1

# Movements of ancient human endogenous retroviruses

# detected in SOX2-expressing cells

4

5

6  **Kazuaki Monde[a,1], Yorifumi Satou[b], Mizuki Goto[e], Yoshikazu Uchiyama[c], Jumpei Ito[f, g],**
7  **Taku Kaitsuka[d], Hiromi Terasawa[a], Shinya Yamaga[a], Tomoya Matsusako[a], Fan-Yan**
8  **Wei[d], Ituro Inoue[f], Kazuhito Tomizawa[d], Akira Ono[h], Takumi Era[e], Tomohiro Sawa[a],**
9  **Yosuke Maeda[a]**

10

11  [a]Department of Microbiology, Faculty of Life Sciences, Kumamoto University, Chuo-ku,
12  860-8556, Kumamoto, Japan; [b]Joint Research Center for Human Retrovirus Infection,
13  Kumamoto University, Chuo-ku, 860-8556, Kumamoto, Japan; [c]Department of Medical
14  Physics, Kumamoto University, Chuo-ku, 862-0976, Kumamoto, Japan; [d]Department of
15  Molecular Physiology, Faculty of Life Sciences, Kumamoto University, Chuo-ku, 860-8556,
16  Kumamoto, Japan; [e]Department of Cell Modulation, Institute of Molecular Embryology and
17  Genetics (IMEG), Kumamoto University, Chuo-ku, 860-0811, Kumamoto, Japan; [f]National
18  Institute of Genetics Laboratory of Human Genetics, Mishima, 411-8540, Shizuoka, Japan;
19  [g]Division of Systems Virology, Department of infectious Disease Control, International
20  Research Center for infectious Diseases, Institute of Medical Science, The University of
21  Tokyo, 108-8639, Tokyo, Japan; [h]Department of Microbiology and Immunology, University
22  of Michigan Medical School, Ann Arbor, Michigan, 48109, United States of America.

23
24
25  [1]Correspondence: monde@kumamoto-u.ac.jp

26

27  Key words; HERV-K, SOX2, iPS cells, retrotransposon

28
29

1 **Summary**

2   Human endogenous retroviruses (HERVs) occupy approximately 8% of human

3 genome. HERVs, which are transcribed in early embryos, are epigenetically silenced

4 in somatic cells, except in pathological contexts. HERV-K is thought to protect the

5 embryo from exogenous viral infection. However, uncontrollable HERV-K expression

6 in somatic cells has been implicated in several diseases. Here, we show that SOX2,

7 which plays a key role in maintaining pluripotency of stem cells, is critical for the

8 transcription of HERV-K LTR5Hs. HERV-K can undergo retrotransposition within

9 producer cells in the absence of Env expression. Furthermore, new HERV-K

10 integration sites were identified in a long-term culture of induced pluripotent stem

11 cells, which express SOX2. Together, these results suggest the possibility that the

12 strict dependence of HERV-K on SOX2 have allowed contribution of HERV-K to the

13 protection of early embryos during evolution while limiting potentially harmful effects

14 of HERV-K retrotransposition on host genome integrity to these early embryos.

15

16

## Introduction

Endogenous retroelements are mobile genetic elements that constitute more than 40% of the human genome. Human endogenous retroviruses (HERVs), which encode the long terminal repeat (LTR)-containing elements, occupy about 8% of the human genome (Bannert & Kurth, 2004; Lander et al, 2001; Venter et al, 2001). For more than 20 million years, HERVs that have persisted in germ-cell lineages have been transmitted vertically from ancestors to descendant (Boeke & Stoye, 1997). At present, almost all HERVs have acquired numerous mutations or deletions. However, HERV-K, a relatively new endogenous retrovirus, apparently encodes intact open reading frames in the human genome (Turner et al, 2001), although no replication-competent HERV-K has been detected (Beimforde et al, 2008; Boller et al, 2008; Lee & Bieniasz, 2007; Stoye, 2012). HERV-K is transcribed during early embryogenesis or exogenous viral infection, producing HERV-K proteins that appear to protect the host cells from viral attack (Grow et al, 2015; Monde et al, 2012; Monde et al, 2017; Terry et al, 2017). HERV-K expression has also been noted in various human diseases, including autoimmune disorders, neurological diseases, infectious diseases, and cancer (Young et al, 2013).

Long interspersed nuclear elements (LINE-1), which are classified among the non-LTR retroelements, are transposition competent (Beck et al, 2010; Brouha et al, 2003; Mills et al, 2007). The transposition of LINE-1 mainly occurs in germ cells during early embryonic development. These transposition events might cause pathogenesis by altering the structures, expression, and functions of genes (Beck et al, 2011; Han et al, 2004; Hancks & Kazazian, 2012). Therefore, transposition is regulated by histone modifications and DNA methylation to avoid the harmful mutations in the genomes (Bourc'his & Bestor, 2004; Levin & Moran, 2011). Recent

1    advances in sequencing technology have allowed the detection of non-reference

2    HERV-K, which is absent from the human genome sequence, in the population

3    (Wildschutte et al, 2016), although HERV-K retrotransposition activity has not yet

4    been reported.

5     HERV-K encodes the 5′-LTR and 3′-LTR at the upstream and downstream of the

6    viral protein ORFs respectively. HERV-K LTR has preserved their promoter activity,

7    and HERV-K is transcribed in embryonic stem cells, several cancer cells, and virus-

8    infected T cells (Grow et al, 2015). The transcription factors Sp1 and Sp3 drive

9    HERV-K transcription in teratocarcinoma cells and melanoma cells (Fuchs et al,

10    2011). The melanoma-specific transcription factor MITF-M is also required for the

11    activation of the HERV-K LTR (Katoh et al, 2011). In virus-infected cells, viral

12    transcription factors Tat and Tax are associated with HERV-K expression (Gonzalez-

13    Hernandez et al, 2012; Toufaily et al, 2011). In embryonic stem cells, DNA

14    hypomethylation and OCT3/4-binding to the HERV-K LTR synergistically facilitate

15    HERV-K transcription (Grow et al, 2015). However, it remains unclear whether these

16    transcription factors are essential for HERV-K activation.

17     Here, we show that SOX2, rather than OCT3/4, is the major factor for activating the

18    transcription of HERV-K LTR5Hs, which is the youngest HERV-K subfamily (Turner

19    et al, 2001). Consistent with this finding, a large amount of HERV-K Gag is

20    expressed in induced pluripotent stem (iPS) cells, which are SOX2-expressing cells.

21    We used next-generation sequencing (NGS) to analyze the genomes of iPS cells

22    and determined the HERV-K integration sites. Surprisingly, we found that new

23    HERV-K insertions into the genome increase in a manner dependent upon the

24    culture period, suggesting that HERV-K retrotransposition occurs in SOX2-

25    expressing cells. Our results suggest that HERV-K is not a harmless fossil left in the

1 human genome; rather, it retains the ability to spread among the human genome by

2 retrotransposition, but is normally repressed due to its dependence on SOX2

3 expression.

4

5 **Results**

6 **SOX2 activates HERV-K transcription**

7     Teratocarcinomas are germ-cell tumors, and teratocarcinoma cells constitutively

8 express HERV-K proteins and release HERV-K particles from their plasma

9 membranes (Bieda et al, 2001; Boller et al, 1983). Several transcription factors,

10 including MITF, MZF1, NF-Y, GATA-2, and OCT3/4, are required to activate the

11 HERV-K LTR (Grow et al, 2015; Katoh et al, 2011; Yu et al, 2005). Wysocka's

12 research group identified the consensus OCT3/4-binding motifs in HERV-K LTR5Hs,

13 and demonstrated the transcriptional activation of HERV-K by OCT3/4 in human

14 preimplantation embryos (Grow et al, 2015). However, it is unknown whether the

15 expression of OCT3/4 is sufficient for the transcriptional activation of HERV-K. Here,

16 we identified the region of HERV-K responsible for the transcription of HERV-K

17 LTR5Hs using deletion mutants of HERV-K LTR5Hs in teratocarcinoma cells (NCCIT

18 cells) (Supplemental Fig. S1A). These results show that the deletion of nucleotides

19 nt 650–700 in LTR5Hs causes the loss of its transactivation activity (Supplemental

20 Fig. S1B and S1C). With the PROMO software (Farre et al, 2003; Messeguer et al,

21 2002), which is used to identify putative transcription factors, we identified 15 SOX2-

22 binding motifs (#1–#12) and two OCT3/4-binding motifs (#13 and #14) in LTR5Hs

23 (Fig. 1A and Supplemental Fig. S1). Some SOX2-binding motifs overlapped with

24 each other (#10 and #11), and therefore these motifs were called each same

25 number. Two OCT3/4-binding motifs (#13 and #14) and three SOX2-binding motifs

5

1    (#9, #10, and #11) occurred in the region nt 650–700 in LTR5Hs (Fig. 1A and

2    Supplemental Fig. S1A). Based on a chromatin immunoprecipitation (ChIP) analysis

3    database in embryonic stem cells, there were two peaks of SOX2 binding at nt 200

4    and 700 of the HERV-K LTR genome (Fig. 1B). OCT3/4-binding peaks were similar

5    to those for SOX2 in the HERV-K LTR genome. To determine whether OCT3/4 is

6    sufficient for the transactivation of HERV-K LTR, as reported previously (Grow et al,

7    2015), we cotransfected HeLa cells with plasmids encoding each transcription factor

8    (OCT3/4, SOX2, KLF4, NANOG) and the HERV-K LTR-Luc (Fig. 1C). Unexpectedly,

9    we found that OCT3/4 was not sufficient to activate the transcription of HERV-K LTR

10    in HeLa cells. The transcription of LTR mutants, with mutations in the OCT3/4-

11    binding motifs, was slightly reduced, but not significantly so, in NCCIT cells

12    (Supplemental Fig. S1D) and when SOX2, KLF4 and OCT3/4 were overexpressed in

13    HeLa cells (Supplemental Fig. S1E). In contrast, SOX2 markedly activated HERV-K

14    LTR transcription (Fig. 1C). In the presence of SOX2, KLF4 slightly increased the

15    transactivation of HERV-K, but OCT3/4 reduced the effect of SOX2 (Supplemental

16    Fig. S1F). In the presence of both SOX2 and KLF4, OCT3/4 increased the

17    transactivation of HERV-K. The transactivation of HERV-K was dose-dependently

18    enhanced by expression of SOX2 alone (Fig. 1D and 1E). However, OCT3/4 alone

19    did not alter the transactivation of HERV-K, even when overexpressed.

20    Because NCCIT cells express large amounts of endogenous SOX2 (Fig. 1E), we

21    examined the binding of endogenous SOX2 to chromosomal HERV-K LTR with a

22    ChIP assay (Fig. 1F). The results showed that endogenous SOX2 binds to the

23    chromosomal HERV-K LTR in NCCIT cells. To confirm that endogenous SOX2

24    drives HERV-K transcription, we established SOX2-knockout NCCIT cells

25    (NCCIT/KOSOX2) (Fig. 1G). Although the genome of the NCCIT/KOSOX2 cells

6

1    encodes four different sequence patterns, no intact *SOX2* gene was detected in the

2    NCCIT/KOSOX2 cells (Fig. 1G right). HERV-K LTR transactivation was dramatically

3    reduced in the NCCIT/KOSOX2 cells, but not completely lost, and was rescued by

4    the transfection of SOX2 (Fig. 1H). The mature HERV-K Gag protein (37 kDa) in the

5    viral particles disappeared from the supernatant of the KOSOX2 cells (Fig. 1I).

6    Together, these results indicate that SOX2 is an essential transcription factor for

7    expression of HERV-K LTR5Hs, and that both OCT3/4 and KLF4 drive HERV-K

8    transcription in the presence of SOX2.

