdivided into septiles, due to visible signal being confined to the top 7th of insertions; those of the top septile were labeled "GRO-seq".

2.5.7 Peak proportion heatmap generation and statistical analysis

Tables with the proportion of solo and 5’ LTRs from a given subfamily positive for select marks (phyloregulatory analysis) were used to generate heatmaps with the R package ggplot (ggheat) (Ginestet, 2011). Those with padj<0.05 (Chi-square Bonferroni correction n=147 tests for a total of 21 marks examined) were considered significantly enriched in 7up1. Enrichment for non-LTR7up subfamilies was not tested. While not all tested marks are displayed in the main text, statistical analysis was performed with all tested marks (n=147) (supp. file 8). For comparing transcribed 7up to untranscribed 7up, 18 pairwise comparisons were made (supp. file 9).

2.5.8 Aggregate signal heatmap generation

GRO-seq (H1 cells - GSE64758), whole-genome bisulfite sequencing (WGBS-seq – H1 cells), and H3K9me3 ChIP-seq (H1 – primed - GSE78099) bams were retrieved from (Estarás et al., 2015), (Dunham et al., 2012), and (Theunissen et al., 2016) respectively. Deeptools (Ramírez et al., 2016) was used to visualize these marks by LTR7 subfamily division in windows 10bp upstream and 8kb downstream of the most 5’ position in the LTR (Figure supplement 2).

2.5.9 Orthologous insertion aging

Human coordinates for 7b, 7c, and 7y and LTR7 used in alignments and tree generation were lifted over (Kent et al., 2002; Raney et al., 2014) from GRCh38/hg38 (Miga et al., 2014) to Clint_PTRv2/panTro6 (Waterson et al., 2005), Kamilah_GGO_v0/gorGor6 (Scally et al., 2012), Susie_PABv2/ponAbe3 (Locke et al., 2011), GGSC Nleu3.0/nomLeu3 (Carbone et al., 2014), or Mmul_10/rheMac10 (Gibbs et al., 2007). Those that were successfully lifted over from human to non-human primate were then lifted over back to human. Only those that survived both liftovers (1:1 orthologous) were counted as present in non-human primates. The proportion of those orthologous to human and total number of orthologous was plotted with ggplot2.

2.5.10 Terminal branch length aging

Terminal branch lengths from the LTR7 phylogenetic tree **(Fig. 2.2.1B)** were extracted and plotted with ggplot2. Similarly aged subfamilies were inferred from means here and from orthologous insertion aging for statistical testing. Three total groups were tested for differences in means (7up1/7up2/7u2 vs. 7d1/7d2/7u1 vs. 7bc/o) via Wilcox rank-sum test with Bonferroni multiple testing correction.

2.5.11 Identification of recombination breakpoints and consensus parsimony tree generation

Major recombination breakpoints were identified by eye from the consensus sequence MSA, where SNPs and structural rearrangements seemed to have different relationships between blocks. Putative block recombination events were identified by looking for shared shapes in the block consensus MSA **(Fig. 2.2.4A)**. To test if these were truly recombination events and could not be explained by evolution by common descent, inter-block sequence relationship differences were tested by generating parsimony trees and comparing these to the overall phylogenetic structure from Fig.

2.2.1A. Parsimony trees were generated in SEAVIEW, treating all gaps as unknown states (except in the case of 2b, where the entire sequence is gaps and gaps were not treated differently than other sequence), bootstrapped 5000 times with the option "more thorough tree search". Differences in block parsimony trees and the overall phylogeny that had bootstrap support were marked in red and included in Fig. 2.2.4D,7.

2.5.12 7up consensus block 2a 2b alignment and parsimony tree

LTR7up blocks 2a and 2b **(Fig. 2.2.4)** appeared to share sequence. To determine if block 2b was the result of a duplication of 2a, we extracted these sequences from the LTR7up1 consensus and aligned them with blastn (NCBI web version) with default settings. To determine the relationship of all HERVH LTR 2a and 2b blocks, we performed a muscle alignment (default settings) of all 2a and 2b from all HERVH LTR consensus sequences and then generated a parsimony tree with 5000 bootstraps with SEAVIEW with the option "more thorough tree search".

2.5.13 New LTR7B/C/Y consensus generation and remasking of human genome

Consensus sequences for LTR7 subfamilies were generated using the tree from figure 1b (see above). For LTR7b/c/y, we used the alignment and tree comprising all HERVH LTR (Figure supplement 5). To do this, we identified nodes with >0.95 ultrafast bootstrap support that were comprised of predominately (>90%) of previously annotated LTR7b, LTR7c, or LTR7y. These sequences were used to generate majority-rule consensus sequences for their respective subfamily. We generated 2 mutually-exclusive LTR7c consensus sequences (LTR7C1 and LTR7C2) due to the high sequence divergence of LTR7C. Both of these subfamilies were merged into "LTR7C" after remasking.

Parsing previously annotated LTR7 into 8 subfamilies and evidence of recurrent recombination events caused concern that HERVH LTRs may be misannotated in the repeat masker annotations. To compensate, we remasked (Smit et al., 2013) GRCh38/hg38 (excluding alt chromosomes) with a custom library consisting of the new consensus sequences for LTR7 subfamilies, new consensus sequences for 7b, 7y, and 7c (see above) based on the HERVH LTR tree from Fig. 2.2.4, and HERVH-int (dfam). We also included annotated consensus sequences from dfam for MER48, MER39, AluYk3, and MST1N2, who we found a HERVH only library also masked to a limited degree (data not shown). With this library, we ran RepeatMasker with crossmatch and "sensitive" settings: -e crossmatch -a -s -no_is. Changes in annotations can be found in (HERVH_LTRremasking.xlsx)

2.5.14 Embryonic HERVH subfamily expression analysis

We downloaded the raw single-cell RNA-seq datasets from early human embryos and embryonic stem cells (GSE36552) and the EPI, PE, TE cells (GSE66507) in sra format. Following the conversion of raw files into fastq format, the quality was determined by using the FastQC. We removed two nucleotides from the ends as their quality scores were highly variable compared with the rest of the sequences in RNA-seq reads. Prior to aligning the resulting reads, we first curated the reference genome annotations using the LTR7 classification, shown in the manuscript. We extracted the genes (genecode V19), and LTR7 subfamilies (see figure 5) genomic sequences and combined them to generate a reference transcriptome. These sequences were then appended, comprising the coding-sequences plus UTRs of genes and locus-level LTR7 subfamilies sequences in fasta format. We then annotated every fasta sequences with their respective genes or LTR7 subfamilies IDs. To guide the transcriptome assembly, we also appended the each of the resulting contigs and modelled them in gtf format that we utilized for the expression quantification. Next, we indexed the concatenated genes and LTR7 subfamilies transcriptome and genome reference sequences using 'salmon' (Patro et al., 2017). Finally, we aligned the trimmed sequencing reads against the curated reference genome. The 'salmon' tool quantified the counts and normalized expression (Transcripts per million (TPM)) for each single cell RNAseq sample. Overall, this approach enabled us to simultaneously calculate LTR7 subfamilies and protein-coding gene expression using expected maximization algorithms. Data integration of obtained count matrix, normalization at logarithmic scale, and scaling were performed as per the "Seurat V.3.7" (http://satijalab.org/seurat/) guidelines. The annotations of cell-types were taken as it was classified in original studies. We calculated differential expression and tested their significance level using Kruskal–Wallis test by comparing cell-types of interest with the rest of the cells. The obtained p-values were further adjusted by the Benjamini-Hochberg method to calculate the False Discovery Rate (FDR). All the statistics and visualization of RNA-seq were performed on R (https://www.r-project.org/).

2.5.15 Motif Enrichment

For each subfamily of LTR7 elements, all re-annotated elements were aligned against the subfamily consensus sequence using MUSCLE (Edgar, 2004). These multiple-sequence alignments were then split based on the recombination block positions in the consensus sequence. The consensus sequence was then removed. Binding motif position-weight matricies were downloaded from HOMER (Heinz et al., 2010) and were used to perform pairwise motif enrichment using the command 'homer2 find'. For LTR7up1 enrichment (**Fig. 2.2.6A** - testing which motifs were enriched in LTR7up1 compared to other subfamilies), enrichment was only calculated for LTR7up1 and the motifs with a -log(p-value) cutoff of 1x10-5 were kept. For enrichment in all subfamilies (supp. files 3,4) – testing all subfamilies against all others), every pairwise subfamily combination within each block was tested and all results are displayed.

2.5.16 SOX2 ChIP-seq signal on LTR7

SOX2 ChIP-seq and whole-cell extract datasets from primed hESCs were downloaded in fastq format from GEO ID GSE125553 (Bayerl et al., 2021). Fastq reads were mapped against the hg19 reference genome with the bowtie2 parameters: *–very-sensitive-local*. All unmapped reads with Phred score < 33 and putative PCR duplicates were removed using *Picard* and *samtools*. All the ChIP-seq narrow peaks were called by MACS2 (FDR < 0.01). To generate a set of unique peaks, we merged ChIP-seq peaks within 50 bp of one another using the *mergeBed* function from bedtools. We then intersected these peak sets with LTR7 subgroups from hg19 repeat-masked coordinates using bedtools *intersectBed* with 50% overlap. LTR7up1 and LTR7up2 were harboring the highest number of peaks compared with the rest of the subgroups. To illustrate the enrichment over the LTR7 subgroups, we first extended 500 basepairs from upstream and downstream coordinates from the left boundary of each LTR7subgroups. These 1KB windows were further divided into 10 bps bins. The normalized ChIP-seq signal over the local lambda (piled up bedGraph outputs from MACS2) was counted in each bin. These counts were then normalized by the total number of mappable reads per million in given samples and presented as signal per million per 10 bps. Finally, these values were averaged across the loci for each bin to illustrate the subfamilies' level of ChIP-seq enrichment. Replicates were merged prior to plotting. Note: Pearson's correlation coefficient between replicates across the bins was found to be r > 0.90.

2.5.17 Luciferase reporter assay

The inserts (LTR7 variants or EF1a promoter) with restriction enzyme overhangs were ordered from Genewiz and cloned into pGL3-basic plasmid upstream of the firefly reporter gene (E1751, Promega). Minipreps were prepared with QIAprep Spin Miniprep kit (Qiagen). Plasmids were sequenced to ensure the correct sequence and directionality of the insert. 24 h before transfection, human iPSC WTC-11 (Coriell Institute) cells were plated on Vitronectin (Thermo Fisher Scientific) coated 12-well plates in Essential 8 Flex medium (Thermo Fisher Scientific) with E8 supplement (Thermo Fisher Scientific), Rock inhibitor and 2.5% penicillin-streptomycin. Cells were co-transfected with 800 ng of plasmid of interest and 150 ng plasmid containing EF1a upstream of GFP for normalization with Lipofectamine Stem transfection reagent (Thermo Fisher scientific) according to manufacturer's instructions. 48 h after transfection, cell pellet was harvested and luciferase activity was measured with Luciferase Reporter Assay kit (Promega) on Glomax (Promega) according to instructions. Transfection efficiency and cell count was normalized with GFP.

1. 7down:

GCTAGCTGTCAGGCCTCTGAGCCCAAGCTAAGCCATCATATCCCCTGTGAC
CTGCACGTACACATCCAGATGGCCGGTTCCTGCCTTAACTGATGACATTCC
ACCACAAAAGAAGTGAAAATGGCCTGTTCCTGCCTTAACTGATGACATTAT
CTTGTGAAATTCCTTCTCCTGGCTCATCCTGGCTCAAAAGCTCCCCTACTGA
GCACCTTGTGACCCCCACTCCTGCCCGCCAGAGAACAACCCCCCTTTGACT
GTAATTTTCCTTTACCTACCCAAATCCTATAAAACGGCCCCACCCCTATCTC
CCTTCGCTGACTCTCTTTTCGGACTCAGCCCGCCTGCACCCAGGTGAAATA
AACAGCTTTATTGCTCACACAAAGCCTGTTTGGTGGTCTCTTCACACGGAC
GCGCATGCTCGAG

2. LTR7upcons:

GCTAGCTGTCAGGCCTCTGAGCCCAAGCCAAGCCATCGCATCCCCTGTGAC
TTGCACGTATACGCCCAGATGGCCTGAAGTAACTGAAGAATCACAAAAGA
AGTGAATATGCCCTGCCCCACCTTAACTGATGACATTCCACCACAAAAGA
AGTGTAAATGGCCGGTCCTTGCCTTAAGTGATGACATTACCTTGTGAAAGT
CCTTTTCCTGGCTCATCCTGGCTCAAAAAGCACCCCCACTGAGCACCTTGC

GACCCCCACTCCTGCCCGCCAGAGAACAAACCCCCTTTGACTGTAATTTTC

CTTTACCTACCCAAATCCTATAAAACGGCCCCACCCTTATCTCCCTTCGCTG

ACTCTCTTTTCGGACTCAGCCCGCCTGCACCCAGGTGAAATAAACAGCCAT

GTTGCTCACACAAAGCCTGTTTGGTGGTCTCTTCACACGGACGCGCATGCT

CGAG

5. LTR7upcons_AAAGAAG_deletion:

GCTAGCTGTCAGGCCTCTGAGCCCAAGCCAAGCCATCGCATCCCCTGTGAC

TTGCACGTATACGCCCAGATGGCCTGAAGTAACTGAAGAATCACAAAAGA

AGTGAATATGCCCTGCCCCACCTTAACTGATGACATTCCACCATTGTAAAT

GGCCGGTCCTTGCCTTAAGTGATGACATTACCTTGTGAAAGTCCTTTTCCT

GGCTCATCCTGGCTCAAAAAGCACCCCCACTGAGCACCTTGCGACCCCCAC

TCCTGCCCGCCAGAGAACAAACCCCCTTTGACTGTAATTTTCCTTTACCTA

CCCAAATCCTATAAAACGGCCCCACCCTTATCTCCCTTCGCTGACTCTCTTT

TCGGACTCAGCCCGCCTGCACCCAGGTGAAATAAACAGCCATGTTGCTCA

CACAAAGCCTGTTTGGTGGTCTCTTCACACGGACGCGCATGCTCGAG

5'NheI highlighted in Yellow

3'XhoI highlighted in Cyan

REFERENCES

Babaian A, Mager DL. 2016. Endogenous retroviral promoter exaptation in human cancer. *Mobile DNA* **7**:24. doi:10.1186/s13100-016-0080-x

Bailly-Bechet M, Haudry A, Lerat E. 2014. "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. *Mobile DNA* **5**:13. doi:10.1186/1759-8753-5-13

Bandelt HJ, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution* **16**:37–48. doi:10.1093/oxfordjournals.molbev.a026036

Bannert N, Kurth R. 2004. Retroelements and the human genome: New perspectives on an old relation. *PNAS* **101**:14572–14579. doi:10.1073/pnas.0404838101

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**:11. doi:10.1186/s13100-015-0041-9

