Mosaic cis-regulatory evolution drives transcriptional partitioning of HERVH endogenous retrovirus in the human embryo

Thomas A Carter[1], Manvendra Singh[1], Gabrijela Dumbović[2,3], Jason D Chobirko[1], John L Rinn[2,4], Cédric Feschotte[1]*

[1]Department of Molecular Biology and Genetics, Cornell University, Ithaca, United States; [2]BioFrontiers Institute, University of Colorado Boulder, Boulder, United States; [3]Institute for Cardiovascular Regeneration, Goethe University Frankfurt, Frankfurt am Main, Germany; [4]Department of Biochemistry, University of Colorado Boulder, Boulder, United States

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 eight previously unrecognized subfamilies that have been active at different timepoints in primate evolution and display distinct expression patterns during human embryonic 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 a mixture of mutational processes, including point mutations, duplications, and 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.

Editor's evaluation

Transposons achieve evolutionary success only upon self-replication in the germline, where novel genomic insertions are passed to the next generation. To define the genetic determinants of transposon specialization to the germline and specifically, specialization to pluripotent embryonic stem cells, this article applies evolutionary analysis, cell type-specific transcriptomics, and expression reporter assays to the human endogenous retrovirus type-H, HERV. Previous links between HERV and the maintenance of pluripotency make the discoveries consequential for the mobile element and genome evolution communities along with those engaged in stem cell biology and regenerative medicine.

***For correspondence:** cf458@cornell.edu

Competing interest: The authors declare that no competing interests exist.

Funding: See page 22

Preprinted: 08 July 2021
Received: 14 December 2021
Accepted: 17 February 2022
Published: 18 February 2022

Reviewing Editor: Mia T Levine, University of Pennsylvania, United States

Introduction

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 (OWMs), and New World monkeys more than 40 million years ago (mya) (**Goodchild et al., 1993**; **Izsvák et al., 2016**; **Mager and Freeman, 1995**). To date, no polymorphic HERVH insertions have been found in the human genome (**Thomas et al., 2018**).

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**; **Jern et al., 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**; **Jern et al., 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**). These inconsistent results have failed to clarify which expressed HERVH loci, if any, are necessary for the maintenance of pluripotency.

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.

Results

LTR7 consists of eight 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 (**Figure 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 (**Figure 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 eight subfamilies (**Figure 1B**).

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 eight newly defined LTR7 subfamilies (7o, 7bc, 7up, etc.). The consensus sequences were deposited at https://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 (**Figure 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 (**Figure 1C**). Together these results indicate that the LTRs of HERVH elements can be divided into additional subfamilies than those previously recognized.

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/

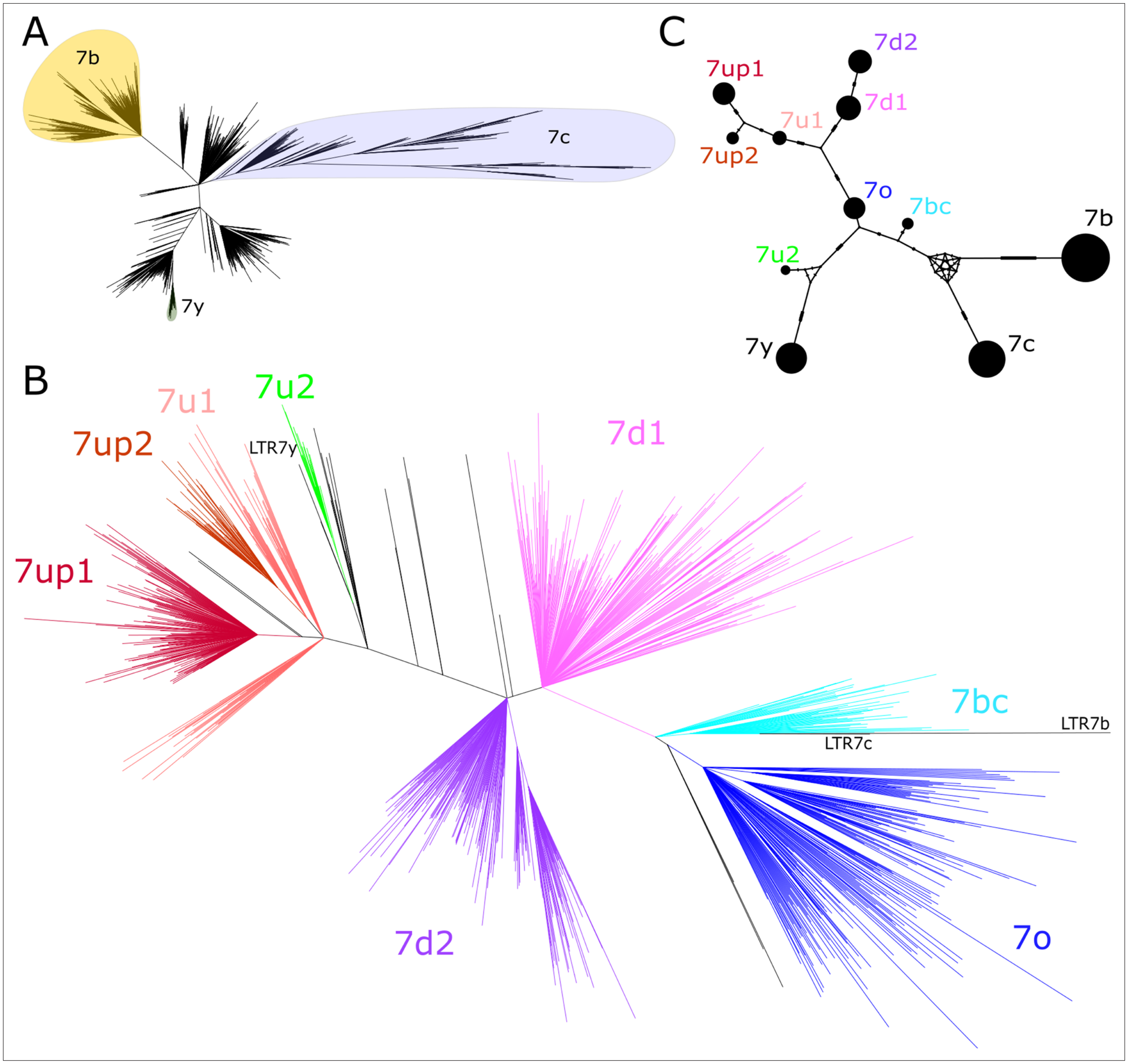

Figure 1. Phylogenetic analysis of LTR7 sequences. (**A**) Unrooted phylogeny of all solo and 5′ LTR7 subfamilies from LTR7, 7b, 7c, and 7y. Colors denote clades consisting of previously annotated 7b, 7c, and 7y with >95% concordance. (**B**) Unrooted phylogeny of all solo and 5′ LTR7 sequences. All nodes with ultrafast bootstraps (UFbootstraps) > 0.95, >10 member insertions, and >1.5 substitutions/100 bp (~6 base pairs) are grouped and colored (see Materials and methods). Previously listed consensus sequences from 7b/c/y were included in the alignment and are shown in black. (**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 Materials and methods) are included while 7b/c/y counts are from dfam.

The online version of this article includes the following figure supplement(s) for figure 1:

Figure supplement 1. Phylogenetic tree from LTR7 reverse transcriptase (RVT) domains.

