Human endogenous retrovirus H protects the genome of human embryonic stem cells from mutagenic retroelements activity

Inaugural-Dissertation

to obtain the academic degree

Doctor of Philosophy (Ph.D.)

submitted to the Department of Biology, Chemistry, Pharmacy

of Freie Universität Berlin

by

Aleksandra Kondrashkina, M. Sc

2022

This work was prepared from November 2016 to October 2022 under the supervision of Dr. Zsuzsanna Izsvak. All the experiments were conducted in the laboratory of Dr. Izsvak.

1st Reviewer:

Dr. Zsuzsana Izsvak

Max-Delbrück Center for Molecular Medicine Berlin, Germany

2nd Reviewer:

Prof. Dr. Katja Nowick

Institute of Biology - Zoology

Freie Universität Berlin, Germany

date of defense: 07.03.2023

Acknowledgments

I would like to thank my supervisor, Dr. Zsuszanna Izsvak, for always believing in me, supporting me even when my understanding of the project was not agreeing with hers, and helping me when I couldn't move forward. I highly appreciate the work of Dr. Manvendra Singh, whose observations and computational analysis started this study. Our collaborator, Prof. Laurence Hurst, had always looked at this research from a different perspective and given the evolutionary outlook to it. I appreciate the input by Prof. Katja Nowick and her research group, which not only helped me scientifically but also allowed me to shortly be back to the university atmosphere, which I missed.

I wouldn't manage to follow this path without the help of my amazing colleagues, who became my friends: Dr. Angelica Garcia-Perez, Bertrand Teneng, Chigozie Samuel, Felix Lundberg, Dr. Katarina Stevanovic, Dr. Rabia Anwar, Dr. Yuliang Qu, Dr. Zhimin Zhou. Even though we worked in different fields, you were always helpful and supportive, when the most needed. And the special thanks to my lab (and all-time) soulmates Kathrin Radscheit and Karam Ibrahim, who were always there to cheer me up, help me with cells on a weekend and tell their opinion about my thesis.

I'm grateful for the MDC graduation program for the opportunity to do my PhD and the chance to meet amazing scientists. Cristina, Matthias, Carlos, Oscar, and Stefan, with whom we endured the Ph.D. journey together, and Masha, Zhenya, Oleg, and Olya, thank you for the support day and night, all these years. My university friends, the girl power, Dr. Shiriaeva, and soon-to-become Dr. Cherepkova, Dr. Lomert, Dr. Gnedina, and Lena you keep inspiring with the endless desire to discover new. Friends, who stayed with me for more than 10 years and saw my emerging interest in science from the school desk, Alyona, Sasha, Dasha and Egor, I'm grateful for your support. I also want to thank all my former teachers and supervisors, who taught me critical thinking and patience.

And last but not least, thank you to my family, who believed in me and never questioned my goals. To my amazing husband, Jorge Martins, for his professional and editorial input and just for being the best. Tu és a minha melhor metade. And my mother, father, and grandma. Спасибо за то, что всегда верили в меня и поддерживали во всех начинаниях!

Declaration of Independence

Herewith I certify that I have prepared and written my thesis independently and that I have not used any sources and aids other than those indicated by me. I also declare that I have not submitted the dissertation in this or any other form to any other institution as a dissertation.

Outline

Summary

Human pluripotent stem cells (hPSCs), which include human embryonic stem cells (hESCs) and human induced pluripotent stem cells (hiPSCs), infinitely self-renew, and can differentiate into any cell type on the human body [1–3]. hESCs are derived from early human embryos and became widely used to study the molecular pathways specific to human embryogenesis [1, 4–8]. Considering the ethical challenge in using embryo-derived cells and the possible immune rejection, hiPSCs are currently more common for regenerative therapies [3, 9–11]. hiPSCs are reprogrammed from a somatic cell line of a patient, genetically modified, and then differentiated to the desired lineage to transplant them back to the patient. hiPSCs are the future of personalized medicine, but not every hiPSC line can differentiate to every given cell type, as a result of cell heterogeneity. To reduce this heterogeneity, a naïve cell state might be a solution [3].

Whereas cultured hPSCs reside in a primed state, the cells of pre-implantation embryos resemble naïve pluripotency [12–16]. By adjusting culture conditions, it is possible to support hPSCs in a naïve state, similar in gene expression signature to early embryos [4, 5, 17–19]. The similarity is reflected as well in transcripts of some of the L1, Alu, and SVA retroelements (REs) [5]. These REs are phylogenetically young and still active human transposons, which might be detrimental for the integrity of the genome [20–27]. Our research group had previously derived the different types of naïve cells, resembling the later stages of pre-implantation development and highly expressing human endogenous retrovirus H (HERVH) [6]. HERVH is a phylogenetically older endogenous retrovirus, which was transposing following New- and Old-World monkey separation [28–30]. Now, HERVH can't mobilize, but its transcripts were shown to support pluripotency in later stages of human embryogenesis, reprogramming, and in cultured primed hPSCs [6, 7, 31, 32].

Here I show that HERVH controls the transposition of young REs. In HERVH-depleted hESCs, L1 transposition increases, which is measured by two transposition assays. The active L1 elements drive the transposition of non-autonomous REs, resulting in the accumulation of *de novo* Alus and SVAs integrations, shown by whole-genome sequencing of cells undergoing stable HERVH knock-down.

A subgroup of HERVH has the potential to control L1 transposition. These HERVHlin loci contain *lin* motif, two tandem LIN28A binding sites [33]. HERVHlin is supposedly evolutionary younger than the other HERVH. There are around 100 of HERVHlin sequences in

the human, chimp, and gorilla genomes, while less exist in orangutans, and none in other primates. Based on the analysis of the previously published CLIP-seq data [33] and performed RIP-qPCRs, the *lin* motif allows LIN28A to bind HERVHlin more efficiently than other HERVH transcripts. LIN28A is known to inhibit the maturation of let-7 microRNA [34–37], which in turn controls the transposition of L1 [38]. HERVHlin sponging LIN28A to allow let-7-mediated inhibition of L1 might be the molecular mechanism of HERVH-controlled transposition of young REs. The supporting experiment shows that a let-7 independent L1-ORFeus reporter does not change the transposition activity in HERVH-depleted cells.

HERVHlin embedded itself in a previously conservative pluripotency-specific LIN28A-let-7 pathway to protect the genome of hESCs from the mutagenic activity of REs. This is an example of a new evolutionary event where the selfish transposon HERVH evolved to compete with other transposable elements, which could be harmful to the host.

Zusammenfassung

Humane pluripotente Stammzellen (hPSZ), zu denen humane embryonale Stammzellen (hESZ) und humane induzierte pluripotente Stammzellen (hiPSZ) gehören, können sich unbegrenzt selbst erneuern und sich in jeden Zelltyp des menschlichen Körpers differenzieren [1–3]. HESZ werden aus frühen menschlichen Embryonen gewonnen und wurden in großem Umfang zur Untersuchung der molekularen Pfade verwendet, die für die menschliche Embryogenese spezifisch sind [1, 4–8]. In Anbetracht der ethischen Herausforderung bei der Verwendung von aus Embryonen gewonnenen Zellen und der möglichen Abstoßung durch das Immunsystem werden hiPSZ derzeit häufiger für regenerative Therapien verwendet [3, 9–11]. HiPSZ werden aus einer somatischen Zelllinie eines Patienten reprogrammiert, genetisch modifiziert und dann in die gewünschte Zelllinie differenziert, um sie dem Patienten zurückzupflanzen. HiPSZ sind die Zukunft der personalisierten Medizin. Aber nicht jede hiPSZ-Linie kann sich zu jedem bestimmten Zelltyp differenzieren, was auf die Heterogenität der Zellen zurückzuführen ist. Um diese Heterogenität zu verringern, könnte ein naiver Zellzustand eine Lösung sein [3].

Während sich kultivierte hPSZ in einem geprimten Zustand befinden, ähneln die Zellen von Präimplantationsembryonen der naiven Pluripotenz [12–16]. Die angepassten Kulturbedingungen können hPSZ in einem naiven Zustand halten, der in der Genexpressionssignatur den frühen Embryonen ähnelt [4, 5, 17–19]. Diese Ähnlichkeit zeigt sich auch in den Transkripten einiger L1-, Alu- und SVA-Retroelemente (RE) [5]. Diese RE sind phylogenetisch jung und im Menschen noch aktiv, was sich nachteilig auf die Integrität des Genoms auswirken könnte [20–27]. Unsere Forschungsgruppe hat zuvor verschiedene Arten von naiven Zellen gewonnen, die den späteren Stadien der Präimplantationsentwicklung ähneln und Humanes Endogenes Retrovirus H (HERVH) in hohem Maße exprimieren [6]. HERVH ist ein phylogenetisch älteres endogenes Retrovirus, das nach der Trennung von Neu- und Altweltaffen transponiert wurde [28–30]. Jetzt kann HERVH nicht mehr mobilisiert werden, aber es wurde gezeigt, dass seine Transkripte die Pluripotenz in späteren Stadien der menschlichen Embryogenese, der Reprogrammierung und in kultivierten geprimten hPSZ unterstützen [6, 7, 31, 32].

In dieser Studie zeige ich, dass HERVH die Transposition junger RE kontrolliert. In HERVH-depletierten hESZ nimmt die L1-Transposition zu, was mit zwei unterschiedlichen Transpositionstests gemessen wird. Die aktiven L1-Elemente treiben die Transposition von

nicht-autonomen RE an, was zu einer Anhäufung von de novo Alu- und SVA-Integrationen führt, wie die Sequenzierung des gesamten Genoms von Zellen zeigt, die einem stabilen HERVH-Knockdown unterzogen wurden.

Insbesondere eine Untergruppe von HERVH hat das Potenzial, die L1-Transposition zu kontrollieren. Diese HERVHlin-Loci enthalten das *Lin*-Motiv und zwei Tandem-LIN28A-Bindungsstellen [33]. HERVHlin ist vermutlich evolutionär jünger als andere HERVHs. Es gibt etwa 100 HERVHlin-Sequenzen im Genom von Menschen, Schimpansen und Gorillas, weniger in Orang-Utans und keine in anderen Primaten. Die Analyse der zuvor veröffentlichten CLIP-seq-Daten [33] und der durchgeführten RIP-qPCRs ergab, dass das *Lin*-Motiv es LIN28A ermöglicht, HERVHlin effizienter zu binden als andere HERVH-Transkripte. Es ist bekannt, dass LIN28A die Reifung von let-7 microRNA hemmt [34–37], die wiederum die Transposition von L1 kontrolliert [38]. HERVHlin, das LIN28A wie ein Schwamm bindet, um die let-7-vermittelte Hemmung von L1 zu ermöglichen, könnte der molekulare Mechanismus zur Kontrolle der HERVH-Transposition von jungen REs sein. Das unterstützende Experiment zeigt, dass der von let-7 unabhängige L1-ORFeus-Reporter die Transpositionsaktivität in HERVH-depletierten Zellen nicht verändert.

HERVHlin hat sich in einen zuvor konservativen Pluripotenz-spezifischen LIN28A-let-7-Weg eingebettet, um das Genom von hESZ vor der mutagenen Aktivität von REs zu schützen. Dies ist ein Beispiel für ein neues evolutionäres Ereignis, bei dem sich das egoistische Transposon HERVH entwickelt hat, um mit anderen Transposons zu konkurrieren, die für den Wirt schädlich sein könnten.

1. Introduction

1.1. Human embryo and pluripotency

1.1.1. Human pre-implantation embryo development

1.1.1.1. Stages of development

The first event in human embryogenesis is fertilization, a fusion of a spermatozoid and an oocyte, which results in the formation of a zygote. A zygote is a unique totipotent cell, it will develop to all the extra- and intra-embryonic tissues of a future human body. The large zygote undergoes several rounds of divisions. The daughter cells do not grow in size, forming smaller cells – blastomeres [39]. Up to 8-cell, each stage during this process is named according to the number of cells, composing the embryo. At the 8-cell stage, round blastomeres become flattered and develop inside-outside polarity. Further, in the morula stage, the embryo consists of 16-32 cells. Cells of morula undergo compaction – they become convex to the outside and concave to the inside [40]. On the fifth day after zygote formation, a large cavity is formed inside the embryo, which is called the blastocyst cavity or blastocoel [39] (Figure 1, bottom panel).

Further, the embryo separates into trophectoderm, which lays on the periphery and inner cell mass (ICM), consisting of centrally placed blastomeres. This stage is called a blastocyst [39]. Trophectoderm cells will further differentiate into trophoblast and contribute to the placenta. The ICM then develops into an epiblast and primitive endoderm. Primitive endoderm will become parietal and visceral endoderm, two extra-embryonic membranes crucial for transporting nutrients to the developing embryo [41]. All the tissues of the future fetus will be formed from the pluripotent cells of the epiblast. During the epiblast-primitive endoderm specification, the embryo implants into the endometrium of the mother's body. The process from zygote formation to embryo implantation lasts approximately seven days [42] (Figure 1, bottom panel).

1.1.1.2. Comparison with the mouse pre-implantation development and embryonic genome activation

Compared with the human embryo, a mouse embryo progresses through development faster, with all the stages taking around five days before implantation [43]. Morula compacts earlier, approximately at the 16-cell stage [44]. The switch between transcription from maternally inherited RNAs to the expression from embryonic genes,

termed embryonic genome activation (EGA) happens at the 2-cell stage in mice [45] and around the 4- to 8-cell stage for the major transcription wave in human [46–48] (Figure 1, top panel). Recently, low-magnitude transcription was detected as early as at the one-cell stage embryo [49].

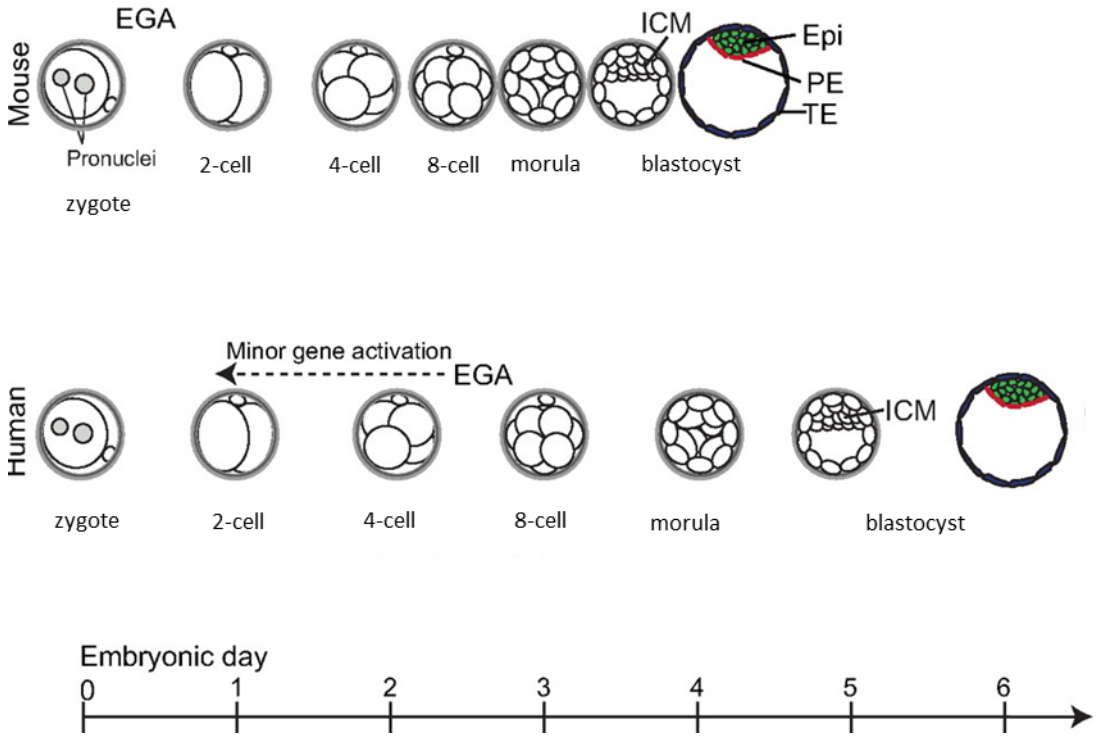


Figure 1. Human and mouse pre-implantation embryo development (figure adapted from [42]). Human and mouse embryos pass through similar pre-implantation stages: zygote, 2-cell, 4-cell, 8-cell stages, morula, and blastocyst. The blastocyst contains trophectoderm (TE) and inner cell mass (ICM), which then develops into epiblast (Epi) and primitive endoderm (PE). The mouse embryo develops faster, taking around 5 days before implantation. The human embryo implants only after 6 days from zygote formation. Embryonic genome activation (EGA) in mouse embryos happens earlier, between the zygote and 2-cell stage. The human genome activates around an 8-cell embryo, with reported minor EGA waves before.

1.1.2. Human pluripotent cells

1.1.2.1. Human embryonic stem cells derivation

Before the actual derivation of human pluripotent stem cells (hPSCs), which would be able to differentiate into any lineage of a human body, embryonal carcinoma cells were

studied as the first cell line with pluripotent features. This cell type was isolated from a germ cell tumor called teratoma [50], where a single embryonal carcinoma cell could form multidifferentiated tumors, and some carcinoma cell types could contribute to chimeras. However, most cells were not stable in cell culture due to karyotypic abnormalities [51].

First pluripotent cells were derived through an attempt to culture the whole pre-implantation human blastocyst [1]. The derivation had low efficiency and human embryonic stem cells (hESC) were challenging to maintain, due to high sensitivity to single cell dissociation in comparison to mouse cells. The possibility to inhibit apoptosis with Rho kinase inhibitor and the importance of the activin and beta fibroblasts growth factor (bFGF) signaling pathways for hESC self-renewal facilitated more reliable hESC culturing [52]. These chemicals were included in good manufacturing practice (GMP) conditions, the culturing protocol, that was developed to maintain hPSCs for further use in regenerative therapy.

Pluripotent cells in general and specifically hESC have an unlimited potential to self-renew in culture and could theoretically differentiate into any cell type of the human body. The application of hPSCs in regenerative therapy usually includes gene engineering when a non-functional gene is corrected. Due to the carcinogenic potential of hESC, they must then be differentiated to the desired cell type and only then being transplanted to a patient. But several challenges to the usage of hESCs in therapy exist.

1.1.2.2. Human embryonic stem cells issues

For successful therapy with hESC, there are two main issues to consider. First is the immune rejection of hESC by a recipient body. The immune response might be triggered when the allogeneic cells are used. Allogenic cells are cells that originated from a different organism to which their derivates will be transplanted. The major histocompatibility complex (MHC) are cell-surface receptors, polymorphic in human populations. Cells with a non-matching MHC type II could cause a massive T-cell-mediated immune response and a rejection of the transplanted allogeneic cells [53]. Even though hESCs express only MHC type I, which does not directly activate T-cells, further differentiation of hESC in MHC type II expressing linages or indirect activation of T-cells could cause transplant rejection [54, 55].

The second issue is an ethical dilemma about whether it is morally acceptable to research novel therapies at the expense of destroying a human embryo [9]. To restrict the usage of embryos or models of human embryonic development, the ‘14-day rule’ is used in

science policy, limiting research to a maximum period of 14 days after their creation [10]. On the 15th day of embryo development, the primitive streak is formed, and gastrulation begins by the differentiation of three layers. The embryo is then defined and can no longer become a twin [56]. Many countries have banned the derivation of new hESCs lines from human embryos. For research purposes, it is allowed to use only the cell lines, derived before the year 2001 [11]. Considering the aforementioned issues, an appreciable alternative to hESCs is induced pluripotent cells (iPSCs), which possess similar therapeutic potential as hESC, and could be derived from regular somatic cells through reprogramming.

1.1.2.3. Induced pluripotent stem cells discovery

The background for the discovery of cell reprogramming was a nuclear transplantation experiment, performed by Briggs and King, where they found that nuclei from a blastocyst of a *Rana pipiens* frog after the transfer to an enucleated oocyte could generate tadpoles [57]. But more specialized cells from the gastrulation developmental stage lost their potential to form tadpoles [58]. John Gurdon later used *Xenopus laevis* tadpoles to perform nuclear transfer from cultured intestine cells to form mature fertile animals [59]. The cloning of Dolly the sheep showed that even when fully specialized cells are used for the nuclear transfer, it is possible to create an entire organism [60].

The Noble Prize-awarded discovery of fibroblasts reprogramming to iPSC was made by Shinya Yamanaka. Firstly, pluripotent stem cells were derived from mouse embryonic fibroblasts or tail-tip fibroblasts through an exclusion screen, which started from 24 gene candidates and resulted in four major transcription factors *Oct3/4*, *Sox2*, *c-Myc*, and *Klf4*, sufficient to convert somatic fibroblasts to stem cells [61]. The four transcription factors were further referred to as Yamanaka factors. Reprogramming of human dermal fibroblasts was reported a year after, utilizing the same cocktail of transcription factors [2, 62].

One of the challenges to using reprogramming as a part of regenerative therapy is the low derivation efficiency for human iPSCs, ranging between 0.1-1% from the starting somatic cell number [63]. Reprogramming methods are being upgraded, increasing up to 4.4% efficiency when using mRNA transfection and adding *LIN28* to the reprogramming factors cocktail [64]. It was shown that some miRNAs could reprogram cells at high efficiency without the Yamanaka factors. Transfections of the seed sequences of the miR302/367 allowed to generate iPSCs from fibroblasts with around 10% efficiency [65]. Nevertheless, there have

been no published reports from other research groups, reproducing the results of the miRNAs method [63].

1.1.2.4. Induced pluripotent stem cells applications

Being one of the main discoveries in the light of personalized medicine, reprogramming is now widely used in clinical trials as a part of substituting regenerative therapy. The ability of hPSCs (both hESC and hiPSC) to differentiate to nearly any functional tissue is a potential tool for the replacement of damaged with healthy donor tissue. iPSC-derived tissue replacement has been uncovered in multiple studies, with some of them transforming into clinical trials [66].

The very first clinical trial started in 2014 and used hESC-derived retinal pigmented epithelial cells (OpRegen®) for treatment of age-related macular degeneration, sponsored by Hoffmann-La Roche (registration number of the study: NCT02286089). So far, the report of the phase 1/2A clinical trial has been published, showing no cases of transplant rejection or inflammation, and preliminary evidence of improvement in visual function [67]. The larger, controlled clinical study would need to uncover the optimal disease stage for intervention, surgical procedure for subretinal delivery, and target delivery location of OpRegen. This research is a great example of fundamental science discoveries transferred to personalized medicine, helping patients at the moment.

The most advanced clinical trials nowadays are active in Japan being in phase III of trials, sponsored by the Kyoto University Hospital [68]. The study used iPSC-derived dopaminergic progenitors for transplantation into the corpus striatum of Parkinson's patients. The trial is still active without results report yet (jmaCCT Clinical Trial Registry portal, registration number: JPRN-JMA-IIA00385).

Even though the progress of regenerative therapy, involving iPSC or hESC usage is undeniable, the cell dose required to treat patients on a commercial scale is not yet achievable with the current culturing practices. The aforementioned GMP conditions for stem cells increase the price of a cell line generation from 25,000 to 800,000 US dollars [69], which makes it challenging to use hPSCs-based therapy for personalized medicine for now.

One of the most promising applications of human iPSCs nowadays could be a test system. In a recently published work, Sequiera with co-authors has used iPSCs cultures, derived from a patient with an ultra-rare mutation, causing an unknown disease with

symptoms, similar to the Leigh syndrome. They tested a panel of drugs on the patient's iPSCs [70]. Earlier, the patient had participated in two clinical trials, showing no response to treatments, possibly due to an unknown mechanism of the syndrome, caused by the mutation. The efficacy and safety the three drugs were confirmed in the iPSCs test platform and after 3 years of treatment, the drugs were effective in shifting the metabolic profile of this patient toward healthy control [70]. This research brings us one step closer to the future of personalized medicine.

1.1.2.5. Induced pluripotent stem cells issues

Even though the iPSC-based regenerative therapy appears to be a very promising new tool in personalized medicine, iPSC usage does conceal some issues. One of them is tumorigenicity, caused by factors, described below [3].

Incomplete differentiation of initial pluripotent cells could result in the presence of potentially tumorigenic cells, transplanted to a patient. Additionally, differentiation protocols often contain intermediate stages with multipotent highly proliferative cells, which grow in a tumor-like fashion when injected *in vivo* [71]. Several strategies were successfully implemented for either positive cell selection via sorting for differentiation markers [72] or negative selection, based on pluripotency-specific cell surface markers [73].

The expression of reprogramming factors itself is a risk factor for cancer formation since one of the main functions of Yamanaka factors is to provide self-renewal support for cells. It has been shown that mice that received transplantations of iPSCs, generated through retroviral transfection of *Oct3/4*, *Sox2*, *Klf4*, and *c-Myc*, often developed tumors [74]. The solution would be a temporal expression of reprogramming factors with excessive validations for the absence of their integration.

Tumorigenicity might also be caused by genetic abnormalities of iPSCs, differentiated to a demanded cell type [3]. Chromosomal aberrations, copy number variation, and single nucleotide mutations could be a consequence of culturing cells for *in vitro* expansion [75]. The major abnormalities like chromosomal aberrations are usually monitored by karyotyping after which the problematic clones are not used further in research. The more complicated to determine are smaller genetic alterations, like copy number variation (CNV) and single nucleotide variation (SNV). For example, hPSCs accumulate SNVs in cancer-related genes, like the tumor suppressor gene *TP53* [76]. Before clinical usage of iPSC-derived cells, researchers

frequently perform whole genome sequencing (WGS) to detect genomic abnormalities [77, 78].

The other challenge in using pluripotent cells for cell therapy is immune rejection. Even though iPSCs created from the patient's own cells provide an opportunity to perform autologous transplantation, the immunogenicity of iPSCs has been controversial [3]. The mouse work has shown rejection of iPSCs, transplanted in the identical strain from which they were reprogrammed [79]. Newly obtained mutations in mitochondria could be a potential source of immune reaction to autologous iPSCs in both mice and humans [80]. Nevertheless, two other research did not show any sign of rejection for differentiated cells, derived from iPSC of the same strain [81, 82]. No obvious signs of rejection were observed after two years of surgical transplantation of autologous iPSC-derived retinal cells during the clinical study for macular degeneration [77].

The last challenge in using iPSC for cell therapy is heterogeneity. Cells of each line have different morphology, growth speed, degree of gene expression, and efficiency to differentiate into various cell types [3]. For example, 17 hESCs lines showed more than hundred-fold differences in the expression levels of lineage-specific markers. Therefore, some lines differentiated better to pancreatic linage, another were prone to differentiate to cardiomyocytes [83]. Later the variation between iPSC cell lines was also shown during cardiac differentiation protocols [84]. The genetic background seems to be the biggest determinant of heterogeneity [85, 86].

To reduce heterogeneity, some researchers have attempted to convert primed pluripotent cells into a naïve state.

1.1.3. States of pluripotency

1.1.3.1. Naive pluripotency

Naïve state was first mentioned by Nichols and Smith, while describing the two phases of pluripotency in mice [12]. Ground, or, naïve state, attributed to early epiblast, whereas primed state corresponded to a later developmental structure – embryonic disc, which consists of epithelial-like but still pluripotent stem cells [12]. Mouse ESCs are maintained in the naïve state by adding leukemia inhibitory factor (Lif) to the culturing media, with a distinct dome-shaped morphology, expressing specific naïve markers *Rex1*, *NrOb1*, and *Fgf4*, in addition to the known pluripotency-specific *Oct4*, *Nanog*, *Sox2*, *Klf2*, *Klf4*. Both X

chromosomes are active in these cells and supplemented bFgf causes differentiation. The primed-state mouse ESCs, also called EpiSCs, in contrast, self-renew in response to bFGF and differentiate in LIF-containing media (Figure 2). One X chromosome of these cells is inactive [12].

