Pluripotency and the endogenous retrovirus HERVH: Conflict or serendipity?

Izsvák, Z., Wang, J., Singh, M., Mager, D.L., Hurst, L.D.

This is the final version of the accepted manuscript, which was published in final edited form in:

Bioessays
2016 JAN ; 38(1): 109-117
2015 DEC 28 (first published online)
doi: 10.1002/bies.201500096

Publisher: Wiley

Pluripotency and the endogenous retrovirus HERVH: conflict or serendipity?

Zsuzsanna Izsvák[1], Jichang Wang[1], Manvendra Singh[1], Dixie L. Mager[2], Laurence D. Hurst[3]

[1]Max-Delbrück-Center for Molecular Medicine (MDC), Robert-Rössle-Strasse 10, 13125 Berlin, Germany.
[2] Terry Fox Laboratory, British Columbia Cancer Agency and Dept. of Medical Genetics, University of British Columbia, 675 West 10th Avenue, Vancouver BC, Canada V5Z 1L3.
[3]University of Bath, Department of Biology and Biochemistry, Bath, Somerset, UK, BA2 7AY.

Keywords: endogenous retrovirus, ERV, host defence, LBP9, naïve, pluripotency, primate,

Summary/abstract

Remnants of ancient retroviral infections during evolution litter all mammalian genomes. In modern humans, such endogenous retroviral (ERV) sequences comprise at least 8% of the genome. While ERVs and other types of transposable element undoubtedly contribute to the genomic "junk yard", functions for some ERV sequences have been demonstrated, and there is growing evidence that ERVs can be important players in gene regulatory processes. Here we review one particular large family of human ERVs, termed HERVH, which several recent studies suggest play a key regulatory role in human pluripotent stem cells. Remarkably, this is not the first instance of an ERV controlling pluripotency. We speculate as to why this convergent evolution might have come about, suggesting that it may reflect selection on the virus to extend the time available for transposition. Alternatively it may reflect serendipity alone.

Introduction

Over the course of evolution there have been multiple invasions of mammalian genomes by different retroviruses. The most "successful" retroviruses can acquire an intracellular lifestyle and become endogenous retroviruses (ERVs). During the endogenization process, retroviruses gradually lose their ability to leave their host cells, and function as transposable elements. Similarly to other retrotransposons, ERVs use a ''copy-and-paste'' mechanism: that facilitates their increase in copy number if the duplication events occur within the germline [1]. Similarly to other transposable elements (TEs), the vast majority of ERVs ultimately become inactive for transposition and, in the modern human genome none appears to retain the ability to retrotranspose. At first glance, the remnants of previous retroviral invasions seem to be prime candidates for "functionless" DNA. That so much of our genome is indeed retroviral in various states of decay is consistent with the view that our genome is large not because everything is functional, but rather that selection on removal of "dead" sequence is too weak to be effective. This does not, however, preclude the possibility that some insertions might *become* functional. In this

review we address recent evidence that one family of ERVs, HERVH, might have evolved functionality associated with pluripotency. Given that this is not the first time that an ERV has assumed a function regulating cell fate decisions (MERVL has a role in murine totipotency) [2], we ask how such convergence might be explained.

HERVH is an abundant primate-specific ERV family

ERVs in humans can be classified into many different families or groups, and intact members have the expected structure of an integrated retrovirus: namely the canonical retroviral genes, *gag*, *pol* and *env,* flanked by long terminal repeats (LTRs), which contain all the transcriptional regulatory signals necessary for retroviral expression. HERVH is a primate-specific ERV family discovered in the human genome in 1984 [3]. Most members of this family of HERVs have a Histidine tRNA primer-binding site, hence the name, HERV(H). Phylogenetic studies place it in the Class I/Type C or gamma-retroviral group [4-6]. While a low number of HERVH-like elements occur in New World monkeys, the major expansion of this family occurred in the Old World branch (Fig. 1), and most HERVH insertions in human have orthologous loci in Old World monkeys such as *Rhesus macaque* [7, 8]. Thus, over 80% of HERVH integrations into the human genome occurred within the last 30MY [9]. Notably, a common, partially deleted form, which amplified to several hundred copies in Old World monkeys, and which is associated primarily with LTRs annotated as LTR7 or LTR7B (originally termed Type I and Type II LTRs, [7, 10]), was estimated by genomic Southern analysis to be present in no more than 50 copies in two species of New World monkeys [8]. A later expansion in hominoids of approximately 100 elements with a variant LTR (termed Type Ia and annotated LTR7Y in Repbase [11], has also been documented [7, 10]. In comparison with LTR7 or LTR7B, LTR7Y is represented by fewer copies in the human genome, and has higher promoter activity in transfection assays [10, 12]. Unlike with HERVK-HML2 (the youngest family of HERVs) [13], there is no evidence of HERVH integrations after the divergence of human and chimpanzee. Such studies and the calculated divergence between related elements indicate that the insertional activity of HERVH ceased over 10 million years ago [9].

While ~100 elements are close to full-length, with a full complement of retroviral genes, all have numerous mutations or deletions, and there is no evidence of replication competence [6, 14, 15]. A few copies have an open reading frame (ORF) coding for the immunosuppressive domain of the *envelope* gene, but no evidence of protein production has been published [16]. Indeed, most elements with two LTRs are roughly 5.7 kb in size and share several common deletions [15], suggesting that this partly deleted form became the favoured substrate for retrotransposition using proteins provided by related full-length elements. HERVH activity over time has generated a total of ~2000 copies (including solitary LTRs), which are fixed in the modern human genome. Solitary LTRs are generated by homologous recombination between the close-to identical LTR sequences, and not by retrotransposition. The number of solitary HERVH LTRs (~1000) roughly equals the number of full-length or partly full-length elements. Such a ratio is highly unusual among ERV families, which typically have much higher numbers of solitary LTRs, due to the

tendency for recombination between 5' and 3' LTRs of proviruses over evolutionary time [17](Fig. 2). The reason for this unusual ratio is unclear, but it is tempting to speculate that maintaining full-length HERVH elements is selectively advantageous, possibly because HERVH RNA has a function.

Several transposable element (TE) families are transcriptionally active during early embryonic development

The global epigenomic reprogramming that takes place during early development is associated with massive transcriptional reactivation of TEs [18]. TEs exhibit a developmental stage-specific pattern that might be taken as hallmarks for certain stages of early development. MERVL-associated transcripts expressed at 2-cell stage are thought to be crucial for totipotency [19] in mice [2]. Reactivation of MERVL in mouse induces many 2-cell-specific transcription products. Many of these transcripts are directly promoted by LTRs of MERVL or related MaLR elements, or are chimeric transcripts derived from MERVL/MaLR LTRs. As in mice, TE reactivation occurs in well-defined waves during human embryogenesis, still involving a distinct set of TEs [19, 20] (Fig. 3A). At morula and blastocyst stages most of the TE-derived transcripts derive from HERVH [21] (Fig. 3). Notably, HERVH elements are driven by several variants of their LTRs, the large LTR7 group. While LTR7B-HERVH elements peak at the eight-cell stage, LTR7-HERVHs are expressed in the blastocyst, and also in early passage, *in vitro* embryonic stem cell cultures [20] (Fig. 3). The youngest subset, LTR7Y-HERVHs, are expressed from the eight-cell to blastocyst-stage embryos (Fig. 3). SVA and Line1 (L1) non-LTR elements, some still capable of retrotransposition in the human genome [22], are also reactivated during early development. The human specific SVA elements (SVA-D, F, F) are expressed at eight-cell stage, and peak in the morula, but are still expressed in the blastocyst [23]. The transcription level of primate-specific L1s reflects their evolutionary age: the younger L1s are more highly expressed than the older ones [23, 24]. The L1HS and L1PA2 (< 7.5 MYA old) subfamilies are expressed at the blastocyst stage [25, 26].

TEs harbour key developmental transcription factor binding sites

The most successful ERVs (in terms of copy number) are those that can replicate/be transcribed in pluripotent stem cells, which eventually allows for germ line transmission and fixation. A subset of LTR7s of HERVH is a very potent promoter/enhancer in human pluripotent stem cells (hPSCs). In these cells, around 400 of the LTR7 driven HERVH loci are transcriptionally activated [21]. HERVH activation is inversely correlated with the DNA methylation status [21, 27], and the activated copies are marked with transcriptionally active histone marks (H3K4me1/2/3, H3K9ac, H3K36me3 and H3K79me2), while the repressive marks (H3K9me3 and H3K27me3) are rare [21, 28, 29].

In order to drive transcription in pluripotent stem cells, TEs should contain transcription factor (TF) binding sites for factors expressed in these cells. The genomic expansion of HERVH over time

was likely facilitated by the presence of core key pluripotent TF binding sites in the LTR. The pluripotency factors NANOG, OCT4, KLF4 and LBP9/Tfcp2l1 all bind to active LTR7s of HERVH and drive transcription of HERVH-derived transcripts [21, 28-30] (Fig. 4). Furthermore, active HERVHs are also enriched for binding sites of the histone acetyl transferase P300, the pluripotency regulators/modifiers CHD1 and Myc/Max [21, 31-33]. Identification of these TF binding sites in a subset of LTR7 sequences indicates that they were likely present upon insertion or gained shortly after initial invasion. The high copy number of HERVH LTRs provides a large pool of stem-cell specific promoters/enhancers distributed in the genome.