Bayerl J, Ayyash M, Shani T, Manor YS, Gafni O, Massarwa R, Kalma Y, Aguilera-Castrejon A, Zerbib M, Amir H, Sheban D, Geula S, Mor N, Weinberger L, Naveh Tassa S, Krupalnik V, Oldak B, Livnat N, Tarazi S, Tawil S, Wildschutz E, Ashouokhi S, Lasman L, Rotter V, Hanna S, Ben-Yosef D, Novershtern N, Viukov S, Hanna JH. 2021. Principles of signaling pathway modulation for enhancing human naive pluripotency induction. *Cell Stem Cell* S1934-5909(21)00158–2. doi:10.1016/j.stem.2021.04.001

Bergsland M, Ramsköld D, Zaouter C, Klum S, Sandberg R, Muhr J. 2011. Sequentially acting Sox transcription factors in neural lineage development. *Genes Dev* **25**:2453–2464. doi:10.1101/gad.176008.111

Blakeley P, Fogarty NME, del Valle I, Wamaitha SE, Hu TX, Elder K, Snell P, Christie L, Robson P, Niakan KK. 2015. Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development* **142**:3151–3165. doi:10.1242/dev.123547

Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, Gifford DK, Melton DA, Jaenisch R, Young RA. 2005. Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. *Cell* **122**:947–956. doi:10.1016/j.cell.2005.08.020

Bruno M, Mahgoub M, Macfarlan TS. 2019. The Arms Race Between KRAB–Zinc Finger Proteins and Endogenous Retroelements and Its Impact on Mammals. *Annual Review of Genetics* **53**:393–416. doi:10.1146/annurev-genet-112618-043717

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**:1972–1973. doi:10.1093/bioinformatics/btp348

Carbone L, Alan Harris R, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, Meyer TJ, Herrero J, Roos C, Aken B, Anaclerio F, Archidiacono N, Baker C, Barrell D, Batzer MA, Beal K, Blancher A, Bohrson CL, Brameier M, Campbell MS, Capozzi O, Casola C, Chiatante G, Cree A, Damert A, de Jong PJ, Dumas L, Fernandez-Callejo M, Flicek P, Fuchs NV, Gut I, Gut M, Hahn MW, Hernandez-Rodriguez J, Hillier LW, Hubley R, Ianc B, Izsvák Z, Jablonski NG, Johnstone LM, Karimpour-Fard A, Konkel MK, Kostka D, Lazar NH, Lee SL, Lewis LR, Liu Y, Locke DP, Mallick S, Mendez FL, Muffato M, Nazareth LV, Nevonen KA, O'Bleness M, Ochis C, Odom DT, Pollard KS, Quilez J, Reich D, Rocchi M, Schumann GG, Searle S, Sikela JM, Skollar G, Smit A, Sonmez K, Hallers B ten, Terhune E, Thomas GWC, Ullmer B, Ventura M, Walker JA, Wall JD, Walter L, Ward MC, Wheelan SJ, Whelan CW, White S, Wilhelm LJ, Woerner AE, Yandell M, Zhu B, Hammer MF, Marques-Bonet T, Eichler EE, Fulton L, Fronick C, Muzny DM, Warren WC, Worley KC, Rogers J, Wilson RK, Gibbs RA. 2014. Gibbon genome and the fast karyotype evolution of small apes. *Nature* **513**:195–201. doi:10.1038/nature13679

Chambers I, Smith A. 2004. Self-renewal of teratocarcinoma and embryonic stem cells. *Oncogene* **23**:7150–7160. doi:10.1038/sj.onc.1207930

Chang N-C, Rovira Q, Wells JN, Feschotte C, Vaquerizas JM. 2021. A genomic portrait of zebrafish transposable elements and their spatiotemporal embryonic expression. *bioRxiv* 2021.04.08.439009. doi:10.1101/2021.04.08.439009

Charlesworth B, Langley CH. 1986. THE EVOLUTION OF SELF-REGULATED TRANSPOSITION OF TRANSPOSABLE ELEMENTS. *Genetics* **112**:359–383.

Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology* **65**:997–1008. doi:10.1093/sysbio/syw037

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* **18**:71–86. doi:10.1038/nrg.2016.139

Chuong EB, Elde NC, Feschotte C. 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**:1083–1087. doi:10.1126/science.aad5497

Cordaux R, Hedges DJ, Batzer MA. 2004. Retrotransposition of Alu elements: how many sources? *Trends in Genetics* **20**:464–467. doi:10.1016/j.tig.2004.07.012

Corsinotti A, Wong FC, Tatar T, Szczerbinska I, Halbritter F, Colby D, Gogolok S, Pantier R, Liggat K, Mirfazeli ES, Hall-Ponsele E, Mullin NP, Wilson V, Chambers I. 2017. Distinct SoxB1 networks are required for naïve and primed pluripotency. *eLife* **6**:e27746. doi:10.7554/eLife.27746

Cosby RL, Chang N-C, Feschotte C. 2019. Host–transposon interactions: conflict, cooperation, and cooption. *Genes Dev* **33**:1098–1116. doi:10.1101/gad.327312.119

Deniz Ö, Ahmed M, Todd CD, Rio-Machin A, Dawson MA, Branco MR. 2020. Endogenous retroviruses are a source of enhancers with oncogenic potential in acute myeloid leukaemia. *Nat Commun* **11**:3506. doi:10.1038/s41467-020-17206-4

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee B-K, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, Simon JM, Song L, Trinklein ND, Altshuler RC, Birney E, Brown JB, Cheng C, Djebali S, Dong X, Dunham I, Ernst J, Furey TS, Gerstein M, Giardine B, Greven M, Hardison RC, Harris RS, Herrero J, Hoffman MM, Iyer S, Kellis M, Khatun J, Kheradpour P, Kundaje A, Lassmann T, Li Q, Lin X, Marinov GK, Merkel A, Mortazavi A, Parker SCJ, Reddy TE, Rozowsky J, Schlesinger F, Thurman RE, Wang J, Ward LD, Whitfield TW, Wilder SP, Wu W, Xi HS, Yip KY, Zhuang J, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M, Pazin MJ, Lowdon RF, Dillon LAL, Adams LB, Kelly CJ, Zhang J, Wexler JR, Green ED, Good PJ, Feingold EA, Bernstein BE, Birney E, Crawford GE, Dekker J, Elnitski L, Farnham PJ, Gerstein M, Giddings MC, Gingeras TR, Green ED, Guigó R, Hardison RC, Hubbard TJ, Kellis M, Kent WJ, Lieb JD, Margulies EH, Myers RM, Snyder M, Stamatoyannopoulos JA, Tenenbaum SA, Weng Z, White KP, Wold B, Khatun J, Yu Y, Wrobel J, Risk BA, Gunawardena HP, Kuiper HC, Maier CW, Xie L, Chen X, Giddings MC, Bernstein BE, Epstein CB, Shoresh N, Ernst J, Kheradpour P, Mikkelsen TS, Gillespie S, Goren A, Ram O, Zhang X, Wang L, Issner R, Coyne MJ, Durham T, Ku M, Truong T, Ward LD, Altshuler RC, Eaton ML, Kellis M, Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Batut P, Bell I, Bell K, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena HP, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Li G, Luo OJ, Park E, Preall JB, Presaud K, Ribeca P, Risk BA, Robyr D, Ruan X, Sammeth M, Sandhu KS, Schaeffer L, See L-H, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Hayashizaki Y, Harrow J, Gerstein M, Hubbard TJ, Reymond A, Antonarakis SE, Hannon GJ, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR, Rosenbloom KR, Sloan CA, Learned K, Malladi VS, Wong MC, Barber GP, Cline MS, Dreszer TR, Heitner SG, Karolchik D, Kent WJ, Kirkup VM, Meyer LR, Long JC, Maddren M, Raney BJ, Furey TS, Song L, Grasfeder LL, Giresi PG, Lee B-K, Battenhouse A, Sheffield NC, Simon JM, Showers KA, Safi A, London D, Bhinge AA, Shestak C, Schaner MR, Ki Kim S, Zhang ZZ, Mieczkowski PA, Mieczkowska JO, Liu Z, McDaniell RM, Ni Y, Rashid NU, Kim MJ, Adar S, Zhang Z, Wang T, Winter D, Keefe D, Birney E, Iyer VR, Lieb JD, Crawford GE, Li G, Sandhu KS, Zheng M, Wang P, Luo OJ, Shahab A, Fullwood MJ, Ruan X, Ruan Y, Myers RM, Pauli F, Williams BA, Gertz J, Marinov GK, Reddy TE, Vielmetter J, Partridge E, Trout D, Varley KE, Gasper C, The ENCODE Project Consortium, Overall coordination (data analysis coordination), Data production leads (data production), Lead analysts (data analysis), Writing group, NHGRI project management (scientific management), Principal investigators (steering committee), Boise State University and University of North Carolina at Chapel Hill Proteomics groups (data production and analysis), Broad Institute Group (data production and analysis), Cold Spring Harbor U of G Center for Genomic Regulation, Barcelona, RIKEN, Sanger Institute, University of Lausanne, Genome Institute of Singapore group (data production and analysis), Data coordination center at UC Santa Cruz (production data coordination), Duke University E University of Texas, Austin, University of North Carolina-Chapel Hill group (data production and analysis), Genome Institute of Singapore group (data production and analysis), HudsonAlpha Institute C UC Irvine, Stanford group (data production and analysis). 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**:57–74. doi:10.1038/nature11247

Ecco G, Imbeault M, Trono D. 2017. KRAB zinc finger proteins. *Development* **144**:2719–2729. doi:10.1242/dev.132605

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**:1792–1797. doi:10.1093/nar/gkh340

Eickbush TH, Malik HS. 2002. Origins and Evolution of Retrotransposons. *Mobile DNA II* 1111–1144. doi:10.1128/9781555817954.ch49

Estarás C, Benner C, Jones KA. 2015. SMADs and YAP Compete to Control Elongation of β-Catenin:LEF-1-Recruited RNAPII during hESC Differentiation. *Molecular Cell* **58**:780–793. doi:10.1016/j.molcel.2015.04.001

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, Waki K, Hornig N, Arakawa T, Takahashi H, Kawai J, Forrest ARR, Suzuki H, Hayashizaki Y, Hume DA, Orlando V, Grimmond SM, Carninci P. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **41**:563–571. doi:10.1038/ng.368

Fernandes JD, Zamudio-Hurtado A, Clawson H, Kent WJ, Haussler D, Salama SR, Haeussler M. 2020. The UCSC repeat browser allows discovery and visualization of evolutionary conflict across repeat families. *Mobile DNA* **11**:13. doi:10.1186/s13100-020-00208-w

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**:397–405. doi:10.1038/nrg2337

Fort A, Hashimoto K, Yamada D, Salimullah M, Keya CA, Saxena A, Bonetti A, Voineagu I, Bertin N, Kratz A, Noro Y, Wong C-H, de Hoon M, Andersson R, Sandelin A, Suzuki H, Wei C-L, Koseki H, Hasegawa Y, Forrest ARR, Carninci P. 2014. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nature Genetics* **46**:558–566. doi:10.1038/ng.2965

Gafni O, Weinberger L, Mansour AA, Manor YS, Chomsky E, Ben-Yosef D, Kalma Y, Viukov S, Maza I, Zviran A, Rais Y, Shipony Z, Mukamel Z, Krupalnik V, Zerbib M, Geula S, Caspi I, Schneir D, Shwartz T, Gilad S, Amann-Zalcenstein D, Benjamin S, Amit I, Tanay A, Massarwa R, Novershtern N, Hanna JH. 2013. Derivation of novel human ground state naive pluripotent stem cells. *Nature* **504**:282–286. doi:10.1038/nature12745

Gemmell P, Hein J, Katzourakis A. 2019. The Exaptation of HERV-H: Evolutionary Analyses Reveal the Genomic Features of Highly Transcribed Elements. *Front Immunol* **10**. doi:10.3389/fimmu.2019.01339

Gemmell P, Hein J, Katzourakis A. 2015. Orthologous endogenous retroviruses exhibit directional selection since the chimp-human split. *Retrovirology* **12**. doi:10.1186/s12977-015-0172-6

Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, Batzer MA, Bustamante CD, Eichler EE, Hahn MW, Hardison RC, Makova KD, Miller W, Milosavljevic A, Palermo RE, Siepel A, Sikela JM, Attaway T, Bell S, Bernard KE, Buhay CJ, Chandrabose MN, Dao M, Davis C, Delehaunty KD, Ding Y, Dinh HH, Dugan-Rocha S, Fulton LA, Gabisi RA, Garner TT, Godfrey J, Hawes AC, Hernandez J, Hines S, Holder M, Hume J, Jhangiani SN, Joshi V, Khan ZM, Kirkness EF, Cree A, Fowler RG, Lee S, Lewis LR, Li Z, Liu Yih-shin, Moore SM, Muzny D, Nazareth LV, Ngo DN, Okwuonu GO, Pai G, Parker D, Paul HA, Pfannkoch C, Pohl CS, Rogers Y-H, Ruiz SJ, Sabo A, Santibanez J, Schneider BW, Smith SM, Sodergren E, Svatek AF, Utterback TR, Vattathil S, Warren W, White CS, Chinwalla AT, Feng Y, Halpern AL, Hillier LW, Huang X, Minx P, Nelson JO, Pepin KH, Qin X, Sutton GG, Venter E, Walenz BP, Wallis JW, Worley KC, Yang S-P, Jones SM, Marra MA, Rocchi M, Schein JE, Baertsch R, Clarke L, Csürös M, Glasscock J, Harris RA, Havlak P, Jackson AR, Jiang H, Liu Yue, Messina DN, Shen Y, Song HX-Z, Wylie T, Zhang L, Birney E, Han K, Konkel MK, Lee J, Smit AFA, Ullmer B, Wang H, Xing J, Burhans R, Cheng Z, Karro JE, Ma J, Raney B, She X, Cox MJ, Demuth JP, Dumas LJ, Han S-G, Hopkins J, Karimpour-Fard A, Kim YH, Pollack JR, Vinar T, Addo-Quaye C, Degenhardt J, Denby A, Hubisz MJ, Indap A, Kosiol C, Lahn BT, Lawson HA, Marklein A, Nielsen R, Vallender EJ, Clark AG, Ferguson B, Hernandez RD, Hirani

K, Kehrer-Sawatzki H, Kolb J, Patil S, Pu L-L, Ren Y, Smith DG, Wheeler DA, Schenck I, Ball EV, Chen R, Cooper DN, Giardine B, Hsu F, Kent WJ, Lesk A, Nelson DL, O'Brien WE, Prüfer K, Stenson PD, Wallace JC, Ke H, Liu X-M, Wang P, Xiang AP, Yang F, Barber GP, Haussler D, Karolchik D, Kern AD, Kuhn RM, Smith KE, Zwieg AS. 2007. Evolutionary and Biomedical Insights from the Rhesus Macaque Genome. *Science* **316**:222–234. doi:10.1126/science.1139247

Ginestet C. 2011. ggplot2: Elegant Graphics for Data Analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **174**:245–246. doi:https://doi.org/10.1111/j.1467-985X.2010.00676_9.x