It's believed that cultured hESCs are derived originally from a primed stage of a human embryo, due to their epithelial-like morphology, low survival rate during single-cell maintenance, and expression of primed pluripotency markers. Therefore, the hESCs in culture correspond roughly to mouse EpiSCs [13–16]. In recent years, different strategies to convert primed hPSC to a mouse-like naïve state have been established (Figure 2) [4, 17, 19, 87]. Each of the derived naïve human cell lines varies between one another in transcription signatures and culture conditions, due to the use of different signaling pathways inhibitors, but shares at least some characteristics with the preimplantation embryo. The genome-wide distribution of heterochromatin epigenetic marks in naïve hESCs remains similar to that of its parental primed hESCs, but different from the human embryo [88].

1.1.3.2. Formative pluripotency

The other distinct type of culture conditions maintains human cells in a formative state. In the human embryo cells reside shortly in the formative state when the epiblast exits the naive pluripotent state around the time of implantation. Some formative cells differentiate to primordial germ cells (PGCs) (Figure 2) [89, 90].

Recently, the possibility to capture an ‘endogenous’ formative state in self-renewing human and mouse stem cell lines has been reported [91]. These formative stem cells (FSCs) generated by Smith’s group are cultured in low activin conditions, with inhibition of WNT and retinoic acid signaling, and can be derived from but not reverted to the naïve pluripotent stem cells, indicating a distinctive cellular state (Figure 2) [91].

Simultaneously, Yu with co-authors reported the derivation of a cell type, called XPSCs (Figure 2) [92]. This cell line maintains robust expression of naïve pluripotency markers, as well as formative markers. Cells are derived not from the post-implantation, but rather from the preimplantation epiblast. The extent to which XPSCs have irreversibly exited the naïve state has not been reported. However, these cells can directly form PGC-like cells in vitro and exhibit germline transmission following blastocyst injection [92]. The culture conditions for XPSCs and FSCs are nearly opposite to each other, which urges more research in this field.

1.1.3.3. Transposons as markers for a state of pluripotency

A few cultured conditions for naïve cell maintenance, especially 5i and 4i conditions (Figure 2), generate pluripotent cells with a distinct expressional landmark – elevated transcripts level of evolutionary young retroelements (REs) like L1s, SVAs, and Alus [5]. The same upregulated REs expression was detected in morula-to-ICM stages of the human pre-implantation embryo [5, 6, 48, 93].

In our research group, a different naïve-like cell state was described, characterized by elevated expression of only one, evolutionary older RE: HERVH (human endogenous retrovirus H) [6]. Its promoter, LTR7, was used to generate a GFP reporter construct to enrich for cells which can maintain the high expression of HERVH due to the specific epigenetic state. We hypothesize that these HERVHhigh cells resemble later stages of pre-implantation human embryonic development compared with naïve cells, but earlier than primed hESC (Figure 2). While naïve cells correspond to morula, HERVHhigh cells are similar to pre-implantation pluripotent epiblast, contrary to primed hESC, being derived from the post-implantation embryo [6] (Singh et al, unpublished). We have shown that ~4% of hESC support LTR7-GFP expression and, after enrichment for the green fluorescent signal, cells form dome-shaped colonies, similar to mESC and naïve hPSCs. The transcriptional signature of HERVHhigh cells resembles the epiblast of the human embryo closer than other naïve cell cultures [6]. We expect HERVH expression to be crucial for pluripotent cell differentiation towards PGC-like stages.

The figure 2 below summarizes the previously described human and mouse primed, naïve, and formative cell cultures, showing the main developmental stages, attributed to the cell types.

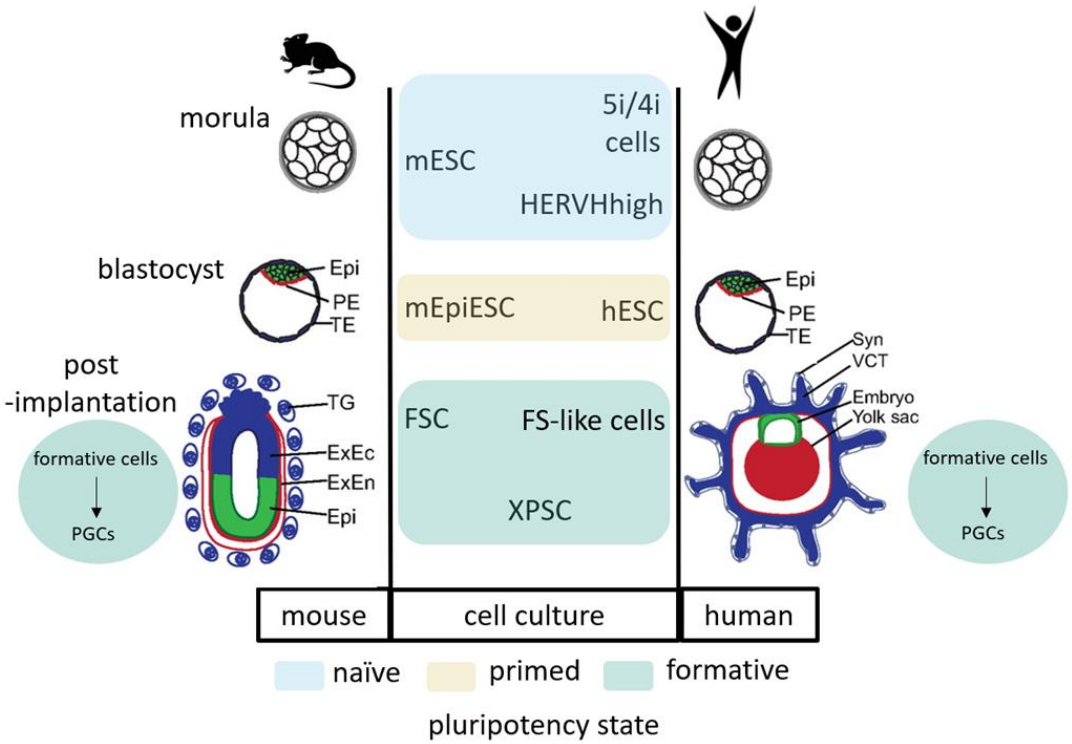


Figure 2. Human and mouse pluripotency *in vivo* and in cell culture (images adapted from [42]). Three major types of pluripotency exist *in vivo* and in cultured cells: naïve, primed, and formative. 5i, 4i, and other naïve cells lines are transcriptionally similar to the morula stage of human pre-implantation embryo due to the expression of young REs, whereas HERVHhigh cells are closer to epiblast cells (Epi), based on the expression of HERVH. The cultured mouse embryonic stem cells (mESC) are reflecting the naïve state of mouse epiblast opposite to converted epiblast-like cells (mEpiESC), resembling the post-implantation epithelial-like epiblast cells. Cultured human embryonic stem cells (hESC) are similar to the post-implantation human embryo, being in a primed pluripotency state as mEpiESC. The naturally occurring later formative cells differentiate into primordial germ cells (PGS) in both, mouse and human post-implantation embryos. FSC and XPSC cell lines resemble formative and PGC-like cell states in cell culture. Additional cell types in human and/or mouse embryo: PE – primitive endoderm, TE – trophectoderm, TG – trophoblast giant cells, ExEc – extra-embryonic ectoderm, ExEn – extra-embryonic endoderm, Syn – syncytium, VCT – villous cytotrophoblast.

1.2. Transposable elements in the human genome

1.2.1. Transposons types

In 1956, Barbara McClintock laid the foundation for transposable elements research through her initial discoveries in maize of what she termed ‘controlling elements’ [94]. The Human Genome Project uncovered that roughly 45% of our genomic sequences are derived from transposons [95]. Transposons or transposable elements (TEs) are DNA sequences, which can reproduce themselves through a DNA intermediate with the cut-and-paste mechanism or RNA intermediate and copy-and-paste mechanism. The latter are also called retroelements (REs). The only fully active retroelement in humans is LINE1 (L1) [96]. It belongs to non-LTR-containing REs, together with Alus and SINE-VNTR-Alus (SVAs) (Figure 3).

The other types of REs are LTR-containing retrotransposons or endogenous retroviruses (ERVs) (Figure 3). LTR serves as a promoter for their DNA sequences. ERVs were active in our ancestors millions of years ago, propagating in a similar way to the human immunodeficiency virus (HIV) now. But ERVs lost their transpositional activity due to mutations, especially in the sequences encoding viral envelopes, and became endogenized, staying in the genomes as DNA reminiscence of the former viruses. The presence of the strong LTR promoter nevertheless allows ERVs sequences to be expressed and play regulatory roles in hosts. HERVH, HERVK and HERVW, and others belong to the LTR retroelements. Several ERVs sequences have been co-opted to function in the human genome. For instance HERVW, whose envelope protein is also known as syncytin-1, mediates cell-cell fusions of cytotrophoblasts into syncytiotrophoblasts and is important for human placenta morphogenesis during pregnancy [97].

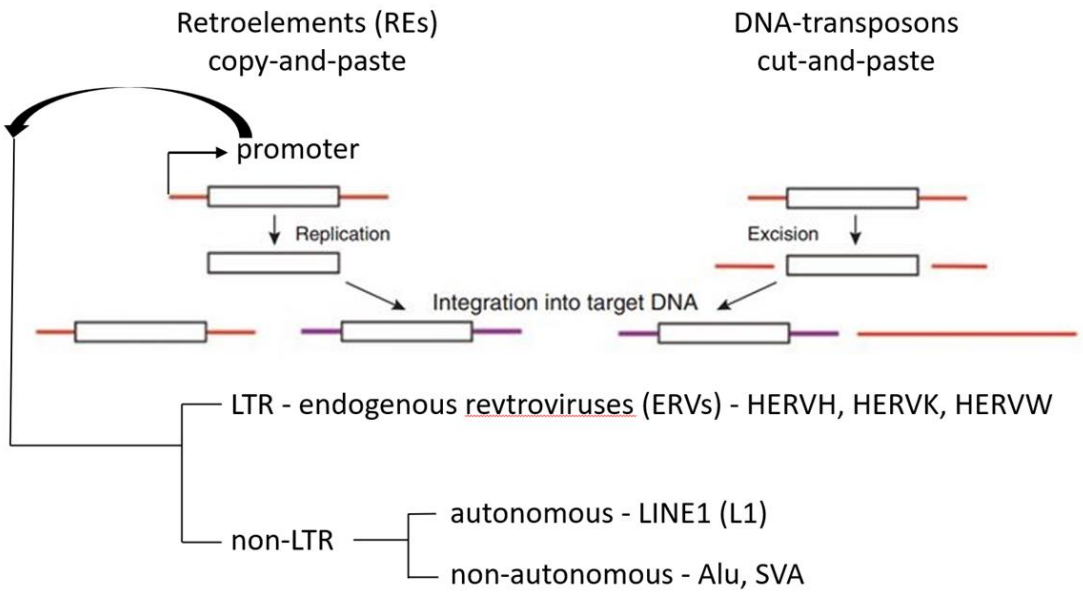


Figure 3. Transposons classes in the human genomes. Transposons are separated into two groups – DNA transposons, which jump through the cut-and-paste mechanism, and RNA transposons or retroelements (REs), which multiply via the copy-and-paste principle. Depending on the promoter type, REs are separated into two groups: LTR retrotransposons, which are also called endogenous retroviruses (ERVs), like HERVH, HERVK, and HERVW, and non-LTR REs. Non-LTR REs could be autonomous, as LINE1 (L1), which can transpose independently, and non-autonomous like Alu and SVA, requiring the active L1 machinery for their transposition in trans. L1, Alu, and SVA families are still active in the human genome.

Here I refer to current actively transposing REs like SVAs, Alus, and L1s as young retrotransposons since most of them are younger in evolutionary age than HERVH, which is not transposing in the human population anymore.

In this study, I discover the new role of HERVH in hESC. Thus, in the next part of the introduction I will concentrate on the HERVH element.

1.2.2. HERVH, and LTR retrotransposon

1.2.2.1. HERVH discovery

HERVH stands for human endogenous retrovirus H, where H is a histidine amino acid abbreviation. HERVH used a histidine tRNA as a primer to start its reverse transcription [28]. HERVH was discovered by Dr. Mager accidentally, while she was trying to clone a region of

the beta-globin gene. HERVH was previously called hsRTVL-H (Homo sapiens retrovirus-like element) [29].

In addition to the beta-globin locus, three other HERVH loci were described by Dr. Mager, showing their full-length structure, LTR7 promoters, and tRNA binding site sequences [98]. Later, HERVH was discovered on all human chromosomes, with enrichment on chromosomes 1 and 7 [99]. A frameshift mutation was detected in the four originally described loci, causing a stop codon in the conserved endonuclease domain of HERVH, likely playing a role in the loss of transposition activity.

The full-length consensus structure of HERVH was described in 2005 by Jern with colleagues [100]. The HERVH provirus was mostly present as a 5' LTR-*gag-pro-pol-env*-3' LTR sequence, where 5' and 3' LTRs were identical at the integration event. A typical LTR has 5' and 3' untranslated regions, separated by a repeat segment. The internal part of HERVH consists of *pre-gag*, *gag*, *pro*, *pol* and *env* (Figure 4). The group-specific antigen (*pre-gag* and *gag*) region includes the matrix, capsid, and nucleocapsid proteins, all important for viral capsid formation. The protease gene (*pro*) is located between the *gag* and the polymerase gene (*pol*), which contains reverse transcriptase, RNaseH, and integrase domains, serving as machinery for transposition to a new genomic position. The envelope gene (*env*) consists of the surface unit, with a signal peptide and a transmembrane unit, which are responsible for binding to the cellular receptor and fusing the viral membrane to the cellular membrane. The primer binding site for reverse transcription is situated between the 5' LTR and *gag*, while the polypurine tract for plus-strand DNA synthesis is located between *env* and the 3' LTR [100].

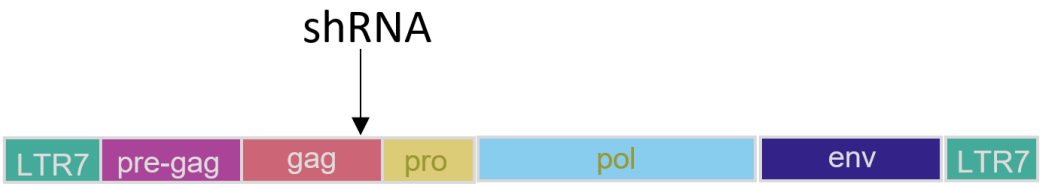


Figure 4. HERVH structure. The typical HERVH sequence contains 5' and 3' LTR7, serving as a promoter or polyA signal respectively. *Pre-gag* and *gag* encode proteins, important for capsid formation, *pro* and *pol* products are crucial during transposition to new genomic locations, and *env* is responsible for interactions with host membranes. *Env* has acquired the highest number of detrimental mutations.

The first homology studies were done by Dr. Mager, showing the *gag* region to be homologous to *gag* regions of other described ERVs like type C baboon endogenous virus, T-cell lymphotropic virus types I and II, and bovine leukemia virus. *Pol* was similar to *pol* sequences of Moloney murine leukemia virus, mammalian type C retroviruses, and murine retrovirus-related sequences [98].

Some HERVH copies in the human genome are full-length but there have been no reported transposition events so far. HERVH is inactive due to acquired mutations, which disrupted function and sometimes translation of the main proteins that were crucial for the activity of the retrovirus.

1.2.2.2. HERVH classifications

Probably around 80% of HERVH had integrated into the genome of a common ancestor of New- and Old-World monkeys around 30 million years ago (MYA) [30]. Most integrations are orthologous between humans, apes, and Rhesus macaque [101, 102].

There are several hypotheses for the origin of HERVH, based on what is known so far about the evolutionary history of the element. HERVH might have originated from an exogenous retrovirus. However, human ERVs are generally more similar to rodent viruses than human infectious retroviruses known so far [98]. Otherwise, HERVH might have evolved directly from other genomic retrotransposons, or, what seems more likely resulted from a recombination product between genomic retrotransposons and exogenous retroviruses, as it was hypothesized for other ERVs. In support of this theory, several regions of homology to different classes of retroviruses were described [102].

Based on the sequence similarity, there are two major groups of HERVH. The consensus of the most abundant HERVH subgroup lacks a part of the *pol* region. The *pol* deletion has probably occurred as a result of homology recombination, following a deletion during reverse transcription [103]. This incomplete HERVH is present in around 800 to 1000 copies in the primate genomes, spreading more after New and Old-World monkey separation [103]. Contrastingly, the other HERVH subgroup has an intact *pol* and is present in 50-100 copies in primate genomes, including less than 50 loci in the New World monkeys. This success in transposition of the incomplete HERVH element was quite surprising, but was explained by the presence of gain-of-function mutations in LTR, which were predicted to allow

higher expression [103]. The *pol*-deleted HERVH elements were hypothesized to integrate using the intact *pol* proteins from other loci. Later the presence of highly active LTR was confirmed [104]. It is interesting to note that the HERVH elements which were the most successful in evolution did lack a part of transposition machinery and integrated through trans-mobilization.

Corresponding to the LTR7 structure, HERVH could be separated into several groups as well. LTR type I (canonical LTR7) and II (LTR7b) expanded after New and Old but before the divergence of apes and Old-World monkeys, 45-30 million years ago. LTR Ia (LTR7y) is a more recent LTR, active around 10-15 million years ago in humans, chimp, and gorilla, and is now present in around 100 copies per genome. LTR7 Ia was a result of recombination between I and II types. This promoter is active between a broad range of cell types and allows expression at higher levels [101].

Recently, LTR7 types were further classified into subgroups based on the evolutionary origin and their spatial-temporal expression pattern in developing embryos [31]. Eight previously unrecognized subtypes of LTR7 were shown to be expressed in distinct patterns in the human pre-implantation embryo and they were active at different time points in primate evolution. For example, almost all the HERVH transcribed in hESC are driven by the LTR7up group, which are also the youngest LTRs, being specific for gorilla and higher primates, including human [31].

1.2.2.3. HERVH promoter is bound by transcription factors

LTR7 Ia is more active than other LTRs due to the presence of negative regulatory elements in LTR7 I and II. Additionally, LTR7 Ia has gained a functional *SP1* binding motif [105, 106]. *SP1* protein belongs to the *SP/KLF* family of transcription factors, expressed in all mammalian cell types, and generally functions to activate transcription [107]. The fact that a conserved and ubiquitous transcription factor drives the expression of HERVH shows the efficiency of HERVH's evolution.

The integrity of the LTR7 is important for the transcription of HERVH, as HERVH transcription consistently correlated with the presence of the first 114bp of the LTR7 consensus [108]. The LTR7 also contains the aforementioned *SP1* binding site followed by the *MYB* binding motif for the MYB family proteins, crucial in supporting stem cell identity [109]. The other stemness and pluripotency-specific transcription factors, binding to the LTR7

promoter are *OCT4*, *NANOG*, *KLF4*, and a naïve pluripotency-specific factor *LBP9* [6, 7, 32]. There are transcription factors, which bind only distinct types of LTR7s, for example, *SOX2/3* activates the expression of only LTR7up [31]. The specific orchestra of pluripotency master-regulators allows HERVH to be expressed during defined stages of early embryonic development and in cultured hPSCs. LTR7 promoter is transcribed together with the internal part of HERVH in hESC [6, 7, 110].

1.2.2.4. HERVH expression pattern and functions

Considering the presence of binding sites for several transcription factors on the LTR7 promoter of HERVH, expression studies were performed to decipher how the element is transcribed. The first expression studies before polymerase chain reaction (PCR) discovery were done by Dr. Mager's group, through Northern blot analysis, and showed HERVH RNAs being expressed in several carcinoma cell lines, HeLa [111] and HEK293 (Human embryonic kidney 293) [112]. Transcripts had different lengths, and some of them were spliced. The highest expression level was detected in the teratocarcinoma line, which has an embryonic origin [113]. LTRs were shown to be promoters not only for HERVH but also for other genes, interacting with cellular enhancers [114].

The first high-throughput data analysis was done by Santoni and co-authors, who discovered that 2% of RNA sequencing reads from hESC were aligning to HERVH sequences [32]. Most of the reads were detected in both LTRs, *gag,* and *pro* regions, with *pol* and *env* having a very low degree of coverage. High expression in human and primate pluripotent cells was detected five years later, showing the presence of HERVH transcripts in iPSC from human, chimpanzee, gorilla, and rhesus [115].

The importance of HERVH for human pluripotency was shown in two studies, published the same year [6, 7]. Lu and co-authors discovered that HERVH depletion causes differentiation of hESC. HERVH transcripts were bound to OCT4, coactivators p300 and Mediator complex. The role of HERVH in the support of pluripotency was hypothesized to be maintained through the establishment of epigenetic modifications [7]. The work of our research group has broadened the knowledge about the role of pluripotency transcription factors and activators, bound to LTR7 promoter (described in 1.2.2.3) [6]. In parallel, it was shown that HERVH depletion causes hPSCs differentiation. LTR7 could serve as a marker for naïve cells in a hPSC culture (described in 1.1.3.3) [6].

Later, both the HERVH family and the other ERVs were described as stage-specific markers of human pre-implantation development, using published single-cell RNA sequencing (RNA-seq) human embryo data [110]. The most specific expression was observed for LTRs, in detail describing four distinct clusters for the LTR7 family, based on their spatio-temporal expression: hESCs and epiblast (LTR7), epiblast and blastocyst (LTR7Y), morula and eight-cell (LTR7B), and early stages (LTR7 with a 38 bp deletion). The Nanog binding to the LTR7 promoter was confirmed and additionally, the H3K4me3 epigenetic activator mark was shown to be enriched [110].

HERVH RNAs have been associated with chromatin remodelers. HERVH transcripts modulated the binding of CHD7 to enhancers, generating expression patterns important for pluripotency maintenance [116]. HERVH interplay with another chromatin remodeler, ARID1A, which is a part of the SWI/SNF family has been detected in colorectal cancer. Depletion of *ARID1A* led to increased transcription at several HERVH loci. This excessively transcribed HERVH colocalizes with BRD4 (Mediator complex) foci in nuclei, supports phase separation, and co-regulates cancer-promoting target genes [117].

HERVH could control the expression of genes not only in association with remodeling factors but also as a transcription-associated domain (TAD) boundary [118]. An association with RNA polymerase II, rather than CTCF binding, defines HERVH-derived TAD boundaries probably via the positioning of cohesin complexes. The effect spreads only upstream of a HERVH locus [118].

HERVH is highly expressed in human pluripotent and cancer cell subtypes. Driven by different LTR7 promoters, transcripts mark stages of human pre-implantation development. HERVH transcripts are also involved in different types of global pluripotency regulation, like TAD formation and chromatin remodeling. Not only groups, but also individual HERVH loci have been shown to play crucial roles in hESC.

1.2.2.5. HERVH chimeric transcripts

HERVH chimeric transcripts – RNA molecules, which include a part or a full-length HERVH transcript and a gene or lncRNA sequence, were primarily described by Dr. Mager's research group. From a normal full-term placenta, a novel *PLT* gene was detected, and its polyadenylation signal was derived from an LTR of HERVH [119]. The other HERVH-derived chimera was driven exclusively by an LTR7 promoter and contained a HERVH sequence,

spliced to a downstream cellular sequence, expressed in a teratocarcinoma cell line. The chimeric transcript had a predicted ORF with two domains of homology to phospholipase A2, and the non-HERVH part was termed *PLA2L* as a novel gene [120, 121].

The first high-throughput analysis was done 19 years later [122], detecting that TEs generally comprised 42% of lincRNA with HERVH providing the highest portion of stem cell-specific lincRNAs. L1s- and Alus-derived lincRNA were less frequent than HERVH-derived [122]. The rest of the published research was concentrated on individual HERVH transcripts. Dr. Rinn's and Dr. Liu's research groups have shown that a non-coding RNA *linc-ROR*, which has a transcription start site in LTR7 and the first two exons transcribed from HERVH [122], shared miRNA seed regions with *NANOG*, *OCT4* and *SOX2*, therefore protecting these core pluripotency transcription factors from miRNA-mediated suppression [123, 124].

The other HERVH-derived chimeric RNAs, *HPAT2* and *HPAT3*, were shown to promote reprogramming, and bind to a subunit of the RNA-induced silencing complex (RISC), which might indicate the importance of the RNAs in miRNA function [8].

Our research group confirmed several HERVH-derived chimeric RNAs by reverse transcription, followed by PCR amplification, specific for hPSCs. *ESRG* lncRNA was shown to be important for the self-renewal of hESCs and increased the efficiency of reprogramming [6]. Contradictory, Dr. Yamanaka's research group found *ESRG* to be dispensable for human pluripotency, due to *ESRG*-depleted pluripotent stem cells not changing their morphology or expression of pluripotency factors, except for *NANOG*. *ESRG* was also not required for somatic cell reprogramming toward pluripotency [125]. The conflicting results urge a follow-up to investigate the crucial differences, explaining variability between the two types of research.

HERVH seems to be involved in several pluripotency-supporting mechanisms. HERVH does not only function through individual transcripts, but also binds master-regulators of transcription, like the Mediator complex. Its expression is controlled by key pluripotency factors and HERVH might be a strong scaffold to promote liquid-liquid phase separation, due to its repetitive nature. The result of the HERVH co-option was the support of human-specific pluripotency.

1.2.3. Non-LTR retroelements

1.2.3.1. Mechanism of L1 transposition

Even though the human genome contains more than 500.000 L1 sequences, most of them are inactive because of rearrangements, point mutations, and 5'-end truncations [20, 21]. An L1 locus is 6kb in length and contains full 5'- and 3'-UTRs, *ORF1*, *ORF2*, and a long polyA tail [126]. ORF1 is a 40kDa protein with RNA binding and chaperone activities [127]. ORF1 protein has to be phosphorylated for transposition [128]. ORF2 is a 150kDa protein with endonuclease [129] and reverse transcriptase [130] activities. L1 is transcribed by RNA Pol II from its promoter, located in the 5'-UTR [131]. Transcription is terminated by a polyA signal in 3'-UTR of the element [132]. After transcription, the L1 RNA is transported to the cytoplasm, where translation and L1 ribonucleoprotein assembly takes place [133]. The L1 transposition complex contains three copies of ORF1 protein, one ORF2, and L1 RNA. Integration then happens with a coupled reverse-transcription, through a mechanism called target-primed reverse-transcription (TPRT). The integration site is determined by endonuclease activity and is represented by 5'-YYYY/RR-3' consensus, where Y stands for pyrimidine, R for purine, and / for the cleavage site [134]. After cleavage, 3'-OH serves as a primer for cDNA synthesis performed by *ORF2*-encoded reverse-transcriptase protein [96]. Due to the start of reverse transcription from the 3'-end, most genomic L1 (>99%) is 5'-end truncated [135] (Figure 5).

L1 is the only autonomous retroelement in the human genome. The other common elements, Alus and SVAs, are called trans-activated REs because they cannot transpose independently and have to hijack L1 machinery for new integrations. Nevertheless, or probably therefore Alu is the most abundant transposon in the human genome [95].

1.2.3.2. Trans-activated retroelements

Although the L1s evolutionary origin is still not confirmed, the Alus source is known. Alu elements evolved from a 7SL RNA [136] with further duplication of monomers [22], which caused higher frequency localization of this non-coding RNA to ribosomes, where it started to hijack L1 protein machinery [137]. Alus are transcribed from an internal RNA polymerase III promoter [138], and for transposition Alus require only proteins, encoded by ORF2 of L1

[139] (Figure 5). These molecular traits could explain why Alus have been so successful in transposition in the human genome.