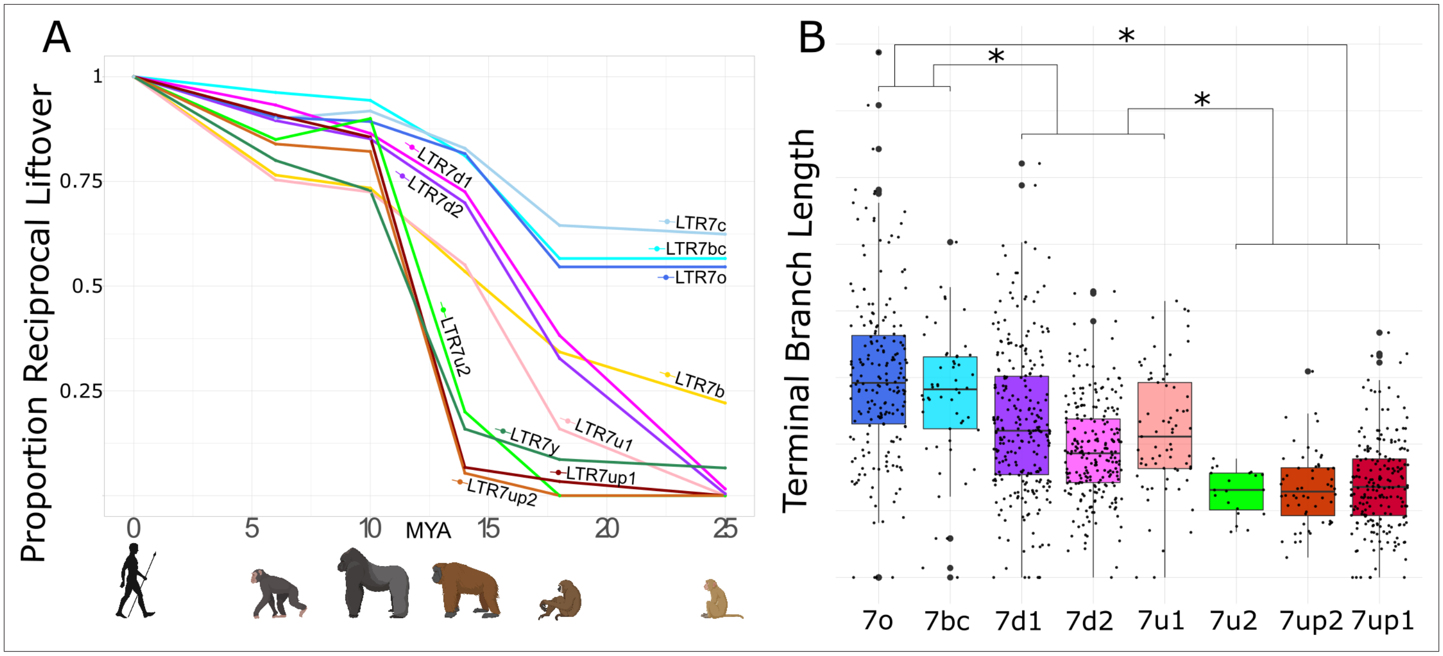

Figure 2. Age analysis of LTR7 subfamilies. (**A**) Proportion of a given subfamily that have 1:1 orthologous insertions between human and other primate species. LTR7 subfamilies are from the tree in ***Figure 1B***; 7b/c/y subfamilies are from RepeatMasker annotations. Non-human primates are spaced 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 ***Figure 1B***. Groups with similar liftover profiles were merged for statistical testing (see Materials and methods). Differences with padj <1e-15 are denoted with * (Wilcoxon rank-sum test with Bonferroni correction).

The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Gross reciprocal liftover from human to non-human primate (NHP).

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 and likely fixed within each 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 (***Figure 2A***). The subfamilies 7o, 7bc, and 7c are the oldest since the majority of their insertions are found at orthologous position in rhesus macaque, an OWM. These three subfamilies share similar evolutionary trajectories, with most of their proliferation occurring prior to the split of OWM and hominoids, ~25 mya (***Figure 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 (***Figure 2A***, ***Figure 2—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 (***Figure 2A***). 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, 14–20 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 (***Figure 2A***). 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 (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 (***Figure 2A***).

As an independent dating method, we used the terminal branch length separating each insertion from its nearest node in **Figure 1B** (**Figure 2**). 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; Wilcoxon 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 three 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.

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 Materials and 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 Materials and 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 sequence divergence or context-specific effects such as local chromatin or cis-regulatory environments. 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 (**Figure 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 3—figure supplement 1**), compared with only 4% of other LTR7 loci (**Figure 3D**, **Figure 3—figure supplement 1**). Half of the loci displaying GRO-seq signal (53/96) also showed evidence of mature RNA product (**Supplementary 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 (**Figure 3A, B, and 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). While KLF4 and NANOG binding is pervasive across the 7up subfamily, OCT4 binds merely 12% of 7up loci. 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 (see **Supplementary file 8** for full statistical analysis of all marks).

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%) (**Figure 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

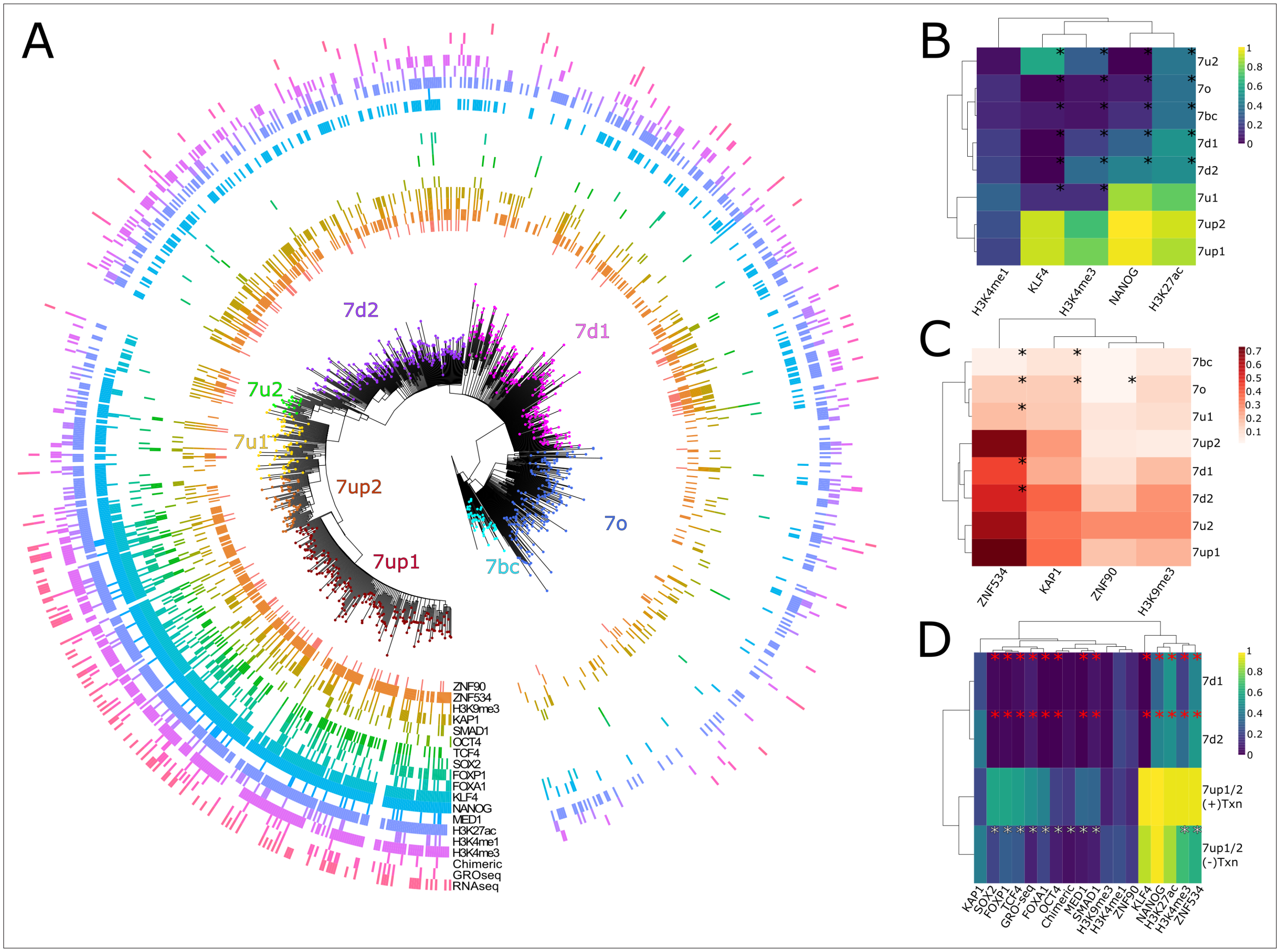

Figure 3. Phyloregulatory analysis of LTR7. (**A**) '"Phyloregulatory' map of LTR7. The phylogenetic analysis to derive the circular tree is the same as for the tree in ***Figure 1B*** but rooted on the 7b consensus. Subfamilies defined in ***Figure 1*** are denoted with dotted colored tips. Positive 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 embryonic stem cell (ESC) except for ZNF90 and ZNF534, which are derived from ChIP-exo data after overexpression of these factors in HEK293 cells (see Materials and 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 Wilcoxon rank-sum with Bonferroni correction). (**C**) Heatmap done in similar fashion to **B** 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.

The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. Heatmap and aggregate signal of two replicates of whole-genome bisulfite sequencing (WGBS), GRO-seq (plus strand), H3K9me3, and SOX2 in H1 cells.

Figure supplement 2. Violin plots visualize the density and distribution of ZNF534 in early embryogenesis and ESCs at passages 0 and 10.

throughout the tree of LTR7 sequences (26% vs. 18%) (***Figure 3A, B***). Thus, promoter marks are primarily restricted to 7up loci, but a broader range of LTR7 loci display putative enhancer marks.

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

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 (KZFPs) that silence the other subfamilies in ESCs. KZFPs 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 (**Figure 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) (**Figure 3C**). This was also the case for CpG methylation, whose presence was not differential between subfamilies (padj > 0.05 Wilcoxon rank-sum with Bonferroni correction) (**Figure 3—figure supplement 1**). 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 (**Figure 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 3—figure supplement 2**) 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 (**Figure 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 (**Figure 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 (**Figure 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.

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 (***Figure 4A***).

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 (***Figure 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 six instances of clades supported in the block parsimony trees that were incongruent with those supported by the overall phylogeny (***Figure 4B and 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 (***Figure 4C***, ***Figure 4—figure supplement 1***), suggestive that it arose via tandem duplication in the common ancestor of these subfamilies.

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.