There are no species with identical repertoire of TEs. Thus, in contrast to the heavily conserved TF binding motifs in the genome, TF binding sites distributed by TEs have the potential to create species-specific regulatory networks. Indeed, LTR7/HERVH provides an intriguing platform for a special combination of pluripotency-specific TF binding sties (OCT4, NANOG, KLF4, LBP9/Tfcp2l1) in primates [34]. In comparison, these key pluripotent TF binding sites are physically not clustered in the mouse genome [35].

In this context, the involvement of one TF, LBP9/Tfcp2l1, is especially interesting. The role of LBP9, in association with HERVH, is to support self-renewal, while transcriptional down-regulation of either LBP9 or HERVH potentiates differentiation [21]. As a key factor in the activity of HERVH in early embryos, LBP9 functions as a switch that regulates HERVH. In mice, the knockdown phenotype of Tfcp2l1 is less severe, and does not affect self-renewal, but rather differentiation potential, suggesting that Tfcp2l1 has a slightly different function in human and mouse [35, 36]. In primates, the HERVH invasion has emerged as a novel challenge for the host defence to deal with. While LBP9 is associated with non-AUG codon usage [37] and pluripotency [35, 36] in mammals, it seems to be engaged with ERV regulation uniquely in primates [21]. Perhaps, LBP9 in conjunction with other members of the LBP/CP2 family [38] has been adopted to defend the host against retroviruses, and now also functions as a repressor of HERVH. If so, then HERVH was recruited to the pluripotency network by modification of a pluripotency factor detailed to defend the cell against it. Consistent with this notion, LBP9, a member of a CP2 family of transcription factors, can form heterodimer complexes with other family members, and LBP9 functions either as a transcriptional activator or a repressor, depending upon the binding partner [21, 39]. Furthermore, LBP9 expression shows specific spatio-temporal regulation, which - combined with its ability to bind LTRs (hence the name LBP9, LTR-binding protein 9) - would have made recruitment of HERVH to the pluripotency network more likely.

The LTR7/HERVH driven transcription adds a primate-specific feature to pluripotency

Simply as a result of their high abundance in human pluripotent stem cells, HERVH-derived transcription products are expected to influence the transcriptome markedly. This feature also opens the possibility that they may affect basic cellular functions. In fact, LTR7/HERVH can

influence the regulation of pluripotency in humans in many ways [21] (Fig. 4). First, activation of LTR7-driven transcription can modulate or seize control from cellular promoters/enhancers. Second, LTR7-driven HERVHs can contribute to new transcriptional units by providing alternative splicing or polyadenylation signals. Nearly 10% of the transcripts driven from HERVH LTRs are annotated as lncRNA [21, 33], which can involve non-HERVH sequences [28]. Strikingly, a subset of the most highly expressed lncRNA transcripts shares conserved core domain (CD) [21]. This domain is predicted to bind RNA-binding proteins, pluripotency factors and pluripotency-associated histone modifiers. One such lncRNA, linc00458, is indeed a known binding partner of SOX2 [40]. LTR7-driven transcription results in generation of canonical HERVH RNAs, but also in production of "chimeric" mRNAs, defined here as those transcripts involving exons of known coding genes [21]. Such chimeric transcripts are not expressed in pluripotent cells from their "native" promoter, and may have novel cellular functions in pluripotency.

An unusual case is the gene, ESRG [41], which is possibly a *de novo* protein-coding gene [21] (Fig. 5). Most of ESRG is derived from HERVH, but its 3' exons capture what is otherwise intronic sequence in the host gene within which ESRG is nested. One putative ORF (80 amino acids long) in this gene has the ATG donated from the 3' end of the LTR7 of HERVH that inserted post the common ancestor with orangutan [21]. Although ESRG is present even in our closest relatives, it might have this coding potential only in human [21]. The ESRG protein appears to have an interacting partner COXII, which is implicated in a pro-apoptotic function in human embryonic stem cells [42]. Besides ESGR, several of the LTR7/HERVH derived transcripts are confirmed to have a role in human pluripotency. The set includes linc-ROR and linc00458 [40, 43] or the chimeric gene, SCGB3A2 [21, 30].

During early embryogenesis the pluripotency factors seem not to be simply activated or silenced, but are expressed in a dynamic range. The epigenetic state of LTR/ERVs is characteristically labile, resulting in a variegation of their expression. While LTR7/HERVHs-derived transcripts inhibit differentiation in pluripotent stem cells, they also poise the cells to diversify their transcriptome prior to differentiation. The LTR/HERV-modulated regulation would provide a more colourful palette of options when differentiation initiates compared with a regular enhancer/promoter. Thus, the LTR7/HERVH-based circuitry seems especially suited to regulate a pluripotency network, and may advantageously contribute to the stochastic nature of the differentiation process [44].

HERVH can define naïve-like cells

During mammalian development the epiblast cells in the inner cell mass (ICM) of blastocysts enter the developmental "ground state", and these pluripotent cells can give rise to all somatic lineages and the germline [45]. Though the ground state pluripotency in the embryo is a transient condition, in principle its nature can be captured indefinitely *in vitro,* through derivation of

embryonic stem cells (ESCs) from the ICM [46, 47]. Naturally, one would expect that a mechanism as important as pluripotency would be conserved in different species of mammals. However, that does not seem to be the case. Thus far, long-term, ground state naïve cell cultures could be successfully established only in mice and rats. Mouse naïve cultures express naïve transcription factors fairly homogenously, and maintain their resemblance to the inner cell mass (ICM) in the long term. Like their mouse counterparts, human ESCs (hESCs) can be also differentiated into three germ layers both *in vitro* and *in vivo* [48]. However, hESCs are more similar to primed mouse epiblast stem cells than to mESCs [24, 25, 49]. Most problematically, the expression profiles of hESC lines are heterogeneous, and are further from those of cells in the ICM.

Over the past decade considerable efforts have been devoted generating or identifying the 'holy grail' of the field, the human ground state, naïve stem cells. In the last few years several naïve-like lineages have been generated, mostly by overexpressing key TFs, and/or improving culture conditions using certain chemical compounds [50-55]. A recent study implicates HERVH as having a potential role in defining some forms of naïve-like stem cells [21]. Notably, an LTR7/HERVH-based GFP reporter system can be used to track and enrich human pluripotent cultures that express HERVH [21] (Fig. 4). These HERVH marked cells resemble the very early, pluripotent state, forming dome shaped colonies, and they uniformly express naïve pluripotent markers, similarly to mouse naïve cells. Importantly, such cells are naturally present in almost all pluripotent stem cells cultures (hESCs and induced pluripotent stem cells, hiPSCs). We have observed that in general around 4% of human PSC cultures exist in HERVH driven green-expressing state. Such cells highly expressing HERVH are currently one of the best available models of ICM in human ESCs [21].

In addition to LTR7, other LTRs also drive expression in pluripotent stem cells. Among these, LTR7Y or LTR5_Hs (HERVK) could also be used as alternative reporters to LTR7, and perhaps mark hPSCs that might be even closer to ICM [19, 20]. Some LTR7-driven transcripts appear to be crucial for human pluripotency, and enriching cells that express the LTR7-diven transcripts are required for maintaining pluripotency. However, it is not clear whether LTR7Y or LTR5_Hs driven transcripts have any cellular function in pluripotency. Thus, future research would need to clarify whether enriching cultures for transcripts derived from LTR7Y or LTR5_Hs would be beneficial for the quality of the cells. Nevertheless, while large numbers of LTR proviruses are expressed in early embryos, it is highly improbable that all of the expressed elements participate in developmental processes. Rather, almost certainly, most are probably passengers that persist because they do no harm.

LTR7/HERVH appears to have evolved a biological function in the acquisition and maintenance of human pluripotency. A specific HERVH expression pattern characteristic of human pluripotent stem cells recapitulates features of the ICM of blastocyst. High-levels of HERVH expression not only mark cells in a naïve-like state, but apparently play a role in maintenance of this state, while

inhibiting differentiation *in vitro* (Fig. 4). However, in hESCs, both the number and the intensity of HERVH-expression are higher in cultured conditions when compared to ICM. In fact, too high a level of HERVH expression in the ICM may inhibit differentiation. Indeed, high level of expression from LTR7/HERVH is associated with a reversible differentiation failure phenotype in human induced pluripotent stem cells (hiPSCs) [56]. However, by tuning down HERVH expression the cells differentiate normally [57]. The lack of diapause behaviour in human embryos suggests that hyperactivation of HERVH/LTR7 after the ICM stage would not be adaptive.

Reporters, such as the LTR7-GFP system, are powerful tools not only in derivation, but also in optimization of naïve-like hPSC culture conditions (Wang et al. *in press*). The next, so far unresolved, task is to maintain these cells in long-term cultures that allow for their proliferation.

How to define human ground state naïveté?