Glinsky GV. 2015. Transposable Elements and DNA Methylation Create in Embryonic Stem Cells Human-Specific Regulatory Sequences Associated with Distal Enhancers and Noncoding RNAs. *Genome Biol Evol* **7**:1432–1454. doi:10.1093/gbe/evv081

Göke J, Lu X, Chan Y-S, Ng H-H, Ly L-H, Sachs F, Szczerbinska I. 2015. Dynamic Transcription of Distinct Classes of Endogenous Retroviral Elements Marks Specific Populations of Early Human Embryonic Cells. *Cell Stem Cell* **16**:135–141. doi:10.1016/j.stem.2015.01.005

Goodchild NL, Wilkinson DA, Mager DL. 1993. Recent Evolutionary Expansion of a Subfamily of RTVL-H Human Endogenous Retrovirus-like Elements. *Virology* **196**:778–788. doi:10.1006/viro.1993.1535

Gouy M, Guindon S, Gascuel O. 2010. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Molecular Biology and Evolution* **27**:221–224. doi:10.1093/molbev/msp259

Haig D. 2016. Transposable elements: Self-seekers of the germline, team-players of the soma. *BioEssays* **38**:1158–1166. doi:10.1002/bies.201600125

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**:576–589. doi:10.1016/j.molcel.2010.05.004

Hermant C, Torres-Padilla M-E. 2021. TFs for TEs: the transcription factor repertoire of mammalian transposable elements. *Genes Dev* **35**:22–39. doi:10.1101/gad.344473.120

Imbeault M, Helleboid P-Y, Trono D. 2017. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**:550–554. doi:10.1038/nature21683

Ito J, Sugimoto R, Nakaoka H, Yamada S, Kimura T, Hayano T, Inoue I. 2017. Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *PLOS Genetics* **13**:e1006883. doi:10.1371/journal.pgen.1006883

Izsvák Z, Wang J, Singh M, Mager DL, Hurst LD. 2016. Pluripotency and the endogenous retrovirus HERVH: Conflict or serendipity? *BioEssays* **38**:109–117. doi:10.1002/bies.201500096

Jacobs FM, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, Paten B, Salama SR, Haussler D. 2014. An evolutionary arms race between KRAB zinc finger genes 91/93 and SVA/L1 retrotransposons. *Nature* **516**:242–245. doi:10.1038/nature13760

Jacques P-É, Jeyakani J, Bourque G. 2013. The Majority of Primate-Specific Regulatory Sequences Are Derived from Transposable Elements. *PLOS Genetics* **9**:e1003504. doi:10.1371/journal.pgen.1003504

Jern P, Sperber GO, Ahlsén G, Blomberg J. 2005. Sequence Variability, Gene Structure, and Expression of Full-Length Human Endogenous Retrovirus H. *Journal of Virology* **79**.

Jern P, Sperber GO, Blomberg J. 2004. Definition and variation of human endogenous retrovirus H. *Virology* **327**:93–110. doi:10.1016/j.virol.2004.06.023

Jetzt AE, Yu H, Klarmann GJ, Ron Y, Preston BD, Dougherty JP. 2000. High rate of recombination throughout the human immunodeficiency virus type 1 genome. *J Virol* **74**:1234–1240. doi:10.1128/jvi.74.3.1234-1240.2000

Johnson WE. 2019. Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat Rev Microbiol* **17**:355–370. doi:10.1038/s41579-019-0189-2

Kalkan T, Smith A. 2014. Mapping the route from naive pluripotency to lineage specification. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**:20130540. doi:10.1098/rstb.2013.0540

Kelley D, Rinn J. 2012. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biology* **13**:R107. doi:10.1186/gb-2012-13-11-r107

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler and D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12**:996–1006. doi:10.1101/gr.229102

Khan H, Smit A, Boissinot S. 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res* **16**:78–87. doi:10.1101/gr.4001406

Kinoshita M, Barber M, Mansfield W, Cui Y, Spindlow D, Stirparo GG, Dietmann S, Nichols J, Smith A. 2021. Capture of Mouse and Human Stem Cells with Features of Formative Pluripotency. *Cell Stem Cell* **28**:453-471.e8. doi:10.1016/j.stem.2020.11.005

Kojima KK. 2018. Human transposable elements in Repbase: genomic footprints from fish to humans. *Mobile DNA* **9**:2. doi:10.1186/s13100-017-0107-y

Kong Y, Rose CM, Cass AA, Williams AG, Darwish M, Lianoglou S, Haverty PM, Tong A-J, Blanchette C, Albert ML, Mellman I, Bourgon R, Greally J, Jhunjhunwala S, Chen-Harris H. 2019. Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. *Nat Commun* **10**:5228. doi:10.1038/s41467-019-13035-2

Krönung SK, Beyer U, Chiaramonte ML, Dolfini D, Mantovani R, Dobbelstein M. 2016. LTR12 promoter activation in a broad range of human tumor cells by HDAC inhibition. *Oncotarget* **7**:33484–33497. doi:10.18632/oncotarget.9255

Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, Chan Y-S, Ng H-H, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics* **42**:631–634. doi:10.1038/ng.600

Lai MM. 1992. RNA recombination in animal and plant viruses. *Microbiol Rev* **56**:61–79.

Lanciano S, Cristofari G. 2020. Measuring and interpreting transposable element expression. *Nat Rev Genet* **21**:721–736. doi:10.1038/s41576-020-0251-y

Leigh JW, Bryant D. 2015. popart: full-feature software for haplotype network construction. *Methods in Ecology and Evolution* **6**:1110–1116. doi:https://doi.org/10.1111/2041-210X.12410

Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM, Yang S-P, Wang Z, Chinwalla AT, Minx P, Mitreva M, Cook L, Delehaunty KD, Fronick C, Schmidt H, Fulton LA, Fulton RS, Nelson JO, Magrini V, Pohl C, Graves TA, Markovic C, Cree A, Dinh HH, Hume J, Kovar CL, Fowler GR, Lunter G, Meader S, Heger A, Ponting CP, Marques-Bonet T, Alkan C, Chen L, Cheng Z, Kidd JM, Eichler EE, White S, Searle S, Vilella AJ, Chen Y, Flicek P, Ma J, Raney B, Suh B, Burhans R, Herrero J, Haussler D, Faria R, Fernando O, Darré F, Farré D, Gazave E, Oliva M, Navarro A, Roberto R, Capozzi O, Archidiacono N, Valle GD, Purgato S, Rocchi M, Konkel MK, Walker JA, Ullmer B, Batzer MA, Smit AFA, Hubley R, Casola C, Schrider DR, Hahn MW, Quesada V, Puente XS, Ordoñez GR, López-Otín C, Vinar T, Brejova B, Ratan A, Harris RS, Miller W, Kosiol C, Lawson HA, Taliwal V, Martins AL, Siepel A, RoyChoudhury A, Ma X, Degenhardt J, Bustamante CD, Gutenkunst RN, Mailund T, Dutheil JY, Hobolth A, Schierup MH, Ryder OA, Yoshinaga Y, de Jong PJ, Weinstock GM, Rogers J, Mardis ER, Gibbs RA, Wilson RK. 2011. Comparative and demographic analysis of orang-utan genomes. *Nature* **469**:529–533. doi:10.1038/nature09687

Loewer S, Cabili MN, Guttman M, Loh Y-H, Thomas K, Park IH, Garber M, Curran M, Onder T, Agarwal S, Manos PD, Datta S, Lander ES, Schlaeger TM, Daley GQ, Rinn JL. 2010. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet* **42**:1113–1117. doi:10.1038/ng.710

Löytynoja A, Goldman N. 2010. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics* **11**:579–579. doi:10.1186/1471-2105-11-579

Lu X, Sachs F, Ramsay L, Jacques P-É, Göke J, Bourque G, Ng H-H. 2014. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nature Structural & Molecular Biology* **21**:423–425. doi:10.1038/nsmb.2799

Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, Firth A, Singer O, Trono D, Pfaff SL. 2012. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**:57–63. doi:10.1038/nature11244

Mager DL, Freeman JD. 1995. HERV-H Endogenous Retroviruses: Presence in the New World Branch but Amplification in the Old World Primate Lineage. *Virology* **213**:395–404. doi:10.1006/viro.1995.0012

Mager DL, Freeman JD. 1987. Human endogenous retroviruslike genome with type C pol sequences and gag sequences related to human T-cell lymphotropic viruses. *J Virol* **61**:4060–4066. doi:10.1128/jvi.61.12.4060-4066.1987

Matsuda E, Garfinkel DJ. 2009. Posttranslational interference of Ty1 retrotransposition by antisense RNAs. *Proc Natl Acad Sci U S A* **106**:15657–15662. doi:10.1073/pnas.0908305106

Miao B, Fu S, Lyu C, Gontarz P, Wang T, Zhang B. 2020. Tissue-specific usage of transposable element-derived promoters in mouse development. *Genome Biol* **21**:1–25. doi:10.1186/s13059-020-02164-3

Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ. 2014. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res* **24**:697–707. doi:10.1101/gr.159624.113

Nakamura T, Yamada KD, Tomii K, Katoh K. 2018. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* **34**:2490–2492. doi:10.1093/bioinformatics/bty121

Nichols J, Smith A. 2009. Naive and Primed Pluripotent States. *Cell Stem Cell* **4**:487–492. doi:10.1016/j.stem.2009.05.015

Niwa H. 2007. How is pluripotency determined and maintained? *Development* **134**:635–646. doi:10.1242/dev.02787

Niwa H, Nakamura A, Urata M, Shirae-Kurabayashi M, Kuraku S, Russell S, Ohtsuka S. 2016. The evolutionally-conserved function of group B1 Sox family members confers the unique role of Sox2 in mouse ES cells. *BMC Evolutionary Biology* **16**:173. doi:10.1186/s12862-016-0755-4

Ohnuki M, Tanabe K, Sutou K, Teramoto I, Sawamura Y, Narita M, Nakamura Michiko, Tokunaga Y, Nakamura Masahiro, Watanabe A, Yamanaka S, Takahashi K. 2014. Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *PNAS* **111**:12426–12431. doi:10.1073/pnas.1413299111

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**:417–419. doi:10.1038/nmeth.4197

Peaston AE, Evsikov AV, Graber JH, Vries WN de, Holbrook AE, Solter D, Knowles BB. 2004. Retrotransposons Regulate Host Genes in Mouse Oocytes and Preimplantation Embryos. *Developmental Cell* **7**:597–606. doi:10.1016/j.devcel.2004.09.004

Pérez-Losada M, Arenas M, Galán JC, Palero F, González-Candelas F. 2015. Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences. *Infect Genet Evol* **30**:296–307. doi:10.1016/j.meegid.2014.12.022

Pontis J, Planet E, Offner S, Turelli P, Duc J, Coudray A, Theunissen TW, Jaenisch R, Trono D. 2019. Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. *Cell Stem Cell* **24**:724-735.e5. doi:10.1016/j.stem.2019.03.012

Posada D, Crandall KA. 2001. Intraspecific gene genealogies: trees grafting into networks. *Trends in Ecology & Evolution* **16**:37–45. doi:10.1016/S0169-5347(00)02026-7

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841–842. doi:10.1093/bioinformatics/btq033

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* **44**:W160–W165. doi:10.1093/nar/gkw257

Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D, Kent WJ. 2014. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* **30**:1003–1005. doi:10.1093/bioinformatics/btt637

Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu Rev Genet* **46**:21–42. doi:10.1146/annurev-genet-110711-155621

Römer C, Singh M, Hurst LD, Izsvák Z. 2017. How to tame an endogenous retrovirus: HERVH and the evolution of human pluripotency. *Current Opinion in Virology*, Animal models for viral diseases • Paleovirology **25**:49–58. doi:10.1016/j.coviro.2017.07.001

Rossant J, Tam PPL. 2017. New Insights into Early Human Development: Lessons for Stem Cell Derivation and Differentiation. *Cell Stem Cell* **20**:18–28. doi:10.1016/j.stem.2016.12.004

Santoni FA, Guerra J, Luban J. 2012. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology* **9**:111. doi:10.1186/1742-4690-9-111

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, McCarthy S, Montgomery SH, Schwalie PC, Tang YA, Ward MC, Xue Y, Yngvadottir B, Alkan C, Andersen LN, Ayub Q, Ball EV, Beal K, Bradley BJ, Chen Y, Clee CM, Fitzgerald S, Graves TA, Gu Y, Heath P, Heger A, Karakoc E, Kolb-Kokocinski A, Laird GK, Lunter G, Meader S, Mort M, Mullikin JC, Munch K, O'Connor TD, Phillips AD, Prado-Martinez J, Rogers AS, Sajjadian S, Schmidt D, Shaw K, Simpson JT, Stenson PD, Turner DJ, Vigilant L, Vilella AJ, Whitener W, Zhu B, Cooper DN, de Jong P, Dermitzakis ET, Eichler EE, Flicek P, Goldman N, Mundy NI, Ning Z, Odom DT, Ponting CP, Quail MA, Ryder OA, Searle SM, Warren WC, Wilson RK, Schierup MH, Rogers J, Tyler-Smith C, Durbin R. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**:169–175. doi:10.1038/nature10842

Schön U, Seifarth W, Baust C, Hohenadl C, Erfle V, Leib-Mösch C. 2001. Cell Type-Specific Expression and Promoter Activity of Human Endogenous Retroviral Long Terminal Repeats. *Virology* **279**:280–291. doi:10.1006/viro.2000.0712

Simon-Loriere E, Holmes EC. 2011. Why do RNA viruses recombine? *Nat Rev Microbiol* **9**:617–626. doi:10.1038/nrmicro2614

Smit AF, Hubley R, Green P. 2013. RepeatMasker Open-4.0.

Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA* **12**:2. doi:10.1186/s13100-020-00230-y

Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. 2014. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res* **24**:1963–1976. doi:10.1101/gr.168872.113

Sundaram V, Wysocka J. 2020. Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences* **375**:20190347. doi:10.1098/rstb.2019.0347

Svoboda P, Stein P, Anger M, Bernstein E, Hannon GJ, Schultz RM. 2004. RNAi and expression of retrotransposons MuERV-L and IAP in preimplantation mouse embryos. *Developmental Biology* **269**:276–285. doi:10.1016/j.ydbio.2004.01.028

Takahashi K, Nakamura M, Okubo C, Kliesmete Z, Ohnuki M, Narita M, Watanabe A, Ueda M, Takashima Y, Hellmann I, Yamanaka S. 2021. The pluripotent stem cell-specific transcript ESRG is dispensable for human pluripotency. *PLOS Genetics* **17**:e1009587. doi:10.1371/journal.pgen.1009587

Tang F, Barbacioru C, Bao S, Lee C, Nordman E, Wang X, Lao K, Surani MA. 2010. Tracing the Derivation of Embryonic Stem Cells from the Inner Cell Mass by Single-Cell RNA-Seq Analysis. *Cell Stem Cell* **6**:468–478. doi:10.1016/j.stem.2010.03.015

Theunissen TW, Friedli M, He Y, Planet E, O'Neil RC, Markoulaki S, Pontis J, Wang H, Iouranova A, Imbeault M, Duc J, Cohen MA, Wert KJ, Castanon R, Zhang Z, Huang Y, Nery JR, Drotar J, Lungjangwa T, Trono D, Ecker JR, Jaenisch R. 2016. Molecular Criteria for Defining the Naive Human Pluripotent State. *Cell Stem Cell* **19**:502–515. doi:10.1016/j.stem.2016.06.011

Thomas J, Perron H, Feschotte C. 2018. Variation in proviral content among human genomes mediated by LTR recombination. *Mobile DNA* **9**:36. doi:10.1186/s13100-018-0142-3

Thompson PJ, Macfarlan TS, Lorincz MC. 2016. Long Terminal Repeats: From Parasitic Elements to Building Blocks of the Transcriptional Regulatory Repertoire. *Molecular Cell* **62**:766–776. doi:10.1016/j.molcel.2016.03.029

Urusov FA, Nefedova LN, Kim AI. 2011. Analysis of the tissue- and stage-specific transportation of the Drosophila melanogaster gypsy retrotransposon. *Russ J Genet Appl Res* **1**:507–510. doi:10.1134/S2079059711060104

Vargiu L, Rodriguez-Tomé P, Sperber GO, Cadeddu M, Grandi N, Blikstad V, Tramontano E, Blomberg J. 2016. Classification and characterization of human endogenous retroviruses; mosaic forms are common. *Retrovirology* **13**:7. doi:10.1186/s12977-015-0232-y

Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, Cai H, Besser D, Prigione A, Fuchs NV, Schumann GG, Chen W, Lorincz MC, Ivics Z, Hurst LD, Izsvák Z. 2014. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* **516**:405–409. doi:10.1038/nature13804

Wang Z, Oron E, Nelson B, Razis S, Ivanova N. 2012. Distinct lineage specification roles for NANOG, OCT4, and SOX2 in human embryonic stem cells. *Cell Stem Cell* **10**:440–454. doi:10.1016/j.stem.2012.02.016

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**:1189–1191. doi:10.1093/bioinformatics/btp033

Waterson RH, Lander ES, Wilson RK, The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**:69–87. doi:10.1038/nature04072

Wolf G, de Iaco A, Sun M-A, Bruno M, Tinkham M, Hoang D, Mitra A, Ralls S, Trono D, Macfarlan TS. 2020. KRAB-zinc finger protein gene expansion in response to active retrotransposons in the murine lineage. *eLife* **9**:e56337. doi:10.7554/eLife.56337

Yang P, Wang Y, Macfarlan TS. 2017. The role of KRAB-ZFPs in transposable element repression and mammalian evolution. *Trends Genet* **33**:871–881. doi:10.1016/j.tig.2017.08.006

Yu H-L, Zhao Z-K, Zhu F. 2013. The role of human endogenous retroviral long terminal repeat sequences in human cancer (Review). *International Journal of Molecular Medicine* **32**:755–762. doi:10.3892/ijmm.2013.1460

Zhang Y, Li T, Preissl S, Amaral ML, Grinstein JD, Farah EN, Destici E, Qiu Y, Hu R, Lee AY, Chee S, Ma K, Ye Z, Zhu Q, Huang H, Fang R, Yu L, Izpisua Belmonte JC, Wu J, Evans SM, Chi NC, Ren B. 2019. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nature Genetics* **51**:1380–1388. doi:10.1038/s41588-019-0479-7

CHAPTER 3

LTR RECOMBINATION ACTS AS A CIS-REGULATORY SWITCH FOR HERVH TRANSCRIPTION IN PLURIPOTENT STEM CELLS[2]

3.1 ABSTRACT

The long terminal repeats (LTRs) of endogenous retroviruses (ERVs) are potent cis-regulatory units. Full-length ERV have 5' and 3' flanking LTRs which function as the promoter and transcript termination sites, respectively. Due to their identical sequence, the 5' and 3' LTR can recombine, where they lose one LTR and their internal portion and exist as a solitary (solo) morph. For most ERV families in the human genome, about 90% of insertions exist as the solo morph and only 10% remain full-length. The human endogenous retrovirus type-H (HERVH) family is an exception, with 50% of insertions existing as solo LTRs. The LTR7up subfamily of HERVH is even more remarkable. More than 70% of insertions exist in the full-length morph and more than 33% of loci produce mature RNA in human embryonic stem cells (ESC). Some of these transcripts seem to have roles regulating ESC stemness and differentiation. The coincidence of robust ESC transcription with an abnormally high proportion of full-length insertions has led to speculation that full-length HERVH might be selectively maintained for its transcription and/or expression. If this is true, full-length HERVH may have a greater capacity for transcription than solo HERVH. I formally test this by comparing transcription, expression, transcription factor binding, and histone

[2] This chapter is under preparation for publication and will be submitted to bioRxiv as "Thomas A. Carter and Cédric Feschotte (2021) LTR recombination acts as a cis-regulatory switch for HERVH transcription in pluripotent stem cells". The author contributions are as follows: Carter TA designed and performed all experiments, created all figures, and prepared the manuscript. Feschotte C aided in experimental design, project definition, and manuscript preparation.

modification profiles on solo and full-length HERVH/LTR7up. I find that only full-length 7up exhibit robust promoter activity, but solo loci retain hallmarks of enhancers. Furthermore, I show that HERVH recombination is a recurrent and ongoing feature of great ape evolution. I propose the HERVH recombination may constitute a promoter to enhancer cis-regulatory switch, which may lead to the modification of the pluripotent regulatory network between species and individuals.

3.2 MAIN

Endogenous retroviruses (ERVs) are derived from exogenous retroviruses with which they share the same structure: an internal (int) coding region flanked by two regulatory units termed long terminal repeats (LTRs) (Eickbush and Malik, 2002). Upon insertion, both LTRs are identical (Coffin et al., 1997), yet they serve different roles. The 5' LTR acts a promoter, while the 3' LTR provides a poly-A signal for viral transcripts (Eickbush and Malik, 2002; Nelson et al., 1996). Whether endogenous or exogenous, a retrovirus's LTR composition must attract the right combination of host transcription factors (TFs) to reproduce. The necessity to bind numerous TF has led to many ERV subfamilies being enriched in specific regulatory contexts (Babaian and Mager, 2016; Glinsky, 2015; Kong et al., 2019; Yu et al., 2013), indicating that they often act as potent promoters.

In human, most ERVs do not exist in the proviral, or "full-length", morph described above (Gemmell et al., 2016). They frequently undergo LTR-LTR recombination, whereby they lost their internal portion and one of their LTRs, existing in the solitary, or "solo", morph. Solo insertions are not replication competent, as they do not have the genes necessary for transposition. But their LTR is the same, presumably still capable of binding host TF and acting as cis-regulatory elements (CREs).

Despite its similar sequence, LTR-LTR recombination has been seen to alter gene expression and host phenotypes. Tarocco and Maro (blood) oranges are derived from Navalinas, where an LTR insertion upstream of a pigmentation gene and its subsequent recombination increase the amount of pigment produced in the flesh of the fruit (Butelli et al., 2012; Lisch, 2013). The same is true of the chardonnay and Okuyama varieties of grape (Cadle-Davidson and Owens, 2008; Kobayashi et al., 2004; Lisch, 2013; Shimazaki et al., 2011). In humans, an LTR insertion upstream of the amylase gene gave rise to salivary amylase expression (Meisler and Ting, 1993; Samuelson et al., 1996, 1990, 1988). Its subsequent recombination reverted one of this gene's copies to pancreatic expression (Samuelson et al., 1996).

LTR-LTR recombination seems to be the most prevalent destination for ERV. Within a few million years of fixation, 90% of ERV insertions will recombine to the solo morph, with only 10% remaining full-length (Gemmell et al., 2016). This is with the exception of human endogenous retrovirus type-H (HERVH). HERVH has evaded recombination with great success. Only 50% of all HERVH exist in the solo morph (Belshaw et al., 2007; Gemmell et al., 2016). Among its youngest and most transcriptionally active subfamilies (LTR7up), fully 70% of its insertions exist in the full-length morph (Carter et al., 2021).

In addition to its curious solo:full ratio, HERVH is famous for having colonized much of the preimplantation embryo (Carter et al., 2021; Göke et al., 2015) and its high expression in human embryonic stem cells (ESCs). Numerous studies have shown that family-wide and locus-specific knockout or knockdown of HERVH RNA results in the loss of pluripotency (Fort et al., 2014; Gemmell et al., 2015; Izsvák et al., 2016;

Kelley and Rinn, 2012; Loewer et al., 2010; Römer et al., 2017; Santoni et al., 2012), and transcriptionally active HERVH have been observed providing topologically associated domains (TAD) in pluripotent stem cells that may alter the phenotypes of daughter cells (Zhang et al., 2019). Combining both its unique propensity to exist in the full-length morph and its potential functions in pluripotent stem cells has led some to postulate that the preservation of HERVH may be due to selection for roles in early embryogenesis (Gemmell et al., 2016; Izsvák et al., 2016). Indeed, most ERV have similar transcript structures, all of which involve genes from the internal region (Wilkinson et al., 1990). Without transcription, HERVH loci cannot create TADs or lncRNAs. To my knowledge, no systematic comparison of the regulatory capacities of HERVH morphs has been performed. I hypothesize that solo and full-length HERVH are differentially regulated, which may have led to the preponderance of full-length HERVH through purifying selection.

3.2.1 PROVIRAL HERVH ARE MORE LIKELY TO BE TRANSCRIBED AND PRODUCE MATURE RNA THAN SOLO HERVH

Under this model, full-length HERVH would contribute more to HERVH-derived transcripts than solo HERVH. To test this, I compared the relative contribution of solo and full-length LTR from 7up and 7d insertions to “highly transcribed” mature RNA (Wang et al., 2014) (FPKM >2) in H1 cells. 7up insertions were previously described as having robust H1 promoter activity, while 7d did not (Carter et al., 2021), so I expected solo and full 7up to equally contribute to mature RNA while 7d would contribute less. I found that while 7d contributed less to RNA, the difference between both subfamilies’ full and solo insertion was even greater **(Fig. 3.2.1a)**. Neither 7up nor 7d solo loci contributed to mature RNA whatsoever.

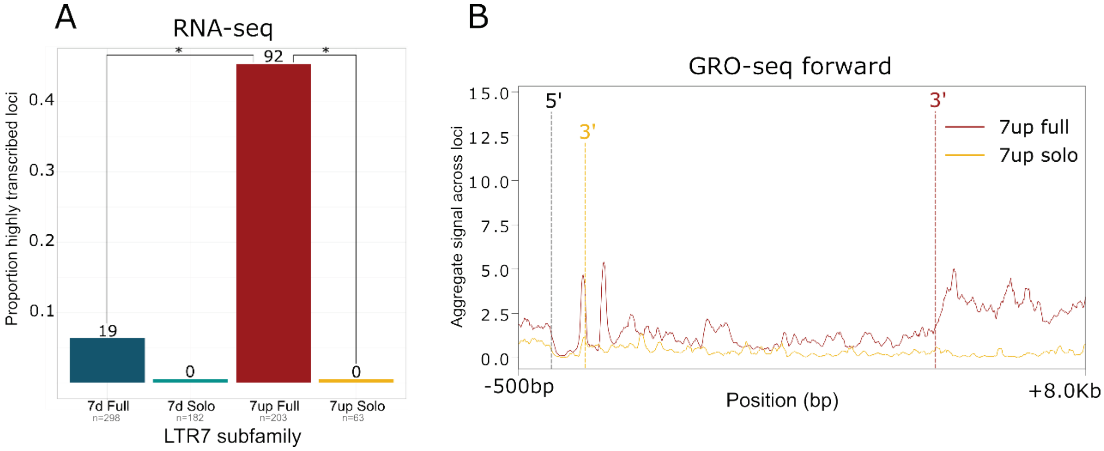


Figure 3.2.1: Full-length, and not solo, 7up loci exhibit transcription in human ESCs.
(a) Bar chart showing the contribution of full-length and solo LTR7 subfamilies 7up and 7d to mature RNA. Groups with different proportions of contribution are denoted with * ($p<0.05$ Chi-square with Bonferroni correction). **(b)** Aggregate signal plot of all solo and full-length 7up loci. The borders of solo (gold) and full-length (red) are shown. Normalized signal within the 5' (full-length) and solo LTRs was significantly different ($p<0.05$ Wilcox rank-sum test).

Because full and solo LTRs of a given subfamily are identical (Carter et al., 2021) and presumably harbor the same TFBS, I hypothesized that both solo and full-length loci had the same transcriptional potential and the difference in mature RNA was the result of differential transcript stability. To test this, I plotted aggregate GRO-seq signal (a direct measure of nascent RNA) from solo and full-length 7up insertions **(Fig. 3.2.1b)**. I found that full-length 7up loci exhibited greater GRO-seq signal at the 5'/solo LTR ($p<0.05$ Wilcox rank-sum test). This transcriptional difference was even greater in genomic DNA past the 3' end of the solo LTR or past the 3' end of the 3' LTR of full-length insertions. These data indicate that full-length 7up, and not solo 7up, exhibit stable transcription originating from the 5' LTR throughout the internal region and 3' LTR and into genomic DNA. This transcriptional difference may fully or partially account for differential contributions to mature RNA.

3.2.2 SOLO AND FULL-LENGTH HERVH HAVE DIFFERENTIAL PROMOTER ACTIVITY, SILENCING, AND TF-BINDING PROFILES

Solo-full differential transcription may be explained by differential silencing, TF binding/promoter activity, local CRE, or elongation stability. To differentiate between these potential mechanisms, I first called peaks for an array of locus specific H1 regulatory data and layered them onto solo and full-length 7up and 7d, calling each insertion as positive or negative for a given feature. Some features such as H3K4me3 and contribution to RNA, seemed exclusive to full-length 7up **(Fig. 3.2.2a)**. Others like NANOG, OCT4, and KLF4 seemed to be shared between all 7up, but absent from 7d. Still others like ZNF534 and H3K4me1 seemed to be shared among 7up and 7d full-length but absent in all solos. To test which of these factors were shared between subfamilies and/or morphs, I tested each 7up full vs. 7d full and 7up full vs. 7up solo using a chi-square test with Bonferroni correction (all mentioned differences hereafter were statistically significant $p<0.05$). I found that most pluripotent TF were enriched in 7up and depleted in 7d **(Fig. 3.2.2b)**. Considering both full-length and solo 7up share the same 5' LTR, and thus the same promoter, equivalent TF recruitment was expected. The repressive histone mark H3K9me3 was also depleted in both morphs of 7up, suggesting that the upregulated LTR7 were less silenced than their downregulated counterparts. None of these marks were different between 7up full-length and 7up solo loci.