HERVH subfamilies show distinct expression profiles in the preimplantation embryo

We hypothesized that the mosaic pattern of LTR sequence evolution, which represents a mixture of point mutations, duplications, and inter-element recombination, 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 eight newly defined LTR7 subfamilies plus newly generated consensus sequences for 7b, 7c, and 7y subfamilies redefined from the phylogenetic analysis presented in ***Figure 1B*** (***Figure 5—figure supplement 1***) (see Materials and methods). Our newly generated RepeatMasker annotations (***Supplementary 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 eight 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 (***Figure 1A–C***) and 7y and 7b share a great deal of sequence through recombination events (***Figure 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 preimplantation embryos and RNA-seq data from human ESCs (***Blakeley et al., 2015***; ***Tang et al., 2010***) (see Materials and methods).

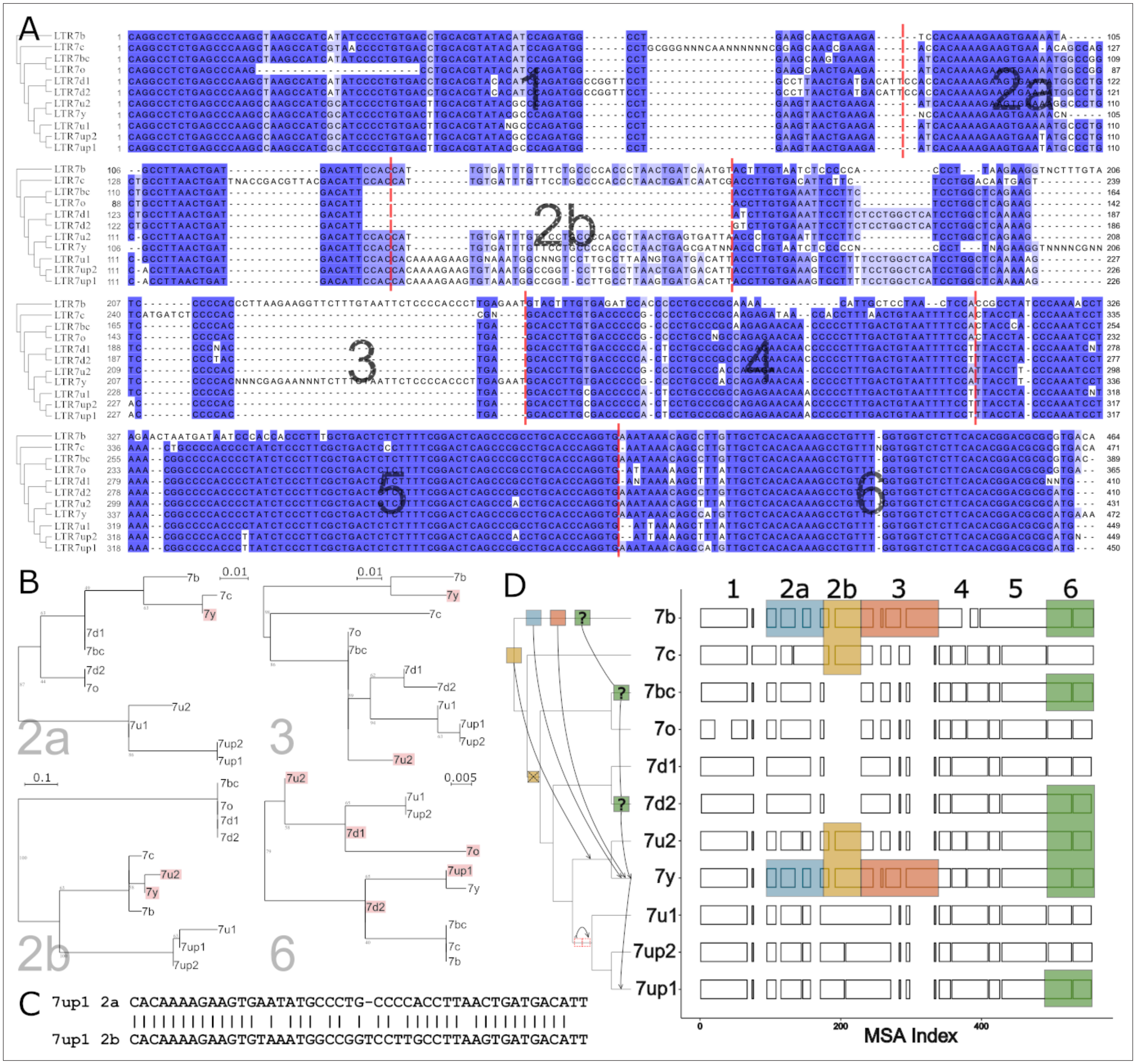

Figure 4. Modular block evolution of LTR7 subfamilies. (**A**) A multiple sequence alignment of LTR7 subfamily consensus sequences. The phylogenetic topology from ***Figure 1*** is shown on the left. The MSA is broken down into sequence blocks (red lines) with differential patterns of relationships. (**B**) Parsimony trees from **A** 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 two red rectangles.

The online version of this article includes the following figure supplement(s) for figure 4:

Figure supplement 1. Parsimony tree of all consensus human endogenous retrovirus type-H (HERVH) long terminal repeats (LTR) blocks 2a and 2b.

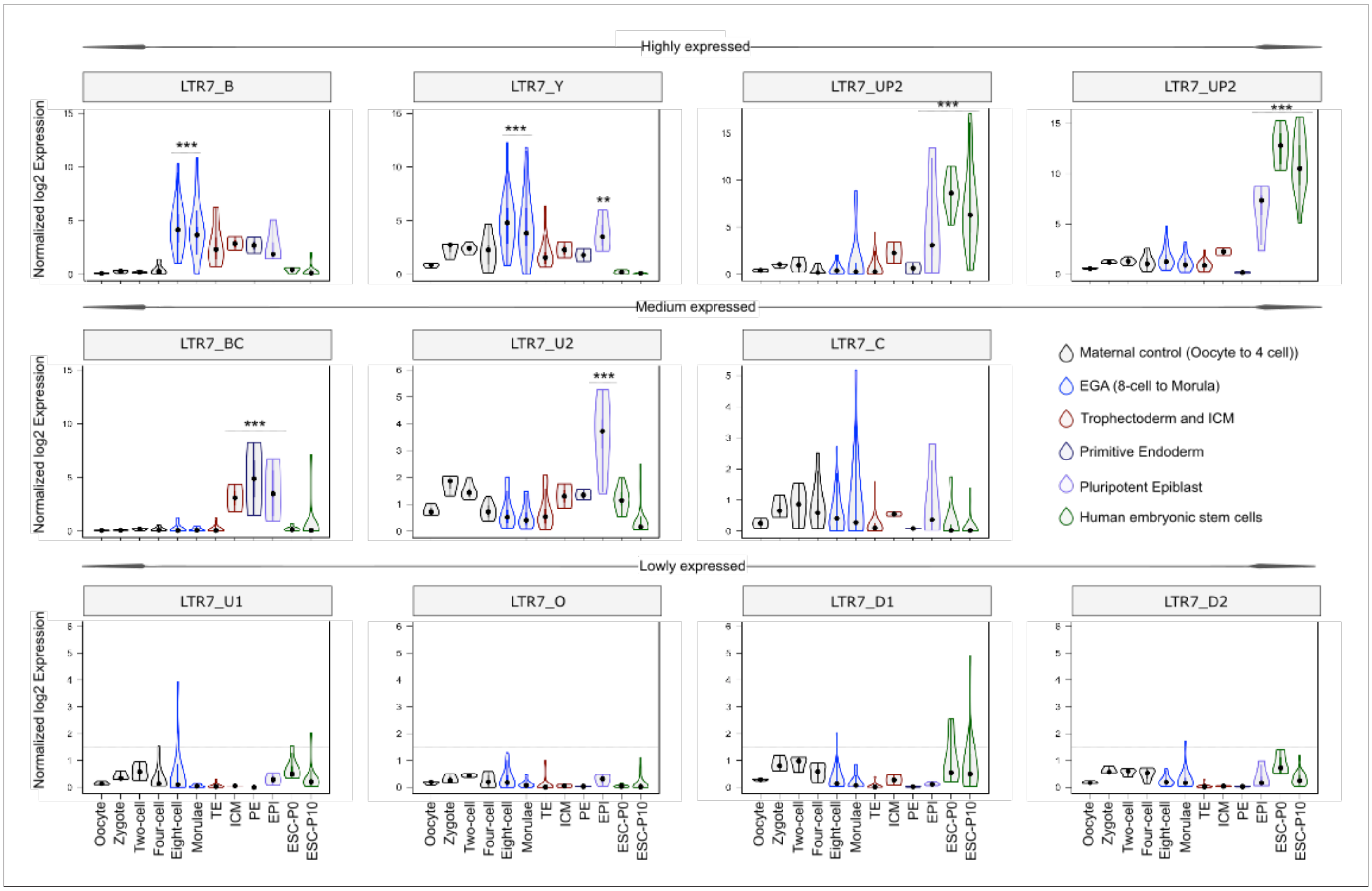

Figure 5. Expression profile of LTR7 subfamilies in human preimplantation embryonic lineages and embryonic stem cells (ESCs). The solid dots and lines encompassing the violins represent the median and quartiles of single cellular RNA expression. The color scheme is based on embryonic stages, defined 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.