9    **Multiple SOX2-binding motifs activate the HERV-K transcription**

10   With the Promo software and the ChIP database, we localized nine of 14 SOX2-

11   binding motifs (#3, #4, #7, #8, #9, #10, and #11) around nt 200 and 700 of the

12   HERV-K LTR genome (Fig. 1A and 1B). Based on Fig. S1C, we speculated that a

13   deletion of the single SOX2-binding motif #9 might abolish the transactivation of

14   HERV-K LTR. To determine the region responsible for HERV-K transactivation by

15   SOX2, HeLa cells were cotransfected with plasmids encoding HERV-K LTR–luc

16   mutants (del#01-#12) and SOX2, KLF4, and OCT3/4. Unexpectedly, any single

17   deletion of a SOX2-binding motif did not reduce the transactivation of HERV-K LTR

18   in HeLa cells (Fig. 2A) or NCCIT cells (Fig. 2B). However, the deletion of all SOX2-

19   binding motifs dramatically reduced HERV-K transactivation in both HeLa and

20   NCCIT cells. Notably, LTR sequences that contain some single deletions showed

21   similar activity to that of WT HERV-K LTR, but other single deletions enhanced the

22   LTR activity, suggesting the redundancy and/or interference between SOX2-binding

23   motifs.  Therefore, we designed mutants of LTR5Hs with multiple deletions of SOX2-

24   binding motifs (Fig. 2C and 2D). Deletion of SOX2-binding motifs #03, #08, #09, and

25   #10 around nt 200 and 700 in LTR5Hs, which correspond to two major SOX2-

7

1    binding regions (Fig. 1A and 1B), reduced HERV-K transactivation to the same

2    degree as the deletion of all the SOX2-binding motifs in both HeLa cells (Fig. 2C)

3    and NCCIT cells (Fig. 2D). These results suggest that SOX2 activates HERV-K

4    transcription even after the accumulation of several mutations in LTR5Hs during its

5    biological evolution.

6    **SOX2 activates chromosomal HERV-K expression**

7       HERV-K genomes have a CpG island between the LTR and the *gag* gene (Fig.

8    3A), which is hypermethylated in HeLa cells (Fig. 3C) compared with NCCIT cells

9    (Fig. 3B). This suggests that HERV-K genomes are packed into heterochromatin and

10    are silenced in HeLa cells. To confirm the modification of the chromatin, we treated

11    HeLa cells with 5-aza-2′-deoxycytidine to hypomethylate the genome. The

12    hypomethylation of the genome enhanced HERV-K Gag mRNA expression when

13    SOX2 was overexpressed in the HeLa cells (Fig. 3D). These results indicate that

14    DNA hypomethylation and SOX2 expression synergistically induce the expression of

15    HERV-K genes in the human genome.

16    **SOX2 activates the 5′ and 3′ LTR5Hs of HERV-K**

17       Because LTR sequences of HERV-K is classified into three major groups (LTR5Hs,

18    5A, and 5B), we cloned 18 different HERV-K LTRs from NCCIT cells and

19    investigated whether HERV-K LTR transactivation by SOX2 is conserved among the

20    three different groups. The LTR sequences of the HERV-K 5Hs group (LTR5Hs) are

21    the part of the most recently integrated sequences (around 9.1 million years ago)

22    (Subramanian et al, 2011). There are two types of LTR5Hs proviruses that are

23    classified based on the presence (type 1) or absence (type 2) of a 292 bp deletion at

24    the pol-env junction. The LTRs of 5A and 5B groups (LTR5A and LTR5B) are

25    associated with proviruses that are mainly classified as type 2 (Subramanian et al,

1    2011). The LTR5B proviruses include the oldest insertions (around 27.9 million years

2    ago), and LTR5A proviruses (around 20.1 million years ago) are originating from

3    LTR5B at an estimated-standard mutation rate of 0.24-0.45% per million years

4    based on the LTR-based and internal-based phylogenies (Subramanian et al, 2011).

5    Interestingly, both 5′- and 3′-LTR of LTR5Hs and three out of four LTR5B were

6    significantly activated by SOX2, whereas three out of four LTR5A were no activated

7    in SOX2-expressing HeLa cells (Fig. 4A) and NCCIT cells (Fig. 4B). A phylogenetic

8    analysis of HERV-K LTRs showed that SOX2-responsive HERV-K LTRs are closely

9    related (Fig. 4C). Both the newest and oldest HERV-K LTRs integrated into genomes

10   retain the capacity for SOX2-dependent transactivation, suggesting that acquiring or

11   maintaining this capacity is advantageous for coexistence between HERV-K and the

12   host.

13   **Reconstructed HERV-K has retrotransposition activity**

14   HERV-K LTR5Hs is expressed in SOX2-expressing cells, such as germ cells, but it

15   is unclear whether HERV-K has retrotransposition activity within these cells. To

16   examine the retrotransposition activity of HERV-K, we designed a HERV-K$_{CON}$

17   construct encoding intron-inserting nanoluciferase (inNanoluc) (Fig. 5A). After the

18   transcription of HERV-K from the *Cytomegalovirus* (CMV) promoter, the orientation

19   of the intact reporter gene was reversed by splicing, and the CMV promoter at the 5′-

20   UTR was then replaced with U3 through reverse transcription (Fig. 5A bottom). The

21   reverse-transcribed HERV-K integrated into genome, and the intact reporter gene

22   was transcribed from the *Simian virus 40* (SV40) promoter. The nanoluciferase

23   values indirectly reflected the retrotransposition activity of HERV-K. HERV-K

24   GagProPol, which encodes full-length *gag, pro,* and *pol*, showed nanoluciferase

25   activity 5 days after transfection, whereas HERV-K deltaGagProPol, which encodes

1   only truncated-*gag,* showed only slight nanoluciferase activity in HeLa cells (Fig. 5B).

2   This suggests that HERV-K protease, reverse transcriptase, and/or integrase is

3   required for HERV-K retrotransposition. Plasmids encoding HERV-K deltaProPol, a

4   protease mutant (D203N), a reverse transcriptase mutant (SIAA), or an integrase

5   mutant (DR1, DR2) with inNanoluc reporter gene were cotransfected into HeLa cells

6   with or without a protein expression plasmid encoding HERV-K GagProPol (Fig. 5C).

7   Although these mutants also showed faint nanoluciferase activity, expression of

8   HERV-K GagProPol rescued the nanoluciferase activity of these mutants. This

9   indicates that HERV-K protease, reverse transcriptase, and integrase are required

10  for the retrotransposition of HERV-K. It also suggests that the assembly of intact

11  proteases, reverse transcriptases, and integrases of different HERV-K origins can

12  complement defective HERV-Ks during HERV-K retrotransposition.

13         To determine the preferred loci for new integration of HERV-K, we analyzed

14  the new integration sites of HERV-K/inBLC, which encodes an intron-containing

15  blasticidin gene (Fig. 6A), in HeLa cells using a ligation-mediated PCR to amplify the

16  host-virus junction (see Method for more details). We identified total 311 HERV-K

17  LTR integration sites in the genome of HeLa cells (Fig. 6B). Nine of these 311

18  HERV-K LTR integration sites (1p13.2, K1; 4p16c, K6; 6p21.32; 6q26, K12;

19  10q24.2b, De12; 11q12.2, K18; 15q13.1; 15q22.2, K24; 19q12, K28) were consistent

20  with previously discovered non-reference HERV-K insertions (Subramanian et al,

21  2011) and were present in HeLa and other cell lines, including fibroblast cells (data

22  not shown). Six of the non-reference HERV-K LTR integration sites (12p13.31,

23  16p12.3, 7p22.1, 11q13.4, 11q22.1, and 19p12) were present in HeLa cells but not

24  in other cell lines, such as fibroblast cells (Table 1 Universal in HeLa, Fig. 6B and

25  6C). One of the six non-reference integration sites was almost identical to one cited

1    in a previous report by John Coffin's group (19p12b, K113), and the others have not

2    yet been reported. Compared to the universal integration sites in HeLa cells, 21 of

3    the new HERV-K/inBLC integration sites occurred in introns, exons, or intergenic

4    regions (Table 1 Specific in HeLa-inBLC, Fig. 6B, 6D-F and Supplemental Fig. S2A).

5    The clone numbers with some HERV-K integration sites (8q24.22, 16p11.2, and

6    Xp11.23) gradually increased during cell culture (Table 1 Specific, rapid-growth, Fig.

7    6D), whereas the clone numbers with the other integration sites did not increase

8    (Table 1 Specific, normal and slow-growth, Fig. 6E, and Supplemental Fig. S2A).

9    This suggests that HERV-K integrations at 8q24.22, 16p11.2, and Xp11.23 might

10   promote cell growth. Seven of the integration sites were observed in HeLa cells but

11   not in HERV-K/BLC-transfected HeLa cells (Fig. 6F, Supplemental Fig. S2B).

12   However, all of these sites in HeLa cells were presented in low clone numbers and

13   were detected in regions of repeated sequence, such as short interspersed nuclear

14   elements (SINEs). Since DNA sequence data we obtained in this study is short

15   reads, it is difficult to argue the reliability of integration sites in the repeated

16   sequences. Some of the integrated HERV-K DNAs were amplified by nested PCR

17   with the indicated primers shown in Fig. 6A. The expected amplification products of

18   ~1000–1500 bp (2F/2R) and ~3000 bp (5F/5R) were confirmed in HERV-K/inBLC-

19   transfected HeLa cells, but those of ~1500–2000 bp (3F/2R) were not (Fig. 6G,

20   Supplemental Fig. S2C). It was possible that HERV-K retrotransposition is

21   dependent on the integration machinery through the 3' poly(A) tail of RNA similar to

22   LINE1 (Doucet et al, 2015). According to our sequencing analysis, HERV-K

23   integrase yielded a 5–6-bp target-site duplication (TSD), which is conserved in the

24   stably integrated provirus, as in alpha-, beta-, gammaretroviruses and lentiviruses

25   but not in LINE1, in the regions flanking the HERV-K integration sites. Moreover, the

11

1    CMV promoter at the 5′-LTR was replaced with U3 in each integrated HERV-K

2    through reverse transcription (Fig. 6H, Supplemental Fig. S2D). Of note, although

3    transient expression of the transfected HERV-K/inBLC construct driven by the CMV

4    promoter allows one round of retrotransposition, subsequent retrotransposition does

5    not occur in HeLa cells because HERV-K LTR is not activated in the absence of

6    SOX2. In summary, these results indicate that reverse-transcribed HERV-K$_{CON}$

7    genomes are preferentially integrated into intron and inter-genes through the

8    retroviral integration machinery and potentially influence the cell proliferation

9    depending on the integration sites.

10    **Endogenous HERV-K retrotransposition occurs in iPS cells**

11    Recently, iPS cells have become potential research models for regenerative

12    medicine. To develop iPS cells, fibroblast cells are reprogrammed by at least three

13    factors: SOX2, OCT3/4, and KLF4. Therefore, we speculated that HERV-K

14    expression might be induced by SOX2 in iPS cells. As expected, large amounts of

15    HERV-K Gag mRNA were detected in iPS cells compared with NCCIT cells

16    (Supplemental Fig. S3A). It is possible that unregulated HERV-K transposes in the

17    genomes of iPS cells. To investigate HERV-K retrotransposition, we analyzed the

18    HERV-K integration sites in fibroblast cells and iPS cells from the same donor with

19    an NGS analysis (See Method for more details) (Fig. 7A). We found six non-

20    reference HERV-K insertions in both the fibroblast cells and iPS cells, which were

21    not found in HeLa cells (Fig. 7B, Table 2 Universal in fibroblast and iPS cells). Two

22    of the six non-reference HERV-K insertions were consistent with HERV-K integration

23    sites previously reported by John Coffin's group (12q12, K20; 13q31.3, K22). Other

24    non-reference HERV-K insertion sites might be unique individual-specific HERV-K

25    sites. On the other hand, we detected non-reference HERV-K integration sites in

1    inter-gene, intron and exon in iPS cells but not in fibroblast cells (Table 2 Specific in

2    iPS cells, Fig. 7C and 7D, Supplemental Fig. S3C). There were four non-reference

3    integration sites in the fibroblast cells that were not present in the iPS cells (Table 2

4    Specific fibroblast and Supplemental Fig. S3B). Compared with the fibroblast cells,

5    the number of non-reference integration sites in the iPS cells increased with time

6    (Table 2 Specific in iPS31, iPS41, Fig. 7C and 7D, Supplemental Fig. S3C). These

7    results indicate that HERV-K has retrotransposition activity and moves within the

8    genomes of iPS cells. However, we found no rapidly growing cells containing non-

9    reference HERV-K integration sites among the iPS-p41 cells (Fig. 7C), indicating that

10    clonal expansion is rare in iPS cells. These results suggest that newly integrated

11    HERV-K is not always advantageous to the cell proliferation, and cellular clonality

12    might depend on the HERV-K integration sites in iPS cells.