The youngest active in the human population retrotransposon is named SVA, after SINE-VNTR-Alu repeats. SVAs are 2kb in length hominid-specific non-coding composite elements [23]. SVAs consist of CCCTCT hexamer repeats, followed by an Alu-like domain, derived from two antisense fragments, a variable number of GC-rich tandem repeats (VNTR), and a SINE-R sequence, homologous to the *env*-LTR region of HERVK, with a polyA tail at the end [140]. SVAs are transcribed by Pol II and require ORF2 of L1 for retrotransposition [141]. After incorporation to the transposition complex, integrations occur much like L1s, since it shares many similarities with SVAs insertions (Figure 5). Nevertheless, SVAs loci are mostly full-length [141].

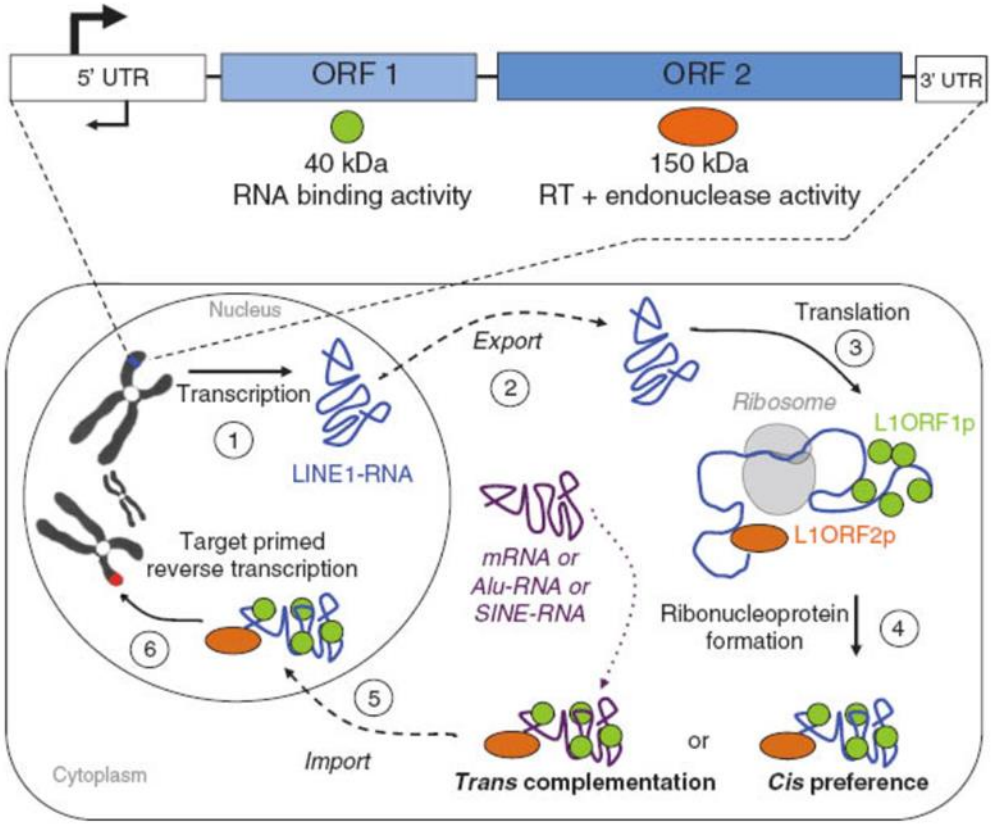


Figure 5. L1 autonomous and Alu, SVA non-autonomous transposition cycle (figure adapted from [142]). L1 is transcribed from its own endogenous 5'-UTR promoter, and the RNA is then exported from the nucleus and translated to L1 ORF1 and ORF2 proteins. L1 RNA then forms ribonuclear particles (RNPs) with ORF1 and 2 proteins in *cis* or the ORF proteins coat Alu or SVA *trans*-mobilizing their RNAs. Alu and SVA RNAs are transcribed from endogenous promoters as well (not shown). Further RNPs are

imported to the nucleus and target-primed reverse transcription happens in parallel with the integration to a new genomic location.

1.2.4. The deleterious impact of young retrotransposons

Young REs such as, L1s, are still active in the human population. A subset of around 80-100 individual L1 loci are hypothesized to be active in any given individual [24, 25]. Each set of active elements is polymorphic in the human population, e.g. any two human beings differ on average by ~285 L1 insertions [26, 27].

New insertions mostly have a damaging effect, as 124 disease-causing insertions are reported to inactivate gene function through insertional mutagenesis or aberrant splicing [96]. For example, when L1 has integrated into an exon or an intron, which is supposed to participate in splicing, a frameshift mutation might occur, which would cause an RNA nonsense-mediated decay [96]. An RE integration might also result in an alternative protein domain composition. In fukuyama muscular dystrophy, where an SVA has integrated into the fukutin gene, it causes mRNA alternative splicing to the SVA sequence and mis-localization of the modified protein from the Golgi complex to the endoplasmic reticulum [143, 144].

REs could cause DNA sequence deletion through non-allelic homologous recombination. This is more frequently observed for Alu elements, likely due to their high copy number. The deletions or inversions might appear after the pairing of two REs sequences on the same strand, usually on homologous chromosomes [145].

L1s might influence humans not only through individual insertions but also as a family of transposons. The negative selection has been shown to exist against the full-length polymorphic members of the human specific Ta1 subfamily of L1s. Because this L1 type is still active in humans, the Ta1 subfamily almost certainly continues to decrease the fitness of modern humans [146].

Not only *de novo* L1 integrations are deleterious due to target DNA sequence disruption but also by a gain-of-function effect. Often, L1 transcription will pass through the polyA of the element in favor of a polyA signal downstream. The downstream sequence is frequently retrotransposed to a new genomic location together with L1 [147]. This shuffling of protein-coding or lncRNA exons could disturb their function but on a larger scale be an evolutionary mechanism.

Activation of REs is an important component of sterile inflammation, which is a hallmark of aging. During cellular senescence, L1 becomes highly transcribed and cytoplasmic L1 cDNA causes activation of interferon type-I response [148, 149].

1.2.5. Defense mechanisms against retroelements

As mentioned above, active REs might be extremely dangerous for a host. Different organisms have evolved a combination of defense mechanisms, which successfully control REs expression and active transposition.

REs are controlled at two levels: through inhibition of their transcription and post-transcriptional modifications to prevent translation. The protective mechanisms against REs activity seem to be intertwined between each other, since an upstream mechanism can use several downstream pathways. The defense pathways could additionally function independently for transposon control (described below in 1.2.5.1).

1.2.5.1. Transcriptional control of retroelements

L1 and other RE DNA sequences are usually methylated by two enzymes, DNMT1, which prefers hemimethylated CpG dinucleotides, and DNMT3, catalyzing *de novo* deposition of 5mC [150]. *DNMT1* knock-out causes global loss of CpG methylation and activation of hominoid-specific L1 elements, while older L1s remain silent [151]. Most elements from a human-specific subfamily L1Hs are controlled by DNMTs in hESC [152]. The other control mechanisms like piwi RNA induction or KAP1 targeting are utilizing DNA methylation as a downstream pathway to control L1 activity [153].

KAP1 – Krüppel-associated box domain (KRAB)-associated protein 1 – is the master cofactor of KRAB-containing zinc finger proteins (KRAB-ZFPs). It is recruited by a KRAB-ZNF after the KRAB-ZNF recognizes a specific sequence motif of an RE. KRAB-ZNFs had evolved to recognize formerly active ERVs and nowadays, there are some proteins with binding motifs at ‘middle-age’ L1s [152]. After a KRAB-ZNF recruits KAP1, it, in turn, could bind any of the epigenetic regulators, like histone methyltransferases (ESET), nucleosome remodeling and deacetylation (NuRD) complex, heterochromatin protein 1 (HP1), human silencing hub (HUSH) complex and DNA methylation enzymes as mentioned above [154–157]. As a result, the chromatin closes at the RE locus, and the expression of an element is not possible.

The correlation between the age of an LTR and the type of methylation was discovered only for ERVs. Young LTRs tend to be CpG-rich, and they are suppressed by DNA methylation, whereas intermediate-age LTRs are associated mostly with histone modifications, particularly H3K9 methylation [158]. A similar age-methylation correlation might exist for non-LTR REs.

1.2.5.2. Post-transcriptional control of retroelements

The major type of post-transcriptional REs inhibition is a PIWI mechanism, during which small piRNAs bind complementary retrotransposons RNAs and form a piRNA-induced silencing complex (piRISC) that specifically cleaves the RNA target through RNase activity [159]. piRNAs can also target genomic REs regions, attracting DNMTs to silence transposons. The PIWI mechanism has been shown only for mouse cells so far and the expression of transposon-targeting piRNAs has been detected in human germ cells [160–162].

However, the existence of the PIWI mechanism in hPSCs is an uncertain matter. One study reported high levels of *PIWIL2* expression in human pluripotent cells in comparison with non-human primates, which did correlate with the higher L1 activity in primates vs humans [163]. PIWIL2 belongs to the Argonaut protein family, which was shown to be important for piRNA maturation [159, 164]. Nevertheless, there has been no other research proving the existence of PIWI mechanism-driven control of REs in human pluripotent cells.

The other type of post-transcriptional control for REs in human is mediated by RNA-editing enzymes like APOBEC, AID, or ADAR. ADAR1 has been shown to control L1s transposition in human cell culture, surprisingly independently of its editing activity [165]. The AID/APOBEC- family members (AID, -1, -3A, -3B, -3C, -3DE, -3F, and -3H) were shown to inhibit both L1s and Alus transposition in cell culture assays [166–177]. Eleven studies from the aforementioned publications showed deaminase-independent control of L1s or Alus transposition by APOBEC proteins, whereas only two articles detected deamination as a crucial step for retrotransposition inhibition. One research group detected deamination-independent activity of APOBEC-1 when L1s are controlled vs LTR retrotransposons being regulated by deamination [170]. Alu-derived RNAs, on the other hand, have been reported to contain deaminated nucleotide residues in different human cell lines and brain tissues, which might be explained by frequent complementation of Alu sequences, forming double-stranded RNAs known to be a substrate for AID/APOBEC enzymes [178–181].

Additional data, confirming the high importance of PIWI and APOBEC mechanisms of REs transposition control is a strong purifying selection, affecting proteins of both families [182]. Certainly, APOBEC proteins do participate in the transposition control of human REs, but the mechanisms behind need to be further investigated.

The major modes of retrotransposition inhibition by a host might crosstalk between expression and the post-transcriptional control of REs activity. First, the repression of active, evolutionary new elements could be achieved by small RNA-induced methylation, followed by KAP1-mediated DNA silencing via a corresponding newly evolved *KRAB-ZNF* [152].

1.2.5.3. Co-option as a defense mechanism

Of all REs, SVAs have the highest ratio between the number of potentially still active elements to the total number of full-length loci per genome, which is nevertheless merely 1.8% [140]. The majority of RE integrations are truncated or mutated, therefore unable to produce functional proteins, crucial for transposition. REs undergo purifying selection [183]. Evolution itself might be considered a defense mechanism against transposons, controlling their activity by directly mutating or co-opting REs to perform new functions in a host.

A well-known example of a co-option event is HERVW and HERVFRD envelop proteins (*env*), which are also called syncytin 1 and 2. Syncytins are expressed in the human placenta and are involved in trophoblast formation. Due to fusogenic features of ancestral env proteins, syncytins contribute to the formation of syncytiotrophoblast formed through cell fusion to support maternal-fetal communication [97]. The other human endogenous retrovirus, HERVE made amylase production possible in the human parotid gland after integration next to the amylase gene [184, 185]. One of the examples of a HERVH element being co-opted is a *PLA2L* gene, which expression is initiated by the 5' part of the HERVH locus [120, 121]. The functionality of *PLA2L* is not yet discovered.

Younger non-LTR REs even though still transpositionally active, might be considered co-opted. *De novo* integrations of L1, Alus, and SVAs have been described in human, mouse, and rat hippocampi and caudate nucleus [186–190]. L1 transposition was providing somatic mosaicism in neurons, although the functional significance of this event was not yet discovered. The fascinating link between behavior and L1 transposition in mice was described by the group of Dr. Gage. Increasing the amount of maternal care blocks the accumulation of

L1 through the activation of DNA-methylation enzymes [191]. Early life experience drives somatic variation in the genome via L1 retrotransposons.

In thid work I report an example of HERVH co-option, involving not just a single transcript or an affected gene, but an actual protective function against actively transposing REs in hESCs.

1.3. Prior knowledge, hypothesis, and aims of the study

Expression of HERVH marks primed hESCs in comparison with chemically induced naïve state of human pluripotency. The transcriptome of cells in the "forced" naïve state had hallmark transcripts, derived from L1, SVA, and Alu elements [5]. A similar pattern of expression was detected in the human pre-implantation embryo, where earlier embryonic stages, 8-cell stage, and morula, had elevated levels of young RE transcripts, whereas pluripotent cells in epiblast mostly expressed HERVH (data from [48, 192], analysis Singh et al, unpublished) (Figure 6, left). L1s, Alus, and SVAs were highly expressed during the early maturation stages of human fibroblasts reprogramming, while later, in stabilization stages LTR7s expression had arisen (data from [193], analysis Singh et al, unpublished) (Figure 6, middle). This expression pattern of HERVH vs young REs was probably not just a correlation due to for example contrasting histone marks, but a HERVH-regulated effect, as HERVH depletion in H1 hESC caused similar upregulation of L1, Alu, and SVA transcripts (data from [7], analysis Singh et al, unpublished) (Figure 6, right).

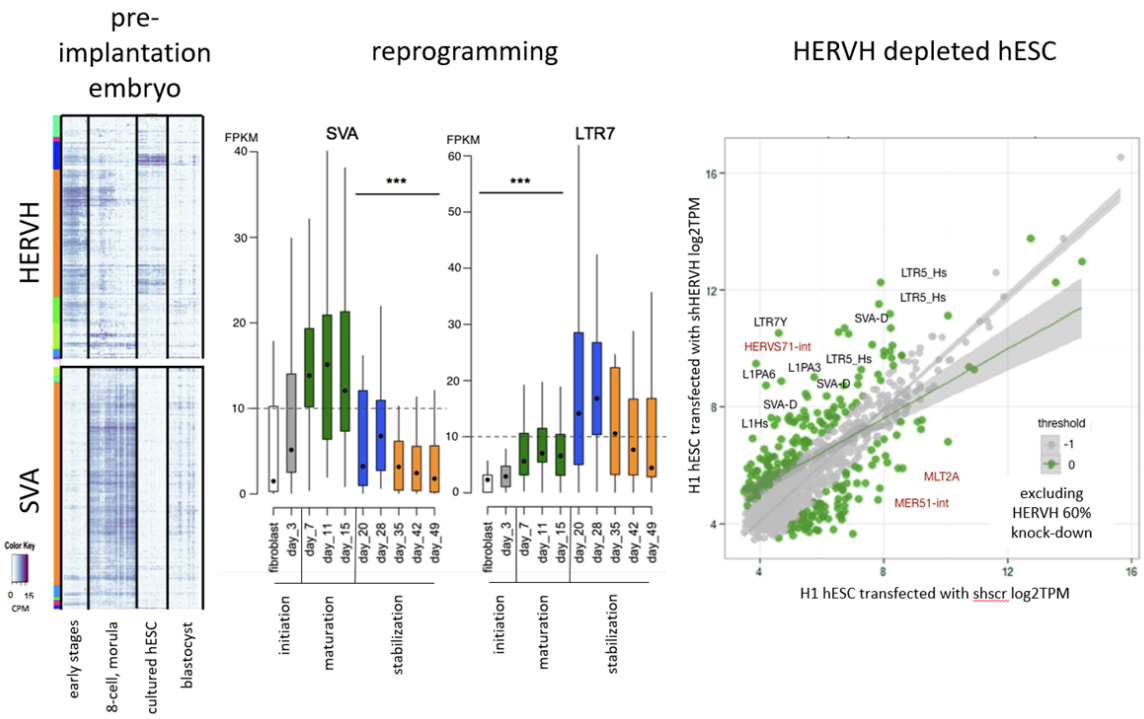


Figure 6. HERVH contrasting expression to young REs (figure and data generated by Dr. Singh). Left panel: in pre-implantation human embryo, SVA transcripts are upregulated during 8-cell and morula stages when HERVH RNAs are underrepresented in single-cell RNA-seq data. Earlier and later stages of development, together with cultured hESCs, have high transcripts number, derived from HERVH loci. Middle panel: during the maturation stage of reprogramming, SVA RNAs are highly present for up to 15 days with the decline after the HERVH transcripts start to be abundant at the beginning of the stabilization stage of reprogramming. Right panel: HERVH depletion in hESC causes elevation of SVA and L1 RNAs. RNA-seq from cells, transfected with an scr control shRNA was compared to shHERVH transfected cells. The significantly changed expression is shown with upregulation for SVA-D, L1Hs, L1PA3, and a few ERVs, for example, LTR5_Hs promoter of HERVK. Except for excluded HERVH depleted transcripts, some other older ERVs are also downregulated based on promoter expression like MLT2A, MER51-int, or full-length HERVS71-int.

These preliminary results allow me to formulate a hypothesis for the research work, which is:

HERVH controls the transposition of L1, SVA, and Alu in human embryonic stem cells.

Based on the hypothesis, I could single out these aims:

1. Assess L1 transposition in HERVH depleted background.
2. Detect *de novo* retroelements integrations in HERVH-depleted human embryonic stem cells.
3. Show the mechanism of HERVH-mediated retrotransposition control.

2. Methods

2.1. Experimental methods

2.1.1. Ethical approval

The work on hESCs was approved by Robert Koch Institute, Berlin, Germany, document number AZ 3.04.02.119-E02.

2.1.2. Human embryonic stem cells maintenance

To confirm experimental results in independent replicates, where specified, two hESC cell lines were used: H1 (WA01 alias, 18-W0260, WiCell) and H9 hESC (WA09 alias, WB67615, WiCell). Both cell lines were derived by Thomson with co-authors from intact or frozen and then thawed IVF embryos [1]. When experiments were performed in one cell line, H9 hESC were used.

H1 hESC were cultured in mTeSR™1 (85870, StemCell Technologies) media with addition of the commercial supplement and the primocin antibiotic (ant-pm-05, Invivogen). The recombinant human vitronectin (A14700, Life Technologies) was used as a coating protein for H1 hESC culturing. H9 on the other hand were cultured in Essential 8 Medium (E8) (A1517001, Life Technologies) with the commercial supplement, primocin and hESC-qualified matrigel (354277, Corning) as a coating agent. For the plasmid selection, 500µg/mL G418 (geneticin) antibiotic was used.

To increase the reproducibility between replicates, first numerous cell bank samples containing 5×10^5 of H1 or H9 hESCs were created and preserved in liquid nitrogen, then for each experiment a sample was thawed and used for further experiments. According to the previously established protocol in our research group, H1 or H9 hESC were thawed to one well of a six-well plate and grown until 70-80% confluency – the percent of the culturing surface, covered by cells. Cells were then passages in clumps using versene (15040066, Thermo Fisher Scientific) and seeded in 1 to 6 dilution ratio to new, matrigel or vitronectin coated plates in E8 or mTESR full media, adding rock inhibitor (Ri) (Y-27632, StemCell Technologies) to inhibit apoptosis and increase the survival rate. The next day Ri was withdrawn from the culturing media. After transfection or colony picking Ri was kept in the culturing media for 24h. Cells were used for transfections or other experiments after passaging twice since the thawing from the cell bank. To collect cells in a single-cell state, accutase (11599686, Thermo Fisher Scientific) was used.

2.1.3. Human embryonic stem cells transfections

Two types of hESC transfections were performed in this research. Transfections of shRNA constructs alone and with L1-EGFP reporter were performed with Neon™ transfection system (MPK5000, Thermo Fisher Scientific), which is an optimized electroporation protocol, where plasmid is delivered based on the high-voltage electric pulse. For each type and combination of plasmids used, the amount of DNA was optimized to obtain a high efficiency transfection and the least possible cell death, since the size of a plasmid would directly correlate with the number of dead cells after a transfection. 1×10^6 H9 hESCs per reaction with a relevant amount of plasmids were dissolved in the R buffer from Neon™ transfection system and using 100µL transfection tip (modified cuvette), electro-stimulated with a single pulse. Immediately after transfection cells were seeded to a warm full E8 media with Ri. For detailed transfection settings see below (Table 1).

Later due to malfunctioning of the Neon transfection machine and precipitate in the R buffer, I had to switch to Xtreme HP reagent (6366244001, MERCK) based transfection. All luciferase-containing plasmids were transfected with this method. The specific composition of the reagent is not disclosed but the general principle of transfection with the Xtreme reagent is based on the positively charged polymer compound, covering negatively charged plasmid DNA and the reagent-DNA complex is then internalized by cells through endocytosis [194]. One day before transfections 5×10^5 H9 hESCs per one well of a six-well plate were seeded in the E8 full media with Ri. The next day Ri was withdrawn, regular culturing media was added, Xtreme HP reagent was mixed with plasmids in OptiMEM media (31985062, Thermo Fisher Scientific) and transfected to cells according to the manufacturer protocol. The E8 media was changed 24 hours after transfection and the cells were maintained in the regular media and after five days collected for further analysis. I had optimized the amount of plasmids and Xtreme HP reagent used for every experiment. For detailed transfection information see below (Table 1).

Due to high toxicity of Xtreme HP reagent for H9 hESC, promptly after Neon™ transfections system issues were resolved, I had switched back to the electroporation protocol. For shRNA cassette integration experiments 5×10^5 cells were used with the same general transfection procedure as mentioned above. The detailed information about transfection settings is shown below (Table 1).

Table 1. Transfections setting, used in the study.

Cell number	Transfection type	Transfection settings/Reagents amount	Plasmids
1×10^6 at the moment of transfection	Neon	1400V 20 msec 1 pulse	8µg shscr
			8µg shHERVH
			8µg shscr + 8µg L1-EGFP reporter
			8µg shHERVH + 8µg L1-EGFP reporter
5×10^5 seeded one day before transfection	Xtreme HP	7.5µL Xtreme HP + 500 µL OptiMEM	2.5µg shscr + 2.5µg JM111 reporter
			2.5µg shscr + 2.5µg CAG-L1 reporter
			2.5µg shscr + 2.5µg L1-ORFeus reporter
			2.5µg shHERVH + 2.5µg JM111 reporter
			2.5µg shHERVH + 2.5µg CAG-L1 reporter
			2.5µg shHERVH + 2.5µg L1-ORFeus reporter
		5µL Xtreme HP + 500 µL OptiMEM	5µg CAG-L1 reporter
			5µg L1-ORFeus reporter
5×10^5 at the moment of transfection	Neon	1400V 20 msec 1 pulse	4µg shscr + 0.2µg piggybac transposase plasmid
			4µg shHERVH + 0.2µg piggybac transposase plasmid

2.1.4. FACS

The L1-EGFP reporter was a gift from Prof. Gerald Schumann, Paul-Ehrlich-Institut, Langen, Germany. Transfected with L1-EGFP reporter cells were analyzed by fluorescent activated cell sorting (FACS). shHERVH with L1-EGFP reporter and shscr with L1-EGFP reporter cells were collected five days after transfection with accutase in the single-cell state and analyzed with BD LSRII flow cytometer (BD Bioscience, India), calibrated, maintained and provided by flow cytometry facility (MDC, Berlin). The voltage selection for every channel, gate application and visualizations were performed with BD FACSDiva™ Software (BD Bioscience, India). The life cell population was selected based on an intensity of the forward and side scatters (FSC-A and SSC-A, respectively), followed by single-cell selection with FSC-H/FSC-W and SSC-H/SSC-W ratios. Next, using "shscr+L1-EGFP" sample as a negative control, the gates for GFP-positive and V-450-negative cell population were established, and applied for the 'shHERVH+L1-EGFP' sample. The detailed voltage set up for each channel is shown below (Table 2).

Table 2. Voltage settings for FACS analysis.

parameter	voltage, V
FSC-A	242
SSC-A	291
GFP-A	363
V-450_50-A	379

To visualize the FACS analysis, FlowJo software was used (BD Bioscience, India).

2.1.5. Luciferase transposition assay

JM111, CAG-L1 and L1-ORFeus luciferase reporters were a gift from Prof. Wenfeng An, Beijing University of Chemical Technology, Beijing, China. For the higher sensitivity luciferase-based transposition assay, five days after transfections with shscr/shHERVH and JM111/CAG-L1/ORFeus cells were collected with accutase in the single-cell state. The high sensitivity Dual-Luciferase® Reporter Assay System (E1980, Promega) was used for the assay. Cell pellets were lysed with 60µL of the 1x passive lysis buffer, provided in the kit and transferred to a black 96-well plate (3603, Corning), 20µL per well, resulting in three technical triplicates for each

sample. First firefly luciferase solution was added to the plate, according to the commercial protocol. The luminescent signal was acquired with the Spark 10M Microplate Reader (Tecan). Then the quenching and *Renilla* luciferase containing solution was added to the plate as described in the Promega protocol and the second luminescent signal was acquired. For both signals 0 msec settle time and 1000 msec integration time was used, showing the count values of the signal. Firefly signal was normalized to *Renilla* luciferase signal for each well independently and a mean value of three technical replicates was calculated for every sample.

2.1.6. Colony picking

To create a homogenous genetic background cell line from stable HERVH depleted and control cells, shHERVH and shscr transfected cells were selected with G418 and single colony picked. Bulk shHERVH or shscr cell lines were seeded to one well of the six-well plate in 1 to 10 ratio from the 80% confluent culture. The next day, using a picking hood in the Pluripotent Stem Cells facility (MDC, Berlin), colonies from each shHERVH and shscr cell cultures were transferred with a 20µL tip to one well of the 96-well plate each. Colonies were grown from 96- to 6-well plate, frozen for the cell bank and collected for the genomic integration validations with DpnI digestion.

2.1.7. Primers and oligonucleotides, used in the study

All oligonucleotides, used in this research, were synthesized by Biotez Berlin (Berlin, Germany). Primers for genomic PCR or qPCRs, designed for this study were created, using Multiple Primer Analyzer online tool (Thermo Fisher Scientific) for 40-60% GC content selection and the absence of predicted intra- or inter-primer dimers. Then the specificity of a primer pair was validated with the University of California Santa Cruz (UCSC) In-Silico PCR online tool from Genome Browser, and the predicted annealing temperature was used as a median value for gradient PCRs. Below the DNA sequences, used in this research are shown (Table 3).

Table 3. Oligonucleotide sequences, used in the study.