What is a naïve stem cell? The central issue here is what precise properties ones considers as adequate to define cells as truly naïve. The stem cell research community often appears to define human naïvety based on mouse criteria [51-55]. Perhaps the most controversial experiments to demonstrate developmental potential are those involving generation of inter-species 'chimeras'. In this strategy, naïve-like cells (e.g. human) are transplanted into the blastocyst of a different species (e.g. mouse), and the fate of the resulting chimera is monitored. In fact, in light of species-specific differences, it is not clear at all how to interpret the hard to replicate (and ethically questionable) inter-species chimaera experiments [58]. Moreover, one should keep in mind that the HERVH-based regulatory network has perhaps exclusive, human-specific, features (e.g. ESRG) that might rule out the chimeric strategy even in primates.

The optimal approach to characterize naïve human cultures would be to avoid the issue of inter-species differences. We suggest an alternative metric: comparing human naïve-like stem cells with the expression pattern of the ICM rather than with mouse naïve cells.

As cells expressing HERVH at a high level cluster near to ICM, HERVH^{high} cells are one of the best current models of human naïve-like status [21]. While LTR7/HERVH appears to have a key biological function in the acquisition and maintenance of human pluripotency, single-cell sequencing data reveal that human ESCs express HERVH at various levels. Thus, human ESCs seem to be a mixture of cells with features of either mouse or primate cells, depending on the expression level of the HERVH-based regulatory network. Intriguingly, approaches to mimicking mouse naïve cells result in naïve-like cultures where the mouse pluripotent features appear as dominant, while the primate ones are suppressed. In fact, global expression profiling identifies two major types of human naive-like ECSs. While one is more similar to mouse naïve cells, HERVH^{high} cells appear to be more primate/human specific [21] (Fig. 6).

Although the HERVH-based regulatory circuitry cannot be completely switched off without compromising pluripotency in human, recent studies suggest that there might be alternative pluripotent states, and perhaps several paths to reach naïve-like pluripotency [59-61]. Notably, such alternative pluripotent states may or may not functionally mimic different phases of pluripotency *in vivo*. Nevertheless, defining naïve cells by their transcriptomic similarity to ICM should consider alternativeness, reflecting a certain degree of heterogeneity within the inner cell mass.

ERV involvement in pluripotency: Serendipity or conflict?

ERVs seem to govern numerous developmental processes [18, 62], often in a clade-specific manner [26]. It is striking that involvement of ERVs in pluripotency is a feature that has evolved convergently more than once [2, 21]. Why might this be?

It could simply be serendipity. Gene expression is relatively easily activated in early embryos, prior to chromatin shut down associated with differentiation and cell specialization. In principle, any potential target site for transcription factors holds the potential for expression, and from there, to possible incorporation into a regulatory network. As Francois Jacob put it, “evolution is a tinkerer”. If so, incorporation of ERVs into a circuitry is more a matter of chance than anything else. The fact that ERVs contain transcription factor binding sites within their LTRs would be consistent with such a model, predisposing them to being expressed -- possibly by accident in the first instance -- although the involvement of LBP9 might, as we suggested, reflect initial suppression mechanisms. Such a model might also in part explain why some ERVs are co-opted in placental and adopt novel functions (e.g. syncytins [63]). Note that placenta too has unusual chromatin environment with extensive demethylation.

Alternatively, we hypothesise that it may be selectively advantageous for a TE, such as an ERV, to control and extend the pluripotent state. To be able to spread in a population and invade a genome, a TE needs to transpose in the germ line. However, such transposition comes at a cost to the host -- a cost outweighed by the increased frequency of the TE in the population. If so, we might conjecture there to be a genetic conflict of interests. To picture this, imagine that in germ line there arises a state where transposition is restricted (for example in males the X is inactivated at some point). To limit potentially deleterious transposition events, the “interest” of the host is to push this "non transposing stage" backwards in time, closer to the zygote stage. By contrast, the “interests” of a TE might be to keep the window of opportunity open as long as possible. Put differently, successful TEs, with the largest genomic copy numbers, might be those that can extend the phase of transposition competency. So how to do this? One possibility is for ERVs to force the cell to be naïve as long as possible. In this model there is in effect a tug of war between host and ERV for control of timing of the phase of when transposition is possible. Such a tug of war could also result in increased numbers of ERV variants that overcome silencing mechanisms by the host. In the case of HERVH, the most recent LTR subtype, LTR7Y, is thought to have arisen via

recombination between older LTR7 and LTR7B elements [7]. As mentioned above, this resulted in the most recent expansion of HERVH elements in hominoids, and such an expansion may have been facilitated by a subtle shift in TF binding sites and in the peak expression window for LTR7Y during pluripotent stages as illustrated in figure 3. Note that this model is not like classic selfish element activity that increases the frequency of the selfish allele at a given locus in a given population (e.g. a meiotic drive gene). Rather it is better considered a modifier of its own transposition rate. That is, a TE that could expand the window of time for transposition, would leave more daughter copies spread throughout the genome. The net effect of such an increase in copy numbers is a likely increase in fixation rates at multiple loci of the TE that controls pluripotency, for no better reason than the effective mutation rate (creation of new TE inserts) has increased and the neutral rate of evolution is the mutation rate. This notion of a selfish element is indeed akin to Doolittle and Sapienza's original intent, they noting that "some divergent copies [of TEs] may be more readily transposed; these will increase in frequency at the expense of others" [64].

In the conflict model, pluripotency implies transposition ability. A key discriminant between the two models is that under the conflict model, ERVs controlling pluripotency must take control of pluripotency during the phase when they are still transpositionally active. If we regularly find that ERVs assume control of pluripotency only after their phase of active transposition, then serendipidity would be the favoured model – i.e. it is mere chance aided by a permissive environment. Knowing how pluripotency is controlled before and after ERV invasion might help distinguish these models. The conflict model, for example, might predict that a naïve phase is more transitory prior (phylogenetically speaking) to ERV control.

Conclusions and outlook

Phylogenetically, the LTR7/HERVH-driven circuitry is young, and exists only in primates. Certain features of pluripotency regulation might even be specific to us humans (e.g. ESRG). Notably, many of the HERVH-derived transcripts might rather reflect a snapshot of the evolutionary reshaping process than a closed scenario. In fact, it is far from clear as to what fraction of the extensively remodelled transcripts is functional, or produces a protein. Similarly, while only a subset of the lncRNAs has the conserved domain, many of these might still represent transcriptional noise and have no regulatory function. Nonetheless, the adoption of ERVs in such a fundamental process as pluripotency control represents a remarkable symbiosis between ancient retroviral invaders and their host.

The HERVH-driven regulatory network enriches the transcriptome of the primate/human blastocysts, improving their options to successfully cope with their subsequent differentiation. This alone could be translates as an evolutionary advantage. Still, the severity of the consequence of disabling the HERVH-driven circuitry might suggest an additional, yet unexplored cellular

function of HERVH, perhaps in host-defence.

Looking forward it would be constructive to consider a general framework to predict or explain which phenotypes and functions ERVs are recruited too, not least because as sources of evolutionary novelty they are an informative case history. A serendipity model might predict that recruitment depends on a suitable genomic environment (such permissive chromatin). Alternatively, or additionally, they might be recruited to phenotypes that are hot spots for selection. For example, are they more likely to be recruited to processes involving antagonistic coevolution? The involvement of ERVs in placental phenotypes (e.g. syncytins) fits with both models.

Acknowledgements

ZIz is funded by ERC-2011-AdG 294742. LDH is funded by Medical Research Grant MR/L007215/1 and ERC grant ERC-2014-ADG 669207.

The authors have no conflict of interest to declare.

Figure legends

Figure 1. Insertional activity of HERVH elements during primate evolution. Arrows mark the windows in time where major expansions occurred with approximate number of copies shown (excluding solitary LTRs). The largest expansion in the Old World branch involved the common, partly deleted form. Data from references [7, 8].

Figure 2. Estimated fractions of total ERV elements that occur as solitary LTRs in the human genome. Copy numbers were taken from estimates provided in Mager DL and Medstrand P (2001) Retroviral Repeat Sequences. In: eLS. John Wiley & Sons Ltd, Chichester. http://www.els.net [doi: 10.1038/npg.els.0005062]. All ERV groups with greater than 500 solitary LTRs are plotted.

Figure 3. A: Left: Schematic representation of the early developmental phases: oocyte, 2-cell, 4-cell 8-cell stages, morula and blastocyst. Epi: Epiblast cells of the blastocyst, TE: Trophoectoderm, PE: primitive endoderm Right: Unbiased clustering of expression values of ERVs transcribed during early development (RPKM>1) using Spearman's rank correlation method. Data are re-analysed from [25]. Arrows highlight the expression pattern of the different LTR7 subfamilies. Note that ESCs are deriving from Passage 0. **B:** The Box-plots show level of expression of the LTR7B/Y subfamilies during early development [20]. Distribution of expression is shown as tag per million (tpm) values for each locus. **C:** The bar-plots show the number of expressed LTR7 subfamilies (RPKM>1).