13

**Discussion**

15    In this study, we demonstrated that HERV-K is capable of retrotransposition in

16    SOX2-expressing cells. The transactivation of HERV-K LTR5Hs and LTR5B by

17    SOX2 is retained even after the accumulation of several mutations in these LTR

18    sequences. Although the physiological roles of HERV-K are still unknown, we found

19    that HERV-K has retrotransposition activity and moves randomly around the host

20    genome. In a blasticidin-selected cell population where SOX2 is not expressed, and

21    hence retrotransposition occurs only once after transfection, the copy number of

22    HERV-K, which is integrated into the intron of a tumor suppressor gene (*NDRG1*)

23    (Table 1), increased. This suggests that HERV-K integration may accelerate cell

24    growth by impairing the host genome and thus can cause several diseases

25    (discussed below). However, in SOX2-expressing cells, only a small number of novel

13

1  HERV-K integration sites were identified. It is possible that the SOX2-expressing

2  cells that have new integration of HERV-K may die or grow slowly during the long-

3  term culture due to the harmful impact of the HERV-K integration on the genome

4  integrity.

5  In addition to the well known role in the maintenance and re-establishment of

6  pluripotency (Avilion et al, 2003) (Takahashi & Yamanaka, 2006), SOX2 is essential

7  for central nervous system (CNS) development and the maintenance of neural stem

8  cells (Pevny & Nicolis, 2010). SOX2 is also expressed in Schwann cells (Le et al,

9  2005) and impairs Schwann cell remyelination and their functional recovery after

10  nerve injury, such as in multiple sclerosis (Roberts et al, 2017). Therefore, it is

11  conceivable that SOX2-induced expression of HERV-K might have impact on CNS

12  development, the maintenance of neural stem cells, remyelination, or recovery from

13  nerve injury. Indeed, HERV-K is implicated in several neural diseases, including

14  multiple sclerosis (Tai et al, 2008). Moreover, the HERV-K LTR integration sites

15  differ slightly among the genomes of individual humans and between human tissues,

16  and HERV-K LTR single-nucleotide polymorphisms (SNPs) are implicated in several

17  neural diseases (Wallace et al, 2018). It is possible that SOX2 might influence the

18  expression of genes adjacent to HERV-K LTR5Hs. Additionally, our results are

19  consistent with the possibility that HERV-K expression, which becomes

20  uncontrollable when the epigenetic regulation of SOX2 is disturbed, disrupts the

21  nervous system through the retrotransposition of HERV-K.

22  SOX2 is known for its association with numerous types of cancer (Weina & Utikal,

23  2014). It regulates the self-renewal and maintenance of cancer stem cell populations

24  by promoting oncogenic signaling (Bareiss et al, 2013; Chen et al, 2012; Laga et al,

25  2011). The expression of HERV-K is considerably higher in malignant tissues, such

1    as germ-cell tumors, melanomas, and ovarian cancers, than in healthy tissues

2    (Buscher et al, 2005; Conrad et al, 1997; Kurth & Bannert, 2010; Wang-Johanning et

3    al, 2007), suggesting a possibility that the transactivation of HERV-K LTR5Hs by

4    SOX2 is involved in numerous malignant tumors. Whether HERV-K expression is

5    involved in the self-renewal and maintenance of cancer stem cells is still unknown,

6    but it is possible that the impairment of the genome by HERV-K retrotransposition

7    may cause the malignancy of tumor tissues.

8        HERV-K is transiently reactivated in early human development to protect cells

9    from the threat of exogenous viral infection (Grow et al, 2015); however, HERV-K

10   retrotransposition entails a risk of genomic impairment in SOX2-expressing cells

11   such as iPS cells. According to our results, such genomic impairment is probably a

12   rare event in iPS cells (Fig. 7). In addition to the possibility that HERV-K

13   retrotransposition causes a defect in cell growth, thereby reducing cells with the

14   genomic impairment, it is possible that HERV-K retrotransposition is prohibited by

15   host restriction factors during its reverse transcription and/or integration. For

16   example, APOBEC3F, a restriction factor in cell-free HERV-K infection (Lee &

17   Bieniasz, 2007), may inhibit HERV-K retrotransposition during the reverse

18   transcription step. In the yeast *Saccharomyces cerevisiae*, Ty1 LTR retrotransposon

19   Gag forms virus-like particle as retrosome for the reverse transcription (Salinero et

20   al, 2018). APOBEC3G interacts with Ty1 Gag in the retrosome and restricts the Ty1

21   retrotransposition (Dutko et al, 2005; Schumacher et al, 2005). However, it is

22   unknown whether HERV-K Gag forms a retrosome, as does Ty1 (Salinero et al,

23   2018), or whether APOBEC3F can access the HERV-K genome in the retrosome. In

24   future, the mechanism of HERV-K retrotransposition must be clarified.

1    Transposable elements such as HERVs often provide new functions to vertebrate

2    hosts, which result in the exaptation (Johnson, 2019). Endogenous retrovirus Env,

3    which is called syncytins, is necessary for fusion of cytotrophoblasts to form the

4    multinucleate syncytiotrophoblast layer of the placenta (Lavialle et al, 2013). The

5    syncytins are involved in the convergent evolution during the changes from oviparity

6    to viviparity because syncytins are originated independently across multiple

7    mammalian lineages and a live-bearing reptile (Cornelis et al, 2017; Cornelis et al,

8    2015). Additionally, the neuronal gene Arc, which is retrotransposon Gag protein,

9    mediates intercellular signaling in neurons and is essential for the animal cognition

10   (Pastuzyn et al, 2018). It suggests that retrotransposon Gag has obtained alternative

11   functions in neurons during evolution. The function of HERV-K is still unclear, but

12   considering the acquisition of SOX2 responsive elements long time ago and the

13   retention of the competent elements in their LTRs since then, it is tempting to

14   speculate that HERV-K play important physiological roles in SOX2-expressing cells.

15

16   **Acknowledgments**

24

25   **Author's Contributions**

1    KM, YM, and TS conceived and coordinated the study. KM, SY, TM and HT

2    performed the experiments. YS and YU supported the NGS analysis. MG, TK, KT,

3    and TE prepared the iPS cells. FW supported the preparation of the knockout cells.

4    JI and II prepared the phylogenetic trees. AO supported the writing of the manuscript.

5    All authors read and approved the final manuscript.

6

7    **Declaration of Interests**

8

9    The authors declare that they have no competing interests.

10

11    **Plasmids**

12    Full-length HERV-K$_{CON}$ was kindly provided by Paul Bieniasz (Lee & Bieniasz,

13    2007). pHERV-K$_{CON}$ LTR-Luc encodes the luciferase gene, driven by HERV-K LTR.

14    pMXs-SOX2, OCT3/4, KLF4, and NANOG were obtained from Addgene.

15    CHKCinNluc and CHKCinBLC were derived from CHKCP (kindly provided from Paul

16    Bieniasz) (Lee & Bieniasz, 2007). The puromycin N-acetyl-transferase gene was

17    removed from CHKCP (CHKCP/delPuro), and a *NotI* site was inserted. Intron-

18    disrupted Nanoluc (inNluc) and blasticidin (inBLC) were designed as previously

19    reported (Xie et al, 2011). The inNluc and inBLC cassettes encode the SV40 early

20    enhancer/promoter and SV40 late poly(A) signal, respectively. These cassettes were

21    introduced into CHKCP/delPuro at the *NotI* site in an antisense orientation.

22    **Cells**

23    HeLa cells were cultured in Dulbecco's modified Eagle's medium (DMEM; Sigma)

24    supplemented with 5% fetal bovine serum (FBS). NCCIT cells (ATCC® CRL-2073™)

25    were cultured in RPMI1640 medium with 10% FBS, 1 mM sodium pyruvate and

26    Glutamax™ (Teshima et al, 1988). Human iPS cells were generated from human-

27    skin- derived fibroblasts and were cultured with mitomycin-C-treated mouse

17

1    embryonic fibroblast feeder cells, as described previously (Fusaki et al, 2009; Soga

2    et al, 2015); Eto et al., 2018).

3    **HERV-K retrotransposition assay**

4      HeLa cells were seeded in six-well plates at a density of $2 \times 10^5$ cells/well. The

5    cells were transfected with Lipofectamine 3000 Reagent (Invitrogen), according to

6    the manufacturer's protocol. The cells were harvested 1– 6 days after transfection,

7    and the nanoluciferase activities in the cells were measured with Nano-Glo®

8    Luciferase Assay Reagent (Promega).

9    **Measurement of dual-luciferase luminescence**

10    Luminescence was measured with the Dual-Luciferase Reporter Assay System

11    (Promega), according to the manufacturer's instructions. The cell lysate was mixed

12    with Luciferase Assay Reagent II. Firefly luciferase activity was measured with a

13    luminometer. *Renilla* luciferase activity was read after the cell lysate containing

14    Luciferase Assay Reagent II was mixed with Stop & Glo Reagent.

15    **Bisulfite sequencing**

16      EpiTect Plus Bisulfite conversion kit (Qiagen) was used as described previously

17    (Grow et al, 2015). The PCR fragments were inserted into pCR-BluntII-Topo vector

18    (Invitrogen). Approximately 10 clones in HeLa and NCCIT cells were Sanger

19    sequenced for quantifying the CpG methylation.

20    **ChIP assay**

21      NCCIT cells were fixed with 1% formaldehyde and lysed with 20% NP-40 (10 mM

22    HEPES-KOH pH 7.9, 1.5 mM $MgCl_2$, 10 mM KCl, 0.5 mM DTT, and 20% NP-40 with

23    protease inhibitor cocktail [Roche]). The chromatin in the lysates was fragmented to

24    320 bp after digestion with micrococcal nuclease. After further lysis with 10% SDS

25    (50 mM Tris-HCl pH 8.1, 0.2 mM EDTA, 10% SDS, with protease inhibitor cocktail

18

1    [Roche]), the chromatin was sonicated. The DNA–protein complexes were

2    precipitated overnight by incubation with an anti-SOX2 antibody (BioLegend),and

3    then incubated with ChIP-Grade Protein G Magnetic Beads (9006, Cell Signaling

4    Technology) for 2 h. The abundance of HERV-K LTR in the precipitated-DNA was

5    analyzed with quantitative PCR using primers: HERV-K LTR primer F 5′-

6    AGCACTGAGATGTTTATGTG-3′ and R 5′-TGTGGGGAGAGGGTCAGC -3′ and

7    SYBR Premix ExTaq II (Takara Bio Inc.). The signal intensity was quantified with the

8    ABI 7900HT Fast Real-Time PCR System (Applied Biosystems).

9    **Linker-mediated (LM)-PCR**

10    The HERV-K integration sites were analyzed with LM-PCR and high-throughput

11    sequencing, as previously described (Gillet et al, 2011; Satou et al, 2017). To

12    analyze HERV-K integration site, the junction between the 3′-LTR of HERV-K and

13    the host genomic DNA was amplified with a primer targeting the 3′-LTR and the

14    linker. The first forward primer targeting the 3′-LTR was B3-K1: 5′-

15    CCTCCATATGCTGAACGCTGGTT-3′; the second forward primer targeting the 3′-

16    LTR was P5B5-K2: 5′-

17    AATGATACGGCGACCACCGAGATCTACACCCAAATCTCTCGTCCCACCTTACGA

18    GAAACACCCACAGG-3′. The DNA libraries were sequenced as paired-end reads

19    with Illumina MiSeq, and the resulting fastq files were analyzed. The sequencing

20    primer targeting the 3′-LTR was Seq-K1: 5′-

21    ACACCCACAGGTGTGTAGGGGCAACCCACC-3′. The flanking host genome

22    sequences were used to determine viral integration sites. The resulting short reads

23    were cleaned using an in house script, which extracts reads with a high index-read

24    sequencing quality (Phred score > 20) in each position of an 8-bp index read. The

25    clean sequencing reads were aligned with HERV-K LTR sequences (CCTACA and

1    CCTTCA) and were mapped to human genome using the BWA-MEM algorithm (Li &

2    Durbin, 2009). Further data processing and cleanup, including the removal of reads

3    with multiple alignments and duplicated reads, were performed using Samtools (Li &

4    Durbin, 2009) and Picard (http://broadinstitute.github.io/picard/). The clone numbers

5    with each HERV-K integration site were quantified as previous reports (Gillet et al,

6    2011; Satou et al, 2017).