Experiments, sequences are used for	Name	Annealing temperature, C°	Sequence	source
HERVH depletion, section 3.1	shHERVH	NA	F:GATCCCCCTAAAGGCATAGTCAAGGTTATTCAAGAGATAACCTTGACTATGCCTTTAGTTTTTA R:CGTAAAAACTAAAGGCATAGTCAAGGTTATCTCTTGAATAACCTTGACTATGCCTTTAGGGG	[7]
	shscr	NA	F:GATCCCCGCGAAGTACGAATAGTTATCATTCAAGAGATGATAACTATTCGTACTTCGCTTTTTA R:CGTAAAAAGCGAAGTACGAATAGTTATCATCTCTTGAATGATAACTATTCGTACTTCGCGGG	[7]
Knock-down validation, sections 3.1	HERVHgag	60	F: ACGCTTTACAGCCCTAGACC R: GTCGGGAGCAGATTGGGTAA	[6]
	S18	60	F: GATGGTAGTCGCCGTGCC R: GCCTGCTGCCTTCCTTGG	[195]
HERVH depleted cells expression profile, section 3.1.1	*NANOG*	60	F: CCAAAGGCAAACAACCCACTT R: CGGGACCTTGTCTTCCTTTTT	[6]
	OCT4	60	F: CGACCATCTGCCGCTTTG R: GCCGCAGCTTACACATGTTCT	[6]
	PAX6	60	F: GTCCATCTTTGCTTGGGAAA R: TAGCCAGGTTGCGAAGAACT	[6]
	SOX1	60	F: CTGGCTGTGGCAAGGTCTTC R: CAGCCCTCAAACTCGCACTT	[6]
	BMP4	60	F: GAAGAATAAGAACTGCCGTCGC R: CACCTTGTCATACTCATCCAGG	[196]
	LMO2	60	F: ACTTCCTGAAGGCCATCGACCAG	[6]

			R: CACCCGCATTGTCATCTCATAGGC	
	AFP	60	F: AGCTTGGTGGTGGATGAAAC R: TCTGCAATGACAGCCTCAAG	[6]
	GATA6	60	F: GAGGGTGAACCCGTGTGCAATG R: TGGAAGTTGGAGTCATGGGAATGG	[6]
	ELF5	60	F: CGTGGACTGATCTGTTCAGCAATGA R: CAGGGTGGACTGATGTCCAGTATGA	[197]
	hCGA	60	F: ACCGCCCTGAACACATCCTGC R: GCGTGCATTCTGGGCAATCCTGC	[198]
shRNA cloned construct Sanger sequencing	PB_H1_seq	NA	F: CGTCATCAACCCGCTCCAAG	Dr. Izsvak group
shRNA cassette integration validation, section 3.3.2	PB_H1_ITR	55	F:TTAACCCTAGAAAGATAATCATATTGT R: TTAACCCTAGAAAGATAGTCTGCG	Dr. Izsvak group
LIN28A RIP qPCRs, section 3.6.2	*CDK4*	60	F: ATGTTGTCCGGCTGATGGA R: CACCAGGGTTACCTTGATCTCC	RIP kit, 03-105, MERCK
	HNRNPF	60	F: CAACAGAAACCAGTCCTGC R: GGAGACACTTCTGGATGGT	designed for this study
	U1	60	F: CCATGATCACGAAGGTGGTTT R: ATGCAGTCGAGTTTCCCACAT	[199]
	chrX locus	60	F: CAGACCGACCAGCCCAAGG R: CCACGGATAAAACGTGTCTCC	designed for this study
	chr1 locus	60	F: CAGACCAACCAGCCCAAGG R: CGGAGTTTTGGGTCCACGGAC	
	Chr4 locus 1	52	F: GAAACATCTCACCAATTTC R: GGGTCCACGGATAAAACA	

	Chr4 locus 2	60	F: ATCTTGGCGCCACACTT R: GGATTAAATACCAAGGGAAG	
	ESRG	60	F: GACATTGTCCTTCCAACTCT R: CTTTTATGCATTGGCTTGTT	
De novo Alu integrations validation, section 3.3.4	short39116a3e	43	F: GTACATACAAACCCAAACTT R: AATAAAGCCAAGATACTCTC	
	long39116a3e	60	F: AAGCCTGATCTAAAAATCA R: AAAGACATCAAGGCCG	
	short4da958be	48	F: AGATATACTGGAAGCAAAC R: GCCTCAATTATTATTATTAC	
	long4da958be	60	F: GTGTTTCCTGTTATTTTACTC R: AAGAATAATGTCCTCCACA	
	short10ef1146	63	F: TGAAAAACCTGGAACAGGCATG R: CCCGGCCGCTAGTTCTTTT	
	long10ef1146	60	F: GCCCAGGCTGTGTTTTT R: AAATGTTGCTAGTTGTTTTCTTT	
	short49c1b582	58	F: CTATGCAAAGAGATTTTGTGTC R: ACCTGCTACTCAATCCAGCT	
	long49c1b582	55	F: CAAATTGCCCAAACTGCT R: TTAGGTCAGAAGCTGAGCAT	
	short29e355c6	63	F: GGATTAAGAAGGATTCTTTTGTG R: CAGGGATGAAAAAAAATCAAAT	
	long29e355c6	56	F: TCCAGCCATCTAGGATACAA R: GATACTAAGTCTGTTTTGTTTTGC	
	shortdd84f645	70	F: GACCCTCAGCTCTACGCAAGCAG R: CTTTGATTAGACCCAAGCTCCTCA	
	longdd84f645	64	F: TGAGCAAACTGAGTCCTTTCC R: CTCTTCCTTGAAGCCCTCAG	

	short681c4984	63	F: CTGTAGTGGGCTGATGGTC R: AGTGACTGCAGGTCAGAGC	
	long681c4984	65	F: TTTGCATCACACACACTGG R: AATCTGTCCCAAATGTCCTG	
	shortcd47b46a	68	F: CTAATCTGAAGTGCTGAAGCTCA R: AATGACTTACTTGGATTTTGCTTG	
	longcd47b46a	65	F: TAACTAGCTGCGGAACTGTGA R: CATTCATATTCCAGTAACCAAGTAC	
	short7e4f627d	70	F: TTGGAGCTCGTGACCCACTC R: TCTCACTGGCCTGGCTTCACT	
	long7e4f627d	64	F: CGAACAGAAACGTGTCACCA R: TGCTCCAGGCTTCTCCCA	
	shortbd2c13a7	63	F: GGCAAACAGTTGTCTTATTA R: CTAAGATATTCTACTCCCAGTT	
	longbd2c13a7	56	F: CAATTTATCAGGACAAACAG R: CTATCTGAAGTGTTGTGCTAG	
	short8203ea94	63	F: CCTTCCAAGTTATTCTCTG R: GCCTGATCTTATTAAACCT	
	long8203ea94	57	F: GTACCATAAAGAACAAGTTG R: GATACATAGGCAGAGAACC	
	short30012f02	62	F: TCAGGCCTGTGGAATCC R: CTTCCAGGAAAGTTCAAGGA	
	long30012f02	56	F: TATTATGGTCTGGGTGATGAT R: CAGCTAGAAACCACTATTAATATG	

Description: F stands for “forward primer”, R for “reverse primer”. Both primer sequences are shown 5’ to 3’ end. Names of primers for *de novo* Alu integrations validation: “short” stands for the second, shorter fragment of a nested PCR, “long” for the longer, first fragment of the nested PCR, letters, and numbers – the first part of the integration’s names, see the supplementary I.

2.1.8. RNA isolation and quantitative PCR analysis

To validate efficiency of a knock-down, cells were collected five days after transfection or after selection with antibiotic for stable shRNA integration. RNA was isolated with Trizol (15596026, Thermo Fisher Scientific) and further purified with Quick-RNA Miniprep Kit (R1055, Zymo Research). Due to repetitive nature of amplified sequences, RNA was excessively treated with DNAse, to avoid any genomic DNA cross-contamination. First, according to the commercial protocol, the samples were DNAse treated in column during the RNA isolation and then, additionally with TURBO DNA-free™ Kit (AM1907, Thermo Fisher Scientific) according to the manufacturer protocol.

The RNA samples of cell lines, undergoing differentiation or established cell lines, used as positive controls for expression of differentiation markers in the section 3.1.1. were kindly provided by my colleagues from the research group of Dr. Izsvak and Dr. Gouti, Max-Delbruck Center for Molecular Medicine, Berlin, Germany.

RNA was reverse-transcribed to cDNA with High-Capacity RNA-to-cDNA™ Kit (4387406, Thermo Fisher Scientific) corresponding to the commercial protocol. One reverse-transcription reaction contained 2000ng RNA. The synthesized cDNA was next used for quantitative PCRs (qPCRs), performed with Power SYBR™ Green PCR Master Mix (4367659, Thermo Fisher Scientific), which consists of SYBR green fluorescent dye, activated upon binding to DNA, and polymerase with nucleotides to perform PCR reaction. Typically, 10ng of cDNA were used per reaction with 10pmol/µl of each forward and reverse primer. 384-well or 96-well plates (12680985, Thermo Fisher Scientific or 1845098, Biorad) were used to perform qPCR reaction on 7900HT Fast Real-Time PCR System with 384-well block module (4329001, Thermo Fisher Scientific) or CFX96 Touch Real-Time PCR Detection System (1845096, Biorad) with 96-well block module, respectively. The data were analyzed in the programs, corresponding to qPCR systems. The signals detection was determined with the automatic threshold values (for the 7900HT machine) or the regression model (for the CFX96 machine). Two ways of quantification were further used: calibration curves normalization or $2^{(-\Delta Ct)}$ method.

2.1.8.1. Calibration curve normalization

Calibration curves were separately built for every primer pair. 1/10 of synthesized cDNA from every sample, which belong to one experiment, were mixed, creating a calibration

curve master mix. The master mix was then two-fold serial diluted and five samples ranging from 50ng to 3,125ng were used for calibration curves. The mean of technical triplicates was calculated, to reduce the high variability of qPCR. Calibration curves were the functions of Ct value (y axis) to the amount of cDNA, used in the reaction (x axis). Ct value is a serial number of the PCR cycle, when a DNA product becomes detectable above a certain threshold. Based on the linear regression model the trendline for each calibration curve was calculated. The generated formula was used to calculate the amount of cDNA in a sample of interest.

The residual cDNA of every sample was diluted up to 10ng/µL and used in the qPCR, 10ng per reaction. A mean Ct value of three technical replicates was calculated and used in the previously established trendline formula for every primer pair, obtaining a relative amount of the product per sample. Then a target gene product quantity was divided by s18 reference gene value. Several independent replicates were analyzed to quantify the expression values, relative to s18, normalized with calibration curve.

2.1.8.2. $2^{\wedge(-\Delta Ct)}$ normalization method

The $2^{\wedge(-\Delta Ct)}$ qPCR quantification method is based on the protocol, described by Livak and Schmittgen [200]. From a Ct value of a target gene the s18 reference gene Ct was subtracted, resulting in ΔCt value. Then $2^{\wedge(-\Delta Ct)}$ was calculated, showing the relative amount of a target gene product to the reference gene.

2.1.9. PCR, gel electrophoresis, and DNA fragments isolation

PCR was performed using 10ng of genomic DNA or 1ng of a plasmid DNA as a template, if not mention else, and KAPA HiFi plus dNTPs kit (KK2102, Roch) with HF buffer. 10pmol/µl of each forward and reverse primer were applied to every reaction. For gradient PCRs, five reactions were used for every primer pair, with the median temperature being predicted by UCSC In-Silico PCR online tool and 1° difference for every ascending and descending temperature values. The cycling amplification program was designed according to the KAPA HiFi polymerase commercial protocol with 30 sec of extension time per one kilobase of a predicted product length and 30 PCR cycles for every reaction. PCR products were analyzed with 1% agarose gel electrophoresis, performed according to the conventional protocol in Tris-Borate-EDTA buffer (TBE) [201]. DNA was visualized with a safe alternative to ethidium bromide, Midori Green (MG04, NIPPON Genetics Europe) dye, which was added to the

agarose gel, before polymerization. After the agarose gel polymerization, PCR products were mixed with 10X Orange Loading Dye (927-10100, LI-COR Biosciences) and added to the wells of the agarose gel, immersed into TBE buffer. GeneRuler DNA Ladder Mix (SM0331, Thermo Fisher Scientific) served as a marker for the length of DNA fragments. The electrophoresis was normally performed at 100V for 30 min and the DNA fragments were then visualized at ~300nm UV wavelength with the ChemiDoc Imaging System (Biorad). The desired size DNA fragments were cut out from the agarose gel and isolated with Zymoclean Gel DNA Recovery Kit (D4008, Zymo Research), according to the manufacturer protocol.

2.1.10. Genomic DNA isolation

For genomic DNA isolation, H9 hESC were cultured to 80% confluency, collected with accutase, lysed according to the DNeasy Blood & Tissue Kit protocol (69504, Qiagen) and further processed as in the commercial protocol, excluding the high-temperature lysis step. Obtained DNA samples were further sent to BGI Genomics (Hong Kong, China) for whole-genome sequencing or for DpnI digestion and PCR amplification, to confirm integrations of shHERVH or shscr cassettes.

2.1.11. DpnI analysis

A shscr plasmid template for the reaction optimization or genomic DNA, isolated from H9 hESC with integrated shRNAs was digested with DpnI restriction enzyme (R0176S, New England Biolabs). 500ng of the plasmid or genomic DNA were incubates for 3.5 hours with the DpnI at 37°C, followed by 20 min inactivation at 80°C. The digestion products were isolated from the solution with the Monarch® PCR & DNA Cleanup Kit (T1030S, New England Biolabs). 1ng of the products served as a PCR template, amplifying the shRNA cassette with the primers, aligning to inverted terminal repeats (ITRs) of the piggybac transposon. 15 rounds of amplification were performed for the plasmid template and 23 for the genomic DNA samples. PCR and agarose gel electrophoresis were conducted as described above.

2.1.12. Alu integrations validation

2.1.12.1. Annealing temperature optimization

Five gradient PCR reactions were performed for each of the 12 primer pairs for the first round of nested PCRs, named “long” in the Table 3. 10ng genomic DNA from wild type H9

hESC served as a template. PCR was performed as described above with following agarose gel electrophoresis and DNA products were isolation from the gel. 100ng of each product was used for the Sanger sequencing with amplification primers. If an amplified product corresponded to the predicted genomic location, 10ng of the product per one reaction were used as a template for the second round of nested gradient PCR with the primers named "short". The same principle of the gradient PCR and products validation was applied as mentioned above.

2.1.12.2. Amplification in the stable knock-down clones

50ng of genomic DNA from scr control clone 1 and HERVH depleted clone 1 and clone 3 were used for the first round of nested PCRs with previously optimized annealing temperatures for every primer pair. PCR products were analyzed with 1% agarose gel, the desired size fragments and up to 500bp above were isolated from the gel and the half of the reaction was used for the second round of nested PCR. The potential Alu products were analyzed with agarose gel.

2.1.13. Molecular cloning

To achieve HERVH depletion in H9 hESC, shRNA sequence, previously described by Lu and co-authors [7], targeting the *gag* region of HERVH consensus (figure 4) and a scrambled shRNA sequence were used. Both hairpin sequences are shown in the table 3. The direct orientation and reverse-complement oligonucleotides (100pmol/µL each) with overhanging unpaired nucleotides of BglII/ClaI restriction sites were annealed in T4 DNA Ligase Reaction Buffer (B0202S, New England Biolabs) at 95°C for 5 min followed by gradually descending temperature in the turned off thermocycler (DYAD DUAL Gradient 48x48 Well Block PCR Thermal Cycler, Biorad). Thus, full hairpins with active BglII/ClaI restriction sites were annealed and then ligated to linearized PB_H1 vector. PB_H1 is a vector for shRNA expression, under H1 promoter, which contains ITRs of piggybac transposon for further integration of the shRNA cassette to a genome. The detailed plasmid map of the vector is shown in figure 7.

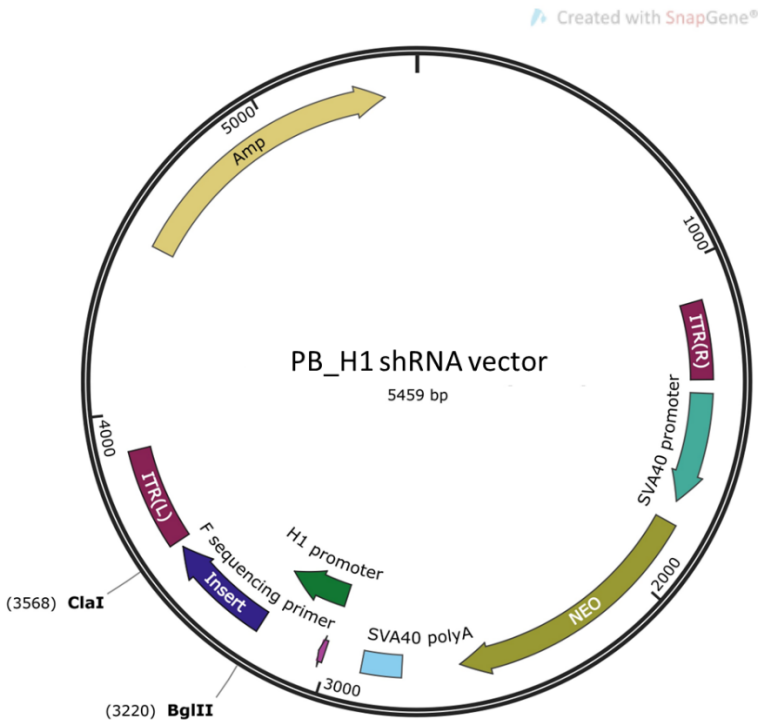


Figure 7. PB_H1 shRNA expression vector map. The plasmid map is generated with the SnapGene Viewer. shRNA cassette is cloned at BglII-ClaI restriction sites, cutting out the insert (blue). shRNA expression is driven by H1 promoter (green), to which a forward (F) sequencing primer (pink) anneals. shRNA cassette is integrated via mobilization with piggybac ITRs (purple) and selected for with G418, resistance provided by neomycin gene (NEO, yellow-green), expressed independently due to its own SVA40 promoter (cyan) and SVA40 polyA signal (blue). To select in bacteria, ampicillin (Amp, yellow) resistance is used.

In parallel to hairpins annealing, PB_H1 vector (2µg) was digested with BglII and ClaI enzymes (R0144S, R0197S, New England Biolabs), each 5 units per reaction, in NEBuffer™ 3.1 (B7203, New England Biolabs) for one hour in 37°C. Then the digestion reaction was analyzed with agarose gel electrophoresis and the 5111bp DNA fragment was isolated from the gel. The ligation of PB_H1 fragment and the annealed shHERVH or shscr hairpins was performed with Quick Ligation™ Kit (M2200S, NEB) in the 1:3 vector-to-insert ratio for 5 minutes at room temperature. 1/10 of the reaction mix was transformed to Mix-and-Go competent bacterial cells, prepared in our research group with Mix & Go E.coli Transformation Kit (T3002, Zymo Research) from DH5α strain. The selection for ligated plasmids was performed by plating the

reaction mix to LB agar (22700041, Thermo Fisher Scientific) plates, containing 50 µg/mL final ampicillin concentration (11593027, Thermo Fisher Scientific). The next day after transformation, three bacterial colonies for each ligation were transferred to 3mL of the liquid LB media (X968.1, Roth), supplemented with the same concentration of ampicillin and grown with agitation (200rpm) for 16 hours. Bacteria were always grown at 37°C.

Plasmids were isolated with NucleoSpin Plasmid Mini kit (740588.250, MACHEREY-NAGEL) according to the protocol. shRNA ligation was confirmed by Sanger sequencing (LGC group, Berlin), using a forward primer, annealing to the H1 promoter of the construct. The positive for shHERVH or shscr ligated construct bacterial clones were preserved in a 25% glycerol freezing solution to further use plasmids from the same clonal genetic background.

To efficiently transfect plasmids in hESC, highly concentrated endotoxin-free plasmids were prepared with NucleoBond Xtra Midi kit (740410.100, MACHEREY-NAGEL). Transfections were performed as mentioned above, section 2.1.3.

2.1.14. LIN28A RIP-qPCRs

Two 90% confluent 6-well plates of H1 or H9 hESC were washed with PBS and collected with accutase in the single cell state. The cells were lyzed and further processed according to the commercial protocol for Magna RIP RNA-binding protein immunoprecipitation kit (17-700, Merck). 10% of the input reaction was saved before LIN28A or IgG precipitation for the further analysis. The RIP-grade antibodies, RIPAb+ Lin28 kit (03-105, Merck), were used to precipitate LIN28A. The isolated from the input, LIN28A and IgG control precipitation reactions RNA was reverse transcribed as described in the section 2.1.8. cDNA was ten times diluted and used for qPCR. The Ct values of 1% input were adjusted to the 6.64 dilution factor ($\log_2 100$), enrichment to the adjusted input of Ct values for LIN28A and IgG precipitations were calculated, and IgG enrichment was subtracted subtracted from LIN28A, to calculate the ΔCt value. The data were compared as $2^{\wedge(-\Delta Ct)}$.

2.2. Computational part

2.2.1. Integrations selection in R

TEBreak annotation and integrations selection resulted in 83 for depleted HERVH clone 1 and 180 for clone 3 predicted genomic locations, where Alus, SVAs or L1s had potentially

landed. The integrations were supported by the number of discordant reads – the reads which are derived from the same fragment in the sequencing data but align to different positions in the reference genome [202]. To select integrations for PCR validation, quantiles of discordant reads numbers were calculated with “quantiles” the basic R function (R version 4.2.1). For validation three integrations were selected from 0-25% quantile and three from the 75-100% for each depleted clone. The corresponding integrations coordinates were manually addressed in UCSC Genome Browser, human genome version hg19 and primers for predicted integrated region were designed as described above. In cases when the region of a potential integration was highly repetitive and efficient PCR primers was challenging to design, the next in discordant reads count predicted integration was addressed.

2.2.2. HERVH loci analysis

Below are the examples of commands, used in this study. First the data names and purpose of the analysis are described, then # stands for comments to executed commands, and ## for the environment, a command will be executed in. The grey highlights mark the actual commands. The commands are shown only for one dataset, for example, coordinates of control loci. All the actions were then repeated on other datasets, used for the specific analysis.

2.2.2.1. HERVH antagonistic loci coordinates retrieval

HERVH loci expression was analyzed by Dr. Singh in pre-implantation development data [48, 192], during reprogramming [193] and in HERVH knock-down cells [7]. HERVH expression in development and reprogramming clustered in several groups (see section 3.4.1). The full coordinates of HERVH loci for each separate cluster were provided to me as bed files. The commands below were used to discover HERVH loci coordinates, antagonistic to young REs.

Data used: HERVH coordinates from development, number - expression cluster: HH_dev(1-7).bed; HERVH coordinates from reprogramming, mat - maturation stage cluster, stab - stabilization stage cluster: HH_mat.bed, HH_stab.bed; HERVH coordinates in knock-down: HH_kd.bed

##in R studio, dplyr package

#strand filtering: all coordinates containing data were separated in two parts: (-) strand and (+)

```
HH_kd_min <- HH_kd %>% filter(strand=='-')
```

Control HERVH loci: HERVH coordinates from maturation stage of reprogramming and HERVH, expressed together with young REs in development.

##in R studio (2022.02.3+492 or older version), basic R functions

#all control loci - from maturation stage and development, excluding duplicates

```
df <- rbind(HH_dev(4-7),HH_mat)
HHcon <- df[!duplicated(df), ]
```

Next: Specific HERVH loci, expressed in negative correlation with young REs and depleted in HERVH knock-down

##in Linux terminal, bedtools toolset

#common loci for knock-down and stabilization stage

```
bedtools intersect -a HH_kd_min.bed -b HH_stab_min.bed -wa > kd_stab_HH_min.bed
```

#strand filtering with dplyr, done as previously described

##in Linux terminal, basic function

#sorting both datasets

```
sort -k1,1 -k2,2n HHcon_min.bed > srt_HHcon_min.bed
sort -k1,1 -k2,2n kd_stab_HH_min.bed > srt_kd_stab_HH_min.bed
```

HERVH antagonistic loci (HHant): common for stabilization stage of reprogramming and depleted in HERVH knock-down, but not present in HERVH control group.

##in Linux terminal, bedtools toolset

#excluding control loci from stabilization-knock-down common dataset

```
bedtools intersect -a srt_kd_stab_HH_min.bed -b srt_HHcon_min.bed -v > HHant_min.bed
```

2.2.2.2. Sequence tailoring and alignment

HERVH antagonistic loci were aligned to control HERVH loci, expressed in parallel with young REs. The size selection was performed, as sequences-outliers could disturb the

alignment. DNA sequences were downloaded from hg19 version of the human genome. The alignment itself was done with Muscle algorithm [203].

First, the length distribution of sequences was addressed, and the homogeneous subset was selected.

```
##in R studio, basic R functions
#combining both strands, same done for control loci
HHant <- rbind(HHant_min,HHant_pls)
#loci length calculated, same done for control loci
size_HHant <- c(HHant$end-HHant$start)
HHant_size <- cbind(HHant, size_HHant)
#visualizing the loci length distribution
df <- HHant_size[order(HHant_size$size),]
plot(df$size)
```

The length distribution is shown in the figure 31. Next, HERVH control and HERVH antagonistic loci were selected by loci length.

```
##in R studio, basic R functions
#subsetting by loci length
HHant_sub <- subset(HHant_size, HHant_size$size>2500 & HHant_size$size<3500)
```

Further, the DNA sequences were obtained from hg19 version of the human genome.

```
##in Linux terminal, bedtools toolset
#retriving fasta sequences
bedtools getfasta -fi hg19.fa -bed HHant_sub -fo HHant.fasta
#combining in one file
cat HHant.fasta HHcon.fasta > HHant_HHcon.fasta
```

The memory-demanding alignment was done with the use of MAX cluster (MDC, Berlin) computational power.

```
##Max cluster
muscle -in HHant_HHcon.fasta -out HERVHaln.afa -maxiters 2
```

2.2.3. *lin* motif genome-wide alignment

To detect how is the lin motif distributed in the human genome, a short read aligner Bowtie was used [204]. Based on the previously performed alignment with Muscle, the 16bp *lin* motif was recorded as a quality score containing sequence file (fastq), where the two GGAGA binding sites were assigned the high quality and the six nucleotides between the sites the low score. The whole 16bp motif was serving as a seed during alignment, one mismatch per seed was allowed and all aligned genomic coordinates were reported. As the result the alignment algorithm preferred mismatches between the GGAGA binding sites.

Data: the full sequence of human genome, hg19 assembly, file: hg19.fasta. *lin* motif with quality control scores, file: lin_motif_q.fastq. *Lin* motif coordinates in hg19 genome, file: lin_motif_hg19_b1n1q.bed.

##Max cluster

#human genome indexing

```
bowtie-build --threads 4 -f hg19.fasta hg19_index
```

#aligning the motif to the genome

```
bowtie -q lin_motif_q.fastq -n 1 -l 16 --all reads_b1n1_hg19 --best -x hg19_index
```

2.2.4. Transposons annotation

After obtaining the coordinates of *lin* motif in the human genome, repetitive elements, residing in these coordinates were annotated with Repeat Masker [205]. The Repeat Masker annotations were downloaded from the UCSC Table Browser and used in a bed file format.

Data: transposons annotations with RepeatMasker, file: hg19_rmsk.bed; *Lin* motif coordinates in hg19 genome, file: lin_motif_hg19_b1n1q.bed.

#obtained coordinates were extracted from the output lin_motif_hg19_b1n1q.bed and separated to (+) and (-) strands in R studio with dplyr package

#repeat masker annotations (hg19_rmsk.bed) were separated to (+) and (-) strands in R studio with dplyr package

##in Linux terminal, bedtools toolset

#transposons coordinates, overlapping with lin motif

```
bedtools intersect -a min_hg19_rmsk.bed -b lin_motif_min_hg19_b1n1q.bed -wa
> TEs_lin_b1n1q_min.bed
##in R studio, dplyr package
#count number of coordinates per each transposon family
TEs_freq_min<-
data.frame(table(unlist(strsplit(tolower(TEs_lin_b1n1q_mi$TEname), " "))))
```

The data were saved as tables and further analysis with graphical representation was done in Excel, MS Office.

Similar function was used to calculate the number of HERVHlin and HERVH control loci, residing in each chromosome of the human genome.

Data: coordinates of HERVHlin loci in the human genome, (-) strand, file: HERVHlin_min.bed

```
#count number of HERVHlin or HERVHcon loci per each chromosome
chr_min <- data.frame(table(unlist(strsplit(tolower(HERVHlin_min$chr), " "))))
```

2.2.5. *lin* motif alignment to primate genomes

The similar method as in the section 2.2.3 was used to discover the *lin* and control motives distribution in genomes of chimpanzee, gorilla, orangutan, gibbon, rhesus macaque and marmoset. Bowtie was used to align *lin* or con sequences with assigned quality scores to these versions of the genomes: panTro6, gorGor6, PonAbe2, nomLeu3, rheMac10, calJac4. Then the transposons were annotated by finding overlaps between Repeat Masker and *lin* or con motif coordinates. Repeat Masker annotations were downloaded from UCSC Table Browser, corresponding to the used genome version.