Figure 4. A: LTR7/HERVH clusters naïve TF binding sites, NANOG, OVT4, KLF4 and LBP9. Transcription from LTR7/HERVH forces diversification of transcripts in hPSCs. Activated HERVHs generate numerous novel, stem cell specific alternative gene products. HERVH incorporates a set of regulatory lncRNAs into the network and defines novel pluripotent genes through alternative splicing or alternative non-AUG usage. HERVH inhibits differentiation, while HERVH-derived products contribute to maintain pluripotency. HERVH regulation might have evolved in conjunction with host defence. **B:** Schematic of the reporter construct, pT2-LTR7-GFP comprising an LTR7 region amplified from the ESRG locus flanked by inverted terminal repeats (ITRs) of the *Sleeping Beauty* transposon-based integration vector [65]. A reporter GFP signal marks a naïve-like cell population in heterogeneous human pluripotent stem cell cultures. Representative pictures of reporter-marked, FACS sorted hESC_H9s single colony are shown.

Figure 5. Integrated Genome Viewer (IGV) snapshot shows the distribution of RNA-seq reads over the ESRG gene during early development. Arrows indicate the direction of the transcription. HERVH is exonized, and contributes with two exons to ESRG. RNA-seq reads help to refine previously predicted exon-intron boundaries. The RNA-seq reads are shown on the scale of 10-5,000, except for ESC, where the scale of 10-15,000 is used for better visualisation. Data are re-analysed from [25]. Note that the last two exons of ESRG are composed of other ERV remnants.

Figure 6. Global expression cluster dendrogram using the average distance method on the dataset of mouse–human orthologous gene expression between GFP-marked hESC_H9 ($GFP^{high,}$ $GFP^{(+)}$; GFP^{low}) and various mouse/human primed/naïve embryonic stem cells (hESCs) [66]. While GFP^{low} cells cluster with primed ESCs, $GFP^{(+)}$ groups with various naïve-like ESCs of both mouse or human origin. Note that certain human naïve lineages clearly form a human-specific cluster, involving GFP^{high} ESCs.

References:

1. **Magiorkinis G, Gifford RJ, Katzourakis A, De Ranter J, et al.,** 2012. Env-less endogenous retroviruses are genomic superspreaders*. Proc Natl Acad Sci U S A,* **109**: 7385-90.
2. **Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, et al.,** 2012. Embryonic stem cell potency fluctuates with endogenous retrovirus activity*. Nature,* **487**: 57-63.
3. **Mager DL and Henthorn PS,** 1984. Identification of a retrovirus-like repetitive element in human DNA*. Proc Natl Acad Sci U S A,* **81**: 7510-4.
4. **Mager DL and Freeman JD,** 1987. Human endogenous retroviruslike genome with type C pol sequences and gag sequences related to human T-cell lymphotropic viruses*. J Virol,* **61**: 4060-6.
5. **Faunes F, Hayward P, Descalzo SM, Chatterjee SS, et al.,** 2013. A membrane-associated beta-catenin/Oct4 complex correlates with ground-state pluripotency in mouse embryonic

stem cells. *Development,* **140**: 1171-83.
6. **Jern P, Sperber GO, and Blomberg J,** 2004. Definition and variation of human endogenous retrovirus H. *Virology,* **327**: 93-110.
7. **Goodchild NL, Wilkinson DA, and Mager DL,** 1993. Recent evolutionary expansion of a subfamily of RTVL-H human endogenous retrovirus-like elements. *Virology,* **196**: 778-88.
8. **Mager DL and 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.
9. **Magiorkinis G, Blanco-Melo D, and Belshaw R,** 2015. The decline of human endogenous retroviruses: extinction and survival. *Retrovirology,* **12**: 8.
10. **Anderssen S, Sjottem E, Svineng G, and Johansen T,** 1997. Comparative analyses of LTRs of the ERV-H family of primate-specific retrovirus-like elements isolated from marmoset, African green monkey, and man. *Virology,* **234**: 14-30.
11. **Bao W, Kojima KK, and Kohany O,** 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA,* **6**: 11.
12. **Nelson DT, Goodchild NL, and Mager DL,** 1996. Gain of Sp1 sites and loss of repressor sequences associated with a young, transcriptionally active subset of HERV-H endogenous long terminal repeats. *Virology,* **220**: 213-8.
13. **Medstrand P and Mager DL,** 1998. Human-specific integrations of the HERV-K endogenous retrovirus family. *J Virol,* **72**: 9782-7.
14. **Hirose Y, Takamatsu M, and Harada F,** 1993. Presence of env genes in members of the RTVL-H family of human endogenous retrovirus-like elements. *Virology,* **192**: 52-61.
15. **Wilkinson DA, Goodchild NL, Saxton TM, Wood S, et al.,** 1993. Evidence for a functional subclass of the RTVL-H family of human endogenous retrovirus-like sequences. *J Virol,* **67**: 2981-9.
16. **de Parseval N, Casella J, Gressin L, and Heidmann T,** 2001. Characterization of the three HERV-H proviruses with an open envelope reading frame encompassing the immunosuppressive domain and evolutionary history in primates. *Virology,* **279**: 558-69.
17. **Belshaw R, Watson J, Katzourakis A, Howe A, et al.,** 2007. Rate of recombinational deletion among human endogenous retroviruses. *J Virol,* **81**: 9437-42.
18. **Rowe HM and Trono D,** 2011. Dynamic control of endogenous retroviruses during development. *Virology,* **411**: 273-87.
19. **Grow EJ, Flynn RA, Chavez SL, Bayless NL, et al.,** 2015. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature,* **522**: 221-5.
20. **Goke J, Lu X, Chan YS, Ng HH, et al.,** 2015. Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell,* **16**: 135-41.
21. **Wang J, Xie G, Singh M, Ghanbarian AT, et al.,** 2014. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature,* **516**: 405-9.
22. **Huang CR, Burns KH, and Boeke JD,** 2012. Active transposition in genomes. *Annu Rev Genet,* **46**: 651-75.
23. **Guo H, Zhu P, Yan L, Li R, et al.,** 2014. The DNA methylation landscape of human early embryos. *Nature,* **511**: 606-10.
24. **Smith ZD, Chan MM, Humm KC, Karnik R, et al.,** 2014. DNA methylation dynamics of the human preimplantation embryo. *Nature,* **511**: 611-5.
25. **Yan L, Yang M, Guo H, Yang L, et al.,** 2013. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol,* **20**: 1131-9.

26. **Chuong EB, Rumi MA, Soares MJ, and Baker JC,** 2013. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat Genet,* **45**: 325-9.
27. **Xie H, Ye M, Feng R, and Graf T,** 2004. Stepwise reprogramming of B cells into macrophages. *Cell,* **117**: 663-76.
28. **Kelley DR and Rinn JL,** 2012. Transposable elements reveal a stem cell specific class of long noncoding RNAs. *Genome Biol,* **13**: R107.
29. **Santoni FA, Guerra J, and Luban J,** 2012. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology,* **9**: 111.
30. **Kunarso G, Chia NY, Jeyakani J, Hwang C, et al.,** 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet,* **42**: 631-4.
31. **Gaspar-Maia A, Alajem A, Polesso F, Sridharan R, et al.,** 2009. Chd1 regulates open chromatin and pluripotency of embryonic stem cells. *Nature,* **460**: 863-8.
32. **Chappell J, Sun Y, Singh A, and Dalton S,** 2013. MYC/MAX control ERK signaling and pluripotency by regulation of dual-specificity phosphatases 2 and 7. *Genes Dev,* **27**: 725-33.
33. **Lu X, Sachs F, Ramsay L, Jacques PE, et al.,** 2014. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat Struct Mol Biol,* **21**: 423-5.
34. **Dunn SJ, Martello G, Yordanov B, Emmott S, et al.,** 2014. Defining an essential transcription factor program for naive pluripotency. *Science,* **344**: 1156-60.
35. **Martello G, Bertone P, and Smith A,** 2013. Identification of the missing pluripotency mediator downstream of leukaemia inhibitory factor. *EMBO J*.
36. **Ye S, Li P, Tong C, and Ying QL,** 2013. Embryonic stem cell self-renewal pathways converge on the transcription factor Tfcp2l1. *EMBO J*.
37. **Zhou W, Clouston DR, Wang X, Cerruti L, et al.,** 2000. Induction of human fetal globin gene expression by a novel erythroid factor, NF-E4. *Mol Cell Biol,* **20**: 7662-72.
38. **Parada CA, Yoon JB, and Roeder RG,** 1995. A novel LBP-1-mediated restriction of HIV-1 transcription at the level of elongation in vitro. *J Biol Chem,* **270**: 2274-83.
39. **To S, Rodda SJ, Rathjen PD, and Keough RA,** 2010. Modulation of CP2 family transcriptional activity by CRTR-1 and sumoylation. *PLoS One,* **5**: e11702.
40. **Ng SY, Johnson R, and Stanton LW,** 2012. Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *EMBO J,* **31**: 522-33.
41. **Zhao M, Ren C, Yang H, Feng X, et al.,** 2007. Transcriptional profiling of human embryonic stem cells and embryoid bodies identifies HESRG, a novel stem cell gene. *Biochem Biophys Res Commun,* **362**: 916-22.
42. **Shi J, Ren C, Liu H, Wang L, et al.,** 2015. An ESRG-interacting protein, COXII, is involved in pro-apoptosis of human embryonic stem cells. *Biochem Biophys Res Commun,* **460**: 130-5.
43. **Loewer S, Cabili MN, Guttman M, Loh YH, et al.,** 2010. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet,* **42**: 1113-7.
44. **Schlesinger S and Goff SP,** 2015. Retroviral transcriptional regulation and embryonic stem cells: war and peace. *Mol Cell Biol,* **35**: 770-7.
45. **Stadtfeld M and Hochedlinger K,** 2010. Induced pluripotency: history, mechanisms, and applications. *Genes Dev,* **24**: 2239-63.
46. **Evans MJ and Kaufman MH,** 1981. Establishment in culture of pluripotential cells from mouse embryos. *Nature,* **292**: 154-6.