**Western blotting analysis**

8    Cells and virus were lysed with 1% Triton X lysis buffer (50 mM Tris-HCl pH 7.5

9    containing 0.5% Triton X-100, 300 mM NaCl, 10 mM iodoacetamide, and protease

10   inhibitor cocktail [Roche]). After 2 x SDS sample buffer was added, the SOX2,

11   OCT3/4, GAPDH and HERV-K Gag proteins were detected with immunoblotting

12   using an anti-SOX2 antibody (Merck Millipore), an anti-OCT3/4 antibody (BD

13   Biosciences), an anti-GAPDH antibody (Sigma) and an anti-HERV-K Gag antibody

14   (Austral Biologicals), respectively, as the primary antibodies. Horseradish peroxidase

15   (HRP)-conjugated anti-mouse Ig antibody (Jackson ImmunoResearch) was used as

16   the secondary antibody. The HRP-conjugated secondary antibody was detected with

17   Chemi-Lumi One L (Nacalai Tesque).

**Reverse transcription–quantitative PCR analysis**

19   Total RNA was purified with the RNeasy Mini Kit (Qiagen). The mRNA was reverse

20   transcribed with *Murine leukemia virus* (MLV) reverse transcriptase after it was

21   annealed with a poly(T) primer. HERV-K *gag* DNA was amplified with primers:

22   HERV-K Gag CA forward primer 5′- CAAGACCCAGGAAGTACCT-3′ and reverse

23   primer 5′-ACACTCAGGATTGGCGTT-3′. All qPCR assays were performed with

24   SYBR Premix ExTaq II (Takara Bio Inc.). The data for the target genes were then

25   normalized to the expression level of glyceraldehyde 3-phosphate dehydrogenase

20

1　(*GAPDH*; housekeeping gene), amplified with GAPDH forward primer 5′-

2　CGCTCTCTGCTCCTCCTGTT-3′ and reverse primer 5′-

3　ACAAAGTGGTCGTTGAGGGC-3′.

4　**Flow cytometry analysis**

5　　Total DNA was extracted from NCCIT cells with the DNeasy Blood & Tissue kit

6　(Qiagen). HERV-K LTR was amplified with HERV-K LTR primers (5Hs forward

7　primer: 5′-CCAAAAGCCATCGATTGTGGGGAAAAGCAAGAGAG-3′; 5Hs 5′LTR

8　reverse primer: 5′-TTCCATCTCGAGTGAAGTGGGGCCAGCCCCTCCACACCT-3′;

9　5Hs 3′LTR reverse primer: 5′-

10　TTCCATCTCGAGTGTAGGGGTGGGTTGCCCCTCCACACC-3′; 5A forward primer:

11　5′-AAAGCCATCGATTGTAGGGAAAAGAAAGAGAGATCAGAC-3; 5A 5′LTR

12　reverse primer: 5′-TTCCATCTCGAGTGAAGGGGTGGCCTGCCCCTCCA-3′; 5A

13　3′LTR: reverse primer 5′-TTCCATCTCGAGCTCCACACCTGTGGGTAT-3′; 5B

14　forward primer: 5′-AAAGCCATCGATTGTAGGGAAAAGAAAGAGAGATCAG-3′; 5B

15　5′LTR reverse primer: 5′-

16　TTCCATCTCGAGTGAAGTGGGGCCAGCCCCTCCACACCT-3′; 5B 3′LTR reverse

17　primer: 5′-TTCCATCTCGAGCTCCACACCTGTGGGTATTTCT-3′). The HERV-K

18　LTR was inserted upstream from the yellow-fluorescent-protein-encoding gene

19　(HERV-K LTR-Venus). HeLa and NCCIT cells were cotransfected with pHERV-K

20　LTR-Venus and pMXs-SOX2. Two days after transfection, the fluorescent signals

21　were analyzed with flow cytometry.

22

23　**Figure legends**

21

1    **Fig. 1.** SOX2 contributes to the promoter function of HERV-K LTR. (A) SOX2 and

2    OCT3/4 binding motifs were identified in the HERV-K LTR with the PROMO

3    software, which is used to identify transcription factor binding motifs. (B) Binding

4    sites of SOX2, OCT3/4 and NANOG on respective LTR5Hs copies in human ES

5    cells were collected from ENCODE ChIP-Seq dataset, and the positions on the

6    consensus sequence of LTR5Hs are shown. (C) HeLa cells were cotransfected with

7    the plasmid pHERV-K$_{CON}$ LTR-Luc, the indicated plasmids, and the *Renilla*-Luc

8    plasmid. The firefly and renilla luciferase activities were measured. (D) HeLa cells

9    were cotransfected with pHERV-K LTR-Luc and different amounts of the indicated

10    plasmids. The luciferase activity was measured. (E) Amounts of OCT3/4, SOX2, and

11    GAPDH proteins were measured with western blotting. (F) Chromatins in HeLa and

12    NCCIT cells were extracted, and SOX2-binding DNA fragments were

13    immunoprecipitated with the indicated antibodies. HERV-K LTRs in the

14    immunoprecipitated DNA were measured with qPCR. (G) Amounts of SOX2 and

15    GAPDH proteins in NCCIT and SOX2-knockout NCCIT (NCCIT/KOSOX2) cells were

16    measured with western blotting. The sequences of SOX2 in each cell were analyzed.

17    (H) NCCIT and NCCIT/KOSOX2 cells were cotransfected with pHERV-K LTR-Luc

18    and SOX2-encoding plasmids. The luciferase activity was measured. (C, D, F, and

19    H) Data from three independent experiments are shown as means ± standard

20    deviations. *P* values were determined with Student's *t* test. *$P$ < 0.01; **$P$ < 0.001;

21    ***$P$ < 0.0001; n.s., not significant. (I) Amounts of mature HERV-K Gag in the

22    supernatants of NCCIT and NCCIT/KOSOX2 cells were measured with western

23    blotting.

24

1    **Fig. 2.** Multiple SOX2-binding motifs contribute to HERV-K transcription. HeLa (A

2    and C) and NCCIT cells (B and D) were cotransfected with pHERV-K LTR mutants

3    and the indicated plasmids. The luciferase activity was measured. (A, B, C, and D)

4    Data from three independent experiments are shown as means ± standard

5    deviations. *P* values were determined with Student's *t* test. *$P < 0.01$; **$P < 0.001$;

6    ***$P < 0.0001$; n.s., not significant.

7

8    **Fig. 3.** HERV-K genome is hypermethylated in HeLa cells. (A) There is likely to be a

9    CpG island (11 CG nucleotides) between HERV-K LTR and *gag*. (B and C) DNA was

10   extracted from NCCIT (B) and HeLa cells (C). The sequences of nine HERV-K

11   genomes between LTR and Gag in NCCIT cells and 12 HERV-K genomes in HeLa

12   cells were analyzed after the DNAs were treated with bisulfite to convert cytosine

13   residues to uracil. White circles indicate unmethylated nucleotides and black circles

14   indicate methylated nucleotides in the CpG island. (D) HeLa cells were treated with

15   5-aza-2′-deoxycytidine for 1 day, and then transfected with a plasmid encoding

16   SOX2. Two days after transfection, the amounts of HERV-K Gag mRNA were

17   measured with RT–qPCR. Data from three independent experiments are shown as

18   means ± standard deviations. *P* values were determined using Student's *t* test. *$P <$

19   $0.01$; **$P < 0.001$; n.s., not significant.

20

21   **Fig. 4.** HERV-K transactivation by SOX2 is conserved among HERV-K LTR5Hs.

22   HERV-K LTR series were amplified with PCR from genome into NCCIT cells and

23   inserted upstream from the *YFP* gene. HeLa cells (A) were cotransfected with

24   plasmid encoding SOX2 and the pHERV-K LTR–YFP series. NCCIT (Parent) and

25   NCCIT/KOSOX2 cells (B) were transfected with the pHERV-K LTR-YFP series. The

23

1  yellow fluorescent protein (YFP)-positive cells were analyzed with flow cytometry. (C)

2  Neighbor-joining tree was constructed based on the aligned nucleotide sequences

3  corresponding to HERV-K LTRs within NCCIT cells.

4

5  **Fig. 5.** HERV-K has retrotransposition activity in HeLa cells. (A) The construction of

6  pHERV-K GagProPol/inNanoluc is described. The 5′ U3 region was replaced with

7  the CMV promoter. Intron-disrupted Nanoluc (inNluc) and SV40 were introduced into

8  the Env region in an antisense orientation. (B) HeLa cells were transfected with

9  pHERV-K GagProPol/inNanoluc or pHERV-K del/GagProPol/inNanoluc.

10  Nanoluciferase activity was measured each day with a nanoluciferase reporter

11  assay. (C) HeLa cells were transfected with pHERV-K mutants/inNanoluc alone or

12  cotransfected with pHERV-K mutants/inNanoluc and HERV-K GagProPol. Five days

13  after transfection, spliced nanoluciferase activity within the retrotransposed HERV-K

14  was measured as the nanoluciferase activity. (B and C) $P$ values were determined

15  with Student's $t$ test. *$P < 0.01$; **$P < 0.001$; ***$P < 0.0001$; n.s., not significant.

16

17  **Fig. 6.** New integration sites for HERV-K appeared in HERV-K-transfected HeLa

18  cells. (A) Construction of pHERV-K GagProPol/inBlasticidin (inBLC) is described.

19  inNluc was replaced with inBlasticidin (inBLC) in the Env region of HERV-K. HeLa

20  cells were transfected with pHERV-K inBLC. Blasticidin-resistant cells were selected

21  2 weeks after transfection, and the HERV-K DNA in the genome was then amplified

22  by PCR and determined with NGS analysis. The primers were designed to bind the

23  outside of repeated sequences (4F, 5F, 1R and 2R arrowheads), (B) HERV-K

24  integration sites that are present in the database are shown in red (Reference). Non-

25  reference HERV-K integration sites in the HeLa and HeLa-inBLC cells, but not in

24

1    fibroblasts, are shown in yellow (passage 0), black (passage 1), green (passage 2),

2    purple (passage 17), and blue (passage 31). The outer most ring is G-band of

3    human chromosome. (C) Universal non-reference HERV-K insertions (Universal)

4    were detected in both HeLa and HeLa-inBLC cells, but not in fibroblast cells. (D and

5    E) Specific non-reference HERV-K insertions (Specific) were detected in HeLa-

6    inBLC-p1, -p2, -p17, and -p31 cells, but not in HeLa-p0. Cell growth rates were

7    classified as rapid (D) or normal (E). (C, D and E) Gray letters indicate HERV-K-

8    integration into a LINE-1, SINE, or repeat elements. These data were collected from

9    two independent samples. (F) The number of specific non-reference HERV-K

10   locations in HeLa and HeLa-inBLC-p1, -p2, -p17, and -p31 cells were determined.

11   (G) HERV-K integration sites (8q24.22 and 16p11.2) in HeLa-inBLC-p1 cells were

12   confirmed with PCR. Arrows (A) indicate primer-binding sites for PCR. The primers,

13   which annealed at the integration sites, were designed to sequences adjacent to

14   HERV-K LTR (A). (H) Sequences between HERV-K LTR and the neighboring HERV-

15   K genomes in HERV-K/inBLC-transfected HeLa cells were analyzed with Sanger

16   sequencing. TSD indicates target-site duplication in the human genome generated

17   by integrase.

18

19   **Fig. 7.** New integration sites for HERV-K appeared in iPS cells. (A) Integration sites

20   of HERV-K were determined with NGS analysis. iPS cells were derived from

21   fibroblasts, then the cells were passaged 31 times (iPS-p31) or 41 times (iPS-p41).

22   HERV-K integration sites that were already present in the database (Reference) are

23   shown in red. Non-reference HERV-K integration sites in fibroblasts and iPS cells but

24   not in HeLa cells are shown in green (fibroblast), purple (iPS cells; passage 31), and

25   blue (iPS cells; passage 41). (B) Universal non-reference HERV-K insertions were

25

1   specifically determined in this donor. (C) Specific non-reference HERV-K insertions

2   were specifically identified in iPS cells. (B and C) Gray letters indicate HERV-K-

3   integration into LINE-1, SINE, or repeat elements. These data were collected from

4   four independent samples. (D) The number of specific non-reference HERV-K

5   locations in fibroblasts and iPS cells were determined.

6

7   **Table 1.** Loci of new HERV-K integration sites in HERV-K-transfected HeLa cells.

8   [a]Universal HERV-K integration sites in HeLa cells.

9   [b]Different HERV-K integration sites between HeLa and HERV-K/inBLC-transfected

10  HeLa cells.

11  [c]Consistent with a previous report from John Coffin's group (Subramanian et al,

12  2011).

13  [d]HERV-K flanking region is in the repetitive sequence.

14

15  **Table 2.** Loci of new HERV-K integration sites in iPS cells.

16  [a]Universal HERV-K integration sites in this donor.

17  [b]Different HERV-K integration sites between fibroblast and iPS cells.

18  [c]Consistent with a previous report from John Coffin's group (Subramanian et al,

19  2011).