2.2.6. CLIP-seq analysis

Cross-linking immunoprecipitation-high-throughput sequencing (CLIP-seq) trimmed data were downloaded from Gene Expression Omnibus (GEO), accession number GSE39873, sample GSM980593 LIN28ES CLIPseq [33] and saved as a fasta file. The previously generated hg19 genome index was used for analysis. The CLIP-seq data were aligned to the hg19 human genome with a 16bp seed region, one allowed mismatch per seed, reporting all hits and

sorting them by the best score in a sam type output. To detect unique coordinates, output reads were merged, reporting the number of reads per coordinate. Then the Repeat Masker annotations were overlapped with unique CLIP-seq coordinates and transposons classes were annotated as before.

Data: LIN28A trimmed CLiP-seq reads, file: Lin28_clip.fasta. Human hg19 genome indexed with Bowtie aligner, file: hg19_index. Coordinates of aligned to hg19 CLIP-seq reads, file: reads_clipn1l16. Repeat Masker annotation, separated by (-) and (+) strand, file: min_hg19_rmsk.bed

##Max Cluster

```
bowtie -f Lin28_clip.fasta -n 1 -l 16 --all --best -S --chunkmbs 128 reads_clipn1l16 -x hg19_index
```

#coordinates separated to (+) and (-) strand in R studio with dplyr package as described above

##in Linux terminal, bedtools toolset

#merging reads from the same coordinates, saving the number of reads per coordinate

```
bedtools merge -i min_clip.bed -c 4,2 -o distinct,count > unique_min_clip_b1n1all.bed
```

#transposons coordinates, overlapping with CLiP-seq unique coordinates, saving the number of reads per coordinate

```
bedtools intersect -a min_hg19_rmsk.bed -b unique_min_clip_b1n1all.bed -wa -wb > TEs_clip_min.bed
```

#full-locus transposon coordinates, saving the number of reads per locus

```
bedtools merge -i TEs_clip_min.bed -c 4,6,12,13,15,2 -o distinct,distinct,distinct,distinct,distinct,count > Tes_clip_min_unq.bed
```

Further the similar method as the section 2.2.4. was used to describe count number of coordinates per each transposon family.

2.2.7. Statistics.

The *lin* motif in antagonistic HERVH loci (HERVHant) sequences vs HERVH control was validated with Fisher's exact enrichment test in Analysis of Motif Enrichment tool (AME), a part of the MEME suit, motif-based sequence online analysis tools [206]. HERVH control sequences functioned as control sequences and the HERVHant sequences as primary sequences. The *lin* motif was typed in and an average odds score was the sequence scoring method.

Fisher's exact test was also used to validate significance for HERVHlin to HERVHcon presence in different chromosomes of the human genome (Figure 36) and *lin* motif enrichment in primate genomes (Figure 37). Easy Fisher Exact Test Calculator from the Social science statistic webpage was employed with the default settings (https://www.socscistatistics.com/tests/fisher/default2.aspx).

For most of the two samples with several replicates types of analyzed data online version of the unpaired t-test from GraphPad by Dotmatics was used (https://www.graphpad.com/quickcalcs/ttest1.cfm). In CLIP-seq analysis (Figure 38) t-test was performed in R studio, the stats (version 3.6.2) package with default settings.

The significance of transposons families enrichment in CLIP-seq data (Figure 38) was validated with one-way ANOVA test via anova basic function in R with default settings.

3. Results

3.1. Transient HERVH knock-down in human embryonic stem cells

To test the hypothesis of HERVH controlling the activity of phylogenetically young retrotransposons, HERVH knock-down was first established in H9 hESC. An shRNA sequence from previously published work (named “shRNA3” in the article, [7], also see the section 2.1.7. and table 3) was used to target the *gag* region of HERVH consensus (Figure 4, section 1.2.2.1) based on sequences of 231 expressed HERVH loci [7].

The detailed protocol for shRNA cloning to the PB-H1 expression vector is described in the section 2.1.13. Briefly, the PB-H1 vector was digested with BglII/ClaI restriction enzymes and then annealed shRNAs against HERVH or non-targeting control were ligated to the vector. The presence of the desired plasmids was confirmed in bacterial clones with Sanger sequencing and a high-yield endotoxin-free plasmid was isolated from one of the clones to use for high efficiency transfection in hESCs [207]. Further, H9 hESC were transfected via electroporation in three independent replicates.

Cells were transfected with shRNA against HERVH (shHERVH) or shscr control, non-targeting shuffled sequence of the original shRNA, named “shRNA3 scramble” in the article by Lu and co-authors [7], (see the section 2.1.7., table 3) using the Neon transfection system. H9 hESC were collected five days after transfection, followed by RNA isolation. The efficiency of HERVH depletion was validated by qPCR. Primers amplifying the *gag* region of HERVH were used (“HERVHgag” primer pair from [6], the section 2.1.7., table 3) to detect a general scope of HERVH transcripts. Based on the UCSC Genome Browser *in silico* PCR prediction, the primer pair amplifies more than 400 HERVH loci. To normalize to the RNA amount and quality, *S18* primer pair amplifying ribosomal protein (RPS18) was used [195]. The normalized HERVHgag expression to s18 with calibration curve is shown below (Figure 8).

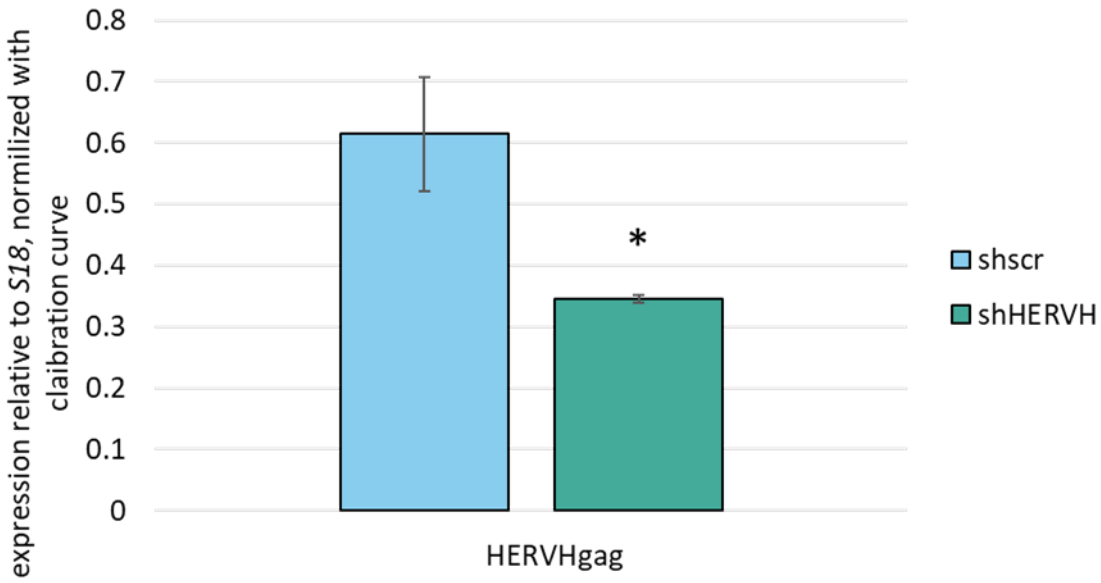
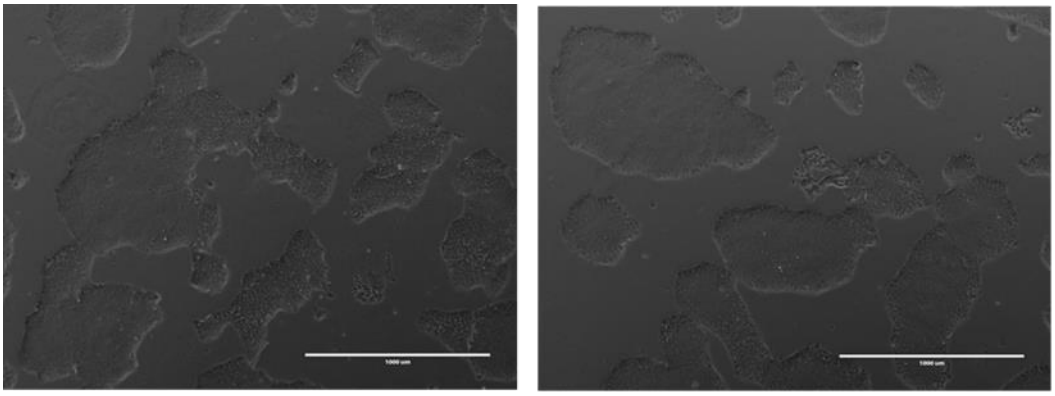


Figure 8. HERVHgag expression is reduced after shHERVH transfection in H9 hESC. HERVHgag expression is normalized to *S18* with the calibration curve, mean of 3 independent replicates is shown, bars represent standard error of the mean (SEM), * – statistical validation with t-test, showing significant differences of HERVHgag expression between shscr and shHERVH samples (p=0.0445).

HERVH transcripts are known to support pluripotency [6–8], which could be affected by shRNA depletion. Therefore, I decided to validate the expression of pluripotency and differentiation markers in HERVH depleted H9 hESC.

3.1.1. Expression profile of HERVH depleted human embryonic stem cells

HERVH depletion has been shown to provoke differentiation of H1 hESC, reflected in morphological changes of colonies and reduced expression of *OCT4*, *SOX2* and *NANOG* pluripotency factors [7]. As no variance in morphology of HERVH depleted H9 hESC was detected (Figure 9), pluripotency and differentiation status after HERVH knock-down was validated with qPCRs.



H9 hESC colonies, 5 days after transfection with shscr

H9 hESC colonies, 5 days after transfection with shHERVH

Figure 9. HERVH depletion does not change morphology of H9 hESC. Phase-contrast light microscopy images acquired five days after transfection with shscr or shHERVH. Scale – 1000um.

First, expression of the pluripotency markers, *NANOG* and *OCT4*, was determined. NANOG is known to recruit a chromatin modification complex to maintain H3K4me3 activating histone marks on genes, which expression is crucial for the maintenance of a core pluripotency network [208]. OCT4 not only controls expression of a wide range of target genes, resulting in differentiation inhibition [209], but also regulates *NANOG* itself by binding to the promoter [210, 211]. Hence *NANOG* and *OCT4* could be considered as sufficient markers to validate pluripotency state changes in HERVH depleted cells.

qPCR was performed with three independent replicates for HERVH knock-down in H9 hESCs and a shscr transfection control, the *S18* gene was used as a reference to calculate expression of HERVHgag with calibration curve method. Wild type H9 hESCs were used as a positive control, and early stages of cell differentiation towards mesodermal and neuro-ectodermal linages served as negative controls (Figure 10).

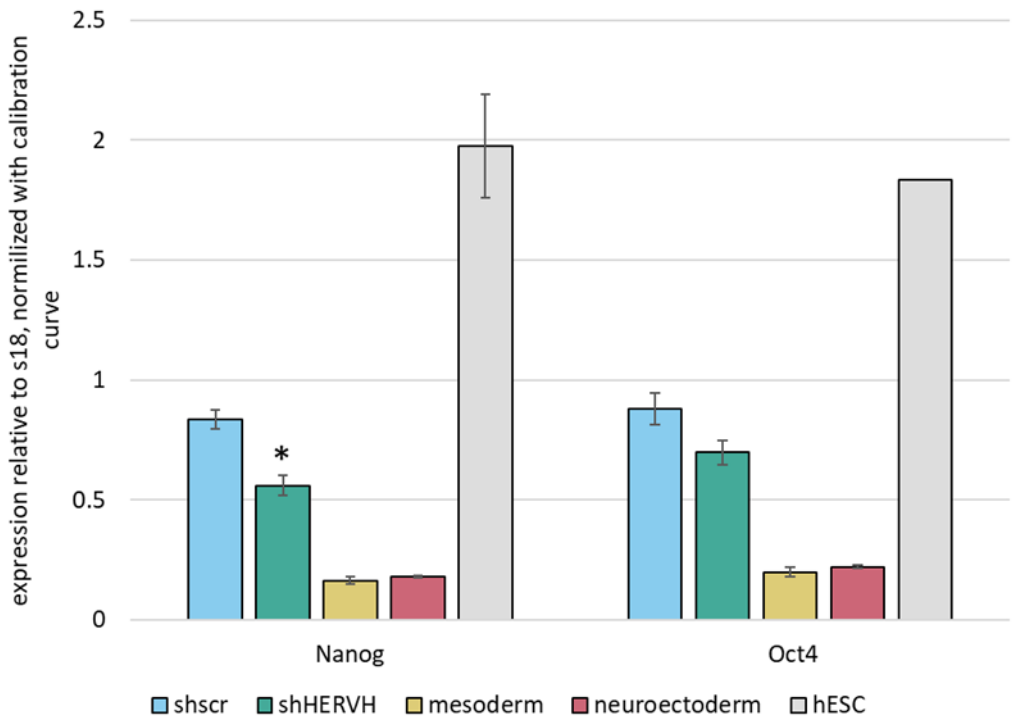


Figure 10. *NANOG* and *OCT4* pluripotency markers expression after HERVH knock-down in hESC. *NANOG* and *OCT4* expression are normalized to *S18* with the calibration curve. The mean of three independent replicates is shown, except for hESC for *NANOG* and *OCT4* in two and one replicates, respectively; bar represents standard error of the mean (SEM), * – statistical validation with t-test, showing significant *NANOG* depletion (p=0.0087) in shHERVH transfected cells.

Contradictory to previously reported results [7], the shHERVH knock-down did not cause depletion of *OCT4*, but resulted in down-regulation of *NANOG* expression, which supports the idea of HERVH being crucial for pluripotency maintenance at least in relation to *NANOG* functionality.

NANOG depletion promotes differentiation of hESC [212], which could cause changes in histone accessibility and, as a result, elevated retrotransposon activity. Thus, an expression of extraembryonic lineage markers and differentiation markers was tested in HERVH knock-down samples.

To test if HERVH depletion causes differentiation of hESC towards neuroectodermal lineage, *PAX6* [213] and *SOX1* [214] expression levels were assessed in shHERVH and shscr samples, with wild type H9 hESC serving as a negative control and cells undergoing neuronal differentiation as a positive control (Figure 11).

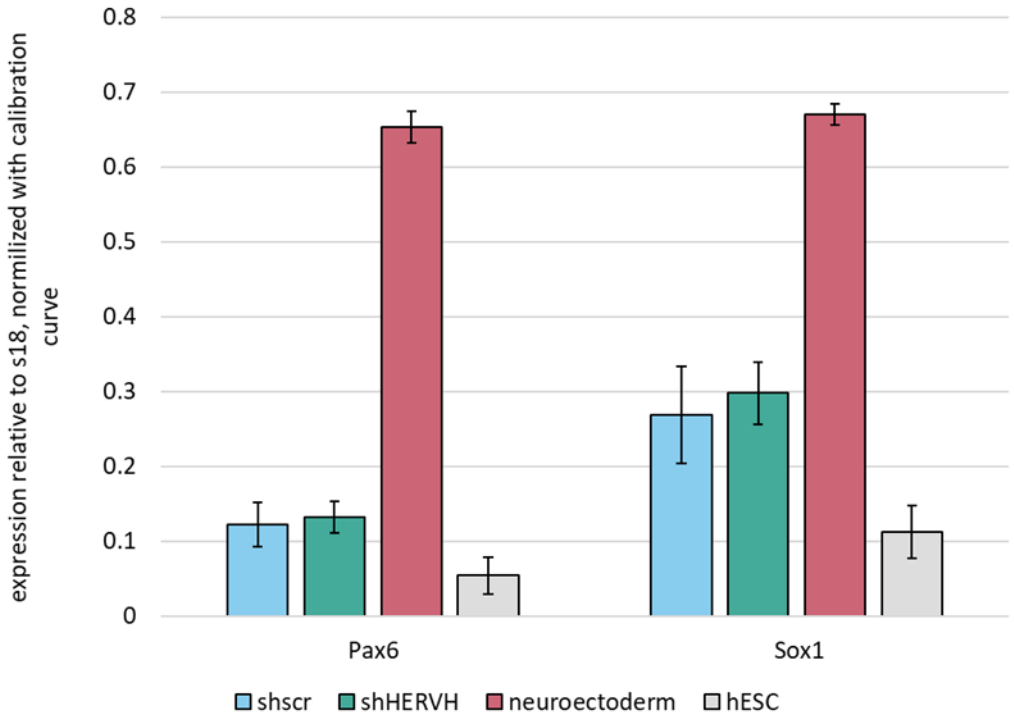


Figure 11. *PAX6* and *SOX1* neuroectoderm markers are uniformly expressed between shscr and shHERVH samples. *PAX6* and *SOX1* expression is normalized to *S18* with the calibration curve. The mean of three independent replicates is shown, except for hESC in two replicates; bar represents standard error of the mean (SEM).

The other possible differentiation direction after HERVH depletion might be mesodermal cell fate. Therefore, *BMP4* [215, 216] and *LMO2* [217, 218] were used as markers for early mesoderm. Wild type H9 hESC samples were used as a negative control and mesodermal progenitor cells as a positive control (Figure 12).

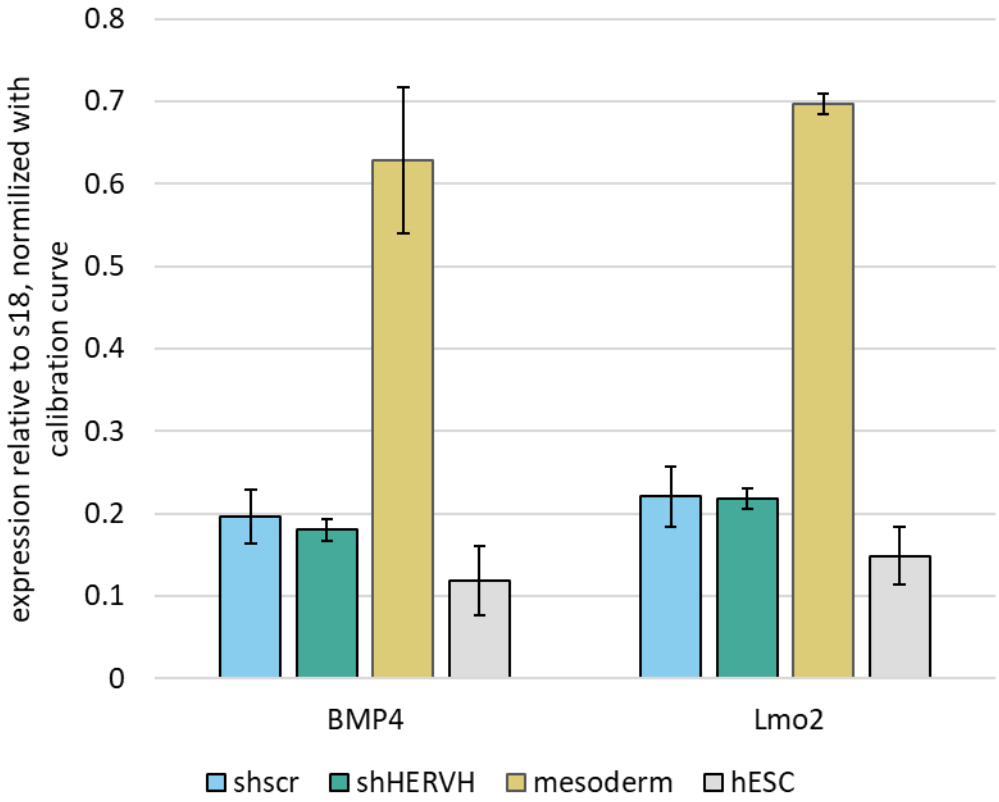


Figure 12. *BMP4* and *LMO2* early mesoderm markers are uniformly expressed between shscr and shHERVH samples. *BMP4* and *LMO2* expression was normalized to *S18* with the calibration curve. The mean of three independent replicates is shown, except for hESC in two replicates; bar represents standard error of the mean (SEM).

NANOG depletion is also known to cause upregulation of extraembryonic endoderm-associated genes and trophectoderm-associated genes in hESCs [219]. Expression of primitive endoderm markers *AFP* and *GATA6* (Figure 13) together with trophectoderm *ELF5* and *hCGA* expressed genes (Figure 14) was validated in HERVH knock-down cells vs control. H9 hESCs were used as a negative control. A clone of H1 hESCs, undergoing beta cells differentiation was used as a positive control since cells pass an endometrial differentiation stage during the intermediate stages of the protocol [220] (Figure 13). As a positive control for trophectoderm markers, an established BeWo trophoblast cell line was implicated (Figure 14).

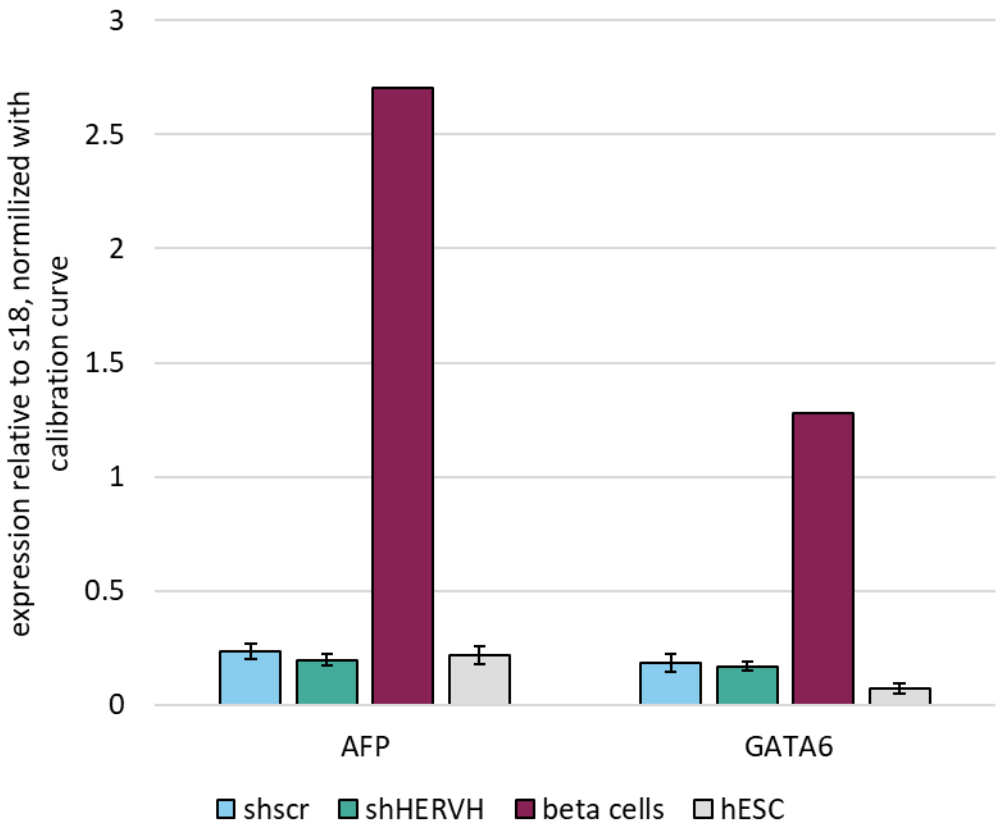


Figure 13. *AFP* and *GATA6* primitive endoderm markers are uniformly expressed between shscr and shHERVH samples. *AFP* and *GATA6* expression is normalized to *S18* with the calibration curve. The mean of three independent replicates is shown, except for hESC in two replicates and beta cells in one replicate; bar represents standard error of the mean (SEM).

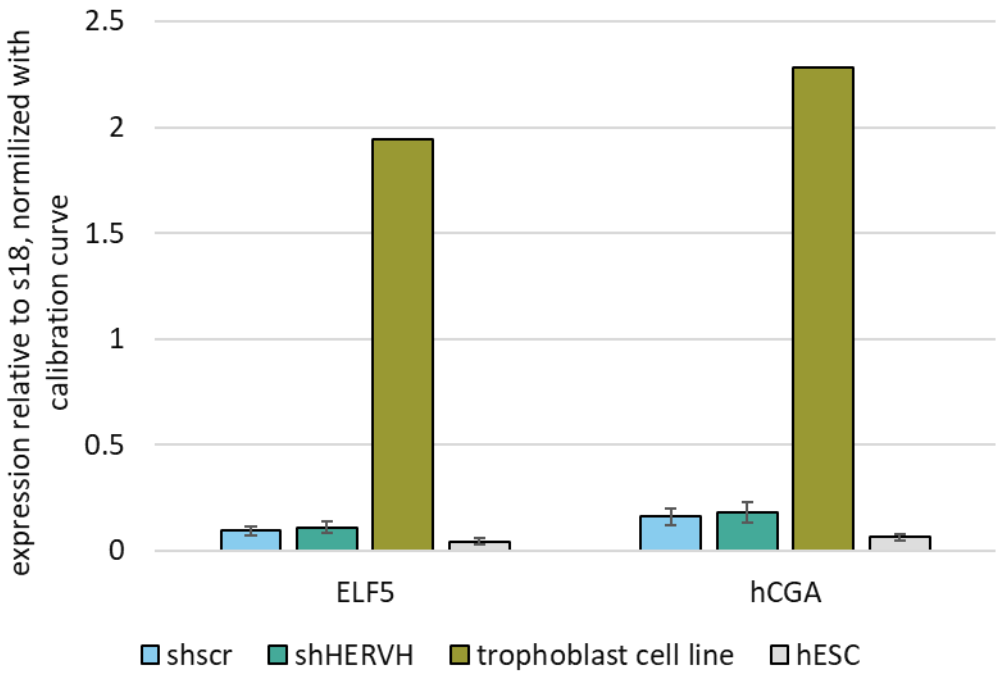


Figure 14. *ELF5* and *hCGA* trophectoderm markers are uniformly expressed between shscr and shHERVH samples. *ELF5* and *hCGA* expression is normalized to *S18* with a calibration curve. The mean of three independent replicates is shown, except for hESC in two replicates and trophoblast cell line in one replicate; bar represents standard error of the mean (SEM).

Based on these results, HERVH depleted cells show reduced levels of *NANOG*, but no changes of any lineage specific genes. Thus, the observed elevated levels of L1 transcripts and other phylogenetically young REs could not be explained by chromatin remodeling or global transcription changes during differentiation, more likely HERVH plays a direct role in controlling retrotransposition. To validate that, I performed several types of transposition assays for the L1 element and a high-throughput analysis to detect *de novo* integrations in HERVH depleted cells.

3.2. Reporter-based L1 transposition

Changes in the amount of L1, SVA and Alu transcripts, observed in HERVH knock-down RNA-seq data [7], could be explained by differential expression or active jumping, since retroelements use RNA intermediates during their transposition. L1 is the only one

autonomous retrotransposon. Detected transposition of L1 would mean mobilization of SVAs [221] and Alus [139].

3.2.1. EGFP-based L1 transposition assay

To decipher if HERVH depletion influences L1 transposition, a previously published L1-EGFP reporter [222] was employed (Figure 15). In this reporter, full length sequence of L1 including 5'-UTR is cloned into the plasmid. *ORF1* and *ORF2* encode nucleic acid-binding proteins and a reverse transcriptase with endonuclease activity, respectively [132]. Closer to the 3' end of the cassette, an enhanced green fluorescence protein (*EGFP*) sequence is located. The sequence is positioned in a reverse orientation, and it contains its own strong promoter and an intron, that disrupts activity of *EGFP* transcribed from the original plasmid. After transcription from the plasmid, splicing and integration to the genome, *EGFP* could be transcribed as an intact product and its fluorescent signal can be detected with FACS or similar method.