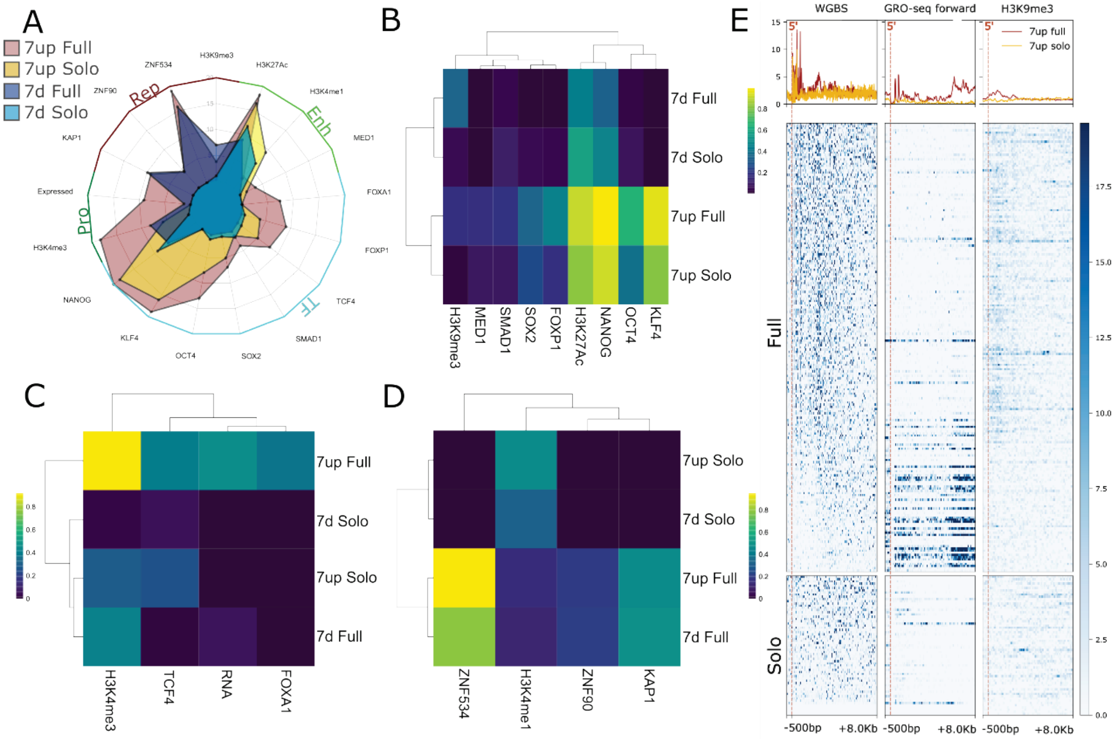


Figure 3.2.2: Full-length and solo 7up have different epigenetic profiles. (a) Spider plot showing the general epigenetic profiles of full-length and solo 7up and 7d. All data is from the H1 cells, save for ZNF534 and ZNF90, which hare from HEK293. Marks are sorted into promoter (PRO), repressed (REP), enhancer (ENH), and transcription factor (TF) groups. **(b)** Heatmap of tested marks significantly different between 7up and 7d loci (p<0.05 Chi-square with Bonferroni correction). **(c)** Heatmap of tested marks significantly different between solo (7up/d) and full-length (7up/7d) loci. **(d)** Heatmap of tested marks enriched at only 7up full-length loci. **(e)** Heatmap (bottom) and aggregate signal (top) of whole genome bi-sulfite sequencing (WGBS), GRO-seq, and H3K9me3 at solo and full-length 7up loci. Differences in aggregate signal intensity determined by Wilcox rank-sum test (p<0.05) at the 5' and solo LTR only.

As expected from Fig. 3.2.1, 7up full-length loci were enriched for contribution to mature RNA **(Fig. 3.2.2c)**. Congruent with this earlier expectation, 7up full-length insertions were also enriched for the promoter mark H3K4me3. Unexpectedly, the pluripotent transcription factors FOXA1 and TCF4 were also enriched in 7up full-length loci. This contrasts with the other pluripotent TFs which were enriched in both 7up morphs, potentially indicating different TF recruitment mechanisms.

The enhancer mark H3K4me1, the potential 7up repressors ZNF534 and ZNF90 (HEK293 cells) (Imbeault et al., 2017), and the silencing complex KAP1 were enriched in both 7d and 7up full-lengths and depleted in all solos **(Fig. 3.2.2c)**. Remarkably, no solo insertion was positive for ZNF534, ZNF90, or KAP1 binding. This suggests that the internal portion of full-length is required for the targeting and silencing of LTR7/HERVH in embryonic stem cells and the loss of this region may free a locus from such silencing.

The co-occurrence of potentially repressive ZNFs and KAP1, along with H3K4me3 and RNA production at full-length 7up presented a paradox. How were these loci both more repressed and more transcribed? To test if repressed loci were being transcribed, I generated a heatmap displaying all reads from WGBS sequencing (a measure of DNA methylation) with forward strand GRO-seq, and H3K9me3 at full-length and solo 7up loci in an 8.5kb window **(Fig. 3.2.2e)**. DNA methylation was more prevalent at full-length 7up than solo 7up and bifurcated with GRO-seq signal (loci with most GRO-seq signal had the least WGBS signal). And despite not showing any difference in the proportion of H3K9me3 positive loci, 7up full-length insertions had more H3k9me3 reads than their solo counterparts ($p<0.05$ Wilcox rank-sum with Bonferroni correction). Like WGBS signal, H3K9me3 reads were confined to those full-length insertions with the least GRO-seq signal. These data indicate that full-length 7up, but not solo, insertions are marked with H3K9me3 and DNA methylation. However, these repressive marks are not present at transcribed loci.

Together, these data indicate that some full-length 7up, and not solo 7up, exhibit many hallmarks of robust activity, including the prevalence of H3K4me3, transcription, and mature RNA production. However, non-transcribed full-length insertions seem to be silenced, a trait not shared with solo insertions. And while solo loci do not have strong promoter activity, they seem to share some hallmarks of enhancer activity, such as a prevalence of H3K4me1 and H3K27ac, and seem to have altogether escaped the silencing marks that coincide with full-length insertions. Together, these data suggest that solo insertions may act as enhancers through maintaining TF recruitment and evading host silencing mechanisms.

3.2.3 SOLO LTR7UP DO NOT HAVE ROBUST PROMOTER ACTIVITY BUT MAY ACT AS ENHANCERS

While LTR-LTR recombination seems to ablate promoter activity, solo HERVH show signatures of enhancer activity **(Figure 3.2.2)**, suggesting that LTR-LTR recombination may constitute a promoter to enhancer cis-regulatory switch. To test this, I used dREG (Danko et al., 2015; Wang et al., 2019) to find signatures of divergent transcription seen at promoters and enhancers. As its GRO-seq signal suggested **(Figure 3.2.1b)**, the 5' LTRs of full-length 7up exhibited a strong promoter/enhancer signature **(Figure 3.2.3a)**. Despite solo 7up showing no evidence of transcription **(Figure 3.2.1b)**, they exhibited a promoter/enhancer signature, albeit a weaker one than 7up 5' LTRs ($p < .00001$ Wilcox rank-sum test) **(Figure 3.2.3a)**. Full-length 7up also showed evidence of additional promoters/enhancers in the

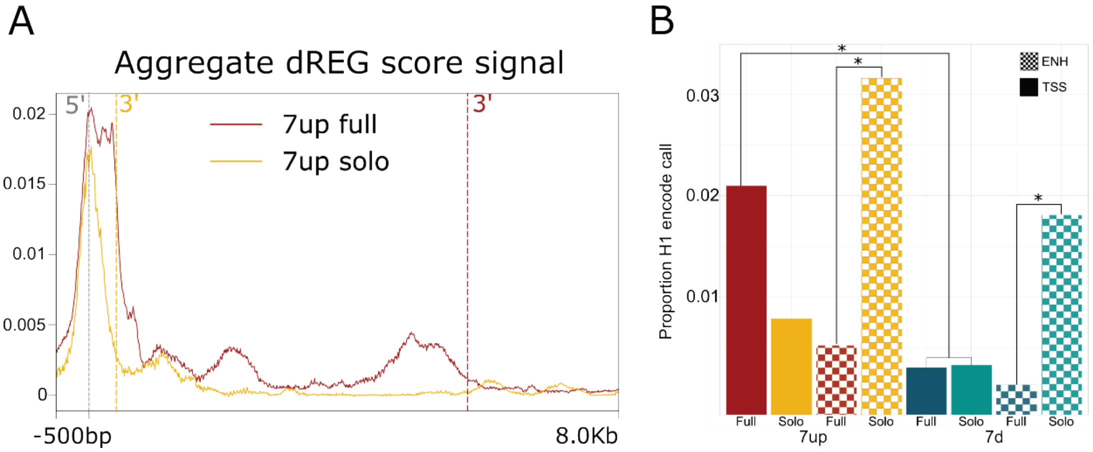


Figure 3.2.3: Solo 7up may function as enhancers.
(a) Aggregate dREG signal from 7up solo and full-length insertions. Differences in aggregate signal intensity determined by Wilcox rank-sum test (p<0.05) at the 5' and solo LTR only. **(b)** Bar chart showing the proportion of solo and full-length 7up and 7d loci were positive for ENCODE H1 ChromHMM 'TSS' and 'enhancer' calls (see methods). Those with different proportions are marked with * (p<0.05 Chi-square with Bonferroni correction).

internal region and at the 3' LTR. LTR7 transcription has been previously observed to originate within the 5' LTR (Nelson et al., 1996; Wilkinson et al., 1990), suggesting that these sites probably function as enhancers.

To help differentiate between promoter and enhancer activity, I layered ChromHMM calls in H1 cells from ENCODE (Ernst et al., 2011; Ernst and Kellis, 2010) onto full-length and solo 7up and 7d. These calls utilize a great number of epigenetic genomic data run through a machine learning algorithm and divide the genome into regions of TSS, enhancer, weak enhancer, transcribed, repressed, etc. Unsurprisingly, full-length 7up overlapped with "TSS" calls more than solo and full-length 7d **(Figure 3.2.3b)**. While lacking clear promoter activity, solos of both 7up and 7d overlapped with "strong enhancer" calls more than their full-length counterparts. Together, these data suggest that LTR-LTR recombination may constitute a promoter to enhancer cis-regulatory switch, with solo LTRs lacking robust transcription but retaining divergent transcription and the binding of most TFs.

3.2.4 HERVH LTR-LTR RECOMBINATION HAS CONTRIBUTED TO LINEAGE-SPECIFIC DNA DIVERGENCE

Our data thus far suggests that HERVH LTR-LTR recombination may ablate 7up transcription, retain enhancer activity, and evade host silencing. But what does this mean for primate evolution? Some combination of mature 7up-derived RNAs seem to be indispensable for pluripotency (Fort et al., 2014; Gemmell et al., 2015; Izsvák et al., 2016; Kelley and Rinn, 2012; Loewer et al., 2010; Römer et al., 2017; Santoni et al., 2012). Additionally, transcribed loci have been observed to create and maintain TAD boundaries, altering proximal gene expression (Zhang et al., 2019). Considering these previous experiments and the data presented above, 7up solo formation may have altered the pluripotent regulatory network. To ascertain the breadth HERVH LTR-LTR recombination may have had on great ape evolution, I surveyed annotated LTR7 insertions in the reference chimpanzee, gorilla, and orangutan genomes. All annotated LTR7 were called as full-length or solo, as was done for the human genome. Next, I used UCSC *liftover* to find orthologous insertions between all great apes and discarded any insertion that was not present in all species. These insertions were then surveyed for LTR-LTR recombination events and I used the principle of parsimony to find the most likely evolutionary timepoint where a recombination event occurred. For example, if an orthologous insertion was ‘solo’ in the human and chimpanzee reference genomes but ‘full-length’ in the gorilla and orangutan genomes, this recombination event most likely occurred after the split of the gorilla and human-chimp lineages but before the split of the pan and homo lineages. Because I did not consider unshared insertions, and parsimony may count multiple events as one if they are congruent with evolution by common descent (in the example above, the human and chimpanzee lineages could have independently undergone the same recombination event), this inter-specific analysis is an underestimate of all the fixed recombination events in the great apes. Even so, I find evidence of 106 fixed LTR7-LTR7 recombination events across 15 million years of great ape evolution **(Figure 3.2.4)**. The occurrence and fixation of these events seems to have occurred at a roughly even and linear rate, with the greatest density of events (events/mya) at the divergence of the African and Asian great ape lineages (4.78 events/mya) and the least density of events occurring in the pan lineage (1.8 events/mya).

LTR7 has been transpositionally dead for approximately 5 million years (Mager and Freeman, 1995). Recent research has indicated that the rate of LTR7-LTR7 recombination has slowed dramatically in recent evolutionary time (Gemmell et al., 2016). However, this research was restricted to recombination events fixed in the human genome. To compare the rate of LTR7-LTR7 recombination within humans to the inter-specific rate, I drew upon recent work (Thomas et al., 2018) which found 69 examples of ‘dimorphic’ solo/full-length HERVH alleles within the Simons Genome Diversity Project (SGDP) **(Figure 3.2.4)**. All these examples were at low allele frequency, indicating that they arose recently. While it is unlikely that any of these rare alleles will reach fixation, these data show that LTR7-LTR7 recombination is an ongoing process, perhaps contributing rare diseases or phenotypes.

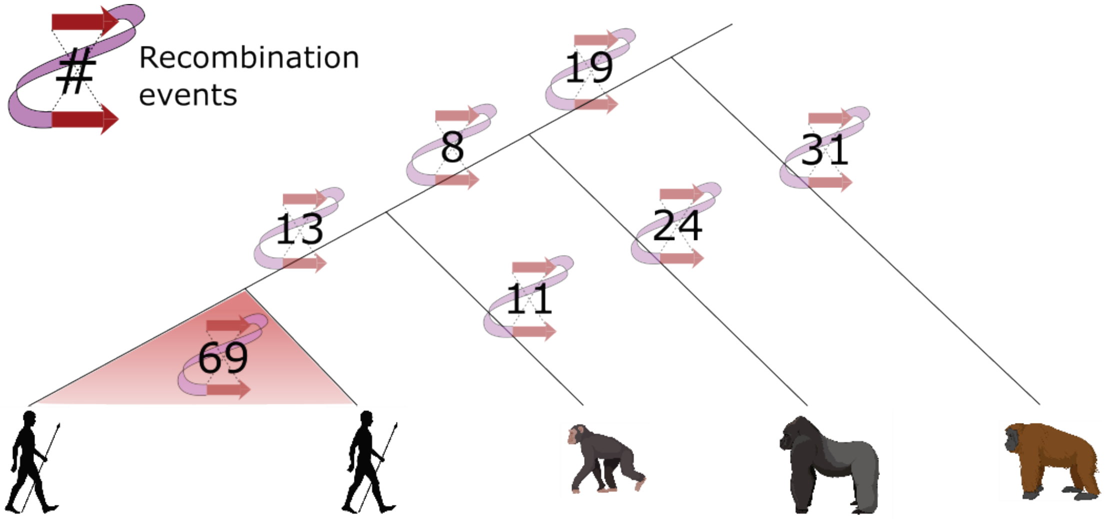


Figure 3.2.4: LTR7-LTR7 recombination is a recurrent feature of great ape evolution.
Fixed orthologous LTR7 recombination events during great ape evolution are shown on appropriate branches. Those in the red triangle are recent events from within the human population (SGDP).