The online version of this article includes the following figure supplement(s) for figure 5:

Figure supplement 1. Majority rule consensus sequences used for remasking of human genome.

Figure supplement 2. LTR7 subfamily expression in ‘primed’ and ‘naïve’ cell lines.

As expected, we found that the 7up subfamilies were highly expressed in the pluripotent epiblast and in ESCs (***Figure 5***). 7up expression was highly specific to these pluripotent cell types, with little to no transcription at earlier developmental timepoints. As previously observed (***Göke et al., 2015***), the 7b subfamily exhibited expression at the 8 cell and morula stages, coinciding with EGA (***Figure 5***). Another remarkable expression pattern was that of 7u2 which was restricted to the pluripotent epiblast (***Figure 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 (***Figure 4B–C***). Despite very similar sequence and age (***Figures 1, 2 and 4A***), 7bc and 7o elements show stark contrast in their expression profiles. 7o elements show no significant transcription at any timepoint in early development, while 7bc elements display RNA expression throughout the blastocyst, including trophectoderm and inner cell mass, primitive endoderm, and pluripotent epiblast (***Figure 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 (***Figure 5***). This pattern may reflect the relatively high level of sequence heterogeneity within this subfamily (***Figure 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.

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 ***Figure 4A*** in a pairwise comparison of each subfamily against each of the other subfamilies (see Materials and methods). The results yielded a slew of TF motifs enriched for each subfamily relative to the others (see ***Figure 6A*** for 7up1 and enrichment for all HERVH subfamilies in ***Supplementary files 3 and 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 (***Figure 6A, B***). Note this motif (and insertion) was also present in 7u1, the closest relative to 7up (***Figure 1***), but absent in all other subfamilies (***Figure 6B***).

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 specifically expressed in ESCs (***Figure 6—figure supplement 1A, 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 (***Figure 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 Materials and 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 (***Figure 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 (***Figure 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 (***Figure 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.

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

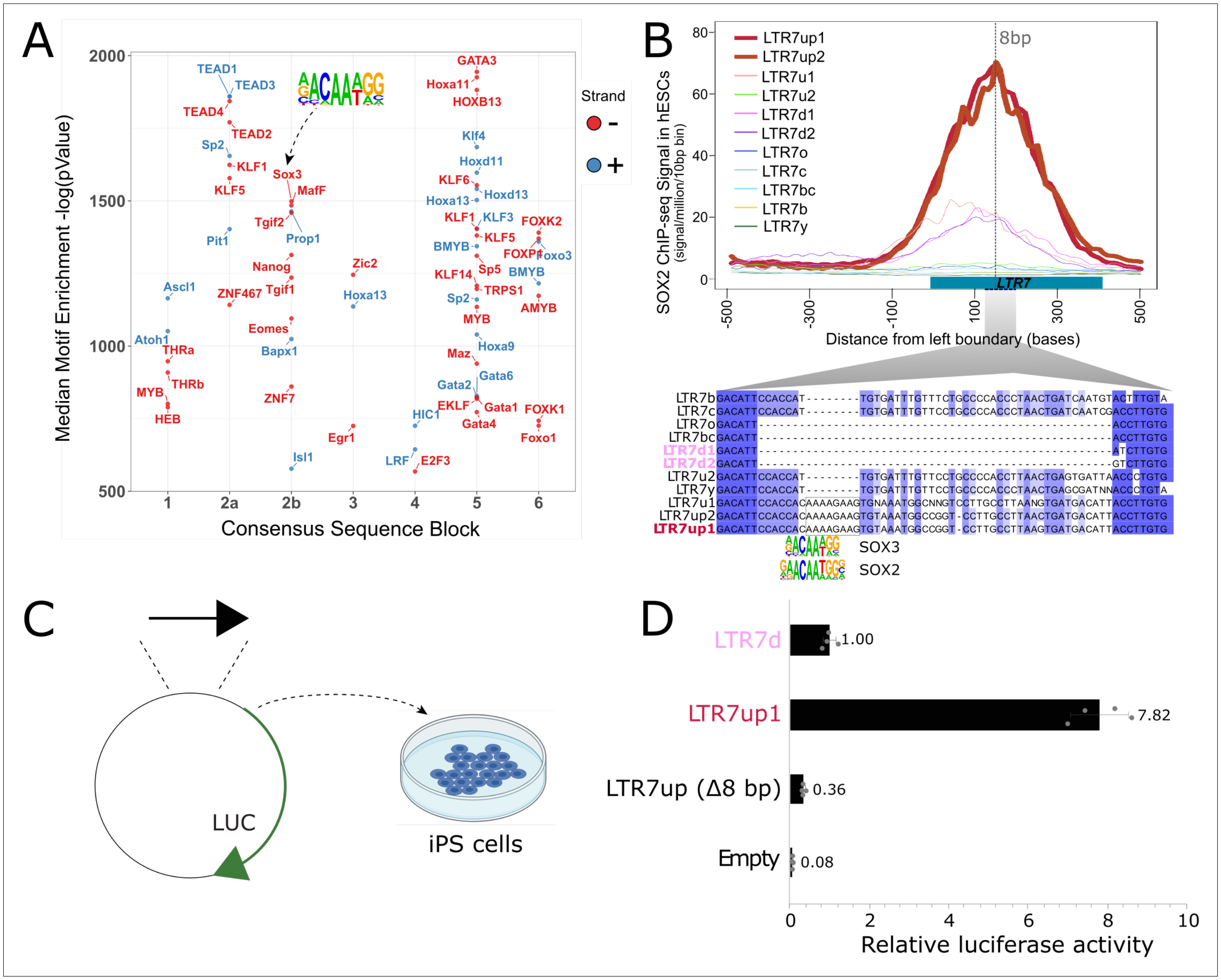

Figure 6. An 8 bp insertion, SOX2/3 binding site necessary for LTR7up transcription. (**A**) (log) p-values > 500 for HOMER motifs enriched in 7up1 insertion's sequence blocks vs. the same blocks from other insertions from other human endogenous retrovirus type-H (HERVH) subfamilies are shown. (**B**) Line plots show SOX2 ChIP-seq signal at LTR7 subfamily loci in human embryonic stem cells (ESCs). Signal from genomic loci was compiled relative to position 0. The 7up/u1 8 bp insertion position is shown with a dotted line. Region 2b harboring SOX2/3 transcription factor binding site (TFBS) is detailed below. (**C**) Scheme of DNA fragments cloned into pGL3-basic vector driving luciferase gene expression (LUC) with identified SOX2/3 motifs. Three constructs were analyzed: Entire LTR7up (7up1), 7d1/2 consensus sequence (approximate ancestral sequence for all LTR7d) and LTR7up with eight 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.

The online version of this article includes the following figure supplement(s) for figure 6:

Figure supplement 1. Transcription factor (TF) expression in the preimplantation embryo and TF binding in embryonic stem cells.

Figure supplement 2. Heatmap showing read pileup from GRO-seq plus strand and SOX2 ChIP-seq on 7up and 7u1 insertions.

Figure supplement 3. Violin plots detailing the distribution of transcription factor (TF) expression shown in ***Figure 6—figure supplement 1*** in early human embryos and embryonic stem cells.

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.

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 among these sequences. Importantly, a phylogeny based on the coding sequence (reverse transcriptase [RVT] domain) of HERVH provided less granularity to separate the subfamilies than the LTR sequences (***Figure 1—figure supplement 1***). The classification of new subfamilies within LTR7 enabled us to discover that they have distinct expression profiles during early embryonic development (***Figure 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 the 7up lineage that is required for promoter activity in pluripotent stem cells. This exact motif is also present in the untranscribed 7u1 subfamily. More distantly related subfamilies also have SOX2/3 binding sites elsewhere on their sequence (***Supplementary file 4***), indicating that the presence of a SOX2/3 site is not sufficient to confer transcriptional activity in ESC. Additionally, the strength of the motif match does not correlate with SOX2 binding or transcription within 7up (***Figure 6—figure supplement 2***). Perhaps SOX2/3 cooperatively bind with other TF to uniquely activate 7up. 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.