47. **Martin GR,** 1981. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proc Natl Acad Sci U S A,* **78**: 7634-8.
48. **Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, et al.,** 1998. Embryonic stem cell lines derived from human blastocysts. *Science,* **282**: 1145-7.
49. **Fang R, Liu K, Zhao Y, Li H, et al.,** 2014. Generation of naive induced pluripotent stem cells from rhesus monkey fibroblasts. *Cell Stem Cell,* **15**: 488-96.
50. **Buecker C, Chen HH, Polo JM, Daheron L, et al.,** 2010. A murine ESC-like state facilitates transgenesis and homologous recombination in human pluripotent stem cells. *Cell Stem Cell,* **6**: 535-46.
51. **Hanna J, Cheng AW, Saha K, Kim J, et al.,** 2010. Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse ESCs. *Proc Natl Acad Sci U S A,* **107**: 9222-7.
52. **Takashima Y, Guo G, Loos R, Nichols J, et al.,** 2014. Resetting transcription factor control circuitry toward ground-state pluripotency in human. *Cell,* **158**: 1254-69.
53. **Theunissen TW and Jaenisch R,** 2014. Molecular control of induced pluripotency. *Cell Stem Cell,* **14**: 720-34.
54. **Chan YS, Goke J, Ng JH, Lu X, et al.,** 2013. Induction of a human pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast. *Cell Stem Cell,* **13**: 663-75.
55. **Ware CB, Nelson AM, Mecham B, Hesson J, et al.,** 2014. Derivation of naive human embryonic stem cells. *Proc Natl Acad Sci U S A,* **111**: 4484-9.
56. **Koyanagi-Aoi M, Ohnuki M, Takahashi K, Okita K, et al.,** 2013. Differentiation-defective phenotypes revealed by large-scale analyses of human pluripotent stem cells. *Proc Natl Acad Sci U S A,* **110**: 20569-74.
57. **Ohnuki M, Tanabe K, Sutou K, Teramoto I, et al.,** 2014. Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proc Natl Acad Sci U S A,* **111**: 12426-31.
58. **Theunissen TW, Powell BE, Wang H, Mitalipova M, et al.,** 2014. Systematic identification of culture conditions for induction and maintenance of naive human pluripotency. *Cell Stem Cell,* **15**: 471-87.
59. **Wu J, Okamura D, Li M, Suzuki K, et al.,** 2015. An alternative pluripotent state confers interspecies chimaeric competency. *Nature,* **521**: 316-21.
60. **Tonge PD, Corso AJ, Monetti C, Hussein SM, et al.,** 2014. Divergent reprogramming routes lead to alternative stem-cell states. *Nature,* **516**: 192-7.
61. **Shakiba N, White CA, Lipsitz YY, Yachie-Kinoshita A, et al.,** 2015. CD24 tracks divergent pluripotent states in mouse and human cells. *Nat Commun,* **6**: 7329.
62. **Robbez-Masson L and Rowe HM,** 2015. Retrotransposons shape species-specific embryonic stem cell gene expression. *Retrovirology,* **12**: 45.
63. **Dupressoir A, Lavialle C, and Heidmann T,** 2012. From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. *Placenta,* **33**: 663-71.
64. **Doolittle WF and Sapienza C,** 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature,* **284**: 601-3.
65. **Mates L, Chuah MK, Belay E, Jerchow B, et al.,** 2009. Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat Genet,* **41**: 753-61.

66. **Gafni O, Weinberger L, Mansour AA, Manor YS, et al.,** 2013. Derivation of novel human ground state naive pluripotent stem cells*. Nature,* **504**: 282-6.

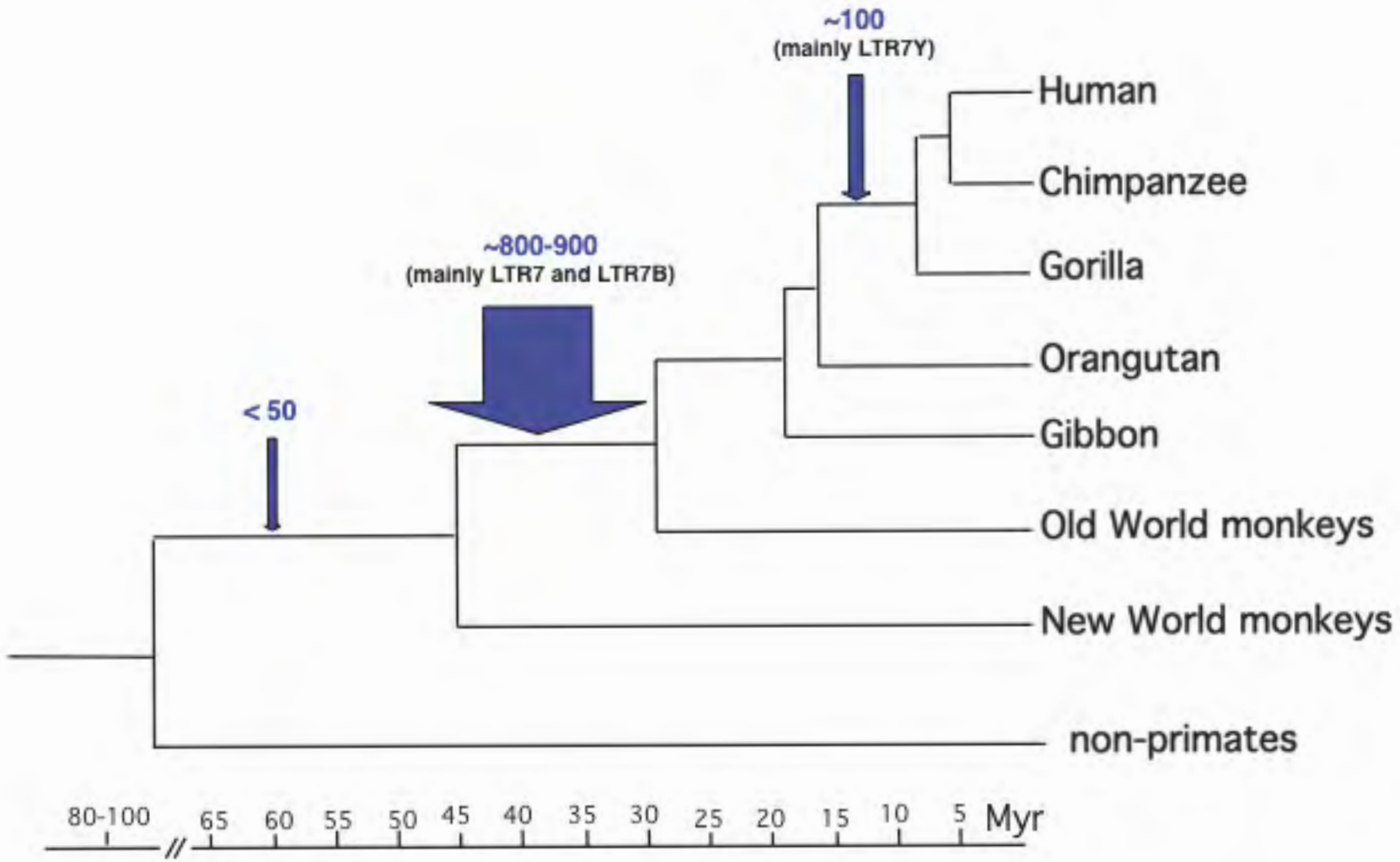

~100
(mainly LTR7Y)
Human
Chimpanzee
~800-900
(mainly LTR7 and LTR7B)
Gorilla
Orangutan
< 50
Gibbon
Old World monkeys
New World monkeys
non-primates
80-100
65
60
55
50
45
40
35
30
25
20
15
10
5
Myr