3.3 DISCUSSION

Previous experiments have indicated HERVH/LTR7 derived and transcribed HERVH/LTR7 (Zhang et al., 2019) functioning in the maintenance of human pluripotency and subsequent differentiation. Additionally, recent data shows that HERVH is unique among HERV in its ability to avoid fixation of LTR-LTR recombination and retain full-length among most of its insertions (Gemmell et al., 2016). Here, for the first time to my knowledge, I demonstrate that solo and full-length HERVH have different regulatory capacities. Only full-length 7up are transcribed and contribute to mature RNA, while solo 7up appear to avoid host silencing and may act as enhancers. Also, recombination events within great ape lineages have generated dozens of species-specific solo morphs, perhaps destroying pluripotency regulating RNA or creating pluripotent enhancers.

The HERVH family has been intensively studied for its regulatory capacity and contribution to lncRNA in human pluripotent cells (Kelley and Rinn, 2012). Perturbation of HERVH loci or transcripts suggest that individual or multiple loci promote pluripotency (Fort et al., 2014; Gemmell et al., 2015; Izsvák et al., 2016; Kelley and Rinn, 2012; Loewer et al., 2010; Römer et al., 2017; Santoni et al., 2012). This has led to speculation that HERVH may have become indispensable to great ape pluripotency or provided some selective advantage (Gemmell et al., 2016; Izsvák et al., 2016; Römer et al., 2017). HERVH's pluripotent transcription is not its only remarkable feature. Most HERV families, including the younger HERVK, exist predominately (90%) in the solo morph, while merely 50% of HERVH are solo. My previous work (Carter et al., 2021) shows that transcribed HERVH belong to a monophyletic subfamily (LTR7up). Among 7up, full-length HERVH is even more conserved, with merely 33% existing as the solo morph. By comparing the promoter activity and regulatory profiles of full-length and solo HERVH among highly similar elements (7up), I have uncovered that these morphs exhibit strikingly different regulatory signatures. This gives further credence to the idea that HERVH may be "unusually" conserved due the inability of solo morphs to contribute to the lncRNA and/or TAD boundaries that may be essential for human pluripotency. Alternatively, this pattern may reflect selection against the solo-morph, which could deleteriously alter proximal gene expression via their evasion of silencing and enhancer potential. Previous work uncovered that the formation of new solo alleles is an ongoing process in humans (Thomas et al., 2018). My work adds to this by showing LTR-LTR recombination has contributed to inter-specific DNA (and likely regulatory) divergence as well. If full-length HERVH truly is necessary for human pluripotency, recombination may alter development or fertility.

However, non-selection mechanisms for HERVH's unusual full-length abundance exist. Both previous annotations and a more recent subdivision of LTR7 (Carter et al., 2021) show as many as 30% of full-length LTR having mis-matched 5' and 3' LTRs (e.g., 5' 7b with 3' 7y). Its possible that HERVH's multiple concurrently active lineages led to a propensity for ectopic recombination between loci, generating mismatched 5'/3' LTRs, which seem to inhibit LTR-LTR recombination(Gemmell et al., 2016; Jordan and McDonald, 1999). However, this 30% is not enough to fully explain the solo:full-length pattern observed in some LTR7 species, leaving the necessity of selection or unknown mechanisms of evading recombination to fully explain this phenomenon.

The transition of full-length to solo coinciding with changes in regulatory signature, especially transcription, presents an opportunity to study the mechanisms governing polymerase loading, pausing, and elongation. The subfamily 7up provides hundreds of closely related loci with strikingly different regulatory signatures **(Figure 3.2.2)**. Though I do not fully elucidate the core processes governing the promoter to enhancer switch that coincide with recombination, these data present some clues as to what modules are and are not essential for TF binding, bidirectional transcription, CRE formation, polymerase pausing, polymerase elongation, and post-transcriptional regulation. While solo 7up maintain bidirectional transcription **(Figure 3.2.3)**, no solo locus shows obvious plus strand GRO-seq signal **(Figure 3.2.1)**. This indicates that most POLII loading can occur at solo loci, but this is insufficient to promote robust POLII elongation. Portions in the HERVH-int or 3' LTR must be necessary for transcription and expression. My data bring up two non-mutually exclusive possibilities **(Figure 3.3.1a)**: 1) The internal and/or 3' LTR provide transcript

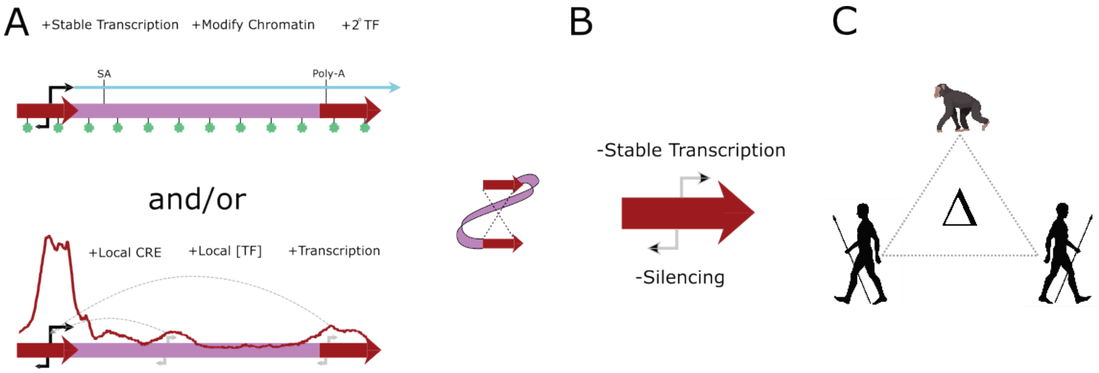


Figure 3.3.1: Multiple possible roles for HERVH-int in LTR7 regulation with consequences for host evolution.
(a) Two compatible models of HERVH-int contribution to LTR7 regulation. In the first, -int provides stable transcription and transcripts via a splice acceptor (SA) and a poly-A site. Increased transcription leads to local chromatin modification and the addition of secondary TFs like FOXA1 and TCF4 **(Fig. 3.2.2c)**. In the second, the -int and 3' LTR provide addition CRE **(Figure 3.2.3a)**, which increase the local TF concentration and lead to more transcription. (b) Either way, when the -int and 3' LTR are lost, so are productive transcription and potential silencing marks (**Figure 3.2.1, 3.2.2**). (c) When these events occur within or between species (**Figure 3.2.4**), changes to the pluripotent regulatory can occur via changes in lncRNA production, enhancer capacity, and TAD boundaries.

stabilization and encourage more transcription by the removal of repressive marks and the deposition of pro-transcriptional marks. 2) The HERVH-int and 3' LTR provide additional CREs which promote transcription from the 5' LTR via an increase in local TF concentration. The lack of CpG methylation and H3K9me3 present at transcribed full-length loci **(Figure 3.2.1e)** support model 1, as do the presence of splice acceptor sites in the internal region (Wilkinson et al., 1990) and the presence of FOXA1 and TCF4 exclusively at full-length HERVH. The presence of dREG peaks **(Figure 3.2.3a)** in the HERVH-int and 3' LTR support that these modules may regulate transcription from the 5' LTR. ChIA-PET in hESC shows that a handful of 5'-3' interactions occur **(data not shown)**, but the poor mappability of this technique render it unable to determine if this occurs at a larger scale. Future studies should seek to better understand the internal and 3' contribution to HERVH transcription. These studies can take advantage of the structural diversity of HERVH insertions (large deletions of internal DNA are relatively common) to find what mutations disallow transcription. Alternatively, and more elegantly, full-length and solo loci can be engineered to take on the opposing morph (or anything in between). Changes in the local regulatory environment should be able to deduce exactly what and where crucial transcriptional regulators are. These experiments can also note changes in adjacent gene expression to see if solo LTR function as enhancers.

No matter the mechanism, the loss of internal and 3' sequence oblate an insertion's promoter capacity while simultaneously evading host silencing **(Figure 3.3.1b)**. When this occurs within or between species, this may lead to regulatory divergence **(Figure 3.3.1c)**. Noting exactly which lncRNA have been destroyed via LTR-LTR recombination may help to clarify the role HERVH has had in pluripotent regulation.

3.4 METHODS

3.4.1 LTR7, LTR7up, and LTR7d sequence identification and solo-full separation

LTR7 insertions for LTR7up and LTR7d were extracted from (Carter et al., 2021) where 5' and solo LTRs for all LTR7 had been previously extracted from the RepeatMasker version 4.0.5 repeat Library 20140131 (Smit et al., 2013) masking of GRCh38/hg38 with alt chromosomes removed. Solo and 5' LTRs were separated with OneCodeToFindThemAll.pl (Bailly-Bechet et al., 2014) followed by rename_mergedLTRelements.pl (Thomas et al., 2018).

3.4.2 Human embryonic stem cell RNA-seq and GRO-seq analysis

In order to quantify the transcription of 5' and solo LTR7up, GRO-seq data from H1 cells (Estaras et al.) (GSE64758) was plotted onto windows 500bp upsteam and 8kb downstream from the 5' most point of the LTR using deeptools (Ramírez et al., 2016). Plus strand reads were summed and averaged for these windows for both 5' (full-length) and solo 7up loci using the deeptools plotProfile --silhouette function. Signal intensity at ONLY the 5' and solo LTRs (~450bp window) were extracted with the multiBigwigSummary function where differences in intensity were calculated with Wilcox rank-sum test. The contribution of 5' and solo LTR7up and LTR7d to mature transcripts was determined by extracting hg19 coordinates of "highly transcribed" (fpkm >2) HERVH loci (GSE54726) (Wang et al., 2014), converting these to hg38 coordinates with UCSC *liftover* (Kent et al., 2002; Raney et al., 2014), and intersecting these with the separated LTR7up/d solo/full-length loci.

3.4.3 Histone and transcription factor ChIP-seq peak calling and enrichment analysis

LTR7 loci positive or negative for a given mark were pulled from (Carter et al., 2021) where histone and TF ChIP-seq data from H1 cells (GSE61475, GSE117395, and GSE78099) were aligned to the human reference genome with bowtie2 and peaks were called with MACS2. These peaks were intersected with 7up/d full/solo loci where any overlap counted as a positive for a given mark. Statistical differences in the proportion of positive loci between 7up full and solo, and 7up full 7d full were determined via Chi-square test with Bonferroni correction (n=17). These statistical tests were used to determine which marks were differentially present between 7up and 7d full-length loci, 7up full-length and 7up solo loci, or both. The resulting groupings are seen in **(Figure 3.2.2b-d)**.

3.4.4 CpG methylation, H3K9me3, and GRO-seq aggregate signal differential analysis

For the direct comparison of 7up full-length and solo CpG methylation, H3K9me3 (H1 – primed - GSE78099), and plus-strand GRO-seq reads (H1 cells - GSE64758), I retrieved reads aligned to the reference human genome from (Dunham et al., 2012), (Theunissen et al., 2016), and (Estarás et al., 2015), respectively. Using deeptools (Ramírez et al., 2016), I plotted all reads for these marks in windows 500bp upstream and 10kb downstream of the 5’ most point of 5’ and solo LTRs, sorting insertions on the number of whole genome bisulfite sequencing reads in this window. Statistical comparison of these marks between full-length and solo loci was done with Wilcox rank-sum test ONLY on reads on the 5' and solo LTRs (~450bp window).

3.4.5 dREG score calculation and comparisons

To determine if LTR7up solo and full-length insertions had differential capacity to act as cis-regulatory units, I ran plus and minus strand GRO-seq (H1 – primed - GSE78099) (Estarás et al., 2015) aligned to the hg38 human genome in windows 500bp upstream and 8kb downstream of each LTR through dREG https://dreg.dnasequence.org/ (Danko et al., 2015; Wang et al., 2019). The infp file was used to calculate the dREG score at each position in the window. Differences between solo and full-length loci were determined by Wilcox rank-sum test of scores for each group ONLY at 5' and solo LTR (~450bp window).

3.4.6 ChrommHMM analysis

ChrommHMM was previously applied to 9 cell types where DNA sequences were called for promoter, enhancer, weak enhancer, repressed, etc. profiles (Ernst et al., 2011; Ernst and Kellis, 2010). Calls for H1 cells were intersected with LTR7up/d full-length/solo insertions. Most LTRs had multiple regulatory calls **(data not shown)**. To ensure LTRs were not double counted, only the call associated with the highest POLII loading was considered:

TSS>PromoterFlanking>Enhancer>WeakEnhancer>Transcribed>Repressed. Only TSS and enhancer (not including weak enhancer) were included in analysis.

3.4.7 Identification of LTR7 recombination events in great apes

To identify fixed LTR-LTR recombination events in the human lineage, I defined orthologous LTR7 insertions between human GRCh38/hg38 (Miga et al., 2014) and non-human great apes using reciprocal UCSC *liftover* (Kent et al., 2002; Raney et al., 2014) to, requiring that all loci have a 1:1 match in human and all target species - Clint_PTRv2/panTro6 (Waterson et al., 2005), Kamilah_GGO_v0/gorGor6 (Scally et al., 2012), Susie_PABv2/ponAbe3 (Locke et al., 2011). OneCodeToFindThemAll.pl (Bailly-Bechet et al., 2014) was run for all insertions in all great apes. Only "solo" and "full-length" insertions were considered for analysis. Recombination events were determined by mis-matches in "solo" and "full-length" status between species. The branch these events occurred on was determined by parsimony. For example, a locus with "full-length" status in human and gorilla and "solo" in chimpanzee would be called in the chimpanzee lineage.

REFERENCES

Babaian A, Mager DL. 2016. Endogenous retroviral promoter exaptation in human cancer. *Mobile DNA* **7**:24. doi:10.1186/s13100-016-0080-x

Bailly-Bechet M, Haudry A, Lerat E. 2014. "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. *Mobile DNA* **5**:13. doi:10.1186/1759-8753-5-13

Belshaw R, Watson J, Katzourakis A, Howe A, Woolven-Allen J, Burt A, Tristem M. 2007. Rate of recombinational deletion among human endogenous retroviruses. *J Virol* **81**:9437–9442. doi:10.1128/JVI.02216-06

Butelli E, Licciardello C, Zhang Y, Liu J, Mackay S, Bailey P, Reforgiato-Recupero G, Martin C. 2012. Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges. *The Plant Cell* **24**:1242–1255. doi:10.1105/tpc.111.095232

Cadle-Davidson MM, Owens CL. 2008. Genomic amplification of the Gret1 retroelement in white-fruited accessions of wild Vitis and interspecific hybrids. *Theor Appl Genet* **116**:1079–1094. doi:10.1007/s00122-008-0737-z

Carter TA, Singh M, Dumbović G, Chobirko JD, Rinn JL, Feschotte C. 2021. Mosaic cis-regulatory evolution drives transcriptional partitioning of HERVH endogenous retrovirus in the human embryo. *bioRxiv* 2021.07.08.451617. doi:10.1101/2021.07.08.451617

Coffin JM, Hughes SH, Varmus HE. 1997. Immune Response to Retroviral Infection, Retroviruses. Cold Spring Harbor Laboratory Press.