20  [d]HERV-K flanking region is in the repetitive sequence.

21

22  **Fig. S1.** SOX2 and OCT3/4 contribute to the promoter function of HERV-K LTR. (A)

23  SOX2- and OCT3/4-binding motifs were identified in the HERV-K LTR with the

24  PROMO software, which is used to identify putative transcription factors. Deletion

25  mutants of LTR were constructed. (B–D) NCCIT cells were transfected with the

1    pHERV-K$_{CON}$ mutants. Two days after transfection, the firefly luciferase activities

2    were measured with a luciferase reporter assay. (E) HeLa cells were cotransfected

3    with pHERV-K$_{CON}$ mutants and the indicated plasmids. Two days after transfection,

4    the firefly luciferase activities were measured with a luciferase reporter assay. (F)

5    HeLa cells were cotransfected with the pHERV-K$_{CON}$ mutants, the indicated plasmids,

6    and the Renilla-Luc plasmid. Two days after transfection, both firefly luciferase and

7    *Renilla* luciferase activities were measured with a dual luciferase reporter assay.

8

9    **Fig. S2.** New integration sites for HERV-K appeared in HERV-K-transfected HeLa

10   cells. (A and B) HeLa cells were transfected with pHERV-K-inBLC. Blasticidin-

11   resistant cells were selected 2 weeks after transfection, and the HERV-K DNA in the

12   genome was then amplified with PCR and analyzed with next-generation

13   sequencing. Specific non-reference HERV-K insertions (Specific) were detected in

14   HeLa (B) or HeLa-inBLC-p1, -p2, -p17, and -p31 cells (A). Cell growth speeds was

15   classified as slow (A). Gray letters indicate HERV-K-integration in LINE-1, SINE, or

16   repeat elements. (C) HERV-K integration sites (1q36.21 and 12p13.33) in HeLa-

17   inBLC-p1 cells were confirmed with PCR. (H) Sequences between HERV-K LTR and

18   the neighboring HERV-K genomes in HERV-K/inBLC-transfected HeLa cells were

19   analyzed with Sanger sequencing. TSD is a target-site duplication in the human

20   genome generated by integrase..

21

22   **Fig. S3.** New integration sites of HERV-K appeared in iPS cells. (A) HERV-K Gag

23   mRNA and GAPDH mRNA expression in NCCIT and iPS cells was measured with

24   reverse transcription–PCR. The integration sites of HERV-K were determined with

25   next-generation sequencing. (B and C) Specific non-reference HERV-K insertions

1     were specifically determined in fibroblasts (B) and iPS-31 cells (C). Gray letters

2     indicate HERV-K-integration in LINE-1, SINE, or repeat elements.

3

4     **Reference**

5

6

7    **Avilion AA, Nicolis SK, Pevny LH, Perez L, Vivian N, Lovell-Badge R (2003)**
8    **Multipotent cell lineages in early mouse development depend on SOX2 function.** *Genes*
9    *& development* **17: 126-140**
10
11   **Bannert N, Kurth R (2004) Retroelements and the human genome: new perspectives on**
12   **an old relation.** *Proceedings of the National Academy of Sciences of the United States of*
13   *America* **101 Suppl 2: 14572-14579**
14
15   **Bareiss PM, Paczulla A, Wang H, Schairer R, Wiehr S, Kohlhofer U, Rothfuss OC,**
16   **Fischer A, Perner S, Staebler A, Wallwiener D, Fend F, Fehm T, Pichler B, Kanz L,**
17   **Quintanilla-Martinez L, Schulze-Osthoff K, Essmann F, Lengerke C (2013) SOX2**
18   **expression associates with stem cell state in human ovarian carcinoma.** *Cancer research*
19   **73: 5544-5555**
20
21   **Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, Badge RM, Moran**
22   **JV (2010) LINE-1 retrotransposition activity in human genomes.** *Cell* **141: 1159-1170**
23
24   **Beck CR, Garcia-Perez JL, Badge RM, Moran JV (2011) LINE-1 elements in structural**
25   **variation and disease.** *Annual review of genomics and human genetics* **12: 187-215**
26
27   **Beimforde N, Hanke K, Ammar I, Kurth R, Bannert N (2008) Molecular cloning and**
28   **functional characterization of the human endogenous retrovirus K113.** *Virology* **371:**
29   **216-225**
30
31   **Bieda K, Hoffmann A, Boller K (2001) Phenotypic heterogeneity of human endogenous**
32   **retrovirus particles produced by teratocarcinoma cell lines.** *The Journal of general*
33   *virology* **82: 591-596**
34
35   **Boeke JD, Stoye JP (1997) Retrotransposons, Endogenous Retroviruses, and the**
36   **Evolution of Retroelements.**
37
38   **Boller K, Frank H, Lower J, Lower R, Kurth R (1983) Structural organization of**
39   **unique retrovirus-like particles budding from human teratocarcinoma cell lines.** *The*
40   *Journal of general virology* **64 ( Pt 12): 2549-2559**
41

Boller K, Schonfeld K, Lischer S, Fischer N, Hoffmann A, Kurth R, Tonjes RR (2008) Human endogenous retrovirus HERV-K113 is capable of producing intact viral particles. *The Journal of general virology* **89:** 567-572

Bourc'his D, Bestor TH (2004) Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431:** 96-99

Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH, Jr. (2003) Hot L1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences of the United States of America* **100:** 5280-5285

Buscher K, Trefzer U, Hofmann M, Sterry W, Kurth R, Denner J (2005) Expression of human endogenous retrovirus K in melanomas and melanoma cell lines. *Cancer research* **65:** 4172-4180

Chen S, Xu Y, Chen Y, Li X, Mou W, Wang L, Liu Y, Reisfeld RA, Xiang R, Lv D, Li N (2012) SOX2 gene regulates the transcriptional network of oncogenes and affects tumorigenesis of human lung cancer cells. *PloS one* **7:** e36326

Conrad B, Weissmahr RN, Boni J, Arcari R, Schupbach J, Mach B (1997) A human endogenous retroviral superantigen as candidate autoimmune gene in type I diabetes. *Cell* **90:** 303-313

Cornelis G, Funk M, Vernochet C, Leal F, Tarazona OA, Meurice G, Heidmann O, Dupressoir A, Miralles A, Ramirez-Pinilla MP, Heidmann T (2017) An endogenous retroviral envelope syncytin and its cognate receptor identified in the viviparous placental Mabuya lizard. *Proceedings of the National Academy of Sciences of the United States of America* **114:** E10991-E11000

Cornelis G, Vernochet C, Carradec Q, Souquere S, Mulot B, Catzeflis F, Nilsson MA, Menzies BR, Renfree MB, Pierron G, Zeller U, Heidmann O, Dupressoir A, Heidmann T (2015) Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. *Proceedings of the National Academy of Sciences of the United States of America* **112:** E487-496

Doucet AJ, Wilusz JE, Miyoshi T, Liu Y, Moran JV (2015) A 3' Poly(A) Tract Is Required for LINE-1 Retrotransposition. *Molecular cell* **60:** 728-741

Dutko JA, Schafer A, Kenny AE, Cullen BR, Curcio MJ (2005) Inhibition of a yeast LTR retrotransposon by human APOBEC3 cytidine deaminases. *Current biology : CB* **15:** 661-666

Farre D, Roset R, Huerta M, Adsuara JE, Rosello L, Alba MM, Messeguer X (2003) Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN. *Nucleic acids research* **31:** 3651-3653

Fuchs NV, Kraft M, Tondera C, Hanschmann KM, Lower J, Lower R (2011) Expression of the human endogenous retrovirus (HERV) group HML-2/HERV-K does

not depend on canonical promoter elements but is regulated by transcription factors Sp1 and Sp3. *Journal of virology* 85: 3436-3448

Fusaki N, Ban H, Nishiyama A, Saeki K, Hasegawa M (2009) Efficient induction of transgene-free human pluripotent stem cells using a vector based on Sendai virus, an RNA virus that does not integrate into the host genome. *Proceedings of the Japan Academy Series B, Physical and biological sciences* 85: 348-362

Gillet NA, Malani N, Melamed A, Gormley N, Carter R, Bentley D, Berry C, Bushman FD, Taylor GP, Bangham CR (2011) The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones. *Blood* 117: 3113-3122

Gonzalez-Hernandez MJ, Swanson MD, Contreras-Galindo R, Cookinham S, King SR, Noel RJ, Jr., Kaplan MH, Markovitz DM (2012) Expression of Human Endogenous Retrovirus Type-K (HML-2) is activated by the Tat protein of HIV-1. *Journal of virology*

Grow EJ, Flynn RA, Chavez SL, Bayless NL, Wossidlo M, Wesche DJ, Martin L, Ware CB, Blish CA, Chang HY, Pera RA, Wysocka J (2015) Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* 522: 221-225

Han JS, Szak ST, Boeke JD (2004) Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* 429: 268-274

Hancks DC, Kazazian HH, Jr. (2012) Active human retrotransposons: variation and disease. *Current opinion in genetics & development* 22: 191-203

Johnson WE (2019) Origins and evolutionary consequences of ancient endogenous retroviruses. *Nature reviews Microbiology* 17: 355-370

Katoh I, Mirova A, Kurata S, Murakami Y, Horikawa K, Nakakuki N, Sakai T, Hashimoto K, Maruyama A, Yonaga T, Fukunishi N, Moriishi K, Hirai H (2011) Activation of the long terminal repeat of human endogenous retrovirus K by melanoma-specific transcription factor MITF-M. *Neoplasia* 13: 1081-1092

Kurth R, Bannert N (2010) Beneficial and detrimental effects of human endogenous retroviruses. *International journal of cancer* 126: 306-314

Laga AC, Zhan Q, Weishaupt C, Ma J, Frank MH, Murphy GF (2011) SOX2 and nestin expression in human melanoma: an immunohistochemical and experimental study. *Experimental dermatology* 20: 339-345

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S,

1   Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH,
2   Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla
3   AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL,
4   Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW,
5   Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N,
6   Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM,
7   Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML,
8   Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A,
9   Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T,
10  Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E,
11  Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee
12  HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu
13  J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel
14  NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR,
15  Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M,
16  Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR,
17  de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind
18  L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti
19  L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey
20  TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H,
21  Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ,
22  Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T,
23  Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF,
24  Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J,
25  Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS,
26  Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P,
27  Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ (2001) Initial sequencing and
28  analysis of the human genome. *Nature* 409: 860-921
29
30  Lavialle C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, Vernochet C, Heidmann
31  T (2013) Paleovirology of 'syncytins', retroviral env genes exapted for a role in
32  placentation. *Philosophical transactions of the Royal Society of London Series B,*
33  *Biological sciences* 368: 20120507
34
35  Le N, Nagarajan R, Wang JY, Araki T, Schmidt RE, Milbrandt J (2005) Analysis of
36  congenital hypomyelinating Egr2Lo/Lo nerves identifies Sox2 as an inhibitor of
37  Schwann cell differentiation and myelination. *Proceedings of the National Academy of*
38  *Sciences of the United States of America* 102: 2596-2601
39
40  Lee YN, Bieniasz PD (2007) Reconstitution of an infectious human endogenous
41  retrovirus. *PLoS Pathog* 3: e10
42
43  Levin HL, Moran JV (2011) Dynamic interactions between transposable elements and
44  their hosts. *Nature reviews Genetics* 12: 615-627
45
46  Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler
47  transform. *Bioinformatics* 25: 1754-1760
48

1  Messeguer X, Escudero R, Farre D, Nunez O, Martinez J, Alba MM (2002) PROMO:
2  detection of known transcription regulatory elements using species-tailored searches.
3  *Bioinformatics* **18: 333-334**

4
5  Mills RE, Bennett EA, Iskow RC, Devine SE (2007) Which transposable elements are
6  active in the human genome? *Trends in genetics : TIG* **23: 183-191**

7
8  Monde K, Contreras-Galindo R, Kaplan MH, Markovitz DM, Ono A (2012) Human
9  endogenous retrovirus K Gag coassembles with HIV-1 Gag and reduces the release
10 efficiency and infectivity of HIV-1. *Journal of virology* **86: 11194-11208**

11
12 Monde K, Terasawa H, Nakano Y, Soheilian F, Nagashima K, Maeda Y, Ono A (2017)
13 Molecular mechanisms by which HERV-K Gag interferes with HIV-1 Gag assembly
14 and particle infectivity. *Retrovirology* **14: 27**