Figure 1

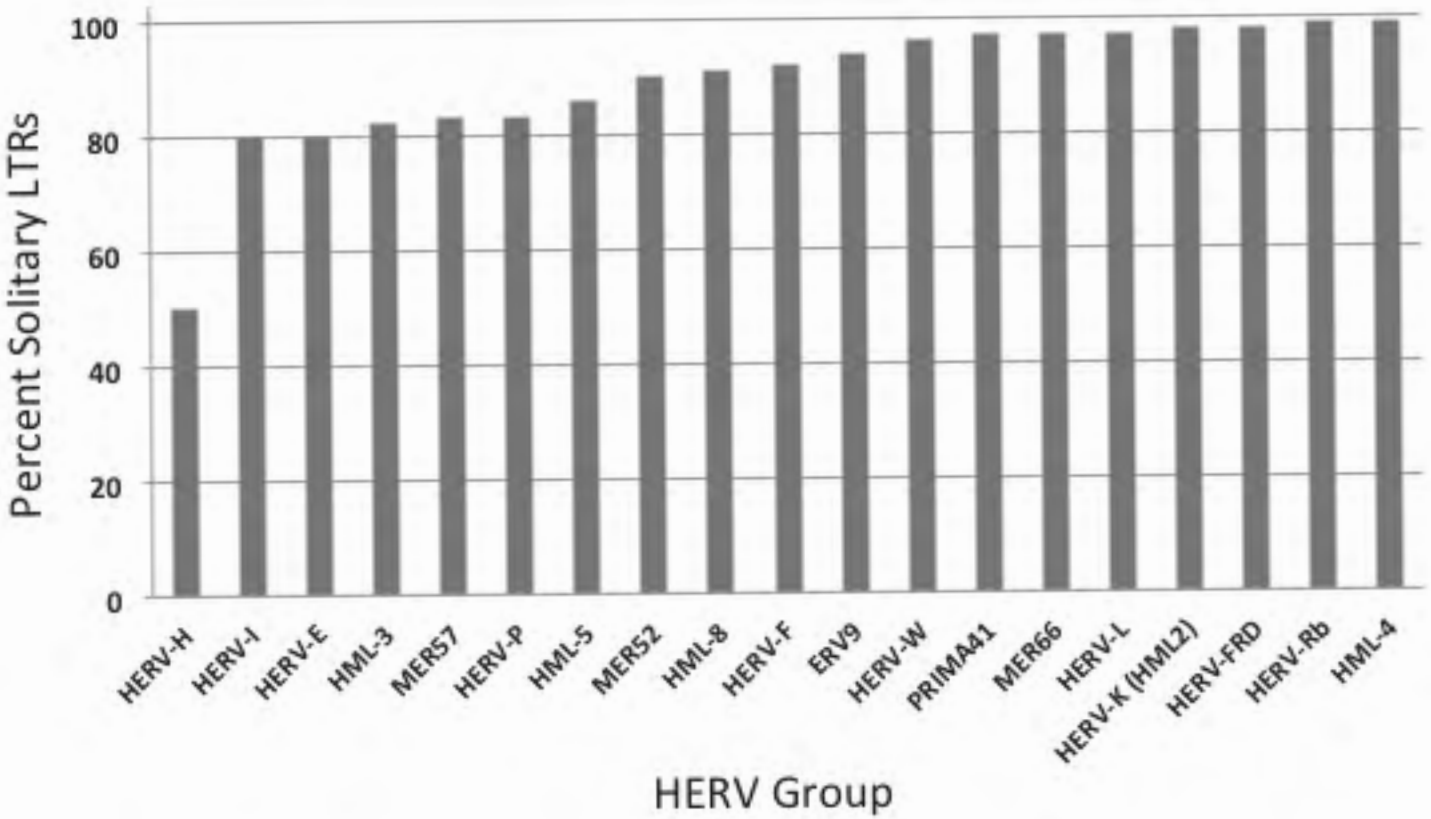
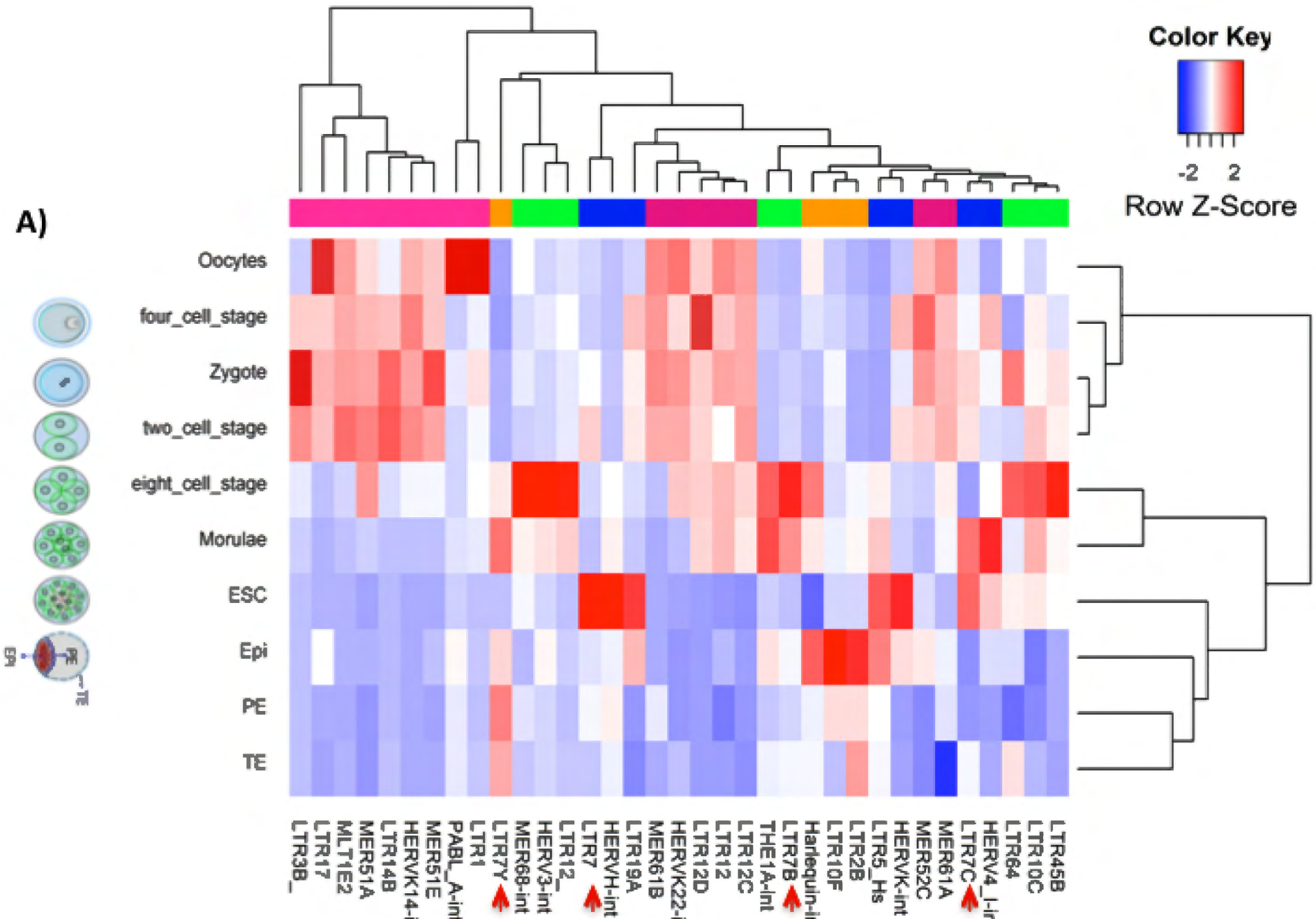
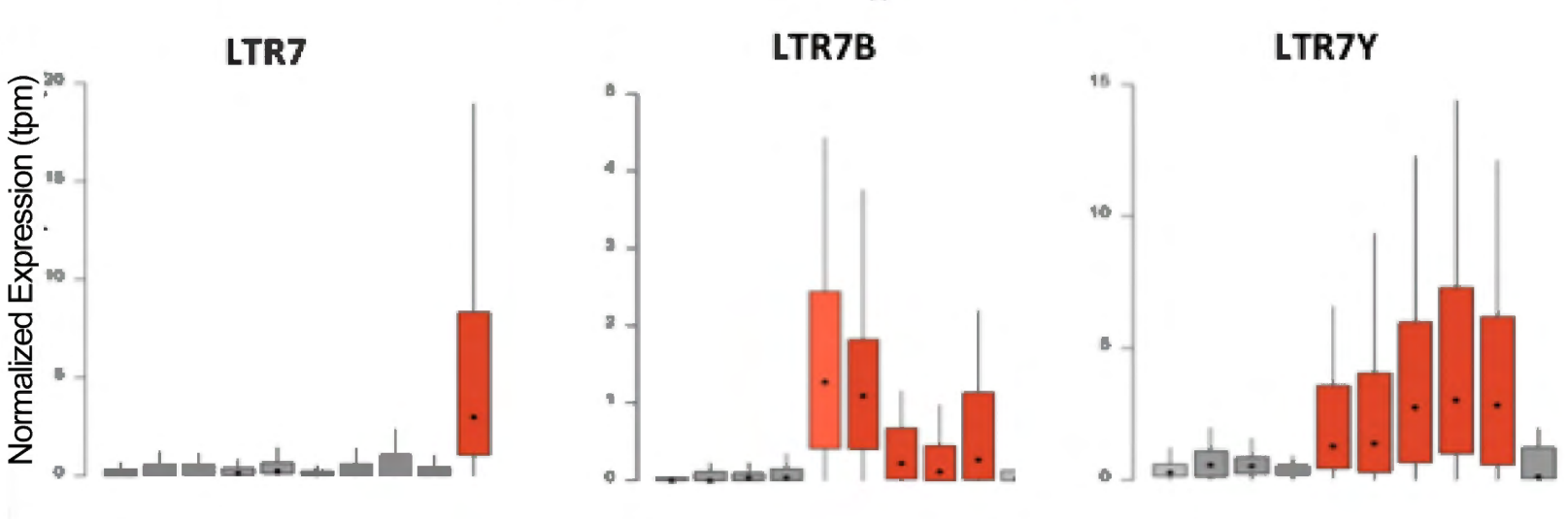