Danko CG, Hyland SL, Core LJ, Martins AL, Waters CT, Lee HW, Cheung VG, Kraus WL, Lis JT, Siepel A. 2015. Identification of active transcriptional regulatory elements from GRO-seq data. *Nat Methods* **12**:433–438. doi:10.1038/nmeth.3329

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee B-K, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, Simon JM, Song L, Trinklein ND, Altshuler RC, Birney E, Brown JB, Cheng C, Djebali S, Dong X, Dunham I, Ernst J, Furey TS, Gerstein M, Giardine B, Greven M, Hardison RC, Harris RS, Herrero J, Hoffman MM, Iyer S, Kellis M, Khatun J, Kheradpour P, Kundaje A, Lassmann T, Li Q, Lin X, Marinov GK, Merkel A, Mortazavi A, Parker SCJ, Reddy TE, Rozowsky J, Schlesinger F, Thurman RE, Wang J, Ward LD, Whitfield TW, Wilder SP, Wu W, Xi HS, Yip KY, Zhuang J, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M, Pazin MJ, Lowdon RF, Dillon LAL, Adams LB, Kelly CJ, Zhang J, Wexler JR, Green ED, Good PJ, Feingold EA, Bernstein BE, Birney E, Crawford GE, Dekker J, Elnitski L, Farnham PJ, Gerstein M, Giddings MC, Gingeras TR, Green ED, Guigó R, Hardison RC, Hubbard TJ, Kellis M, Kent WJ, Lieb JD, Margulies EH, Myers RM, Snyder M, Stamatoyannopoulos JA, Tenenbaum SA, Weng Z, White KP, Wold B, Khatun J, Yu Y, Wrobel J, Risk BA, Gunawardena HP, Kuiper HC, Maier CW, Xie L, Chen X, Giddings MC, Bernstein BE, Epstein CB, Shoresh N, Ernst J, Kheradpour P, Mikkelsen TS, Gillespie S, Goren A, Ram O, Zhang X, Wang L, Issner R, Coyne MJ, Durham T, Ku M, Truong T, Ward LD, Altshuler RC, Eaton ML, Kellis M, Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Batut P, Bell I, Bell K, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena HP, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Li G, Luo OJ, Park E, Preall JB, Presaud K, Ribeca P, Risk BA, Robyr D, Ruan X, Sammeth M, Sandhu KS, Schaeffer L, See L-H, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Hayashizaki Y, Harrow J, Gerstein M, Hubbard TJ, Reymond A, Antonarakis SE, Hannon GJ, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR, Rosenbloom KR, Sloan CA, Learned K, Malladi VS, Wong MC, Barber GP, Cline MS, Dreszer TR, Heitner SG, Karolchik D, Kent WJ, Kirkup VM, Meyer LR, Long JC, Maddren M, Raney BJ, Furey TS, Song L, Grasfeder LL, Giresi PG, Lee B-K, Battenhouse A, Sheffield NC, Simon JM, Showers KA, Safi A, London D, Bhinge AA, Shestak C, Schaner MR, Ki Kim S, Zhang ZZ, Mieczkowski PA, Mieczkowska JO, Liu Z, McDaniell RM, Ni Y, Rashid NU, Kim MJ, Adar S, Zhang Z, Wang T, Winter D, Keefe D, Birney E, Iyer VR, Lieb JD, Crawford GE, Li G, Sandhu KS, Zheng M, Wang P, Luo OJ, Shahab A, Fullwood MJ, Ruan X, Ruan Y, Myers RM, Pauli F, Williams BA, Gertz J, Marinov GK, Reddy TE, Vielmetter J, Partridge E, Trout D, Varley KE, Gasper C, The ENCODE Project Consortium, Overall coordination (data analysis coordination), Data production leads (data production), Lead analysts (data analysis), Writing group, NHGRI project management (scientific management), Principal investigators (steering committee), Boise State University and University of North Carolina at Chapel Hill Proteomics groups (data production and analysis), Broad Institute Group (data production and analysis), Cold Spring Harbor U of G Center for Genomic Regulation, Barcelona, RIKEN, Sanger Institute, University of Lausanne, Genome Institute of Singapore group (data production and analysis), Data coordination center at UC Santa Cruz (production data coordination), Duke University E University of Texas, Austin, University of North Carolina-Chapel Hill group (data production and analysis), Genome Institute of Singapore group (data production and analysis), HudsonAlpha Institute C UC Irvine, Stanford group (data production and analysis). 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**:57–74. doi:10.1038/nature11247

Eickbush TH, Malik HS. 2002. Origins and Evolution of Retrotransposons. *Mobile DNA II* 1111–1144. doi:10.1128/9781555817954.ch49

Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**:817–825. doi:10.1038/nbt.1662

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**:43–49. doi:10.1038/nature09906

Estarás C, Benner C, Jones KA. 2015. SMADs and YAP Compete to Control Elongation of β-Catenin:LEF-1-Recruited RNAPII during hESC Differentiation. *Molecular Cell* **58**:780–793. doi:10.1016/j.molcel.2015.04.001

Fort A, Hashimoto K, Yamada D, Salimullah M, Keya CA, Saxena A, Bonetti A, Voineagu I, Bertin N, Kratz A, Noro Y, Wong C-H, de Hoon M, Andersson R, Sandelin A, Suzuki H, Wei C-L, Koseki H, Hasegawa Y, Forrest ARR, Carninci P. 2014. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nature Genetics* **46**:558–566. doi:10.1038/ng.2965

Gemmell P, Hein J, Katzourakis A. 2016. Phylogenetic Analysis Reveals That ERVs "Die Young" but HERV-H Is Unusually Conserved. *PLOS Computational Biology* **12**:e1004964. doi:10.1371/journal.pcbi.1004964

Gemmell P, Hein J, Katzourakis A. 2015. Orthologous endogenous retroviruses exhibit directional selection since the chimp-human split. *Retrovirology* **12**. doi:10.1186/s12977-015-0172-6

Glinsky GV. 2015. Transposable Elements and DNA Methylation Create in Embryonic Stem Cells Human-Specific Regulatory Sequences Associated with Distal Enhancers and Noncoding RNAs. *Genome Biol Evol* **7**:1432–1454. doi:10.1093/gbe/evv081

Göke J, Lu X, Chan Y-S, Ng H-H, Ly L-H, Sachs F, Szczerbinska I. 2015. Dynamic Transcription of Distinct Classes of Endogenous Retroviral Elements Marks Specific Populations of Early Human Embryonic Cells. *Cell Stem Cell* **16**:135–141. doi:10.1016/j.stem.2015.01.005

Imbeault M, Helleboid P-Y, Trono D. 2017. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**:550–554. doi:10.1038/nature21683

Izsvák Z, Wang J, Singh M, Mager DL, Hurst LD. 2016. Pluripotency and the endogenous retrovirus HERVH: Conflict or serendipity? *BioEssays* **38**:109–117. doi:10.1002/bies.201500096

Jordan IK, McDonald JF. 1999. Tempo and mode of Ty element evolution in Saccharomyces cerevisiae. *Genetics* **151**:1341–1351.

Kelley D, Rinn J. 2012. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biology* **13**:R107. doi:10.1186/gb-2012-13-11-r107

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler and D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12**:996–1006. doi:10.1101/gr.229102

Kobayashi S, Goto-Yamamoto N, Hirochika H. 2004. Retrotransposon-Induced Mutations in Grape Skin Color. *Science* **304**:982–982. doi:10.1126/science.1095011

Kong Y, Rose CM, Cass AA, Williams AG, Darwish M, Lianoglou S, Haverty PM, Tong A-J, Blanchette C, Albert ML, Mellman I, Bourgon R, Greally J, Jhunjhunwala S, Chen-Harris H. 2019. Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. *Nat Commun* **10**:5228. doi:10.1038/s41467-019-13035-2

Lisch D. 2013. How important are transposons for plant evolution? *Nat Rev Genet* **14**:49–61. doi:10.1038/nrg3374

Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM, Yang S-P, Wang Z, Chinwalla AT, Minx P, Mitreva M, Cook L, Delehaunty KD, Fronick C, Schmidt H, Fulton LA, Fulton RS, Nelson JO, Magrini V, Pohl C, Graves TA, Markovic C, Cree A, Dinh HH, Hume J, Kovar CL, Fowler GR, Lunter G, Meader S, Heger A, Ponting CP, Marques-Bonet T, Alkan C, Chen L, Cheng Z, Kidd JM, Eichler EE, White S, Searle S, Vilella AJ, Chen Y, Flicek P, Ma J, Raney B, Suh B, Burhans R, Herrero J, Haussler D, Faria R, Fernando O, Darré F, Farré D, Gazave E, Oliva M, Navarro A, Roberto R, Capozzi O, Archidiacono N, Valle GD, Purgato S, Rocchi M, Konkel MK, Walker JA, Ullmer B, Batzer MA, Smit AFA, Hubley R, Casola C, Schrider DR, Hahn MW, Quesada V, Puente XS, Ordoñez GR, López-Otín C, Vinar T, Brejova B, Ratan A, Harris RS, Miller W, Kosiol C, Lawson HA, Taliwal V, Martins AL, Siepel A, RoyChoudhury A, Ma X, Degenhardt J, Bustamante CD, Gutenkunst RN, Mailund T, Dutheil JY, Hobolth A, Schierup MH, Ryder OA, Yoshinaga Y, de Jong PJ, Weinstock GM, Rogers J, Mardis ER, Gibbs RA, Wilson RK. 2011. Comparative and demographic analysis of orang-utan genomes. *Nature* **469**:529–533. doi:10.1038/nature09687

Loewer S, Cabili MN, Guttman M, Loh Y-H, Thomas K, Park IH, Garber M, Curran M, Onder T, Agarwal S, Manos PD, Datta S, Lander ES, Schlaeger TM, Daley GQ, Rinn JL. 2010. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet* **42**:1113–1117. doi:10.1038/ng.710

Mager DL, Freeman JD. 1995. HERV-H Endogenous Retroviruses: Presence in the New World Branch but Amplification in the Old World Primate Lineage. *Virology* **213**:395–404. doi:10.1006/viro.1995.0012

Meisler MH, Ting C-N. 1993. The Remarkable Evolutionary History of the Human Amylase Genes. *Critical Reviews in Oral Biology & Medicine* **4**:503–509. doi:10.1177/10454411930040033501

Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ. 2014. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res* **24**:697–707. doi:10.1101/gr.159624.113

Nelson DT, Goodchild NL, Mager DL. 1996. Gain of Sp1 Sites and Loss of Repressor Sequences Associated with a Young, Transcriptionally Active Subset of HERV-H Endogenous Long Terminal Repeats. *Virology* **220**:213–218. doi:10.1006/viro.1996.0303

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* **44**:W160–W165. doi:10.1093/nar/gkw257

Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D, Kent WJ. 2014. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* **30**:1003–1005. doi:10.1093/bioinformatics/btt637

Römer C, Singh M, Hurst LD, Izsvák Z. 2017. How to tame an endogenous retrovirus: HERVH and the evolution of human pluripotency. *Current Opinion in Virology*, Animal models for viral diseases • Paleovirology **25**:49–58. doi:10.1016/j.coviro.2017.07.001

Samuelson LC, Phillips RS, Swanberg LJ. 1996. Amylase gene structures in primates: retroposon insertions and promoter evolution. *Molecular Biology and Evolution* **13**:767–779. doi:10.1093/oxfordjournals.molbev.a025637

Samuelson LC, Wiebauer K, Gumucio DL, Meisler MH. 1988. Expression of the human amylase genes: recent origin of a salivary amylase promoter from an actin pseudogene. *Nucleic Acids Research* **16**:8261–8276. doi:10.1093/nar/16.17.8261

Samuelson LC, Wiebauer K, Snow CM, Meisler MH. 1990. Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution. *Molecular and Cellular Biology* **10**:2513–2520. doi:10.1128/mcb.10.6.2513-2520.1990

Santoni FA, Guerra J, Luban J. 2012. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology* **9**:111. doi:10.1186/1742-4690-9-111

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, McCarthy S, Montgomery SH, Schwalie PC, Tang YA, Ward MC, Xue Y, Yngvadottir B, Alkan C, Andersen LN, Ayub Q, Ball EV, Beal K, Bradley BJ, Chen Y, Clee CM, Fitzgerald S, Graves TA, Gu Y, Heath P, Heger A, Karakoc E, Kolb-Kokocinski A, Laird GK, Lunter G, Meader S, Mort M, Mullikin JC, Munch K, O'Connor TD, Phillips AD, Prado-Martinez J, Rogers AS, Sajjadian S, Schmidt D, Shaw K, Simpson JT, Stenson PD, Turner DJ, Vigilant L, Vilella AJ, Whitener W, Zhu B, Cooper DN, de Jong P, Dermitzakis ET, Eichler EE, Flicek P, Goldman N, Mundy NI, Ning Z, Odom DT, Ponting CP, Quail MA, Ryder OA, Searle SM, Warren WC, Wilson RK, Schierup MH, Rogers J, Tyler-Smith C, Durbin R. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**:169–175. doi:10.1038/nature10842

Shimazaki M, Fujita K, Kobayashi H, Suzuki S. 2011. Pink-Colored Grape Berry Is the Result of Short Insertion in Intron of Color Regulatory Gene. *PLOS ONE* **6**:e21308. doi:10.1371/journal.pone.0021308

Smit AF, Hubley R, Green P. 2013. RepeatMasker Open-4.0.

Theunissen TW, Friedli M, He Y, Planet E, O'Neil RC, Markoulaki S, Pontis J, Wang H, Iouranova A, Imbeault M, Duc J, Cohen MA, Wert KJ, Castanon R, Zhang Z, Huang Y, Nery JR, Drotar J, Lungjangwa T, Trono D, Ecker JR, Jaenisch R. 2016. Molecular Criteria for Defining the Naive Human Pluripotent State. *Cell Stem Cell* **19**:502–515. doi:10.1016/j.stem.2016.06.011

Thomas J, Perron H, Feschotte C. 2018. Variation in proviral content among human genomes mediated by LTR recombination. *Mobile DNA* **9**:36. doi:10.1186/s13100-018-0142-3

Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, Cai H, Besser D, Prigione A, Fuchs NV, Schumann GG, Chen W, Lorincz MC, Ivics Z, Hurst LD, Izsvák Z. 2014. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* **516**:405–409. doi:10.1038/nature13804

Wang Z, Chu T, Choate LA, Danko CG. 2019. Identification of regulatory elements from nascent transcription using dREG. *Genome Res* **29**:293–303. doi:10.1101/gr.238279.118

Waterson RH, Lander ES, Wilson RK, The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**:69–87. doi:10.1038/nature04072

Wilkinson DA, Freeman JD, Goodchild NL, Kelleher CA, Mager DL. 1990. Autonomous expression of RTVL-H endogenous retroviruslike elements in human cells. *Journal of Virology* **64**:2157–2167.