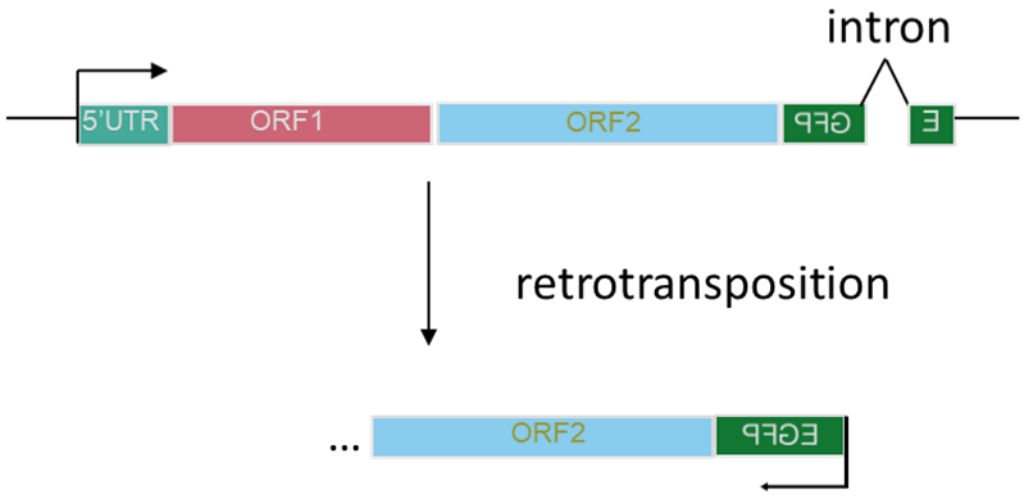


Figure 15. L1-EGFP transposition reporter structure and activity mechanism, adapted from [222]. The reporter contains full-length L1 elements with 5'-UTR serving as a native promoter, *ORF1* and *ORF2* encoding proteins for L1 maturation and retrotransposition. The *EGFP* sequence is encoded downstream to L1 in a reverse orientation, separated with an intron. During transposition, the intron is spliced out and EGFP expression could be driven by its own promoter.

H9 hESC were co-transfected with shHERVH or shscr control constructs and the L1-EGFP reporter plasmid. After five days, cells were collected in a single cell state and the EGFP signal was analyzed with FACS. The selection strategy for the live cells' population and true EGFP positive signal is shown on one replicate of the flow cytometry analysis (Figure 16). Based on the intensity of the forward and side scatters (FSC-A and SSC-A, respectively), a population of alive hESC was selected. Then, only the live population was analyzed for EGFP signal, with V-450 (blue fluorescence) being used to sort out dead cells, which usually are auto fluorescent in the full length of the spectra [223].

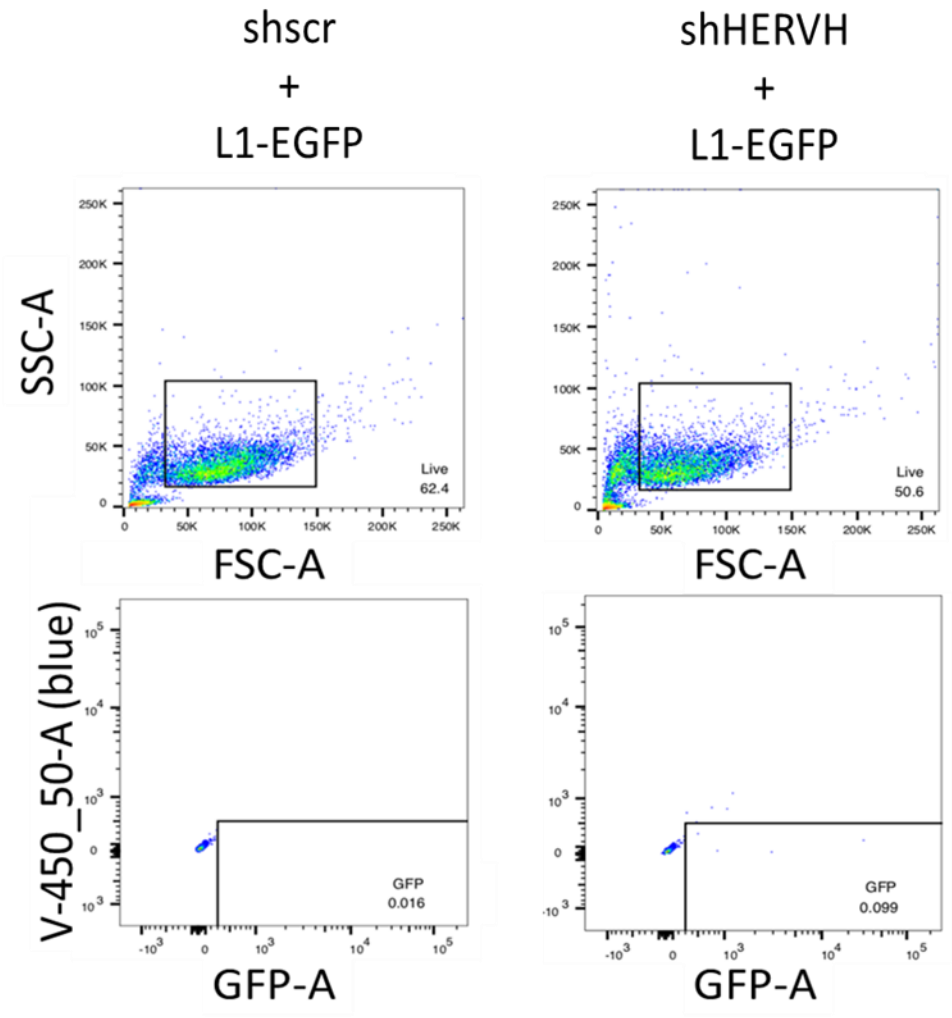


Figure 16. The flow cytometry analysis of L1 transposition in HERVH depleted cells. Marked with black square, top panel: live cells from H9 hESC transfected with L1-EGFP reporter and shscr/shHERVH constructs were selected via SSC-A/FSC-A ratio. Black square, bottom panel: EGFP fluorescence was detected (GFP-A), excluding apoptotic

autofluorescence with the blue part of the spectra (V-450_50-A). 0.099% of life cells are positive for green fluorescence.

To detect if the L1 transposition difference between HERVH depleted and control cells is significant, the experiment was performed in three independent replicates. Below the percent and the number of EGFP-positive cells, together with the life population for each of the three independent replicates is shown (Table 4).

Table 4. L1-EGFP transposition, analyzed by FACS in three replicates.

Sample/gating name	Statistic	Cells number
shHERVH L1-EGFP rep3		10000
shHERVH L1-EGFP rep3/Live	66.8	6685
shHERVH L1-EGFP rep3/Live/GFP	0.19	13
Geometric Mean : GFP-A = 371.155456543	371	
Robust CV : GFP-A = 24.330793381	24.3	
shscr L1-EGFP rep3		10000
shscr L1-EGFP rep3/Live	75.5	7555
shscr L1-EGFP rep3/Live/GFP	0.093	7
Geometric Mean : GFP-A = 320.457794189	320	
Robust CV : GFP-A = 6.863583565	6.86	
shHERVH L1-EGFP rep2		10000
shHERVH L1-EGFP rep2/Live	51.7	5167
shHERVH L1-EGFP rep2/Live/GFP	1.72	89
Geometric Mean : GFP-A = 415.744873047	416	
Robust CV : GFP-A = 20.498947144	20.5	
shscr L1-EGFP rep2		10000
shscr L1-EGFP rep2/Live	48.8	4876
shscr L1-EGFP rep2/Live/GFP	0.84	41
Geometric Mean : GFP-A = 367.650299072	368	
Robust CV : GFP-A = 13.518325806	13.5	
shHERVH L1-EGFP rep1		10000

shHERVH L1-EGFP rep1/Live	50.6	5057
shHERVH L1-EGFP rep1/Live/GFP	0.099	5
Geometric Mean : GFP-A = 1331.624389648	1332	
Robust CV : GFP-A = 0.0	0	
shscr L1-EGFP rep1		10000
shscr L1-EGFP rep1/Live	62.4	6240
shscr L1-EGFP rep1/Live/GFP	0.016	1
Geometric Mean : GFP-A = 394.968048096	395	
Robust CV : GFP-A	n/a	

Description: Robust CV— robust coefficient of variation, Equals 100 * 1/2(Intensity[at 84.13 percentile] – Intensity [at 15.87 percentile]) / Median. The robust CV is not as skewed by outlying values as the CV.

Each sample's percent of EGFP-positive cells was normalized to the mean between two samples inside every replicate, to account for batch-to-batch variation (Figure 17).

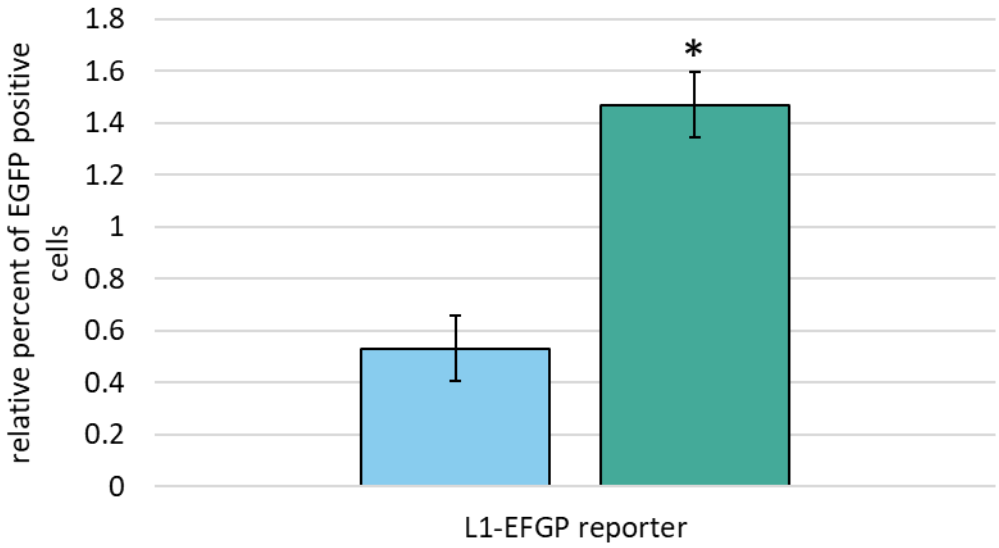


Figure 17. L1 transposition is elevated after HERVH knock-down in H9 hESC. L1 transposition measured with EGFP reporter, shown in percentage of EGFP positive cells. The mean of three independent replicates is shown, each replicate is normalized to the batch variation, bars represent standard error of the mean (SEM), * – statistical validation with t-test, showing significantly different number of EGFP positive cells between shscr and shHERVH samples (p=0.0062).

The retrotransposition experiment shows that the presence of HERVH transcripts is crucial for inhibition of L1 activity from EGFP reporter in hESC. Even though the possibility of "leaking" fluorescence signal of the L1 reporter is quite low, due to the cassette structure (intron presence and reverse orientation), a negative control is needed to exclude false positive results. In addition, sensitivity of the assay is relatively low, reflected in only up to 0.1% of cells being EGFP-positive. To increase the sensitivity of the assay, luciferase-based reporter was used further with an additional negative control for the transposition.

3.2.2. Luciferase-based L1 transposition assay

Luciferase L1 transposition reporter (CAG-L1) has been developed as a more sensitive alternative to EGFP- or antibiotic resistance-based transposition reporters [224]. The reporter was designed for a shorter assay time frame and a broader detection range that allows larger signal acceleration after five days of transfection. Additionally, the assay has a superlative signal-to-noise ratio and the full-length L1 is driven by a strong CAG promoter [224, 225]. A negative control, JM111, that encodes an inactive L1 due to mutations in *ORF1*, was available as well to serve as an additional control for false positive values of transposition activity.

The principle of activity is similar to the EGFP-based reporter, with the firefly luciferase sequence, instead of *EGFP*, driven by its own promoter and being separated by an intron (Figure 18). Only after a full transposition round i.e., transcription, splicing, reverse transcription, and integration of the cassette, could a positive signal be detected. *Renilla* luciferase, used to normalize for transfection efficiency, is encoded in the same plasmid, and is expressed via a strong constitutive promoter (Figure 18).

Figure 18. CAG-L1 transposition reporter structure and activity mechanism, adapted from [224]. The reporter contains full length L1 elements with CAG strong constitutive promoter (CAG, black arrow), *ORF1* and *ORF2* encoding proteins for L1 maturation and retrotransposition. A firefly luciferase sequence (FLuc) is encoded after L1 in reverse orientation, separated with an intron. After transposition, the intron is spliced out and the luciferase expression could be driven by its own promoter (black arrow). *Renilla*

luciferase (RLuc), used to normalize an amount of transfected plasmid, is located on the same plasmid and driven by its own constitutive promoter (black arrow).

H9 hESCs were co-transfected with shHERVH or shscr control constructs and either a CAG-L1 transposition reporter plasmid or a JM111 inactive control. After five days, cells were collected in a single-cell state, and a luciferase assay was performed according to the commercial protocol from Promega in 96 well plates. To reduce the handling error, luminescence in at least two and up to four wells was simultaneously measured (technical replicates). Firefly and *Renilla* values were detected from the same well for the later Firefly-to-Renilla signal normalization. Experiments were performed in four replicates, except for CAG-L1 reporter in two. Each luminescence ratio was normalized to the mean of values inside the replicate, to account for batch-to-batch variation (Figure 19).

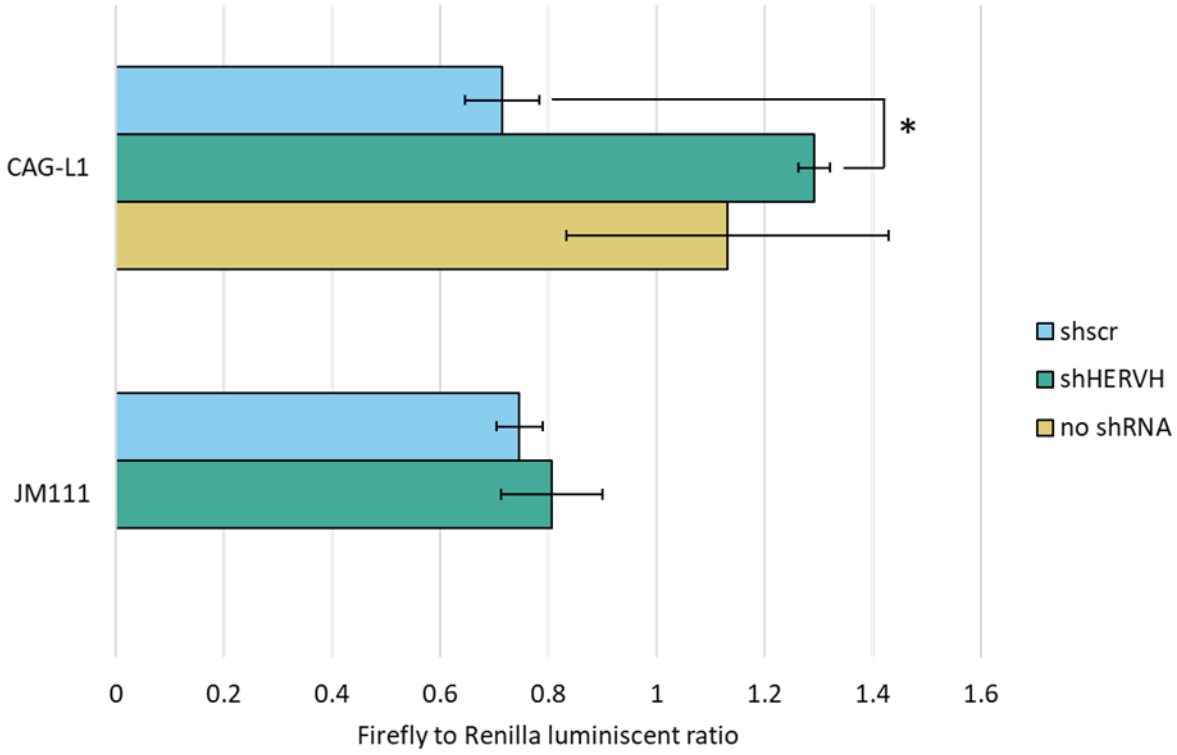


Figure 19. L1 transposition is elevated after HERVH knock-down in H9 hESC. L1 transposition measured with luciferase reporter (CAG-L1), shown as a ratio of Firefly to *Renilla* luminescent signal. JM111 is a transposition impaired reporter and contains both Firefly and *Renilla* luciferases, serving as a negative control. The mean of four independent replicates is shown, except for CAG-L1 only transfection in two. Each replicate is normalized to the batch variation, bar represents standard error of the

mean (SEM), * – statistical validation with t-test, showing the significantly different Firefly to *Renilla* signal ratio between shscr and shHERVH samples (p=0.0002).

These results point to the fact that HERVH seems to be crucial for controlling L1 transposition.

Based on the results of L1 transposition assays, performed with EGFP and luciferase reporters, the first aim of the study to assess L1 transposition in HERVH depleted background, was achieved.

But due to general low values of Firefly luciferase, EGFP signal, and an effect of shscr construct on the CAG-L1 reporter signal, *de novo* L1, Alu and SVA integrations were detected in HERVH depleted cells with a high-throughput method, corresponding to the second aim of the study.

3.3. High-throughput transposition detection

To reach sufficient coverage of new retrotransposons integrations for a high-throughput analysis, HERVH-depleted cells must be cultured for enough time to undergo clonal expansion of cells, where a retroelement has jumped. Based on previous experience, to gain the detectable number of cells with integrations, our collaborator, Dr. Garcia-Perez advised to culture HERVH depleted hESC for at least 10 passages. Previously established transient HERVH knock-down disappears after 7 days of culturing (data not shown) and on average 10 passages of hESC would take up to 2 months [1]. Therefore, I had to establish a stable cell line, constitutively expressing shHERVH. The pipeline of inquired experiments is shown below (Figure 20).

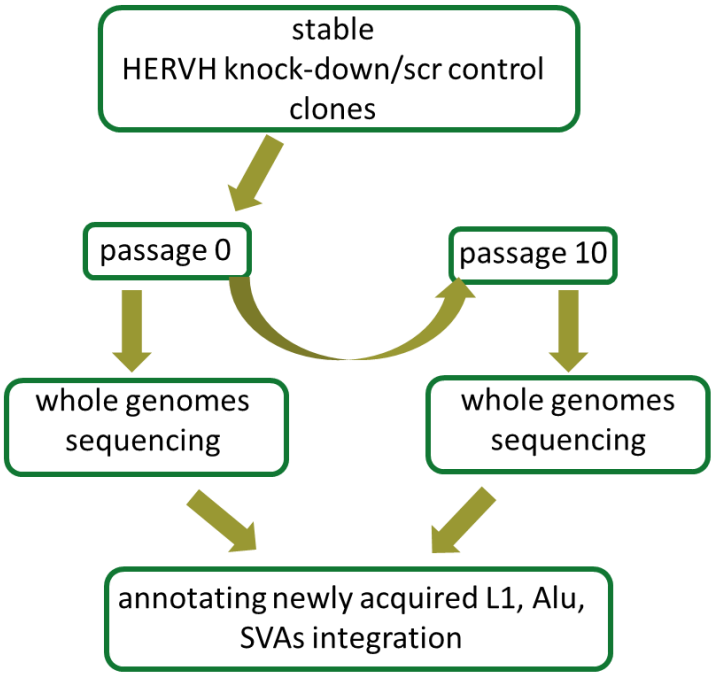


Figure 20. *de novo* integrations detection strategy. Established stable HERVH knock-down clones or control clones, cultured for 10 passages and collected in two time points: early (passage 0) and late (passage 10) for genomic DNA isolation. Whole genome sequencing (WGS) will be performed on all the samples and newly acquired integrations of phylogenetically young REs (L1s, Alus and SVAs) will be annotated from the high-throughput data.

3.3.1. An attempt to generate stable HERVH knock-down

The previously used shRNA vector contained inverted repeats of the piggyBac transposon, which makes it eligible for a genome integration via DNA transposition [226, 227]. The shHERVH and shscr also carry neomycin resistance that allows selection for an integrated cassette in cell culture. H9 hESCs were transfected with shscr or shHERVH plasmids and a vector carrying piggyBac transposase to mobilize shRNA expressing cassettes upon transfection and integrate them in the genome. Three days after the transfection, G418 selection started. G418 is an analog of neomycin sulfate and could be used to select neomycin resistant clones [228]. After ten days G418 was withdrawn, shHERVH – or shscr – transfected cell lines were cultured for three more days. Part of the cultures were collected during passaging to validate HERVH transcripts levels in the established cell lines. HERVHgag expression was normalized to *S18* with calibration curve and reflected a 20% HERVH depletion in the bulk cell population (Figure 21).

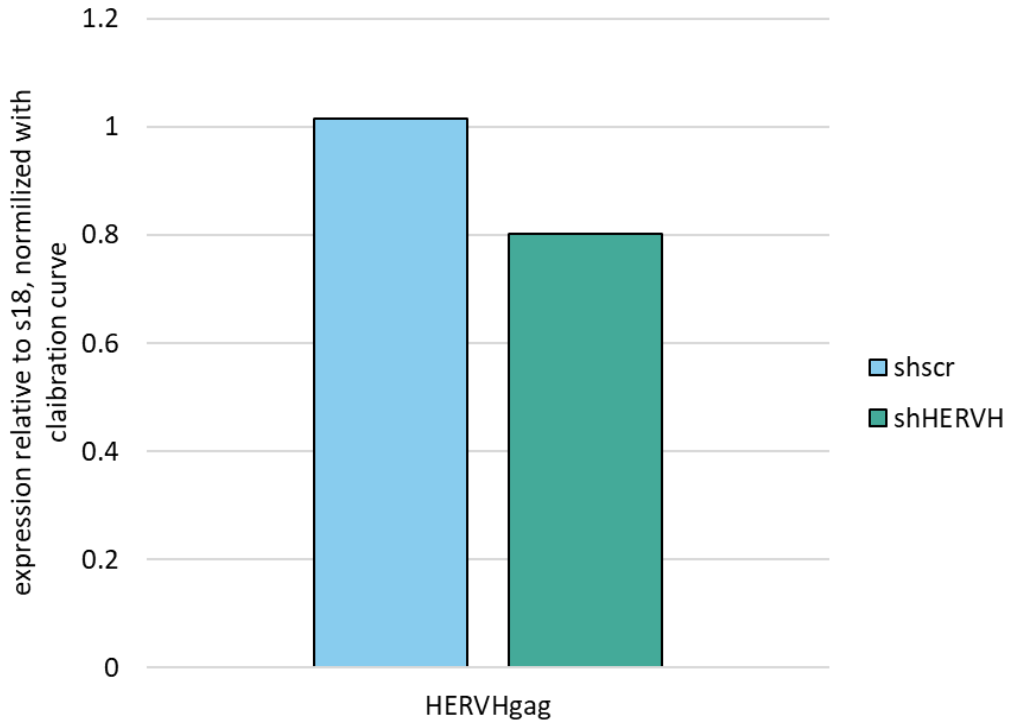


Figure 21. HERVH is depleted in the stable H9 cell line, expressing shHERVH. HERVHgag expression is normalized to *S18* with a calibration curve, one replicate is shown.

To increase the efficiency of HERVH knock-down and create a more homogenous population for higher probability of the *de novo* integrations detection, colonies were selected to establish a clonal HERVH depleted cell line.

Cells were split in 1:10 ratio in clumps and the next day, five colonies from shHERVH, and five from shscr cell lines were manually selected. From that point all the experiments were performed in the clonal cell population. Clones were cultured to expand from 96 well to 6 well plates, and afterwards were passaged three times to collect replicates for qPCR validations of HERVH depletion. HERVHgag expression was normalized to s18 with a calibration curve (Figure 22).

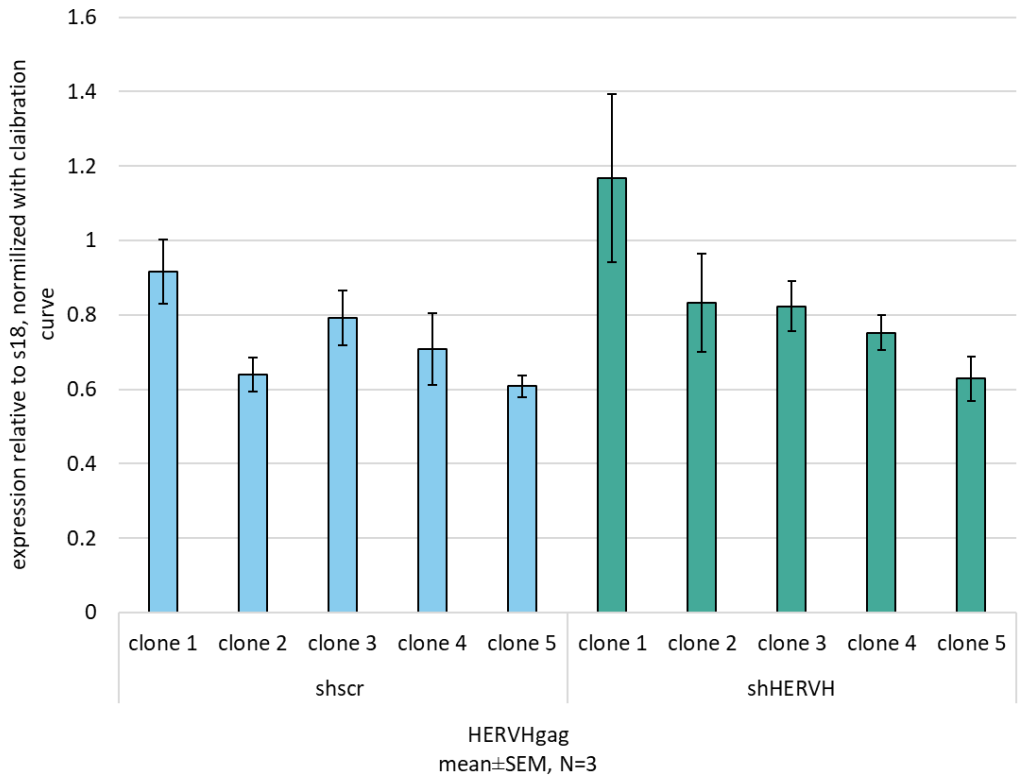


Figure 22. HERVH expression does not change in shHERVH stable clones compared with shscr controls. HERVHgag expression is normalized to s18 with a calibration curve. The mean of 3 independent replicates is shown, bars represent standard error of the mean (SEM).

Surprisingly, HERVH expression did not change in any of the shHERVH-expressing H9 hESC clones. A derivation of the stable cell lines had to be repeated. To avoid the absence of the knock-down effect, cells were cultured continuously on G418 after transfection and clones were tested for the genomic integrations of shRNA expressing cassettes.

3.3.2. Stable HERVH knock-down generation

H9 hESC were transfected with shHERVH/shscr and piggyBac transposase in the same ratio as before. G418 selection started 3 days after and, from that step on, the growing media always contained the selective antibiotic for both bulk and clonal cell maintenance. Bulk cell lines were tested for HERVH expression. HERVHgag was normalized to *S18* reference gene via the $2^{\wedge(-\Delta Ct)}$ method, showing around 40% reduction of HERVH expression in the shHERVH transfected cell line (Figure 23).