Percent Solitary LTRs
100
80
60
40
20
0
HERV-H
HERV-I
HERV-E
HML-3
MER57
HERV-P
HML-5
MER52
HML-8
HERV-F
ERV9
HERV-W
PRIMA41
MER66
HERV-L
HERV-K (HML2)
HERV-FRD
HERV-Rb
HML-4
HERV Group

Figure 2

A)


Color Key
-2
2
Row Z-Score
Oocytes
four_cell_stage
Zygote
two_cell_stage
eight_cell_stage
Morulae
ESC
Epi
PE
TE
EPI
PE
TE
LTR3B_
LTR17
MLT1E2
MER51A
LTR14B
HERVK14-int
MER51E
PABL_A-int
LTR1
LTR7Y
MER68-int
HERV3-int
LTR12_
LTR7
HERVH-int
LTR19A
MER61B
HERVK22-int
LTR12D
LTR12
LTR12C
THE1A-int
LTR7B
Harlequin-int
LTR10F
LTR2B
LTR5_Hs
HERVK-int
MER52C
MER61A
LTR7C
HERV4_I-int
LTR64
LTR10C
LTR45B

B)


LTR7
LTR7B
LTR7Y
Normalized Expression (tpm)

C)

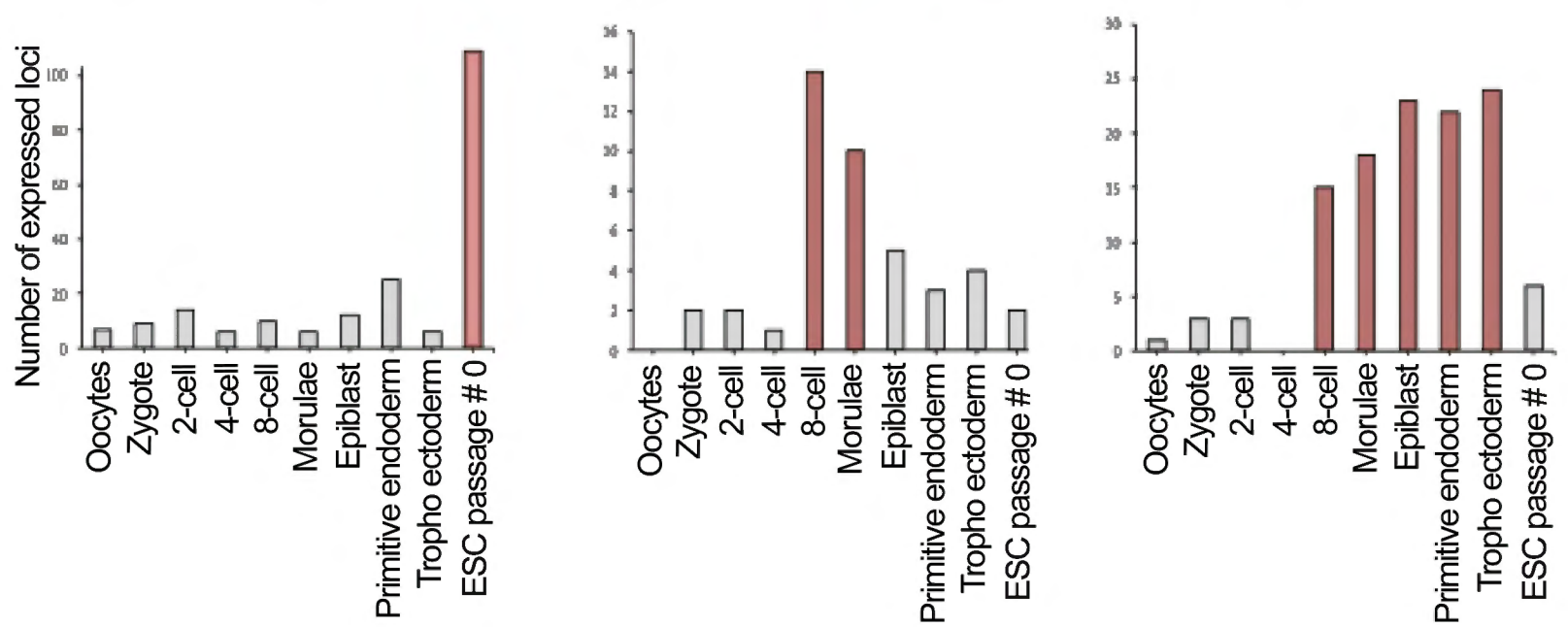
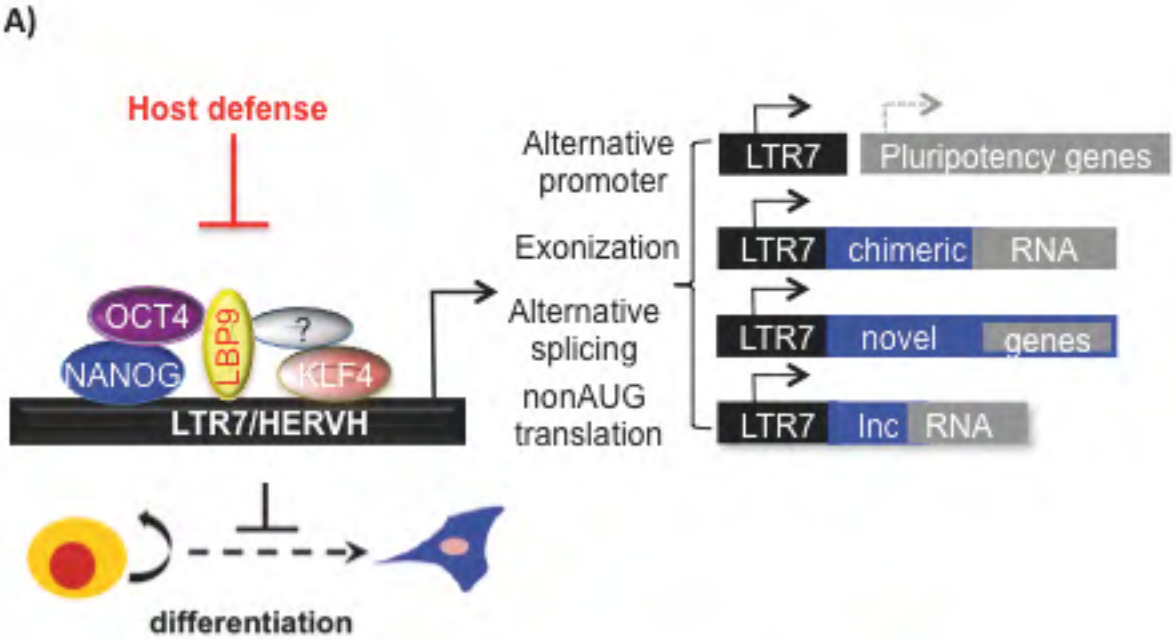

Number of expressed loci
Oocytes
Zygote
2-cell
4-cell
8-cell
Morulae
Epiblast
Primitive endoderm
Tropho ectoderm
ESC passage # 0

Figure 3


A)
Host defense
OCT4
LBP9
?
NANOG
KLF4
LTR7/HERVH
Alternative promoter
Exonization
Alternative splicing
nonAUG translation
LTR7
Pluripotency genes
LTR7
chimeric
RNA
LTR7
novel
genes
LTR7
lnc
RNA
differentiation