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 ERVs (***Vargiu et al., 2016***). Traditional models of recombination describe recombination occurring due to template

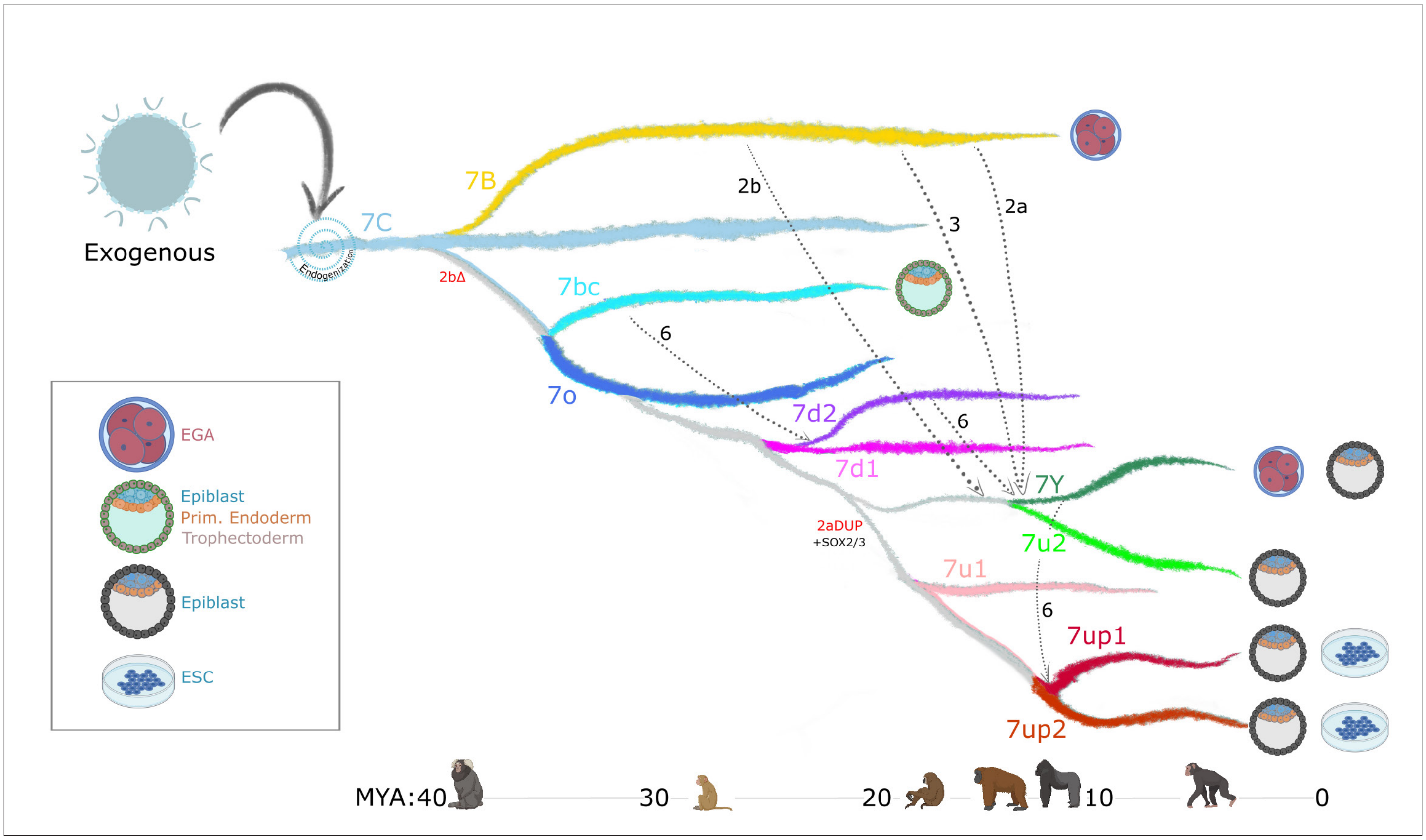

Figure 7. Model of LTR7 subfamily evolution. Estimated LTR7 subfamily transpositional activity in million years ago (mya) is listed with corresponding approximate primate divergence times (bottom). The positioning and duration of transpositional activity are based on analysis from ***Figure 3B***. The gray 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 ***Figure 5A*** are detailed to the right of each corresponding branch.

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 (***Figure 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 (***Figure 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 my of primate evolution (see ***Figure 4D*** and summary model in ***Figure 7***). The coincidence of recombination events with changes in expression profiles (***Figure 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 (***Figures 1 and 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 (*Figure 2*). Thus, it is possible that the formation of new LTR via recombination and their subsequent amplification catalyzed cis-regulatory divergence across primate species.

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 (*Figures 2 and 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 (*Figure 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 (*Figure 3A, D*). Our analysis shows that ZNF534 is most highly expressed in the morula, but dips in human ESC (*Figure 3—figure supplement 2*). 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.

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*). Previous pan-LTR7 knockdown experiments have focused on highly transcribed LTR7 in primed ESCs (*Lu et al., 2014*; *Ohnuki et al., 2014*; *Wang et al., 2012*). As such, these experiments have primarily, if not exclusively, targeted 7up elements. While the role of individual 7up copies in the maintenance of pluripotency of primed cells remains to be clarified (*Takahashi et al.,*

2021), our work highlights potential additional roles of other LTR7 subfamilies in naïve, formative ESCs. 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 (***Figure 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***). Indeed, while 7up shows expression preference in primed cell lines, 7u2 shows preference for naïve lines (***Figure 5—figure supplement 2***). Thus, it might be possible to develop an 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 ERV 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.

Materials and methods

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 >4 kb were combined with and solo LTRs. LTRs > 350 bp were considered for future analysis.

Multiple sequence alignment, phylogenetic tree generation, and LTR7 subdivision

All HERVH LTRs (***Figure 1A*** – ***Supplementary file 5***) or only LTR7s (***Figure 1B*** – ***Supplementary 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-Gutiérrez 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 UFbootstrap 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 (1.5 subs/100 bp) were defined as belonging to a new bona fide LTR7 subfamily (***Figure 1B***).

LTR7 consensus generation and network analysis

Majority rule (51%) was used to generate each LTR7 subfamily at nodes described in ***Figure 1***. Positions without majority consensus are listed as 'N'. Majority rule consensus sequences were aligned with MUSCLE in SEAVIEW (***Supplementary file 7***; ***Edgar, 2004***; ***Gouy et al., 2010***). Alignment was visualized with Jalview2 (***Waterhouse et al., 2009***; ***Figure 4A***) and ggplot2 (***Figure 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***).

RVT domain extraction, alignment, and tree generation

The RVT 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 Materials and methods section).

Peak calling

ChIP-seq datasets representing TFs, histone modifications, and regulatory complexes in human ESCs 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 <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 false discovery rate (FDR) < 1%. ENCODE-defined blacklisted regions were excluded from called peaks. For phyloregulatory analysis (***Figure 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 (***Figure 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).

Phyloregulatory analysis

Peaks from external ChIP-seq datasets were intersected with LTR7 insertions (***Quinlan and Hall, 2010***). LTR7 insertions that intersected with >1 bp of peaks were counted as positive for the respective mark. We repeated this analysis with a range of overlap requirements from extending the LTR 500 bp 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 (***Estarás et al., 2015***) (GSE64758) were created for windows 10 bp upstream and 8 kb downstream of 5' and solo LTR7 (***Ramírez et al., 2016***). The most visible signal was confined to the top 7th of insertions (***Figure 3—figure supplement 1***). 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'.

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) (***Supplementary file 8***). For comparing transcribed 7up to untranscribed 7up, 18 pairwise comparisons were made (***Supplementary file 9***).

Aggregate signal heatmap generation

GRO-seq (H1 cells – GSE64758), whole-genome bisulfite sequencing (WGBS-seq – H1 cells), SOX2 GEO ID GSE125553 (***Bayerl et al., 2021***), and H3K9me3 ChIP-seq (H1 – primed – GSE78099) bams were retrieved from ***Estarás et al., 2015***, ***ENCODE Project Consortium, 2012***, and ***Theunissen et al., 2016***, respectively. Deeptools (***Ramírez et al., 2016***) was used to visualize these marks by LTR7 subfamily division in windows 500 bp upstream and 8 kb downstream of the most 5' position in the LTR (***Figure 3—figure supplement 1***, ***Figure 6—figure supplement 2***).

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 primates (syntenic regions) were then lifted over back to human. Only those that survived both lift-overs (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.

Terminal branch length aging

Terminal branch lengths from the LTR7 phylogenetic tree (**Figure 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 Wilcoxon rank-sum test with Bonferroni multiple testing correction.

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 (**Figure 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 **Figure 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 **Figures 4D and 7**.

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

LTR7up blocks 2a and 2b (**Figure 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’.

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 5—figure supplement 1**). To do this, we identified nodes with >0.95 UFbootstrap 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 two 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 eight subfamilies and evidence of recurrent recombination events caused concern that HERVH LTRs may be misannotated in the RepeatMasker 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 **Figure 4**, and HERVH-int (dfam). We also included annotated consensus sequences from dfam for MER48, MER39, AluYk3, and MST1N2, who we found an 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).

Embryonic HERVH subfamily expression analysis

We downloaded the raw single-cell RNA-seq datasets from early human embryos and ESCs (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, as 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 RNA-seq 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 FDR. All the statistics and visualization of RNA-seq were performed on R (https://www.r-project.org/).

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 matrices were downloaded from HOMER (***Heinz et al., 2010***) and were used to perform pairwise motif enrichment using the command 'homer2 find'. For LTR7up1 enrichment (***Figure 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 1×10^{-5} were kept. For enrichment in all subfamilies (***Supplementary files 3 and 4***) – testing all subfamilies against all others, every pairwise subfamily combination within each block was tested and all results are displayed.

SOX2 ChIP-seq signal on LTR7

SOX2 ChIP-seq and whole-cell extract datasets from primed human ESCs 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*. This analysis was repeated with exclusively uniquely mapping reads which did not affect the outcome of analysis (data not shown). 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 1 kb 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 $r > 0.90$.