Yu H-L, Zhao Z-K, Zhu F. 2013. The role of human endogenous retroviral long terminal repeat sequences in human cancer (Review). *International Journal of Molecular Medicine* **32**:755–762. doi:10.3892/ijmm.2013.1460

Zhang Y, Li T, Preissl S, Amaral ML, Grinstein JD, Farah EN, Destici E, Qiu Y, Hu R, Lee AY, Chee S, Ma K, Ye Z, Zhu Q, Huang H, Fang R, Yu L, Izpisua Belmonte JC, Wu J, Evans SM, Chi NC, Ren B. 2019. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nature Genetics* **51**:1380–1388. doi:10.1038/s41588-019-0479-7

CHAPTER 4
DISCUSSION AND OPEN QUESTIONS

4.1 A PRECISE 5' LTR TFBS REPERTOIRE AND THE PRESENCE OF THE INTERNAL AND/OR 3' LTR REGIONS ARE NECESSARY FOR ERV TRANSCRIPTION

Previous work highlighted a correlation between the binding of pluripotent transcription factors (TFs) and LTR7's pluripotent transcription (Göke et al., 2015; Ito et al., 2017; Kelley and Rinn, 2012; Kunarso et al., 2010; Ohnuki et al., 2014; Pontis et al., 2019; Santoni et al., 2012). However, to my knowledge, no TF has been experimentally shown to regulate LTR7. Our data from Chapter 2 shows that an 8bp insertion was required for robust transcription of LTR7up in induced pluripotent stem cells (iPSCs). The reporter construct lacking these 8bp exhibited 20-fold less reporter signal compared to intact LTR7up. Corroborating motif and ChIP-seq evidence leads us to suspect this was due to the gain of a singular SOX2/3 motif. The gigantic fold change associated with this small change is curious, considering that the closely related LTR7u1 has this motif, but does not bind SOX2 as strongly and is not transcribed in iPSC or human embryonic stem cells (ESCs). These data indicate that other 7up-specific mutations have been necessary for its pluripotent transcription. These mutations may be nearby cooperating transcription factor binding sites (TFBS), or other regulatory sequences contained in the HERVH-int or 3' LTR7up. My data from Chapter 3 show that within LTR7up, only full-length insertions are transcribed and give rise to mature transcripts. While less pronounced, this trend holds for the relatively untranscribed LTR7d, indicating that the full-length morph promotes transcription at loci with less-than-ideal TFBS.

We do not fully elucidate the contribution and interaction of the 5’ LTR, 3’ LTR, and internal region of HERVH/LTR7up. However, our data indicate a complex relationship. Most tested TFs bound an equal proportion of 5’ and solo LTR7up, suggesting that both morphs perform equally well at opening chromatin at the LTR and initiating transcription. Additionally, our promoter assays in iPSCs placed only an LTR upstream of a reporter gene without the presence of the internal portion. In this extra-genomic system, LTR7up is still capable of driving transcription of the downstream luciferase gene. This model is congruent with the traditional role of 5’ LTRs serving as promoters and internal regions and 3’ LTRs serving as transcript stabilizers (Coffin, 1988; Eickbush and Malik, 2002; Nelson et al., 1996; Wilkinson et al., 1990). On the other hand, the TFs TCF4 and FOXA1 exclusively bind full-length 7up. This potentially indicates that there are secondary TFs who bind LTR7up after transcription has been established, or that these TFs are donated by the internal or 3’ LTR regions. This idea is given more credence by my discovery that the 3’ LTR and internal region contain cis-regulatory elements. Future work should seek to determine the precise role of TFBS and transcript stabilization signals in the LTR and internal regions. The most conclusive experiments will likely involve knocking out multiple combinations of TFBS, splice sites, and poly-A sites in a genomic context and using a combination of RNA- and GRO-seq to elucidate the role of these factors in transcript stabilization and transcription.

4.2 EVOLUTION BY TWO TYPES OF RECOMBINATION

Inter-element recombination of coding genes and the formation of mosaic elements is a common feature of ERV (Vargiu et al., 2016). To my knowledge, our data in Chapter 2 represent the first documentation of prolific inter-element recombination of LTRs within an ERV family. These recombination events coincide with the colonization of new niches in human embryogenesis and their subsequent proliferation, indicating that these events may be adaptive and increase the fitness of an ERV. If HERVH has been coopted for host function, inter-element recombination of LTRs was likely instrumental to the emergence of its host functions. Intra-element LTR-LTR recombination cannot be advantageous for ERV, as it results in the death of an element through the loss of indispensable DNA. However, these events may be selected for or against by the host. In Chapter 3, I show that the full-length to solo transition constitutes a cis-regulatory switch, where promoter activity is. Considering that transcribed HERVH have the capacity to demarcate TAD domains (Zhang et al., 2019) and contribute lncRNA (Fort et al., 2014; Gemmell et al., 2015; Izsvák et al., 2016; Kelley and Rinn, 2012; Loewer et al., 2010; Römer et al., 2017; Santoni et al., 2012), any perturbation of transcription has the capacity to alter gene expression. LTR-LTR recombination is accompanied by the loss of host silencing marks on the LTR, while retaining the hallmarks of enhancers. This indicates another possible mechanism by which recombination can affect host gene regulation, possibly leading to regulatory innovation.

4.3 HUMAN SELECTION FOR HERVH

Previous work has speculated on the possible cooption of HERVH transcription or RNA for host function (Gemmell et al., 2016; Izsvák et al., 2016). The most compelling evidence for this hypothesis is 1) differentiation of human ESC upon HERVH knockdown (Loewer et al., 2010; Lu et al., 2014; Ohnuki et al., 2014; Wang et al., 2014), 2) changes in daughter cell gene expression upon TAD-contributing HERVH knockout (Zhang et al., 2019), and 3) the prevalence for full-length HERVH compared to the solo morph (Gemmell et al., 2016). These data suggest that primate hosts repurposed HERVH soon after the first copies integrated into their genome.

Thereafter, many full-length HERVH loci were indispensable for the proper regulation of pluripotency and selection disfavored the appearance of solo HERVH morphs. In this scenario, the most improbable event is HERVH selection for function in pluripotency soon after integration. Pluripotency is a necessary and constrained feature of vertebrate development (Endo et al., 2020; Kuijk et al., 2015). The likelihood that there is room for rapid positive selection in this arena seems unlikely. If anything, it seems more likely that HERVH replaced some existing regulator of pluripotency or the effects of its potential regulatory innovation are in more differentiated cells, as suggested in (Zhang et al., 2019).

Some of the data presented here are compatible with the above model, but others are incompatible. In congruence, I show that there is a strong regulatory difference between full-length and solo morphs within LTR7up. The primary difference between the two morphs is that full-lengths can be transcribed, while solos cannot. Considering that transcription is required for the contribution to HERVH-TADs (Zhang et al., 2019) and to HERVH-lncRNAs, solo elements cannot contribute to pluripotent regulation through either of these mechanisms. This gives further credence to the idea that full-length HERVH may have been preserved for host function. Opposing this model, my granular subdivision of LTR7 shows that pluripotent transcription is confined to a small subset, 7up. Non-7up subfamilies are expressed elsewhere in early development and are unlikely to contribute to pluripotent regulation through transcription or RNA products. These other subfamilies also have a preponderance of full-length morphs when compared to other HERV (Gemmell et al., 2016). Have these other subfamilies had or currently have roles in regulating other cell stages? Possibly, but as of yet there is no evidence for this. Most at odds with this model is the HERVH subfamily LTR7C, who has a high full-length to solo ratio but lack stage-specific expression, instead exhibiting expression in a variety of pre-implantation cell types. Could LTR7C have integrated into the regulatory networks for all these cell types? Considering that cooption of viral products is rare, it seems exceedingly unlikely. At the very least, selection for full-length insertions is insufficient to explain the strange solo:full-length ratio seen across HERVH subfamilies.

Two other explanations may explain the preponderance of HERVH full-length loci: 1) an unknown mechanistic recombination block, such as unequivocal 5' and 3' LTRs (discussed in Chapter 3), or 2) selection against HERVH solo morphs. LTR-LTR recombination results in the loss of silencing marks and enhancer-like hallmarks. The cell-type-specific expression of many HERVH subfamilies and their accompanying TFBS repertoires indicate that they might have the capacity to (mis)regulate nearby genes in the solo morph. Under this model and congruent with the data, all HERVH would have a high proportion of full-length elements, regardless of where they are expressed in development. This view necessitates that solo HERVH are unusually disruptive embryonic regulators, a fact that has not been elucidated, and does not account for the effects of full-length HERVH on pluripotency.

In conclusion, my work here highlights how changes in promoter activity through inter-element recombination allowed an endogenous retrovirus to diversify its expression and colonize new niches. The spread of these elements may have resulted in species-specific cis-regulatory elements, which could be further modified by intra-element LTR-LTR recombination. I believe the inter-specific study of ERV-derived regulatory DNA may yield further insights in genomics, cis-regulatory evolution, and biomedical areas.

REFERENCES

Coffin JM. 1988. Replication of Retrovirus GenomesRNA Genetics. CRC Press.

Eickbush TH, Malik HS. 2002. Origins and Evolution of Retrotransposons. *Mobile DNA II* 1111–1144. doi:10.1128/9781555817954.ch49

Endo Y, Kamei K, Inoue-Murayama M. 2020. Genetic Signatures of Evolution of the Pluripotency Gene Regulating Network across Mammals. *Genome Biology and Evolution* **12**:1806–1818. doi:10.1093/gbe/evaa169

Fort A, Hashimoto K, Yamada D, Salimullah M, Keya CA, Saxena A, Bonetti A, Voineagu I, Bertin N, Kratz A, Noro Y, Wong C-H, de Hoon M, Andersson R, Sandelin A, Suzuki H, Wei C-L, Koseki H, Hasegawa Y, Forrest ARR, Carninci P. 2014. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nature Genetics* **46**:558–566. doi:10.1038/ng.2965

Gemmell P, Hein J, Katzourakis A. 2016. Phylogenetic Analysis Reveals That ERVs “Die Young” but HERV-H Is Unusually Conserved. *PLOS Computational Biology* **12**:e1004964. doi:10.1371/journal.pcbi.1004964

Gemmell P, Hein J, Katzourakis A. 2015. Orthologous endogenous retroviruses exhibit directional selection since the chimp-human split. *Retrovirology* **12**. doi:10.1186/s12977-015-0172-6

Göke J, Lu X, Chan Y-S, Ng H-H, Ly L-H, Sachs F, Szczerbinska I. 2015. Dynamic Transcription of Distinct Classes of Endogenous Retroviral Elements Marks Specific Populations of Early Human Embryonic Cells. *Cell Stem Cell* **16**:135–141. doi:10.1016/j.stem.2015.01.005

Ito J, Sugimoto R, Nakaoka H, Yamada S, Kimura T, Hayano T, Inoue I. 2017. Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *PLOS Genetics* **13**:e1006883. doi:10.1371/journal.pgen.1006883

Izsvák Z, Wang J, Singh M, Mager DL, Hurst LD. 2016. Pluripotency and the endogenous retrovirus HERVH: Conflict or serendipity? *BioEssays* **38**:109–117. doi:10.1002/bies.201500096

Kelley D, Rinn J. 2012. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biology* **13**:R107. doi:10.1186/gb-2012-13-11-r107

Kuijk E, Geijsen N, Cuppen E. 2015. Pluripotency in the light of the developmental hourglass. *Biol Rev Camb Philos Soc* **90**:428–443. doi:10.1111/brv.12117

Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, Chan Y-S, Ng H-H, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics* **42**:631–634. doi:10.1038/ng.600

Loewer S, Cabili MN, Guttman M, Loh Y-H, Thomas K, Park IH, Garber M, Curran M, Onder T, Agarwal S, Manos PD, Datta S, Lander ES, Schlaeger TM, Daley GQ, Rinn JL. 2010. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet* **42**:1113–1117. doi:10.1038/ng.710

Lu X, Sachs F, Ramsay L, Jacques P-É, Göke J, Bourque G, Ng H-H. 2014. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nature Structural & Molecular Biology* **21**:423–425. doi:10.1038/nsmb.2799

Nelson DT, Goodchild NL, Mager DL. 1996. Gain of Sp1 Sites and Loss of Repressor Sequences Associated with a Young, Transcriptionally Active Subset of HERV-H Endogenous Long Terminal Repeats. *Virology* **220**:213–218. doi:10.1006/viro.1996.0303

Ohnuki M, Tanabe K, Sutou K, Teramoto I, Sawamura Y, Narita M, Nakamura Michiko, Tokunaga Y, Nakamura Masahiro, Watanabe A, Yamanaka S, Takahashi K. 2014. Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *PNAS* **111**:12426–12431. doi:10.1073/pnas.1413299111

Pontis J, Planet E, Offner S, Turelli P, Duc J, Coudray A, Theunissen TW, Jaenisch R, Trono D. 2019. Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. *Cell Stem Cell* **24**:724-735.e5. doi:10.1016/j.stem.2019.03.012

Römer C, Singh M, Hurst LD, Izsvák Z. 2017. How to tame an endogenous retrovirus: HERVH and the evolution of human pluripotency. *Current Opinion in Virology*, Animal models for viral diseases • Paleovirology **25**:49–58. doi:10.1016/j.coviro.2017.07.001

Santoni FA, Guerra J, Luban J. 2012. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology* **9**:111. doi:10.1186/1742-4690-9-111

Vargiu L, Rodriguez-Tomé P, Sperber GO, Cadeddu M, Grandi N, Blikstad V, Tramontano E, Blomberg J. 2016. Classification and characterization of human endogenous retroviruses; mosaic forms are common. *Retrovirology* **13**:7. doi:10.1186/s12977-015-0232-y

Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, Cai H, Besser D, Prigione A, Fuchs NV, Schumann GG, Chen W, Lorincz MC, Ivics Z, Hurst LD, Izsvák Z. 2014. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* **516**:405–409. doi:10.1038/nature13804

Wilkinson DA, Freeman JD, Goodchild NL, Kelleher CA, Mager DL. 1990. Autonomous expression of RTVL-H endogenous retroviruslike elements in human cells. *Journal of Virology* **64**:2157–2167.

Zhang Y, Li T, Preissl S, Amaral ML, Grinstein JD, Farah EN, Destici E, Qiu Y, Hu R, Lee AY, Chee S, Ma K, Ye Z, Zhu Q, Huang H, Fang R, Yu L, Izpisua Belmonte JC, Wu J, Evans SM, Chi NC, Ren B. 2019. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nature Genetics* **51**:1380–1388. doi:10.1038/s41588-019-0479-7