15
16 Pastuzyn ED, Day CE, Kearns RB, Kyrke-Smith M, Taibi AV, McCormick J, Yoder N,
17 Belnap DM, Erlendsson S, Morado DR, Briggs JAG, Feschotte C, Shepherd JD (2018)
18 The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that
19 Mediates Intercellular RNA Transfer. *Cell* **173: 275**

20
21 Pevny LH, Nicolis SK (2010) Sox2 roles in neural stem cells. *The international journal of*
22 *biochemistry & cell biology* **42: 421-424**

23
24 Roberts SL, Dun XP, Doddrell RDS, Mindos T, Drake LK, Onaitis MW, Florio F,
25 Quattrini A, Lloyd AC, D'Antonio M, Parkinson DB (2017) Sox2 expression in
26 Schwann cells inhibits myelination in vivo and induces influx of macrophages to the
27 nerve. *Development* **144: 3114-3125**

28
29 Salinero AC, Knoll ER, Zhu ZI, Landsman D, Curcio MJ, Morse RH (2018) The
30 Mediator co-activator complex regulates Ty1 retromobility by controlling the balance
31 between Ty1i and Ty1 promoters. *PLoS genetics* **14: e1007232**

32
33 Satou Y, Katsuya H, Fukuda A, Misawa N, Ito J, Uchiyama Y, Miyazato P, Islam S,
34 Fassati A, Melamed A, Bangham CRM, Koyanagi Y, Sato K (2017) Dynamics and
35 mechanisms of clonal expansion of HIV-1-infected cells in a humanized mouse model.
36 *Scientific reports* **7: 6913**

37
38 Schumacher AJ, Nissley DV, Harris RS (2005) APOBEC3G hypermutates genomic
39 DNA and inhibits Ty1 retrotransposition in yeast. *Proceedings of the National Academy*
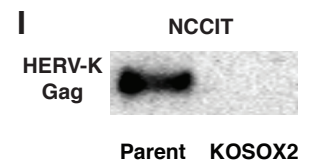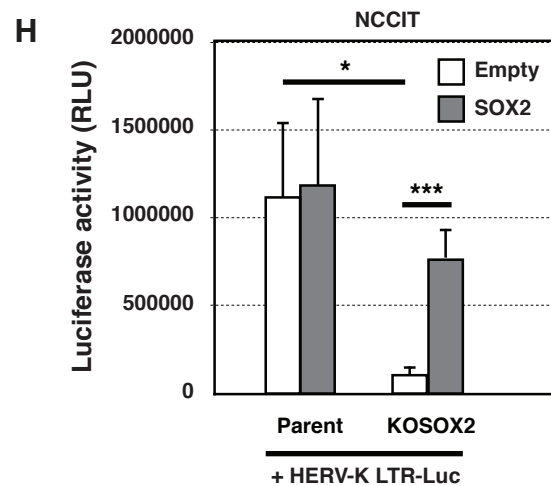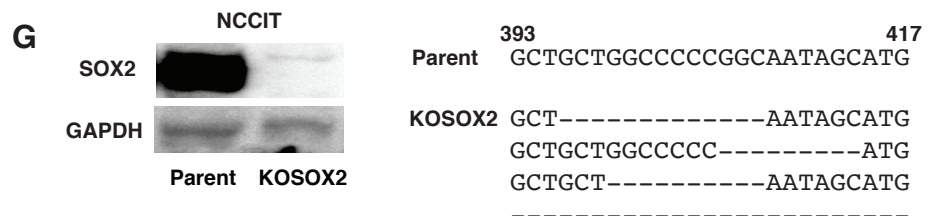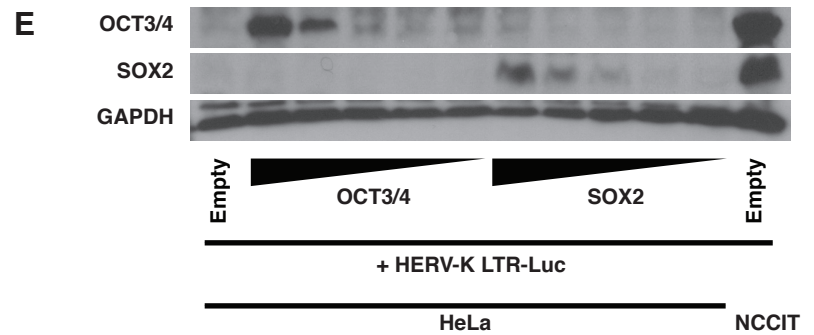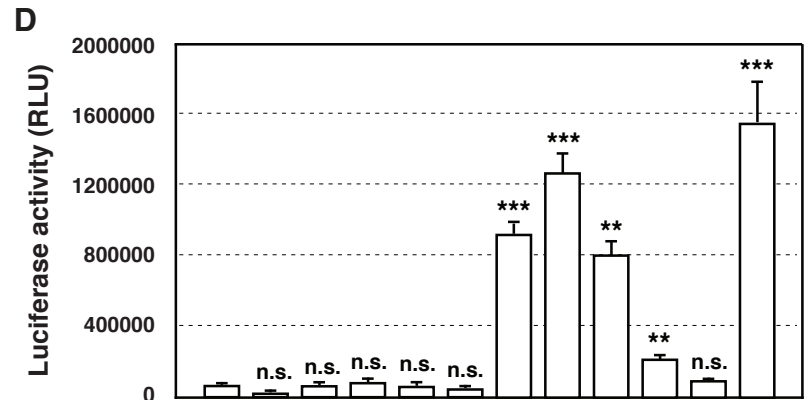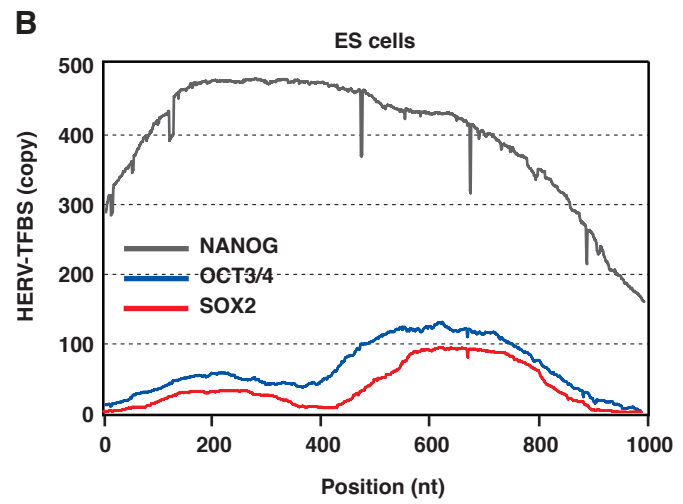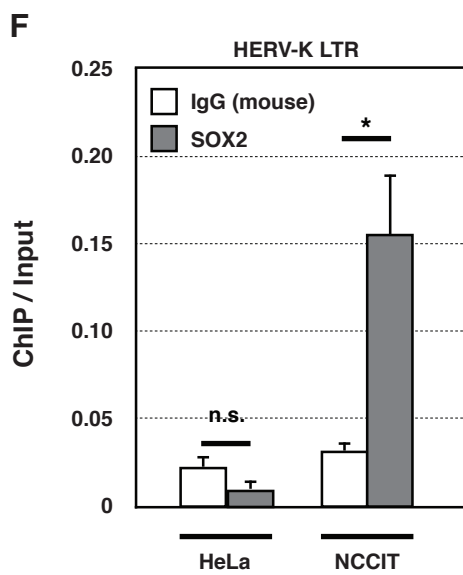40 *of Sciences of the United States of America* **102: 9854-9859**

41
42 Soga M, Ishitsuka Y, Hamasaki M, Yoneda K, Furuya H, Matsuo M, Ihn H, Fusaki N,
43 Nakamura K, Nakagata N, Endo F, Irie T, Era T (2015) HPGCD outperforms HPBCD
44 as a potential treatment for Niemann-Pick disease type C during disease modeling with
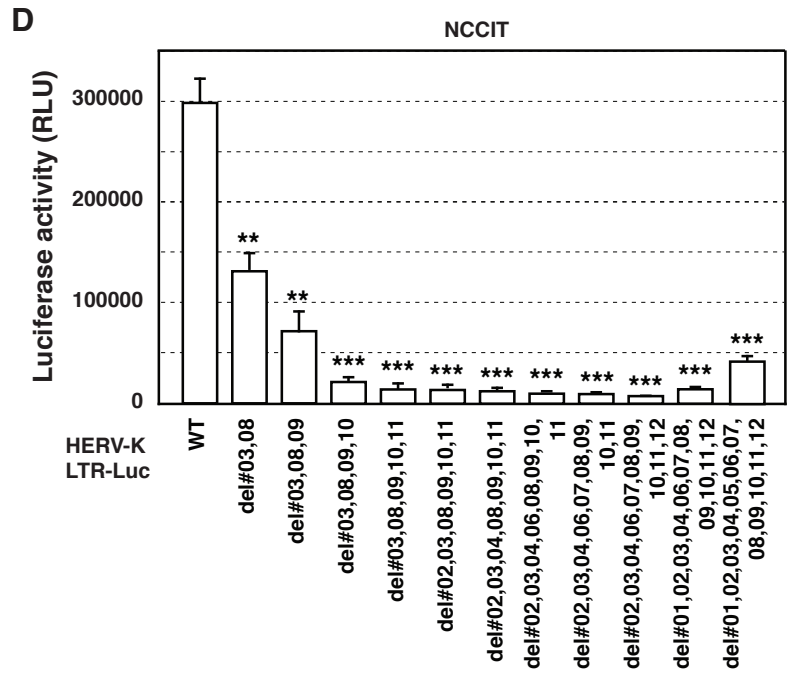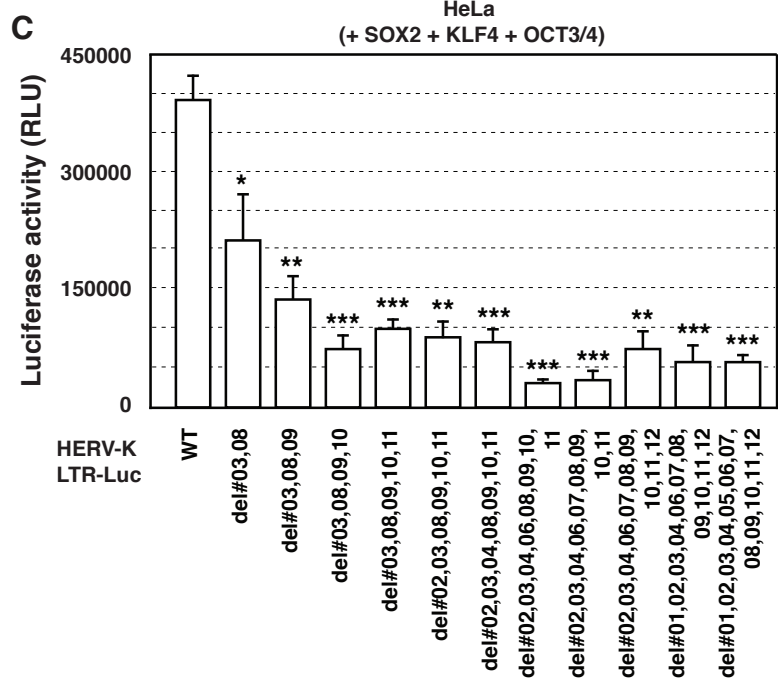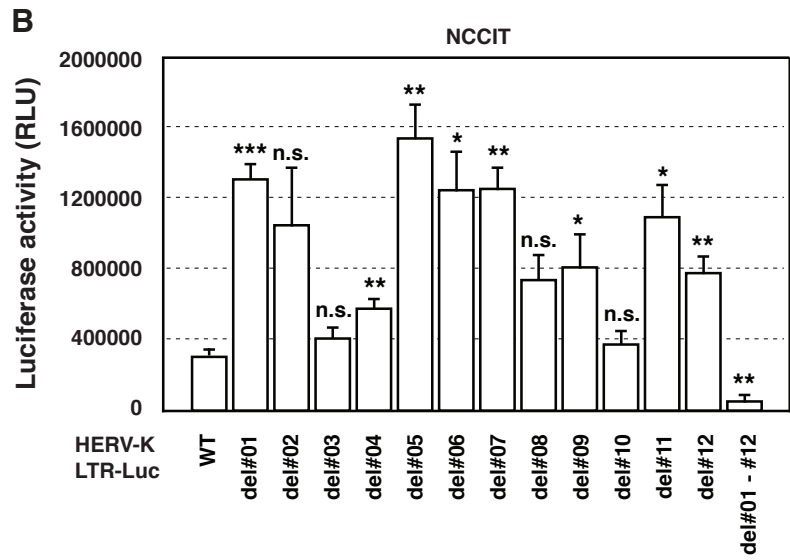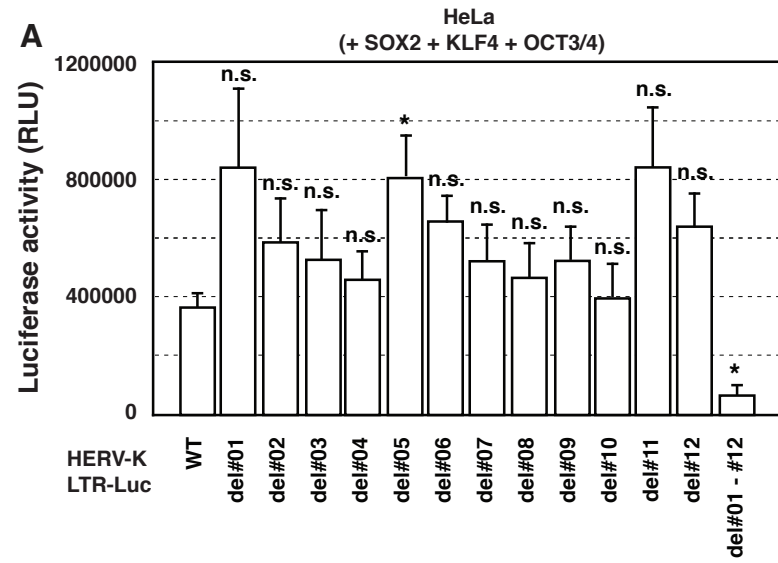45 iPS cells. *Stem cells* **33: 1075-1088**