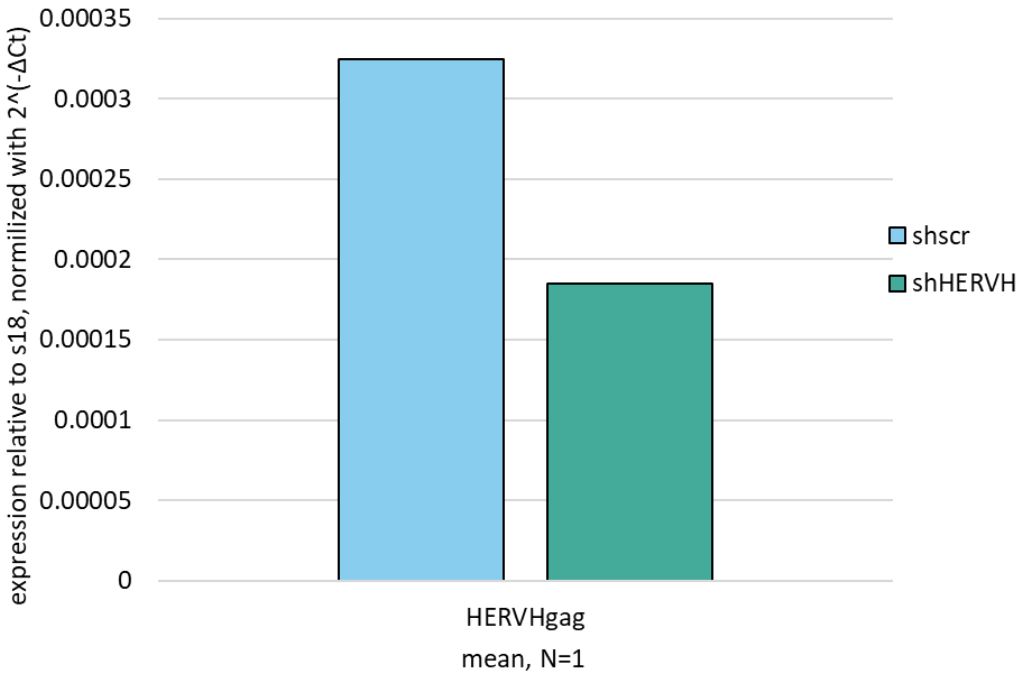


Figure 23. HERVH is depleted in the stable H9 cell line, expressing shHERVH. HERVHgag expression is normalized to s18 with $2^{\wedge(-\Delta Ct)}$ method, one replicate is shown.

Five colonies were selected from shHERVH and five from shscr cell lines, expanded from 96 to 6 well plates, and genomic DNA was isolated. To distinguish PCR products, amplified from plasmid or genomic DNA, DpnI cleavage method is usually used [229]. Due to overlapping CpG methylation, digestion of genomic DNA is blocked, and therefore DpnI degrades only plasmid DNA. The product, amplified in the consequent reaction, would then be synthesized only from the genome.

First, DpnI cleavage protocol was tested on the shscr plasmid. The plasmid was digested according to the NEB commercial protocol, purified from the solution, and used as a template for the PCR amplification with a pair of primers, annealing to the inverted piggyBac repeats of the shRNA cassette (section 2.1.13, Figure 7). The DNA products were analyzed with agarose gel electrophoresis (Figure 24).

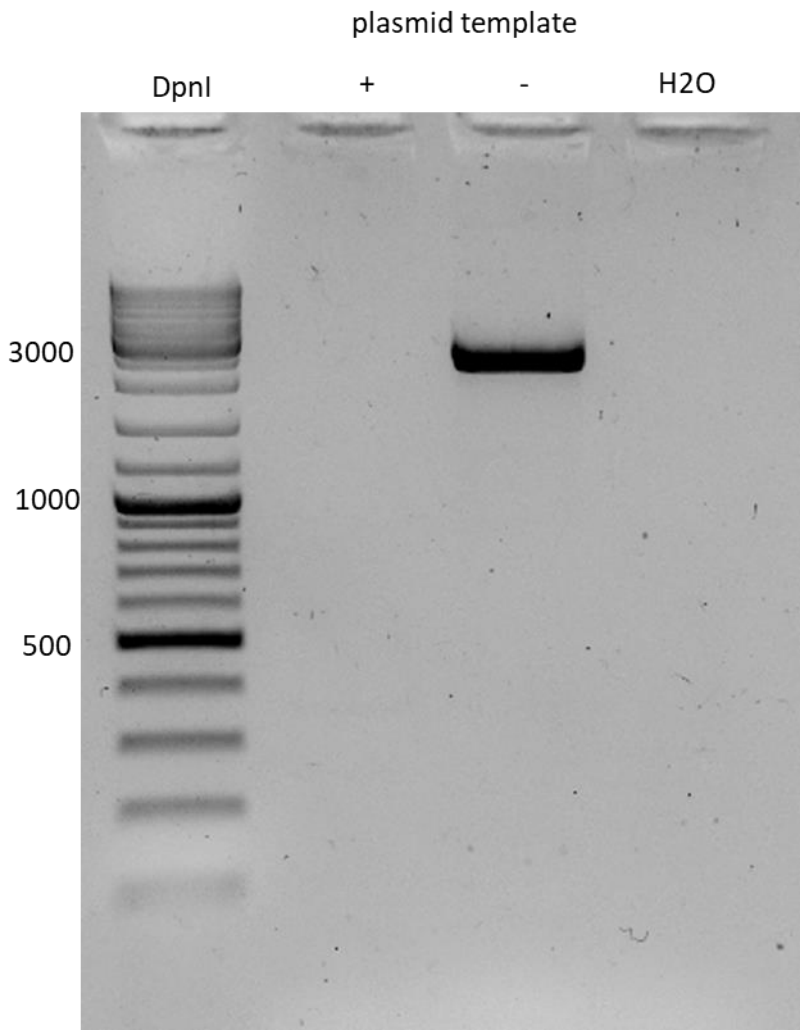


Figure 24. shRNA cassette amplification after the DpnI digestion of shscr plasmid. DpnI treated (DpnI +) or control (DpnI -) samples served as a template. Water (H_2O) was used as a negative control. DNA marker bands are marked according to their size in bp.

DpnI did successfully digest the shscr plasmid, with no detectable PCR product for the shRNA cassette amplification. Following that, the genomic DNA from shHERVH and shsrc transfected clones was treated with DpnI enzyme and used as a template for the subsequent PCR amplification of the possibly integrated shRNA cassette. DNA products were analyzed via an agarose gel (Figure 25).

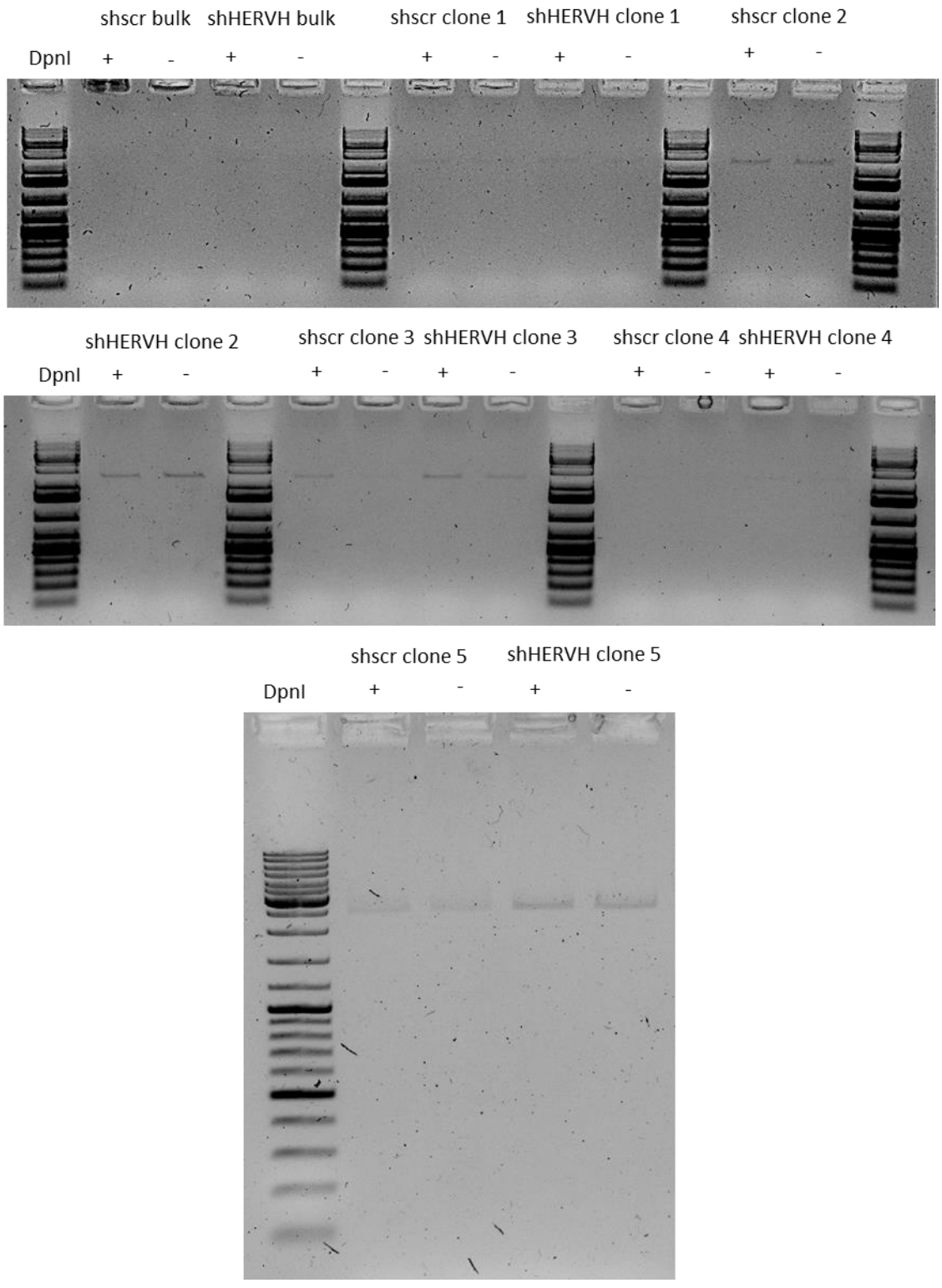


Figure 25. shRNA cassette amplification after DpnI digestion of genomic DNA from shscr and shHERVH transfected clones. DpnI treated (DpnI +) or control (DpnI -) samples served as a template. DNA marker is identical to figure 24.

The DpnI digestion experiment with the following shRNA cassette amplification showed that all shscr and shHERVH transfected clones have the construct integrated in the genome.

To collect replicates for qPCR validation of HERVH expression, clones were cultured for three passages. Unfortunately, due to a technical incubator malfunction, the death of three scr control and two shHERVH transfected clones, as well as damage of the remaining clones occurred. To recover, surviving clones had to be passaged several times, which added time constraints to the experiment. Due to this, qPCR validation of HERVHgag expression was performed on three independent replicates only for shscr clone 1, shHERVH clones 2 and 3 had two replicates and shscr clone 2 and shHERVH clone 1 – only one replicate (Figure 26). HERVHgag expression was normalized to *S18* using the $2^{\wedge(-\Delta Ct)}$ method (Figure 26).

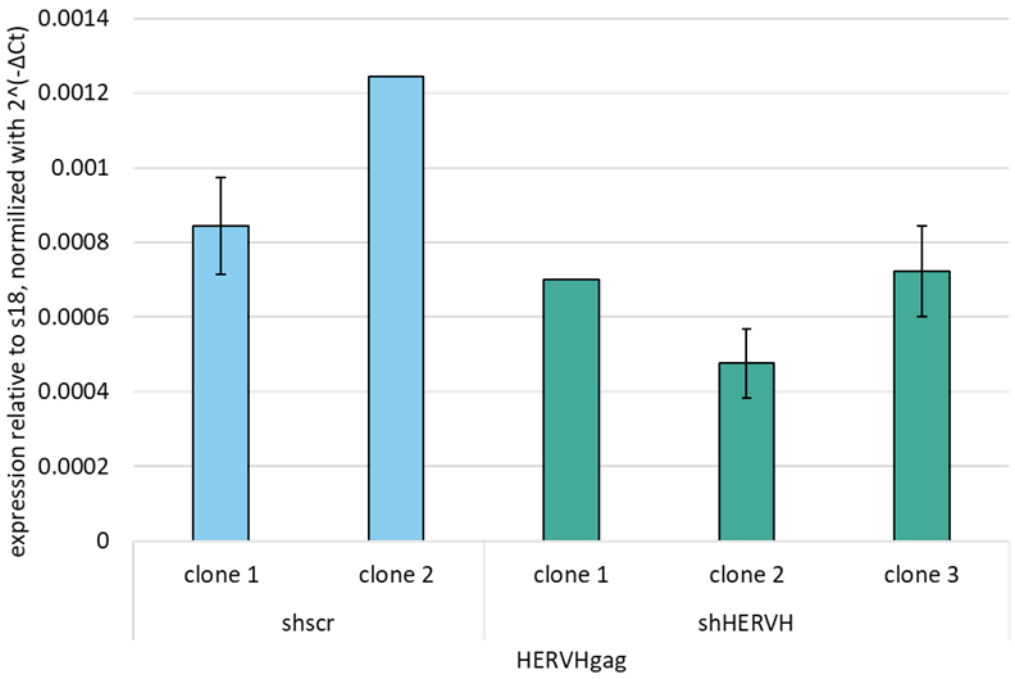


Figure 26. HERVH expression is reduced in stable-transfected shHERVH clones. HERVHgag expression is normalized to *S18* with $2^{\wedge(-\Delta Ct)}$. The mean of three independent replicates for scr clone 1, two – shHERVH clone 2 and 3, shscr clone 2, one – shHERVH clone 1 is shown, bars represent standard error of the mean (SEM), where applicable.

HERVH was depleted to different extents in all the stable shHERVH transfected clones, in comparison with controls. After validating shRNA cassette integrations and levels of HERVH expression, I would refer to shscr transfected stable clones as scr control clones and shHERVH

transfected clones as HERVH knock-down clones. All five clones (2 scr control and 3 HERVH knock-down) were cultured for ten passages, collecting both passage 0 (early time point) and passage 10 (late time point) samples for further genomic DNA isolation. High quality genomic DNA, i.e. consisting of high molecular weight fragments, was isolated according to Qiagen commercial protocol and shipped to BGI Group for sequencing libraries preparation and WGS. From all samples, one scr control clone (clone #2) and two HERVH knock-down (clones #1 and #3) DNA samples passed the quality control, performed by BGI, and were further sequenced.

3.3.3. *de novo* integrations prediction in HERVH depleted cells

After WGS, the original data was analyzed by our collaborator, Alejandro Rubio-Roldan from the research group of Dr. Garcia-Perez, applying the TEBreak pipeline to map new L1, SVA and Alu insertions (https://github.com/adamewing/tebreak). This resulted in primary annotated 4370 possible integration loci for HERVH knock-down clone 1, 3019 for clone 3 and 1163 for scr control clone 2.

Further data was selected by our collaborator Dr. Manvendra Singh based on four main criteria: 1) integrations, present only in the late time point samples; 2) supported by both 5' and 3' sequencing reads coverage; 3) not reported in any previous study; 4) target site duplication in the range of 2-25 nucleotides. Integrations corresponding to these criteria were considered as predicted *de novo* integrations of L1, Alu and SVA elements. For the scr control clone, 0 integrations were detected, while HERVH knock-down clones 1 and 3 had 180 and 83 predicted integrations, respectively (Figure 27). Most of them were due to activity of Alu, with single active L1.

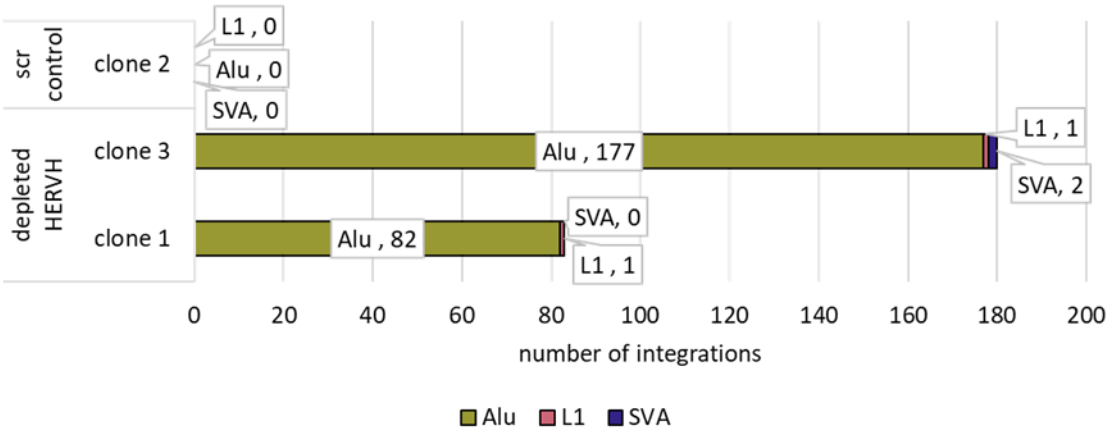


Figure 27. *de novo* integrations in HERVH depleted clones. Number of integrations per each HERVH knock-down or scramble control clone shown. Name of the element is followed by the number of annotated loci, separated by comma.

Predicted integrations were then separated into high and low confidence based on the number of discordant reads. These reads are the sequences which map to different positions of the genome, reflecting a possible new integration. The range of discordant read numbers fell between 4 and 61 for clone 1 and 4 and 58 for clone 3. Three loci were selected from the 0-25% quantile of each range as low confidence predicted integrations and 3 from 75-100% quantile as high confidence for each clone. ID, coordinates of reads, type of integrations and number of discordant reads, supporting each annotation are shown in the supplementary (Supplementary I).

3.3.4. PCR validation of *de novo* integrations

The predicted integrations most probably arose in a single cell of a clone as it was cultured for 10 passages, which would cause a mosaic integration pattern and the presence of a predicted transposon only in a portion of the cultured clonal cells. To increase the chances for detection of newly integrated Alus, 12 pairs of nested PCR primers and 12 pairs of main primers were designed (section 2.1.7, table 3).

Nested PCR was designed to amplify first a longer product, the large window of a predicted integration site, by this increasing the template concentration several dozens of times. Then the shorter product, with primers, annealing to the genomic location close to but outside of a predicted Alu integration, was amplified [230]. 12 nested pairs of long primers were tested on wild type H9 hESCs genomic DNA to detect a suitable condition through the

gradient of annealing temperature during the PCR reactions. Then, each reaction product, when corresponding to the predicted amplification sizes, was isolated from the gel, and used as a template for the next round of gradient PCRs to amplify and analyze short fragments (data not shown).

After detecting the functional conditions for each of the 24 primer pairs, genomic DNA from HERVH knock-down and scr control clones in both early and late time points were used as templates. If the predicted integration was real, the product should have been detectable only in the late passage sample of one of the knock-down clones. From all 12 integrations, only two resulted in a detectable product, corresponding to the mentioned criteria in DNA sample from HERVH knock-down clone 1 (Figure 28).

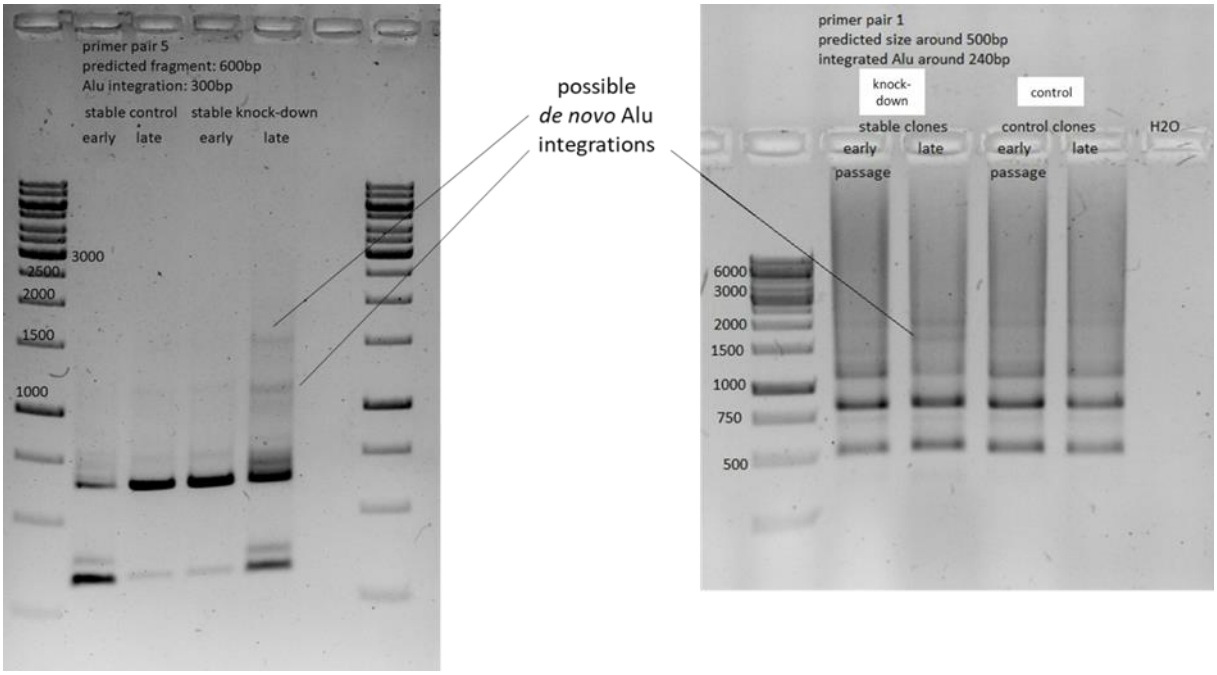


Figure 28. Detection of predicted *de novo* Alu integrations. PCR products from the second round of reactions are shown, marked – possible Alu sequences. The left panel has a product corresponding to the predicted size of the fragment with integrated Alu. The right panel has an additional product with higher molecular weight than expected. DNA marker's bands are marked according to their size in bp.

The product of primer pair 5 was close to the predicted molecular weight of the amplified genomic fragment with integrated Alu (Figure 28, left), but primer pair's 1 product had much higher molecular weight than expected (Figure 28, right). This could be explained

by the repetitive nature of a transposon-containing DNA fragment, which might complicate the movement of a polymerase along a sequence, thus resulting in low precision of the PCR.

To confirm the presence of Alu sequences in the amplified fragments, they were cut from the gel, isolated, and ligated to a pJET vector, designed to clone PCR products. The experiment was unsuccessful, as none of the selected bacteria clones had a corresponding genomic Alu containing sequence. The low efficiency of cloning could be explained by the repetitive nature of the sequences of interest as well.

To conclude, in this segment of the work I showed that HERVH plays a crucial role in the control of L1 transposition. HERVH depletion causes activation of not only autonomous L1 but also non-autonomous Alu and SVA, with *de novo* integrations that could be detected in hESC. The second aim to detect *de novo* retroelements integrations in HERVH-depleted human embryonic stem cells was achieved.

To address the last aim of the thesis, discovering the mechanism of HERVH-mediated retrotransposition control, HERVH loci sequences were analyzed to search for patterns, affiliated to REs transposition control.

3.4. HERVHlin discovery

3.4.1. Uneven expression of HERVH loci

926 loci HERVH loci were described in the human genome [104]. We had shown previously that, among them, 553 elements are highly-to-moderately expressed in hESC [6]. We decided to assess if, from the entire group of the expressed loci, there is a subset, that would be antagonistic to phylogenetically young retroelements. Based on the previously published data (section 1.3), our former postdoctoral researcher Dr. Manvendra Singh had analyzed the expression of the HERVH family in a locus-specific manner. He could describe several clusters of HERVH loci, differentially expressed in human pre-implantation development (data: [48, 192]) (Figure 29). ´

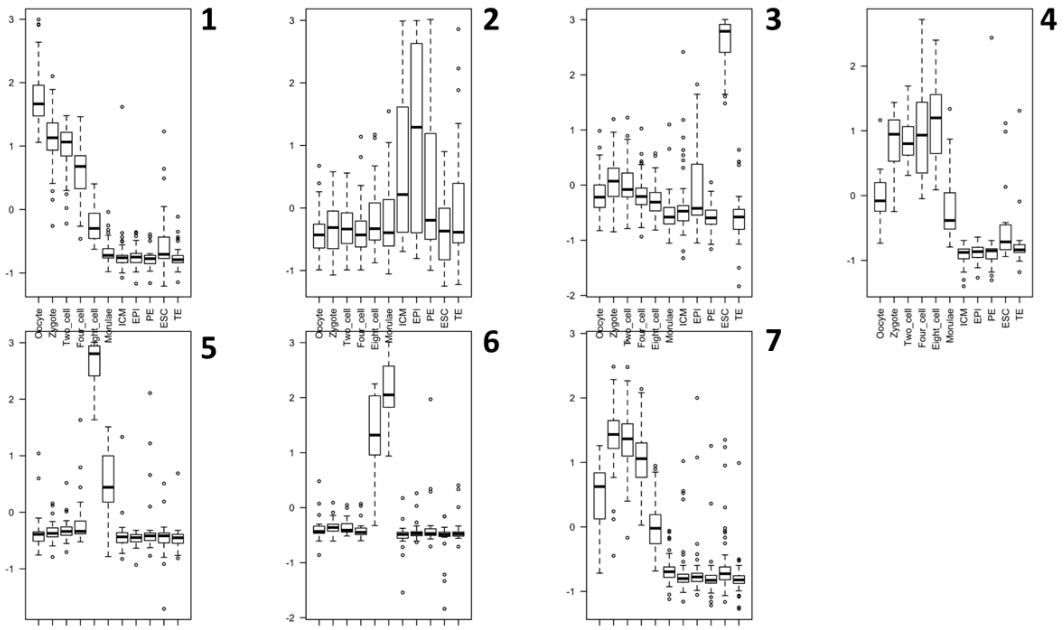


Figure 29. Clusters of HERVH loci based on the spatial-temporal expression in the human pre-implantation development. Single cell RNA-seq data was analyzed for the loci-specific HERVH expression, resulting in 7 distinct clusters. Y axis – FPKM, X axis – development stages: oocyte, zygote, two-cell, four-cell, eight-cell embryo, morula, inner cell mass (ICM), pluripotent epiblast (EPI), primitive endoderm (PE), embryonic stem cells (ESC), trophectoderm (TE). Bar plots showing median and 25-75% quantiles expression, error bars – 0 and 100%, dots – outliers. The figure was generated by Dr. Singh, data from [48, 192].

Based on previously detected antagonistic pattern of HERVH expression to SVAs, Alus and L1s in epiblast and hESC (section 1.3), clusters number 2 and 3 contain HERVH coordinates, which possibly participate in the REs control. Additionally, clusters 2 and 3 are the HERVH loci, expressed specifically in epiblast and cultured hESCs. On the other hand, HERVH from clusters 1, 4, 5, 6, and 7 might have a different biological function.

The antagonistic pattern of expression was observed during reprogramming as well (section 1.3). Here, analyzed by Dr. Singh, HERVH could be separated in 3 major clusters, one specific for a maturation stage of reprogramming, one for a stabilization stage and a cluster expressed through the whole reprogramming process.