SOX3 motif strength correlation with SOX2 binding

All 7up1, 7up2, and 7u1 loci used in other analyses were ranked based on the strength/presence of the SOX3 motif in block 2b. GRO-seq and SOX2 (see above data sources) reads were layered onto these loci in windows 500 bp upstream and 8 kb downstream.

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 (Promega, #E1751). Minipreps were prepared with QIAprep Spin Miniprep kit (Qiagen, #27104). Plasmids were sequenced to ensure the correct sequence and directionality of the insert. Twenty-four hours before transfection, human iPSC WTC-11 (Coriell Institute, GM25256) cells were plated on Vitronectin (Thermo Fisher Scientific, #A14700) coated 12-well plates in Essential 8 Flex medium (Thermo Fisher Scientific, #A2858501) with E8 supplement, Rock inhibitor (STEMCELL Technologies, #72304) 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, #STEM00008) according to manufacturer's instructions. Forty-eight hours after transfection, cell pellet was harvested and luciferase activity was measured with Luciferase Reporter Assay kit (Promega, #E1910) on Glomax (Promega) according to instructions. Transfection efficiency and cell count was normalized with GFP.

1.7down:
GCTAGCTGTCAGGCCTCTGAGCCCAAGCTAAGCCATCATATCCCCTGTGACCTGCACGTA
CACATCCAGATGGCCGGTTCCTGCCTTAACTGATGACATTCCACCACAAAAGAAGTGAAA
ATGGCCTGTTCCTGCCTTAACTGATGACATTATCTTGTGAAATTCCTTCTCCTGGCTCATCCTG
GCTCAAAAGCTCCCCTACTGAGCACCTTGTGACCCCCACTCCTGCCCGCCAGAGAACAAC
CCCCCTTTGACTGTAATTTTCCTTTACCTACCCAAATCCTATAAAACGGCCCCACCCCTATCTC
CCTTCGCTGACTCTCTTTTCGGACTCAGCCCGCCTGCACCCAGGTGAAATAAACAGCTTT
ATTGCTCACACAAAGCCTGTTTGGTGGTCTCTTCACACGGACGCGCATGCTCGAG
2.LTR7upcons:
GCTAGCTGTCAGGCCTCTGAGCCCAAGCCAAGCCATCGCATCCCCTGTGACTTGCACGTA
TACGCCCAGATGGCCTGAAGTAACTGAAGAATCACAAAAGAAGTGAATATGCCCTGCCCC
ACCTTAACTGATGACATTCCACCACAAAAGAAGTGTAAATGGCCGGTCCTTGCCTTAAGT
GATGACATTACCTTGTGAAAGTCCTTTTCCTGGCTCATCCTGGCTCAAAAAGCACCCCCA
CTGAGCACCTTGCGACCCCCACTCCTGCCCGCCAGAGAACAAACCCCCTTTGACTGTAAT
TTTCCTTTACCTACCCAAATCCTATAAAACGGCCCCACCCTTATCTCCCTTCGCTGACTCTCTT
TTCGGACTCAGCCCGCCTGCACCCAGGTGAAATAAACAGCCATGTTGCTCACACAAAGCC
TGTTTGGTGGTCTCTTCACACGGACGCGCATGCTCGAG
5.LTR7upcons_AAAGAAG_deletion:
GCTAGCTGTCAGGCCTCTGAGCCCAAGCCAAGCCATCGCATCCCCTGTGACTTGCACGTA
TACGCCCAGATGGCCTGAAGTAACTGAAGAATCACAAAAGAAGTGAATATGCCCTGCCCC
ACCTTAACTGATGACATTCCACCATTGTAAATGGCCGGTCCTTGCCTTAAGTGATGACAT
TACCTTGTGAAAGTCCTTTTCCTGGCTCATCCTGGCTCAAAAAGCACCCCCACTGAGCAC
CTTGCGACCCCCACTCCTGCCCGCCAGAGAACAAACCCCCTTTGACTGTAATTTTCCTTT
ACCTACCCAAATCCTATAAAACGGCCCCACCCTTATCTCCCTTCGCTGACTCTCTTTTCG
GACTCAGCCCGCCTGCACCCAGGTGAAATAAACAGCCATGTTGCTCACACAAAGCCTGTT
TGGTGGTCTCTTCACACGGACGCGCATGCTCGAG
5'NheI highlighted in yellow
3'XhoI highlighted in cyan

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

(***Figures 2 and 7***) and embryonic cell (***Figures 6 and 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 TAC; and the Cornell Presidential Fellow Program to MS. JLR is supported by the Howard Hughes Medical Institute Faculty Scholar program and NIH (GM099117).

Additional information

Funding

Funder	Grant reference number	Author
National Institutes of Health	GM112972	Cédric Feschotte
National Institutes of Health	HG009391	Cédric Feschotte
National Institutes of Health	GM122550	Cédric Feschotte
Cornell Center for Vertebrate Genomics		Thomas A Carter
Howard Hughes Medical Institute		John L Rinn
National Institutes of Health	GM099117	John L Rinn
Cornell Presidential Fellow Program		Manvendra Singh

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Thomas A Carter, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review and editing; Manvendra Singh, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – review and editing; Gabrijela Dumbović, Jason D Chobirko, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – review and editing; John L Rinn, Cédric Feschotte, Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review and editing

Author ORCIDs

Thomas A Carter http://orcid.org/0000-0001-7081-3259
Manvendra Singh http://orcid.org/0000-0002-8626-5418
Jason D Chobirko http://orcid.org/0000-0001-8495-9152
John L Rinn http://orcid.org/0000-0002-7231-7539
Cédric Feschotte http://orcid.org/0000-0002-8772-6976

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.76257.sa1
Author response https://doi.org/10.7554/eLife.76257.sa2

Additional files

Supplementary files

- Supplementary file 1. List of all LTR7 insertions included in phyloregulatory analysis in hg38 coordinates with accompanying binary regulatory calls.
- Supplementary file 2. OneCodeToFindThemAll.pl calls for remasked LTR7, LTR7B, LTR7C, and LTR7Y (hg38 coordinates) and quantification of subfamily annotation changes between old and new repeat masking.

- Supplementary file 3. Complete statistical support for HOMER motif enrichment for LTR7 subfamilies by sequence block.
- Supplementary file 4. Most enriched HOMER motifs for each human endogenous retrovirus type-H (HERVH) subfamily by sequence block.
- Supplementary file 5. Alignment of all human endogenous retrovirus type-H (HERVH) long terminal repeats (LTRs) used in study (including 7b/c/y).
- Supplementary file 6. Alignment of only LTR7 long terminal repeats (LTRs) used in study.
- Supplementary file 7. Alignment of human endogenous retrovirus type-H (HERVH) long terminal repeats (LTR) consensus sequences.
- Supplementary file 8. Full statistical support for phyloregulatory analysis.
- Supplementary file 9. Statistical support for transcribed vs. non-transcribed LTR7up1/2.
- Transparent reporting form

Data availability

Scripts, data tables, and notes for figures 1-4,6a and figure supplements 1-1,2-1,3-1,4-1,5-1,6-2 - https://github.com/LumpLord/Mosaic-cis-regulatory-evolution-drives-transcriptional-partitioning-of-HERVH-endogenous-retrovirus../ copy archived at swh:1:rev:c5e3c56fdeb74511786be120d262393f18fea185. Scripts and data tables by MS for figures 5,6c and figure supplements 6-1,6-3,5-2 - https://github.com/Manu-1512/LTR7-up copy archived at swh:1:rev:23b4f17bc5c40f2992dcca264ed3b14fed84555b.

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Tsankov AM, Gu H, Akopian V, Ziller MJ, Donaghey J, Amit I, Gnirke A, Meissner A	2015	Transcription factor binding dynamics during human ES cell differentiation	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE61475	NCBI Gene Expression Omnibus, GSE61475
Ji X, Dadon DB, Powell BE, Fan ZP, Borges-Rivera D, Shachar S, Weintraub AS, Hnisz D, Pegoraro G, Lee TI, Misteli T, Jaenisch R, Young RA	2015	3D Chromosome Regulatory Landscape of Human Pluripotent Cells	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE69647	NCBI Gene Expression Omnibus, GSE69647
Pontis J, Planet E, Offner S, Turelli P, Duc J, Coudray A, Theunissen TW, Jaenisch R, Trono D	2019	Hominid-specific transposable elements and KRAB-ZFPs facilitate human embryonic genome activation and transcription in naïve hESCs	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117395	NCBI Gene Expression Omnibus, GSE117395
Imbeault M, Helleboid PY, Trono D	2017	ChIP-exo of human KRAB-ZNFs transduced in HEK 293T cells and KAP1 in hES H1 cells	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE78099	NCBI Gene Expression Omnibus, GSE78099
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	Repeat elements study in pluripotent stem cells	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54726	NCBI Gene Expression Omnibus, GSE54726
Estarás C, Benner C, Jones KA	2015	Wnt3a-Activin A Synergy Induces eRNAPII Pause-Release and Counteracts a Yap1 Elongation Block during hESC Differentiation	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64758	NCBI Gene Expression Omnibus, GSE64758