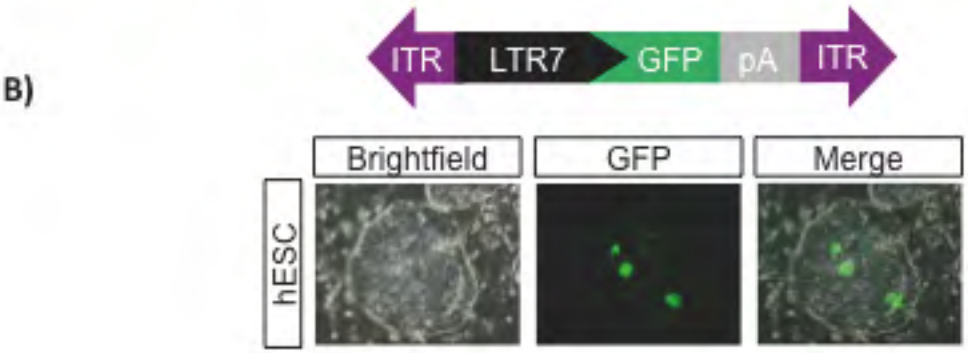

B)
ITR
LTR7
GFP
pA
ITR
Brightfield
GFP
Merge
hESC

Figure 4

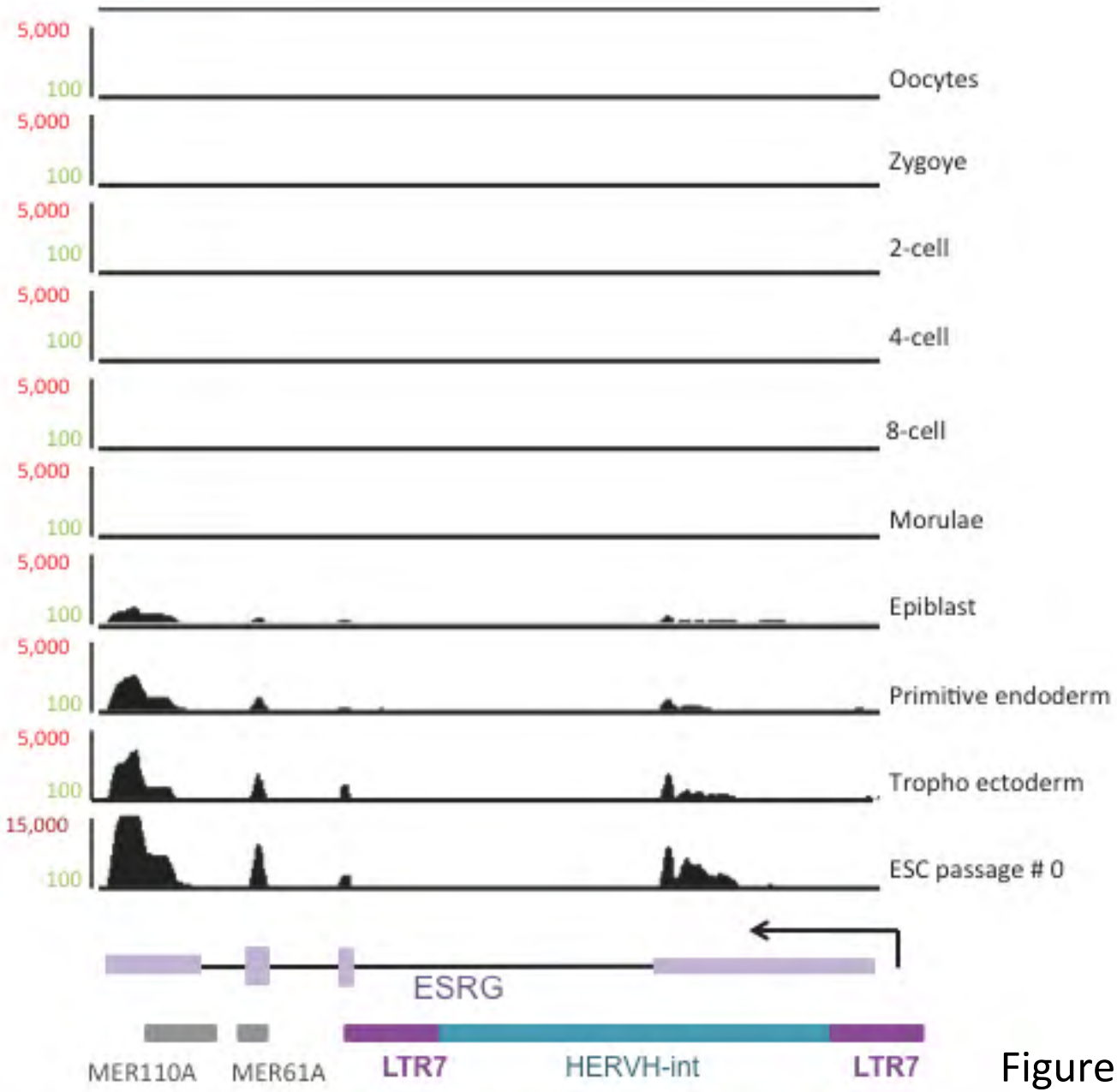

5,000
100
Oocytes
5,000
100
Zygoye
5,000
100
2-cell
5,000
100
4-cell
5,000
100
8-cell
5,000
100
Morulae
5,000
100
Epiblast
5,000
100
Primitive endoderm
5,000
100
Tropho ectoderm
15,000
100
ESC passage # 0
ESRG
MER110A
MER61A
LTR7
HERVH-int
LTR7

Figure 5

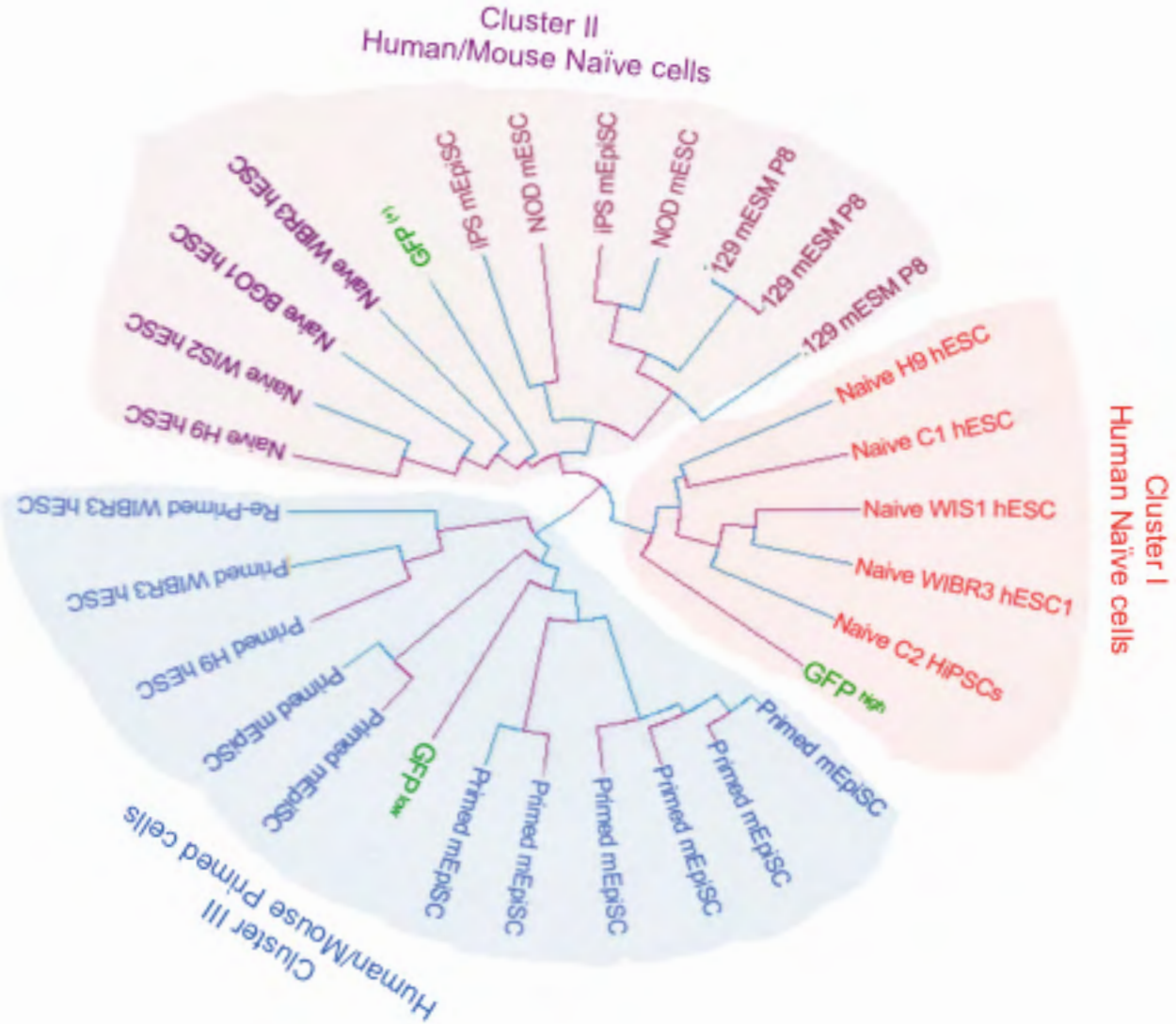

Cluster II
Human/Mouse Naïve cells
Naive H9 hESC
Naive WIS2 hESC
Naive BGO1 hESC
Naive WIBR3 hESC
GFP high
iPS mEpiSC
NOD mESC
iPS mEpiSC
NOD mESC
129 mESM P8
129 mESM P8
129 mESM P8
Naive H9 hESC
Naive C1 hESC
Naive WIS1 hESC
Naive WIBR3 hESC1
Naive C2 HiPSCs
GFP high
Cluster I
Human Naïve cells
Re-Primed WIBR3 hESC
Primed WIBR3 hESC
Primed H9 hESC
Primed mEpiSC
Primed mEpiSC
GFP low
Primed mEpiSC
Primed mEpiSC
Primed mEpiSC
Primed mEpiSC
Primed mEpiSC
Cluster III
Human/Mouse Primed cells

Figure 6