46
47 Stoye JP (2012) Studies of endogenous retroviruses reveal a continuing evolutionary
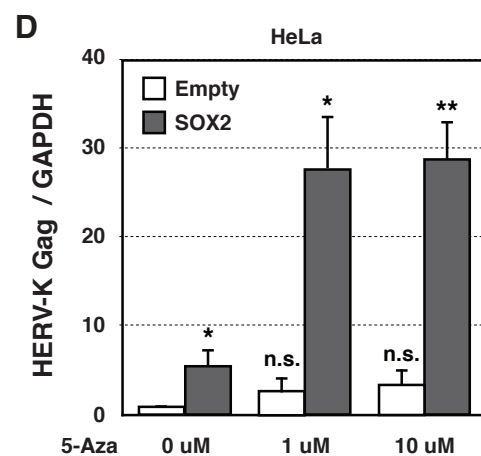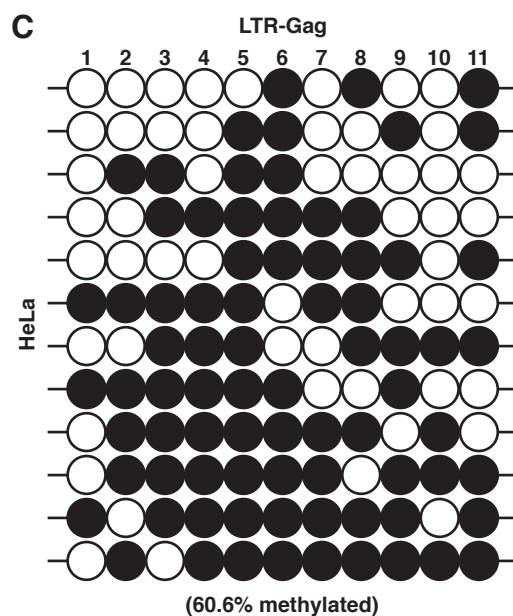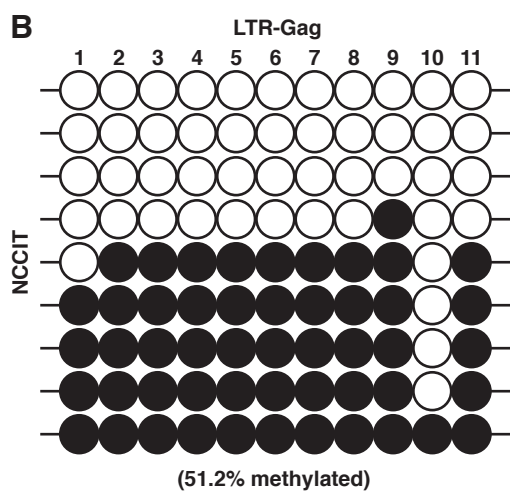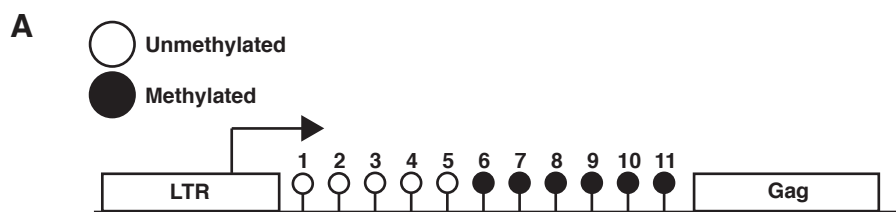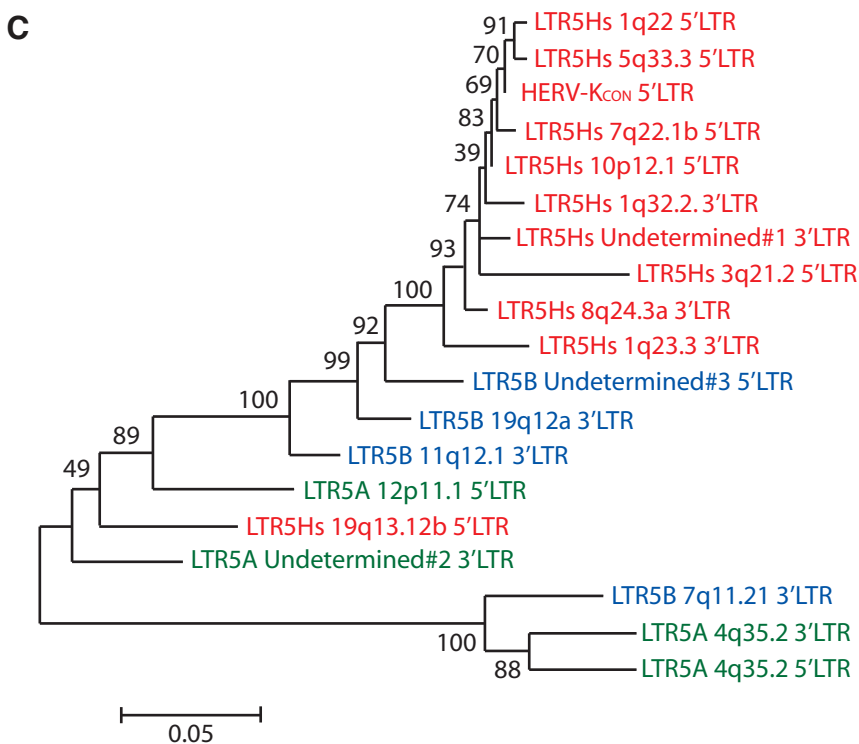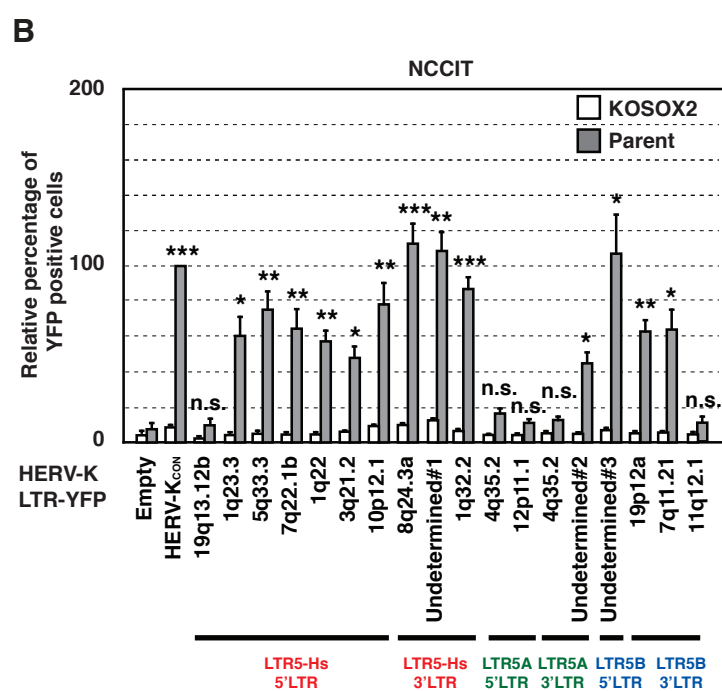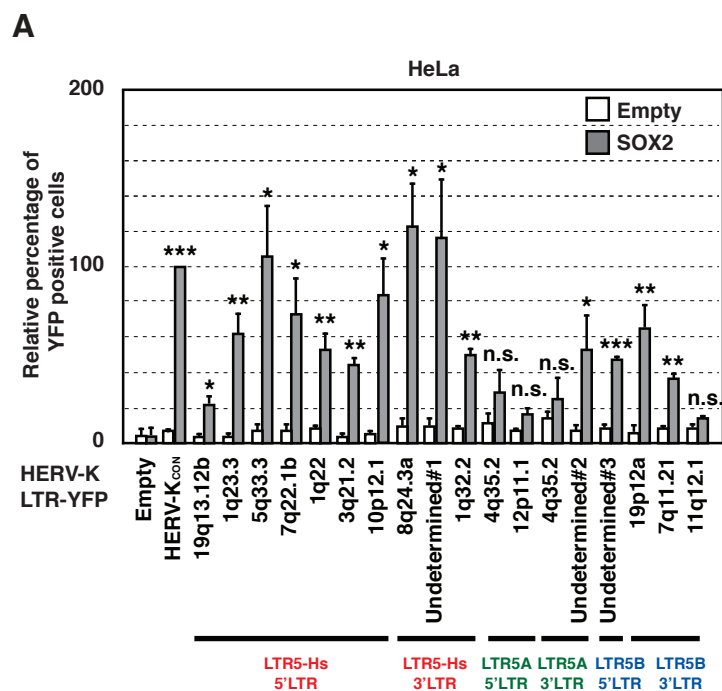48 saga. *Nature reviews Microbiology* **10: 395-406**

49

Subramanian RP, Wildschutte JH, Russo C, Coffin JM (2011) Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology* 8: 90

Tai AK, O'Reilly EJ, Alroy KA, Simon KC, Munger KL, Huber BT, Ascherio A (2008) Human endogenous retrovirus-K18 Env as a risk factor in multiple sclerosis. *Multiple sclerosis* 14: 1175-1180

Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126: 663-676

Terry SN, Manganaro L, Cuesta-Dominguez A, Brinzevich D, Simon V, Mulder LCF (2017) Expression of HERV-K108 envelope interferes with HIV-1 production. *Virology* 509: 52-59

Teshima S, Shimosato Y, Hirohashi S, Tome Y, Hayashi I, Kanazawa H, Kakizoe T (1988) Four new human germ cell tumor cell lines. *Laboratory investigation; a journal of technical methods and pathology* 59: 328-336

Toufaily C, Landry S, Leib-Mosch C, Rassart E, Barbeau B (2011) Activation of LTRs from different human endogenous retrovirus (HERV) families by the HTLV-1 tax protein and T-cell activators. *Viruses* 3: 2146-2159

Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J (2001) Insertional polymorphisms of full-length endogenous retroviruses in humans. *Current biology : CB* 11: 1531-1535