Continued on next page

Continued

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
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 Signalling Pathway Modulation for Enhancing Human Naïve Pluripotency Induction [ChIP-seq]	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE125553	NCBI Gene Expression Omnibus, GSE125553
Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J, Huang J, Li M, Wu X, Wen L, Lao K, Li R, Qiao J, Tang F	2013	Tracing pluripotency of human early embryos and embryonic stem cells by single cell RNA-seq	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36552	NCBI Gene Expression Omnibus, GSE36552
Gerri C, McCarthy A, Alanis-Lobato G, Demtschenko A, Bruneau A, Loubersac S, Fogarty NME, Hampshire D, Elder K, Snell P, Christie L, David L, Van de Velde H, Fouladi-Nashta AA, Niakan KK	2015	Single-Cell RNA-seq Defines the Three Cell Lineages of the Human Blastocyst	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE66507	NCBI Gene Expression Omnibus, GSE66507

References

Babaian A, Mager DL. 2016. Endogenous retroviral promoter exaptation in human cancer. *Mobile DNA* **7**:24. DOI: https://doi.org/10.1186/s13100-016-0080-x, PMID: 27980689

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: https://doi.org/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: https://doi.org/10.1093/oxfordjournals.molbev.a026036, PMID: 10331250

Bannert N, Kurth R. 2004. Retroelements and the human genome: new perspectives on an old relation. *PNAS* **101 Suppl 2**:14572–14579. DOI: https://doi.org/10.1073/pnas.0404838101, PMID: 15310846

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**:11. DOI: https://doi.org/10.1186/s13100-015-0041-9, PMID: 26045719

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, et al. 2021. Principles of signaling pathway modulation for enhancing human naive pluripotency induction. *Cell Stem Cell* **28**:1549-1565.. DOI: https://doi.org/10.1016/j.stem.2021.04.001, PMID: 33915080

Bergsland M, Ramsköld D, Zaouter C, Klum S, Sandberg R, Muhr J. 2011. Sequentially acting Sox transcription factors in neural lineage development. *Genes & Development* **25**:2453–2464. DOI: https://doi.org/10.1101/gad.176008.111, PMID: 22085726

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 (Cambridge, England)* **142**:3151–3165. DOI: https://doi.org/10.1242/dev.123547, PMID: 26293300

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: https://doi.org/10.1016/j.cell.2005.08.020, PMID: 16153702

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: https://doi.org/10.1146/annurev-genet-112618-043717, PMID: 31518518

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)* **25**:1972–1973. DOI: https://doi.org/10.1093/bioinformatics/btp348, PMID: 19505945

Carbone L, Harris RA, 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, et al. 2014. Gibbon genome and the fast karyotype evolution of small apes. *Nature* **513**:195–201. DOI: https://doi.org/10.1038/nature13679, PMID: 25209798

Chambers I, Smith A. 2004. Self-renewal of teratocarcinoma and embryonic stem cells. *Oncogene* **23**:7150–7160. DOI: https://doi.org/10.1038/sj.onc.1207930, PMID: 15378075

Chang NC, Rovira Q, Wells JN, Feschotte C, Vaquerizas JM. 2021. A Genomic Portrait of Zebrafish Transposable Elements and Their Spatiotemporal Embryonic Expression. *Genomics* **10**:439009. DOI: https://doi.org/10.1101/2021.04.08.439009

Charlesworth B, Langley CH. 1986. THE EVOLUTION OF SELF-REGULATED TRANSPOSITION OF TRANSPOSABLE ELEMENTS. *Genetics* **112**:359–383. DOI: https://doi.org/10.1093/genetics/112.2.359

Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology* **65**:997–1008. DOI: https://doi.org/10.1093/sysbio/syw037, PMID: 27121966

Chuong EB, Elde NC, Feschotte C. 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science (New York, N.Y.)* **351**:1083–1087. DOI: https://doi.org/10.1126/science.aad5497, PMID: 26941318

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews. Genetics* **18**:71–86. DOI: https://doi.org/10.1038/nrg.2016.139, PMID: 27867194

Cordaux R, Hedges DJ, Batzer MA. 2004. Retrotransposition of Alu elements: how many sources? *Trends in Genetics* **20**:464–467. DOI: https://doi.org/10.1016/j.tig.2004.07.012, PMID: 15363897

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: https://doi.org/10.7554/eLife.27746, PMID: 29256862

Cosby RL, Chang NC, Feschotte C. 2019. Host-transposon interactions: conflict, cooperation, and cooption. *Genes & Development* **33**:1098–1116. DOI: https://doi.org/10.1101/gad.327312.119, PMID: 31481535

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. *Nature Communications* **11**:3506. DOI: https://doi.org/10.1038/s41467-020-17206-4, PMID: 32665538

Ecco G, Imbeault M, Trono D. 2017. KRAB zinc finger proteins. *Development (Cambridge, England)* **144**:2719–2729. DOI: https://doi.org/10.1242/dev.132605, PMID: 28765213

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**:1792–1797. DOI: https://doi.org/10.1093/nar/gkh340, PMID: 15034147

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

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**:57–74. DOI: https://doi.org/10.1038/nature11247, PMID: 22955616

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: https://doi.org/10.1016/j.molcel.2015.04.001, PMID: 25936800

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, et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nature Genetics* **41**:563–571. DOI: https://doi.org/10.1038/ng.368, PMID: 19377475

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: https://doi.org/10.1186/s13100-020-00208-w, PMID: 32266012

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nature Reviews. Genetics* **9**:397–405. DOI: https://doi.org/10.1038/nrg2337, PMID: 18368054

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, et al. 2014. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nature Genetics* **46**:558–566. DOI: https://doi.org/10.1038/ng.2965, PMID: 24777452

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, et al. 2013. Derivation of novel human ground state naive pluripotent stem cells. *Nature* **504**:282–286. DOI: https://doi.org/10.1038/nature12745, PMID: 24172903

Gemmell P, Hein J, Katzourakis A. 2015. Orthologous endogenous retroviruses exhibit directional selection since the chimp-human split. *Retrovirology* **12**:172. DOI: https://doi.org/10.1186/s12977-015-0172-6

Gemmell P, Hein J, Katzourakis A. 2019. The Exaptation of HERV-H: Evolutionary Analyses Reveal the Genomic Features of Highly Transcribed Elements. *Frontiers in Immunology* **10**:1339. DOI: https://doi.org/10.3389/fimmu.2019.01339

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, et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science (New York, N.Y.)* **316**:222–234. DOI: https://doi.org/10.1126/science.1139247, PMID: 17431167

Ginestet C. 2011. ggplot2: Elegant Graphics for Data Analysis. *Journal of the Royal Statistical 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 Biology and Evolution* **7**:1432–1454. DOI: https://doi.org/10.1093/gbe/evv081, PMID: 25956794

Göke J, Lu X, Chan YS, Ng HH, Ly LH, 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: https://doi.org/10.1016/j.stem.2015.01.005, PMID: 25658370

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: https://doi.org/10.1006/viro.1993.1535, PMID: 8372448

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: https://doi.org/10.1093/molbev/msp259, PMID: 19854763

Haig D. 2016. Transposable elements: Self-seekers of the germline, team-players of the soma. *BioEssays* **38**:1158–1166. DOI: https://doi.org/10.1002/bies.201600125, PMID: 27604404

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. *Molecular Cell* **38**:576–589. DOI: https://doi.org/10.1016/j.molcel.2010.05.004, PMID: 20513432

Hermant C, Torres-Padilla ME. 2021. TFs for TEs: the transcription factor repertoire of mammalian transposable elements. *Genes & Development* **35**:22–39. DOI: https://doi.org/10.1101/gad.344473.120, PMID: 33397727

Imbeault M, Helleboid PY, Trono D. 2017. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**:550–554. DOI: https://doi.org/10.1038/nature21683, PMID: 28273063

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: https://doi.org/10.1371/journal.pgen.1006883, PMID: 28700586

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: https://doi.org/10.1002/bies.201500096, PMID: 26735931

Jacobs FMJ, 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 ZNF91/93 and SVA/L1 retrotransposons. *Nature* **516**:242–245. DOI: https://doi.org/10.1038/nature13760, PMID: 25274305

Jacques PÉ, Jeyakani J, Bourque G. 2013. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLOS Genetics* **9**:e1003504. DOI: https://doi.org/10.1371/journal.pgen.1003504, PMID: 23675311

Jern P, Sperber GO, Blomberg J. 2004. Definition and variation of human endogenous retrovirus H. *Virology* **327**:93–110. DOI: https://doi.org/10.1016/j.virol.2004.06.023, PMID: 15327901

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**:6325–6337. DOI: https://doi.org/10.1128/JVI.79.10.6325-6337.2005, PMID: 15858016

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. *Journal of Virology* **74**:1234–1240. DOI: https://doi.org/10.1128/jvi.74.3.1234-1240.2000, PMID: 10627533

Johnson WE. 2019. Origins and evolutionary consequences of ancient endogenous retroviruses. *Nature Reviews. Microbiology* **17**:355–370. DOI: https://doi.org/10.1038/s41579-019-0189-2, PMID: 30962577

Kalkan T, Smith A. 2014. Mapping the route from naive pluripotency to lineage specification. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **369**:20130540. DOI: https://doi.org/10.1098/rstb.2013.0540, PMID: 25349449

Kelley D, Rinn J. 2012. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biology* **13**:R107. DOI: https://doi.org/10.1186/gb-2012-13-11-r107, PMID: 23181609

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Research* **12**:996–1006. DOI: https://doi.org/10.1101/gr.229102, PMID: 12045153

Khan H, Smit A, Boissinot S. 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Research* **16**:78–87. DOI: https://doi.org/10.1101/gr.4001406, PMID: 16344559

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.. DOI: https://doi.org/10.1016/j.stem.2020.11.005, PMID: 33271069

Kojima KK. 2018. Human transposable elements in Repbase: genomic footprints from fish to humans. *Mobile DNA* **9**:2. DOI: https://doi.org/10.1186/s13100-017-0107-y, PMID: 29308093

Kong Y, Rose CM, Cass AA, Williams AG, Darwish M, Lianoglou S, Haverty PM, Tong AJ, 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. *Nature Communications* **10**:5228. DOI: https://doi.org/10.1038/s41467-019-13035-2, PMID: 31745090

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: https://doi.org/10.18632/oncotarget.9255, PMID: 27172897

Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics* **42**:631–634. DOI: https://doi.org/10.1038/ng.600, PMID: 20526341

Lai MM. 1992. RNA recombination in animal and plant viruses. *Microbiological Reviews* **56**:61–79. DOI: https://doi.org/10.1128/mr.56.1.61-79.1992, PMID: 1579113

Lanciano S, Cristofari G. 2020. Measuring and interpreting transposable element expression. *Nature Reviews. Genetics* **21**:721–736. DOI: https://doi.org/10.1038/s41576-020-0251-y, PMID: 32576954

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, et al. 2011. Comparative and demographic analysis of orang-utan genomes. *Nature* **469**:529–533. DOI: https://doi.org/10.1038/nature09687, PMID: 21270892

Loewer S, Cabili MN, Guttman M, Loh YH, 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. *Nature Genetics* **42**:1113–1117. DOI: https://doi.org/10.1038/ng.710, PMID: 21057500

Löytynoja A, Goldman N. 2010. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics* **11**:579. DOI: https://doi.org/10.1186/1471-2105-11-579, PMID: 21110866

Lu X, Sachs F, Ramsay L, Jacques PÉ, Göke J, Bourque G, Ng HH. 2014. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nature Structural & Molecular Biology* **21**:423–425. DOI: https://doi.org/10.1038/nsmb.2799, PMID: 24681886

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: https://doi.org/10.1038/nature11244, PMID: 22722858

Mager DL, Freeman JD. 1987. Human endogenous retroviruslike genome with type C pol sequences and gag sequences related to human T-cell lymphotropic viruses. *Journal of Virology* **61**:4060–4066. DOI: https://doi.org/10.1128/JVI.61.12.4060-4066.1987, PMID: 2446010

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: https://doi.org/10.1006/viro.1995.0012, PMID: 7491764

Matsuda E, Garfinkel DJ. 2009. Posttranslational interference of Ty1 retrotransposition by antisense RNAs. *PNAS* **106**:15657–15662. DOI: https://doi.org/10.1073/pnas.0908305106, PMID: 19721006

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 Biology* **21**:1–25. DOI: https://doi.org/10.1186/s13059-020-02164-3, PMID: 32988383

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 Research* **24**:697–707. DOI: https://doi.org/10.1101/gr.159624.113, PMID: 24501022

Nakamura T, Yamada KD, Tomii K, Katoh K. 2018. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics (Oxford, England*) **34**:2490–2492. DOI: https://doi.org/10.1093/bioinformatics/bty121, PMID: 29506019

Nichols J, Smith A. 2009. Naive and primed pluripotent states. *Cell Stem Cell* **4**:487–492. DOI: https://doi.org/10.1016/j.stem.2009.05.015, PMID: 19497275

Niwa H. 2007. How is pluripotency determined and maintained? *Development (Cambridge, England*) **134**:635–646. DOI: https://doi.org/10.1242/dev.02787, PMID: 17215298

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: https://doi.org/10.1186/s12862-016-0755-4, PMID: 27582319

Ohnuki M, Tanabe K, Sutou K, Teramoto I, Sawamura Y, Narita M, Nakamura M, Tokunaga Y, Nakamura M, 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: https://doi.org/10.1073/pnas.1413299111, PMID: 25097266

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**:417–419. DOI: https://doi.org/10.1038/nmeth.4197, PMID: 28263959

Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D, Knowles BB. 2004. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Developmental Cell* **7**:597–606. DOI: https://doi.org/10.1016/j.devcel.2004.09.004, PMID: 15469847

Pérez-Losada M, Arenas M, Galán JC, Palero F, González-Candelas F. 2015. Recombination in viruses: mechanisms, methods of study, and evolutionary consequences. *Infection, Genetics and Evolution* **30**:296–307. DOI: https://doi.org/10.1016/j.meegid.2014.12.022, PMID: 25541518

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.. DOI: https://doi.org/10.1016/j.stem.2019.03.012, PMID: 31006620

Posada D, Crandall KA. 2001. Intraspecific gene genealogies: trees grafting into networks. *Trends in Ecology & Evolution* **16**:37–45. DOI: https://doi.org/10.1016/s0169-5347(00)02026-7, PMID: 11146143

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**:841–842. DOI: https://doi.org/10.1093/bioinformatics/btq033, PMID: 20110278

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: https://doi.org/10.1093/nar/gkw257, PMID: 27079975

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 (Oxford, England)* **30**:1003–1005. DOI: https://doi.org/10.1093/bioinformatics/btt637, PMID: 24227676

Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annual Review of Genetics* **46**:21–42. DOI: https://doi.org/10.1146/annurev-genet-110711-155621, PMID: 22905872

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* **25**:49–58. DOI: https://doi.org/10.1016/j.coviro.2017.07.001, PMID: 28750248

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: https://doi.org/10.1016/j.stem.2016.12.004, PMID: 28061351

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: https://doi.org/10.1186/1742-4690-9-111, PMID: 23253934

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, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**:169–175. DOI: https://doi.org/10.1038/nature10842, PMID: 22398555

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: https://doi.org/10.1006/viro.2000.0712, PMID: 11145909

Simon-Loriere E, Holmes EC. 2011. Why do RNA viruses recombine? *Nature Reviews. Microbiology* **9**:617–626. DOI: https://doi.org/10.1038/nrmicro2614, PMID: 21725337

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

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: https://doi.org/10.1186/s13100-020-00230-y, PMID: 33436076

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 Research* **24**:1963–1976. DOI: https://doi.org/10.1101/gr.168872.113, PMID: 25319995

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 of London. Series B, Biological Sciences* **375**:20190347. DOI: https://doi.org/10.1098/rstb.2019.0347, PMID: 32075564

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: https://doi.org/10.1016/j.ydbio.2004.01.028, PMID: 15081373

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: https://doi.org/10.1371/journal.pgen.1009587, PMID: 34033652

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: https://doi.org/10.1016/j.stem.2010.03.015, PMID: 20452321

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, et al. 2016. Molecular Criteria for Defining the Naive Human Pluripotent State. *Cell Stem Cell* **19**:502–515. DOI: https://doi.org/10.1016/j.stem.2016.06.011, PMID: 27424783

Thomas J, Perron H, Feschotte C. 2018. Variation in proviral content among human genomes mediated by LTR recombination. *Mobile DNA* **9**:36. DOI: https://doi.org/10.1186/s13100-018-0142-3, PMID: 30568734

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: https://doi.org/10.1016/j.molcel.2016.03.029, PMID: 27259207

Urusov FA, Nefedova LN, Kim AI. 2011. Analysis of the tissue- and stage-specific transportation of the *Drosophila melanogaster* gypsy retrotransposon. *Russian Journal of Genetics* **1**:507–510. DOI: https://doi.org/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: https://doi.org/10.1186/s12977-015-0232-y, PMID: 26800882

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: https://doi.org/10.1016/j.stem.2012.02.016

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: https://doi.org/10.1038/nature13804, PMID: 25317556

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics (Oxford, England*) **25**:1189–1191. DOI: https://doi.org/10.1093/bioinformatics/btp033, PMID: 19151095

Waterson RH, Lander ES, Wilson RK. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**:69–87. DOI: https://doi.org/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: https://doi.org/10.7554/eLife.56337, PMID: 32479262

Yang P, Wang Y, Macfarlan TS. 2017. The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. *Trends in Genetics* **33**:871–881. DOI: https://doi.org/10.1016/j.tig.2017.08.006, PMID: 28935117

Yu HL, Zhao ZK, 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: https://doi.org/10.3892/ijmm.2013.1460, PMID: 23900638

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, et al. 2019. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nature Genetics* **51**:1380–1388. DOI: https://doi.org/10.1038/s41588-019-0479-7, PMID: 31427791