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J,

1 Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril
2 JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B,
3 Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S,
4 Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L,
5 Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A,
6 Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser
7 K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S,
8 Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A,
9 Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M,
10 Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J,
11 Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M,
12 Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001) The sequence of the human genome.
13 *Science* 291: 1304-1351
14
15 Wallace AD, Wendt GA, Barcellos LF, de Smith AJ, Walsh KM, Metayer C, Costello
16 JF, Wiemels JL, Francis SS (2018) To ERV Is Human: A Phenotype-Wide Scan
17 Linking Polymorphic Human Endogenous Retrovirus-K Insertions to Complex
18 Phenotypes. *Frontiers in genetics* 9: 298
19
20 Wang-Johanning F, Liu J, Rycaj K, Huang M, Tsai K, Rosen DG, Chen DT, Lu DW,
21 Barnhart KF, Johanning GL (2007) Expression of multiple human endogenous
22 retrovirus surface envelope proteins in ovarian cancer. *International journal of cancer*
23 120: 81-90
24
25 Weina K, Utikal J (2014) SOX2 and cancer: current research and its implications in the
26 clinic. *Clinical and translational medicine* 3: 19
27
28 Wildschutte JH, Williams ZH, Montesion M, Subramanian RP, Kidd JM, Coffin JM
29 (2016) Discovery of unfixed endogenous retrovirus insertions in diverse human
30 populations. *Proceedings of the National Academy of Sciences of the United States of*
31 *America* 113: E2326-2334
32
33 Xie Y, Rosser JM, Thompson TL, Boeke JD, An W (2011) Characterization of L1
34 retrotransposition with high-throughput dual-luciferase assays. *Nucleic acids research*
35 39: e16
36
37 Young GR, Stoye JP, Kassiotis G (2013) Are human endogenous retroviruses
38 pathogenic? An approach to testing the hypothesis. *BioEssays : news and reviews in*
39 *molecular, cellular and developmental biology* 35: 794-803
40
41 Yu X, Zhu X, Pi W, Ling J, Ko L, Takeda Y, Tuan D (2005) The long terminal repeat
42 (LTR) of ERV-9 human endogenous retrovirus binds to NF-Y in the assembly of an
43 active LTR enhancer complex NF-Y/MZF1/GATA-2. *The Journal of biological*
44 *chemistry* 280: 35184-35194
45
46
47

**A**

○ Unmethylated
● Methylated

**B** LTR-Gag

NCCIT

(51.2% methylated)

**C** LTR-Gag

HeLa

(60.6% methylated)

**D**

HeLa

□ Empty
■ SOX2

HERV-K Gag / GAPDH

5-Aza    0 uM    1 uM    10 uM

**A** HeLa

**B** NCCIT

**C**

**A**

CMV | R | U5 | GagProPol | SD | SA | SV40 | U3 | R | U5
5`UTR — Luc — Nano — 3`LTR

Transcription, splicing,
reverse transcription and integration

U3 | R | U5 | GagProPol | NanoLuc | U3 | R | U5
5`LTR — 3`LTR

**B**

HeLa

Nanoluciferase activity (RLU)

— pHERV-K GagProPol/inNanoluc
— pHERV-K delta/GagProPol/inNanoluc

n.s.   n.s.   n.s.   *

days

**C**

HeLa

Nanoluciferase activity (RLU)

**           ***          ***          **          ***

| GagProPol | – | + | – | + | – | + | – | + | – | + |
| Mutants/inNanoluc | delta ProPol | | delta Gag ProPol | | Pro D203N | | RT SIAA | | IN DR1,DR2 | |

**A**

CMV
R | U5 | GagProPol | cibin | Blasti | U3 | R | U5
5`UTR | SD | SA | 3`LTR
3F

↓ Transcription, splicing, reverse transcription and integration

4F
5F | 4R
5R | 1F | 2F | 1R
U3 | R | U5 | GagProPol | Blasticidin | U3 | R | U5 | 2R
TSD | 5`LTR | 3`LTR | TSD

**B**



HeLa-p0
HeLa-inBLC-p1
HeLa-inBLC-p2
HeLa-inBLC-p17
HeLa-inBLC-p31
Reference

**C** HeLa-inBLC

Clone number/Total (%)

passage: p0 p1 p2 p17 p31 (repeated)
Locus: **12p13.31** 16p12.3 7p22.1 **11q13.4** 11q22.1 **19p12**

**D** HeLa-inBLC

Clone number/Total (%)

samples: 1st 2nd
passage: p0 p1 p2 p17 p31
Locus: **8q24.22** 16p11.2 Xp11.23

**E** HeLa-inBLC

Clone number/Total (%)

samples: 1st 2nd
passage: p0 p1 p2 p17 p31
Locus: 8p11.1 18p11.21 Xq24

**F**

Number of HERV-K locations

■ Exon
■ Intron
□ Inter-gene

p0 | p1 p2 p17 p31
HeLa | HeLa-inBLC

**G**

chr: 8q24.22 | 16p11.2
Primer: 2F/2R 3F/2R 5F/5R | 2F/2R 3F/2R 5F/5R
1F/1R 4F/4R | 1F/1R 4F/4R

**H**

| | | 5`host genome | HERV-K inBLC | | 3`host genome |
| | | | TSD 5`-U3 | 3`-U5 TSD | |
| 8q24.22 | HeLa | CCCTACAC | CCAAG | | ATTTTGGAGACT |
| | HeLa-inBLC-p1 | CCCTACAC | CCAAG TGTGG......CTACA | CCAAG | ATTTTGGAGACT |
| 16p11.2 | HeLa | ACATCAAT | GCCTG | | GGAAACCTCTCC |
| | HeLa-inBLC-p1 | ACATCAAT | GCCTG TGTGG......CTACA | GCCTG | GGAAACCTCTCC |

| HeLa-inBLC | Locus | Coordinate GRCh38/hg38 | Flanking region | Function |
|---|---|---|---|---|
| Universal[a]<br>HeLa | 7p22.1 | chr7:4,599,432 | ERVK int[d] | |
| | 11q13.4 | chr11:71,800,891 | FAM86C1 exon4,5 | |
| | 11q22.1 | chr11:101,696,031 | ERVK int[d] | |
| | 12p13.31 | chr12:8,231,534 | | |
| | 16p12.3 | chr16:20,180,721 | AluSx[d] | |
| | 19p12[c] | chr19:21,658,739 | | |
| Specific[b]<br>HeLa-inBLC<br>Rapid-growth | 8q24.22 | chr8:133,275,187 | NDRG1 intron 3 | Stress-responsive protein in hormone responses, cell growth, and differentiation. Acts as a tumor suppressor in many cell types. |
| | 16p11.2 | chr16:30,785,310 | MIRc[d]; ZNF629 intron 1 | Maybe involved in transcriptional regulation. |
| | Xp11.23 | chrX:120,549,077 | AluSz6[d] | |
| Specific[b]<br>HeLa-inBLC<br>Normal-growth | 8p11.1 | chr8:43,312,043 | L1PA2[d]; POTEA intron 6 | |
| | 19p11.21 | chr18:15,163,428 | ERVL-MaLRb;<br>AP005901.2 intron 1 | |
| | Xq24 | | CUL4B intron 2 | Core component of multiple culin-RING-based E3 ubiquitin-protein ligase complexes |
| Specific[b]<br>HeLa-inBLC<br>Slow-growth | 1p36.21 | chr1:15,112,068 | KAZN intron 13,<br>TMEM51-AS1 exon 6 | Desmosome assembly, cell adhesion, cytoskeletal organization, and epidermal differntiation. |
| | 2q13 | chr2:111,999,744 | LTR7B[d];MERTK intron 9 | MER/AXL/TYRO3 receptor kinase family, cell survival, migration, differentiation, and phagocytosis of apoptotic cells. |
| | 3p26.1 | chr3:4,366,476 | L1P1[d];SUMF1 intron 8 | Enzyme that catalyzes the hydrolysis of sulfate esters by oxidizing a cysteine residue in the substrate sulfatase to an active site 3-oxoalanine residue. |
| | 6p22.3 | chr6:23,230,703 | AluSc[d] | |
| | 7p21.2 | chr7:16,217,694 | L1PA2[d];ISPD intron 8 | 2-C-metyl-D-erythritol 4-phosphate cytidylyltransferase-like protein |
| | 10p15.1 | chr10:6,263,616 | LTR50[d] | |
| | 10q11.21 | chr10:42,101,508 | | |
| | 12p13.33 | chr12:768,683 | WNK1 intron 1 | Serin/threonine protein kinases. Maybe a key regulator of blood pressure by controlling the transport of sodium and chloride ions. |
| | 12q23.1 | chr12:100,854,613 | L1ME4a[d];ANO4 intron 3<br>AC138360.1 intron 1 | |
| | 13q22.2 | chr13:76,617,585 | L1HS[d];<br>AL136441.1 intron 3 | |
| | 16p13.3 | chr16:1,748,182 | MAPK8IP3 intron 6 | JNK signaling by aggregating specific components of the MAPK cascade. |
| | 16q11.2 | chr16:46,398,910 | | |
| | 17p13.1 | chr17:8,072,100 | | |
| | 20q13.2 | chr20:51,642,183 | AluSx1[d];ATP9A intron 9 | ATPase phospholipid transporting 9A |
| | Xq13.3 | chrX:75,233,189 | L1PA3[d] | |

| iPS | Locus | Coordinate GRCh38/hg38 | Flanking region | Function |
|---|---|---|---|---|
| Universal[a] | 3q13.2 | chr3:113,032,475 | HERV-K int[d]; AC078785.1 intron 1 | |
| fibroblast | 4q25 | chr4:111,308,304 | AluY[d] | |
| iPS31 | 6q14.1 | chr6:77,725,406 | HERV-K int[d]; MEI4 intron 2 | Required for meiotic DNA double-strand break formation. |
| iPS41 | 7p14.3 | chr7:334,071,494 | L1PA3[d]; BBS9 intron 19 | Required for ciliogenesisi but is dispensable for centriolar satellite function. (bardet-biedl syndrome) |
| | 12q12[c] | chr12:43,919,854 | L1MB1[d]; TMEM117 intron 2 | |
| | 13q31.3[c] | chr13:90,090,934 | ; LINC00559 intron 3 | Long non-coding RNA. |
| Specific[b] | 8q13.1 | chr8:66,953,856 | L1PA2[d]; TCF24 intron 3 | Putative transcription factor. |
| fibroblast | 9q22.33 | chr9:99,500,005 | L1PA2[d] | |
| | 15q15.1 | chr15:41,053,309 | AluY[d]; INO80 intron 19 | DNA helicase, transcriptional regulation, DNA replication, DNA repair |
| | 21q11.2 | chr21:13,820,849 | MIR/LTR5Hs[d] | |
| Specific[b] | 1p13.2 | chr1:113,755,874 | AluY[d];PHTF1 intron 2 | Transcription regulation |
| iPS31 | 2q22.3 | chr2:147,700,152 | | |
| | 5p13.3 | chr5:30,822,734 | L1PA2[d] | |
| | 15q11.2 | chr15:23,715,336 | L1PA3[d] | |
| | 20q11.23 | chr20:38,348,370 | AluSc5[d]; LBP intron 1 | Binds to the lipid A moiety of bacterial lipopolysaccharides (LPS), a glycolipid present in the outer membrane of all Gram-negative bacteria. (toxic pneumonitis, mastitis, infective endocarditis, alcoholic hepatitis, mesenteric lymphadenitis, peritonitis, appendictis, bacteremia, endocarditis, acute respiratory distress syndrome) |
| | Xq25 | chrX:126,189,98 | L1PA2[d] | |
| Specific[b] iPS31 iPS41 | 1p13.2 | chr1:111,609,710 | AluSx[d]; RAP1A intron 1 | Induces morphological reversion of a cell line transformed by a Ras oncogene. (Kabuki syndrome 1, leukocyte adhesion deficiency type iii, tuberous sclerosis, babesiosis, central nervous system hemangioma, cerebral cavernous malformations-1, cerebral cavernous malformations-2) |
| | 1p34.3 | chr1:38,842,086 | AluY[d]; RRAGC intron 6 | Has guanine nucleotide-binding activity |
| | 1p35.1 | chr1:32,614,846 | AluSc[d] | |
| | 3q11.2 | chr3:97,642,693 | L1PA2[d]; EPHA6 intron 14 | (Oculoauricular syndrome) |
| | 4q22.1 | chr4:91,979,478 | L1HS[d] | |
| | 5p13.3 | chr5:32,212,988 | AluSx3[d] | |
| | 9q21.13 | chr9:73,987,206 | /LTR5Hs[d] | |
| | 10q24.2 | chr10:99,256,369 | MSTD int[d] | |
| | 11q13.2 | chr11:67,868,810 | LTR5Hs int[d] | |
| | 11q13.3 | chr11:70,471,356 | ; SHANK2 exon 2 | Molecular scaffolds in the postsynaptic density of excitatory synapses. (Autism susceptibility 17, autism spectrum disorder, pervasive developmental disorder, deafness, autosomal recessive 63, secretory diarrhea, autistic disorder) |
| | 15q14 | chr15:36,348,199 | | |
| | 16p12.3 | chr16:18,135,887 | AluSz[d]; AC010601.1 intron 1 | |
| | 21p11.2 | chr21:8,586,021 | MLT1A0[d] | |
| | 21q22.3 | chr21:43,711,450 | AluY[d] | |