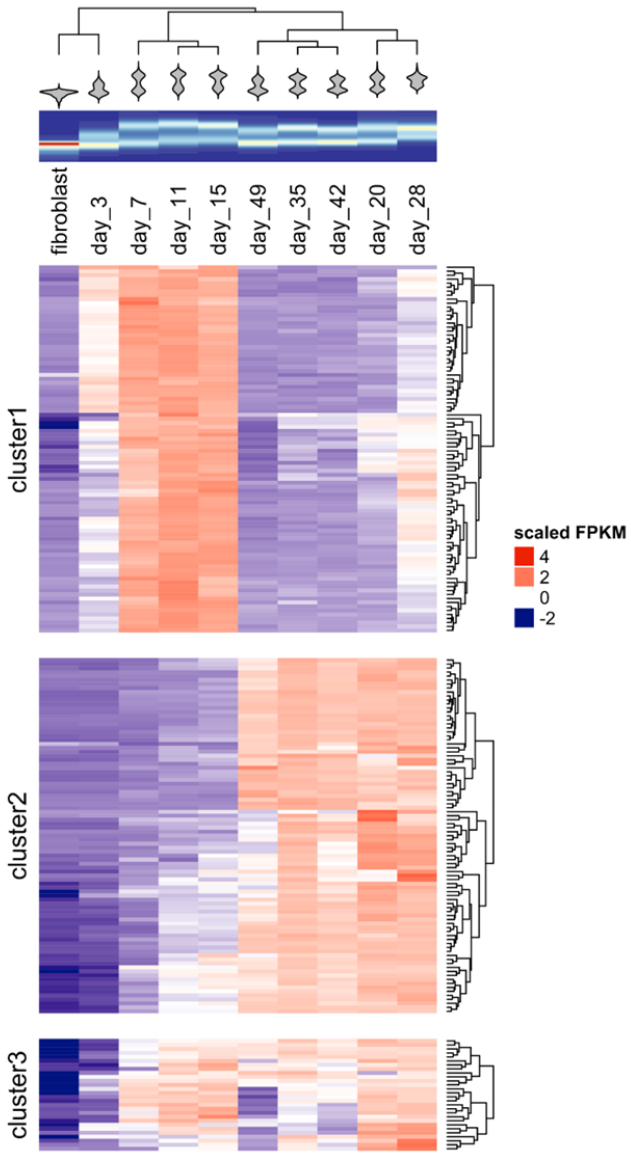


Figure 30. Clusters of HERVH loci based on the spatial-temporal expression during reprogramming of human fibroblasts to induced pluripotent stem cells. Bulk RNA-seq data was analyzed to group loci-specific HERVH expression, resulting in 3 distinct clusters. The heatmap shows the expression in FPKM for each HERVH locus. Data analyzed from fibroblast stage through 49 days of reprogramming to iPSC. The figure was generated by Dr. Singh, data from [193].

During reprogramming SVAs, Alus and L1s were expressed through the maturation stage, together with the cluster 1 HERVH and were reduced in the stabilization stage, when cluster 2 HERVH expression had risen (section 1.3).

From the development- and reprogramming-specific HERVH expression, some HERVH sequences are likely to represent antagonistic to young REs activity.

3.4.2. HERVH loci, antagonistic to young retroelements

The HERVH depletion experiment [7] suggests that antagonistic expression to young REs is not only a correlation but a possible direct effect of HERVH (section 1.3). Additionally to HERVH loci, expressed at specific stages of development and reprogramming, the third group of HERVH loci, knocked down in the work of Lu with co-authors [7], was used for the downstream analysis. RNA-seq from the publication was re-analyzed by Dr. Singh specifically focusing on transposable elements. HERVH loci were annotated and coordinates for loci depleted down to a certain level were reported.

First, the full genomic coordinates of HERVH loci downregulated in HERVH knock-down and loci from cluster 2 of reprogramming (Figure 30) were intersected with bedtools [231] to detect shared HERVH coordinates. Genomic coordinates of HERVH, present in clusters 4-7 in pre-implantation development (Figure 29) were appended with HERVH coordinates from the maturation stage of reprogramming, cluster 1 (Figure 30). That resulted in 266 loci, called HERVH control loci (HERVHcon). After that, all loci, downregulated in HERVH knock-down and expressed in the stabilization stage of reprogramming, but not present in the control dataset were reported (section 2.2.2.1). It amounted to 83 coordinates – the loci called HERVH antagonistic loci (HERVHant). After visual inspection of coordinates, it became clear that some of the annotated loci are much shorter than the rest. To depict the sequence length distribution in both datasets, the size of each HERVH locus was calculated and visualized (Figure 31).

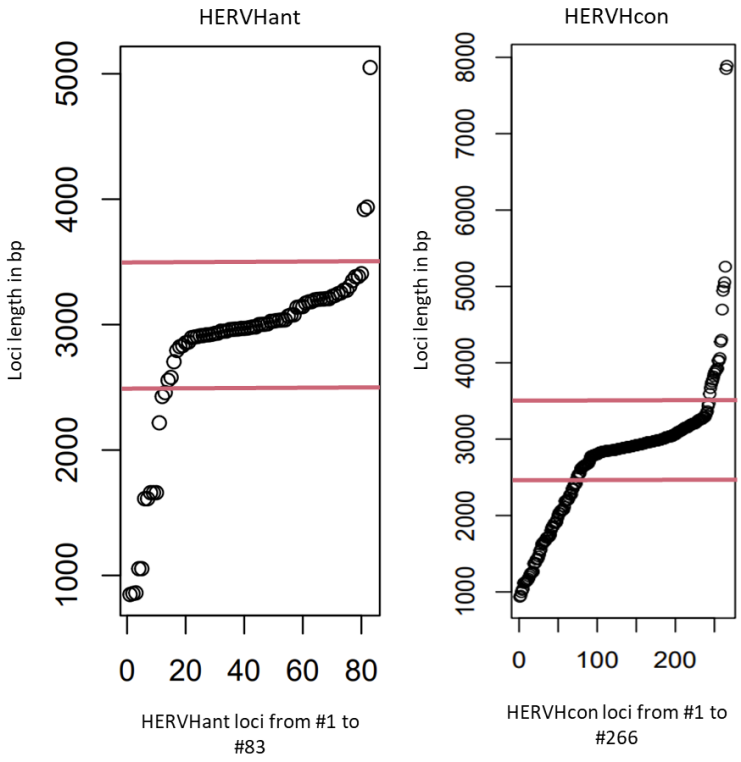


Figure 31. Length distribution of HERVHant and HERVHcon loci. Y axis – sequences length in bp, X axis – each HERVHant or HERVHcon locus. Red lines show the length range of selected loci.

Only few loci from HERVHant dataset are shorter than 2500bp and longer than 3500bp, therefore this range was used as a criterion to filter out HERVHant loci. For HERVHcon to be a suitable control in further analysis, only loci between 2.5kb-3.5kb were selected as well.

3.4.3 HERVH alignment and motif discovery

Next, HERVHant and HERVHcon sequences were retrieved from the hg19 genome annotation and aligned with the Muscle algorithm [203] (section 2.2.2.2). The alignment was manually edited, resulting in 63 HERVH antagonistic loci and 150 HERVH control loci. Edited alignment was visualized with the NCBI alignment viewer, the full alignment is shown in the supplementary (Supplementary II). All 63 HERVHant or 150 HERVHcon loci were also aligned independently to create consensus sequences, which were further analyzed with Muscle and visualized with the NCBI viewer (Figure 32).

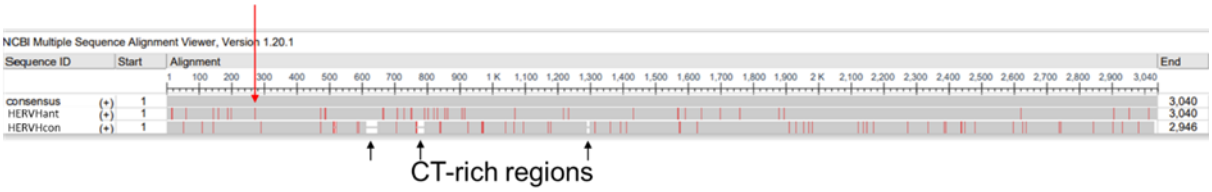


Figure 32. Visualization of HERVHant and HERVHcon consensuses alignment. Missing sequences in HERVHcon consensus are CT-rich. Mismatches are shown as short red lines along alignments. A motif of interest is marked with the red arrow.

CT-rich regions of alignment seem to be a repetitive part of the sequence, which theoretically could play a role in the formation of a secondary RNA structures of antagonistic HERVH transcripts. The sequences of these CT-rich regions were selected and aligned with Muscle, showing some degree of similarity (Figure 33).

```
1251-1300      ---------------YTCCTCCCCAGGCTG----------      15
601-650        CCCAATCSCTTATTTCCRYRCCCCAACCTCTTATCTCTGYR      41
751-800        -------TCTTATCTCTGYGCCCCAACCCCY----------      24
                              *****   *
```

Figure 33. Visualization CT-rich regions alignment, which are present exclusively in HERVHant sequences. Names of sequences are the positions at the HERVHant consensus. * showing conserved nucleotide residues.

After an additional visual inspection of the alignment, a 16bp motif GKAGAGACAAAGGAGA, around 270bp of the consensus was detected, varying between HERVHant and HERVHcon sequences (Figure 34). In the motif K stands for a G or T nucleotide in HERVHant sequences, whereas all HERVHcon have G. Full alignment around the motif is shown in the supplementary (supplementary III), below alignment of consensus is represented (Figure 34).

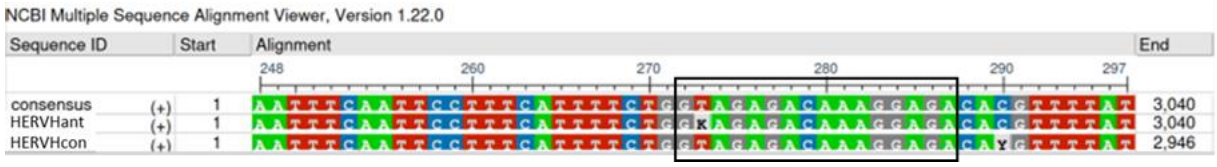


Figure 34. Visualization of HERVHant and HERVHcon consensuses alignment at the motif of interest. Black frame shows the motif, K stands for G or T nucleotide in the position.

The significance of the motif enrichment in HERVHant was validated with an online version of Analysis of Motif Enrichment (AME, [206]) from the MEME suit. HERVHant sequences were uploaded as primary sequences and HERVHcon as control sequences, additionally providing GGAGAGACAAAGGAGA as an input motif for the analysis. Fisher's exact test had shown significance of the motif enrichment in HERVHant dataset with p=0.00902.

Based on the previously described results, HERVH, expressed antagonistically to young retroelements has a significantly enriched GGAGAGACAAAGGAGA motif to compare with other control HERVH loci. The GGAGA sequence, present in the beginning and the end of the motif, has been described as a binding site for LIN28A protein, functional in hESC [33]. Therefore, I will further refer to this motif and similar sequences as *lin* motif, and *con* motif to its counterpart from HERVH control loci consensus. Following this, the genomic distribution of the *lin* motif was addressed.

3.5. *lin* motif and HERVH in human and apes genomes

3.5.1. *lin* motif in the human genome

To detect the positions of the *lin* motif in the genome, the motif sequence was aligned with Bowtie 1 to the hg19 assembly of the human genome. Bowtie 1 is an ultrafast, memory-efficient short read aligner, which is fast and sensitive for reads less than 50bp [204]. First, the hg19 genome was indexed with Bowtie 1 aligner (section 2.2.2.2). Next, to prefer mismatches in the nucleotides between LIN28A binding sites, the *lin* motif was saved as a sequence with attributed qualities for each nucleotide – fastq format, assigning highest quality score (I symbol, ASCII Code 73) to all nucleotides in both LIN28A binding sites (GGAGA) and lower scores (< symbol, ASCII Code 62) to nucleotides in between. The sequence was aligned to the indexed genome with 1 allowed mismatch in the 16bp seed region (the whole motif) and all possible hits to report. Then the reported targets were additionally filtered out for mismatches in GGAGA LIN28A binding site. A similar analysis was performed on the control motif. All reported variations of the *lin* and *con* motifs were used as an input to generate a sequence logo - a graphical representation of multiple sequence alignment [232] (Figure 35).

To detect which repeat families contain either *lin* or *con* motifs, the RepeatMasker [205] annotation of hg19 version of the human genome was uploaded from UCSC Table

browser [233]. Coordinates of *lin*/*con* were extracted from the output files of Bowtie 1 aligner and intersections with annotated repetitive sequences were detected (Figure 35, also section 2.2.3 and 2.2.4).

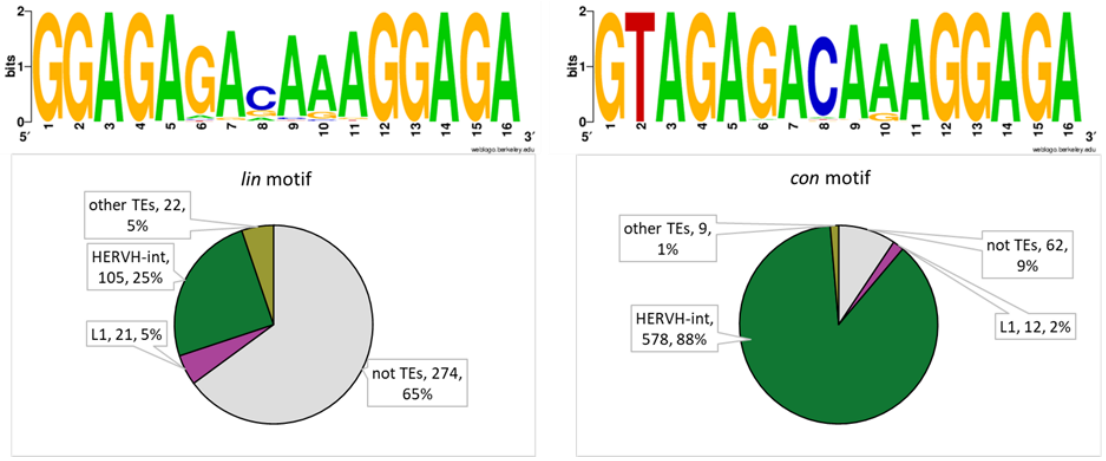


Figure 35. *lin* and *con* motif consensuses and genomic distribution of the motifs. Top: generated in WebLogo graphical representations of *lin* and *con* motif sequences in the human genome. Bottom: distribution of repetitive element types, which have *lin* or *con* motifs in their sequences. Labels: name of a repeat family, number of positions, percent from total amount of mapped reads to the human genome. TEs – transposable elements, HERVH-int – full length HERVH.

Most *lin* sequences are in non-repetitive areas of the human genome, but from all the analyzed transposons, the HERVH family has the highest number of elements (105) with *lin* motifs. The *con* sequence is present mostly in the HERVH family (578 HERVH elements). From this point forward, I would refer to HERVH elements containing the *lin* motif as HERVHlin, and HERVH elements containing the *con* motif as HERVHcon.

3.5.2. HERVHlin chromosomes distribution in the human genome

Next, the HERVHlin and HERVHcon distribution between human chromosomes was analyzed. Proportion of loci from either group was calculated for each chromosome and visualized (Figure 36).

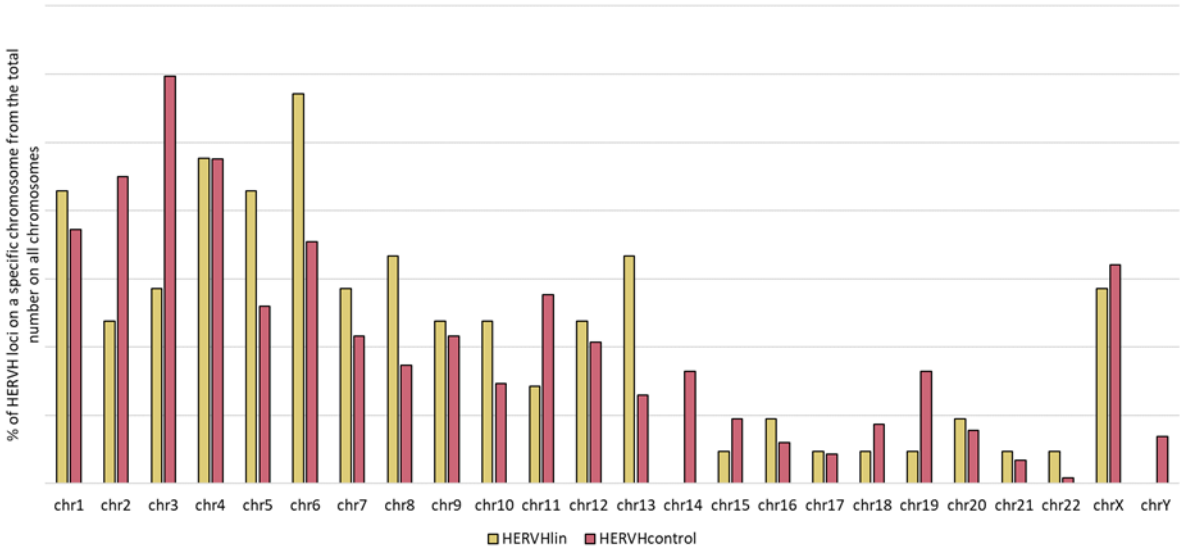


Figure 36. Distribution of HERVHlin and HERVHcon between human chromosomes. Percent of loci from the general number of HERVHlin or HERVHcon loci is shown.

No significant difference, validated with Fisher's exact test, was detected between chromosomal distribution of HERVHlin and HERVHcon.

3.5.3. HERVHlin in primate genomes

Most of HERVH elements had integrated in a genome of a common primate ancestor after New- and Old-World monkeys separation around 30 MYA [102]. Surprisingly, less active HERVH loci are more often present in other primates, whereas the ones active in human pluripotent cells are absent or degraded [108]. An evolutionary age could be an indicator of activity for a HERVH element. Therefore, the presence of the *lin* motif and its localization in HERVH sequences was analyzed in apes, rhesus macaque (Old-World monkey) and marmoset (New-World monkey).

Bowtie 1 was used to align *lin* or *con* motif sequences with quality scores to chimpanzee, gorilla, orangutan, gibbon, rhesus macaque and marmoset genomes, with 1 allowed mismatch in the 16bp seed region and all locations to report (section 2.2.5). The targets were also filtered out for mismatches in the LIN28A binding site. To detect if reported sequences were included in HERVH elements, available versions of RepeatMasker annotation for each genome were intersected with corresponding bowtie 1 output coordinates (Figure 37).

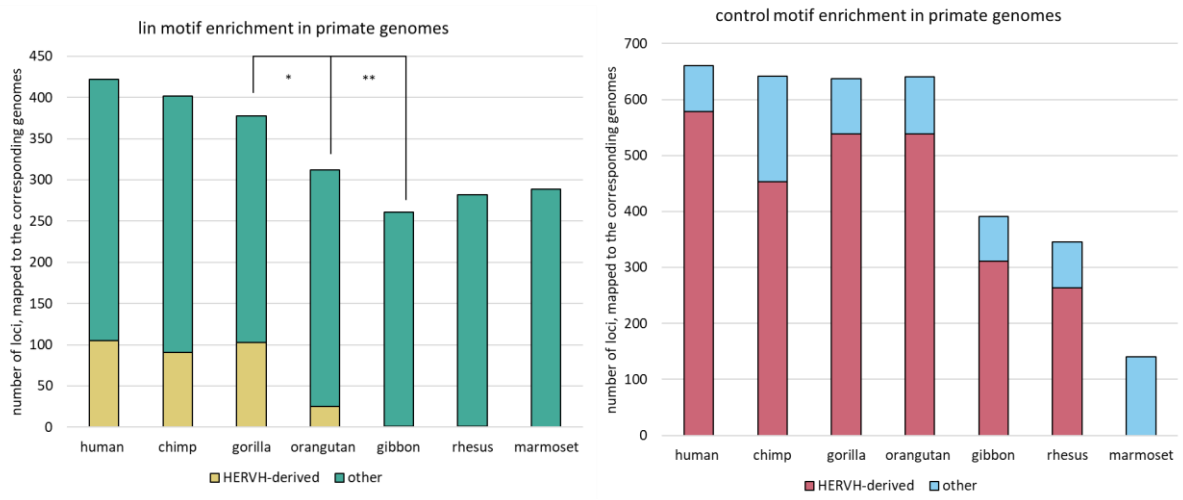


Figure 37. *lin* and *con* motifs distribution in human, apes, New- and Old-World monkeys. The number of loci, containing *lin* or *con* motifs, is shown. When *lin*/*con* motif coordinates overlap with HERVH – motif is "HERVH-derived", whereas *lin*/*con* motif coordinates overlap with any other locations in the genome – motif is "other". *Lin* motif overlaps with HERVH significantly more frequently in gorilla than orangutan (* – Fisher's exact test $p < 0.00001$), and, in turn, in orangutan more than gibbon (** – Fisher's exact test $p < 0.00001$). The difference for the HERVH-derived *con* motif in orangutan to gibbon is not significant (Fisher's exact test $p= 0.0633$).

Lin and *con* motifs are present in a similar pattern in primates, except for the New-world monkey, marmoset. However, HERVHlin exists predominantly in human, chimp, and gorilla and, to a lesser extent, in orangutan. HERVHcon does not show the same correlation, being highly present in human, chimp, gorilla, orangutan and insignificantly dropping in gibbon and rhesus. These results suggest that HERVHlin and HERVHcon integrations had happened in a divergent evolutionary window. It additionally emphasizes the difference between HERVHlin and HERVHcon groups.

3.6. HERVHlin functionality

3.6.1. Analysis of published LIN28A Clip-seq data

Previous work has shown LIN28A binding to GGAGA motif on RNA in H9 hESC through crosslinking and immunoprecipitation, coupled with a high-throughput sequencing (CLIP-seq) experiment [33]. The immunoprecipitation was performed with an antibody, recognizing the

endogenous protein. Nonetheless, sequencing data analysis was designed to ignore reads, which would map to several positions in the genome. Therefore, the published processed dataset could not be used for HERVH analysis and the raw sequencing data must be re-analyzed.

The trimmed 40bp single-end reads CLIP-seq data was downloaded from the GEO NCBI depository in a fasta format. Bowtie 1 alignment to the hg19 version of the human genome was performed (section 2.2.6). It resulted in 69089796 reads, many of which were residing in the same coordinates, reflecting the number of transcripts, bound by LIN28A, expressed from the same genomic position. To discover the number of reads per locus all the repeating coordinates were combined (section 2.2.6). That resulted in 4638958 unique genomic positions with 1 to 247588 reads per position. Then, to discover how many of these targets are transposons, the RepeatMasker annotation was applied to the unique coordinates (section 2.2.6). The outcome of this analysis was 2236915 TEs coordinates. In this output as well, each TE had several reads along its sequence, hence the reported TEs coordinates were combined (section 2.2.6). There were 901309 TEs with 2 to 27203 reads, aligning to each annotated element.

Due to differences in length between annotated TEs, the number of reads per transposon needed to be normalized to the length of a corresponding locus. After this normalization, the values reflected the amount of RNA bound by LIN28A, transcribed from a specific TE locus and range between 0.00022899 and 220.3421. Most present types in the 75-100% quantile were Alu and L1Hs elements (Figure 38, left). The difference between the number of reads per three transposons families is insignificant based on ANOVA analysis ($p=0.4106$).

Following this, the difference in LIN28A binding between HERVHlin and HERVHcon was addressed. From all TEs HERVH coordinates, 1750 were selected. This number is based on the RepeatMasker annotation, which has around 6000 HERVH. From all HERVH bound by LIN28A, 87 were HERVHlin and 401 HERVHcon loci. The range of reads (Figure 38, right), normalized to the element length varied between 0.03097893 and 0.11885467 for HERVHlin and 0.000502513 to 0.10807947 for HERVHcon (Figure 38, right).

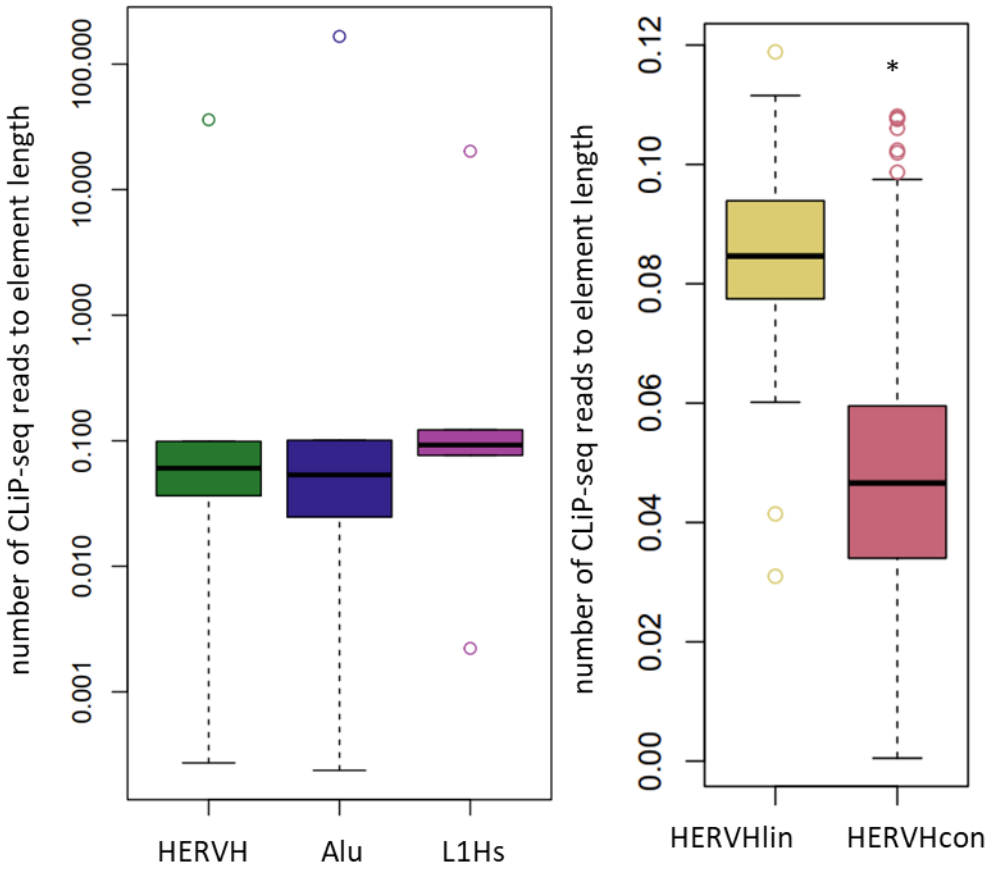


Figure 38. Transposable elements, bound by LIN28A based on CLIP-seq analysis. Left: HERVH, Alu and L1Hs transcripts are bound to LIN28A with the same efficiency (ANOVA p=0.4106). Four quantiles for each group are shown as the number of CLIP-seq reads normalized to the length of every element from a corresponding TE family, values of the axis are logarithmic. Right: more HERVHlin transcripts are bound by LIN28A than HERVHcon (* - t-test p<2.2e-16). All values of CLIP-seq reads, normalized to the length of each element are shown.

In agreement with GGAGA being a functional binding site of LIN28A, more HERVHlin is bound by LIN28A, most probably due to the presence of two LIN28A binding sites in the *lin* motif to compare with the *con* motif. Nevertheless, many HERVH sequences are bound by LIN28A, which likely also has a biological function.

To additionally confirm the functionality of the double LIN28A binding site in HERVHlin loci, immunoprecipitation followed by RNA isolation and qPCR was performed in H1 and H9 hESC.

3.6.2. LIN28A immunoprecipitation followed by qPCR

To confirm binding of LIN28A to HERVHlin in the culture conditions used in the previous experiments, immunoprecipitation of LIN28A, followed by qPCR amplification (RIP-qPCR) of several specific HERVHlin loci was performed. First, pairs of primers for single HERVHlin and HERVHcon loci were designed, tested in qPCRs with RNA-derived cDNA from wild type cells, and each product was Sanger sequenced to confirm the right identity of the fragments (data not shown). It was challenging to find primers, which would be efficient in qPCRs and amplify specifically either a HERVHlin or a HERVHcon single locus. This was due to the repetitive nature of the region and high homology between HERVHlin and HERVHcon sequences, except for the *lin*/*con* motif.

Two pairs of primers, amplifying different HERVHlin loci from chromosome 1 and chromosome X, and two pairs of HERVHcon primers targeting different loci on chromosome 4 were used for the RIP-qPCR experiment. As an additional negative control *ESRG*, a previously described HERVH product was used. The *ESRG* transcript has two GGAGA LIN28A binding sites located in proximity but separated by different number of nucleotides in comparison with *lin* or *con* motifs. U1, small nuclear RNA gene, was used as a general negative control for the experiment and *HNRNPF* as a reported positive control for LIN28A binding [33]. An additional positive control primers to *CDK4* were provided with the kit (section 2.1.14). The difference between two replicates for both H9 and H1 hESC lines was high, therefore values of each gene were normalized to U1 of the corresponding replicate (Figure 39 and 40, respectively).

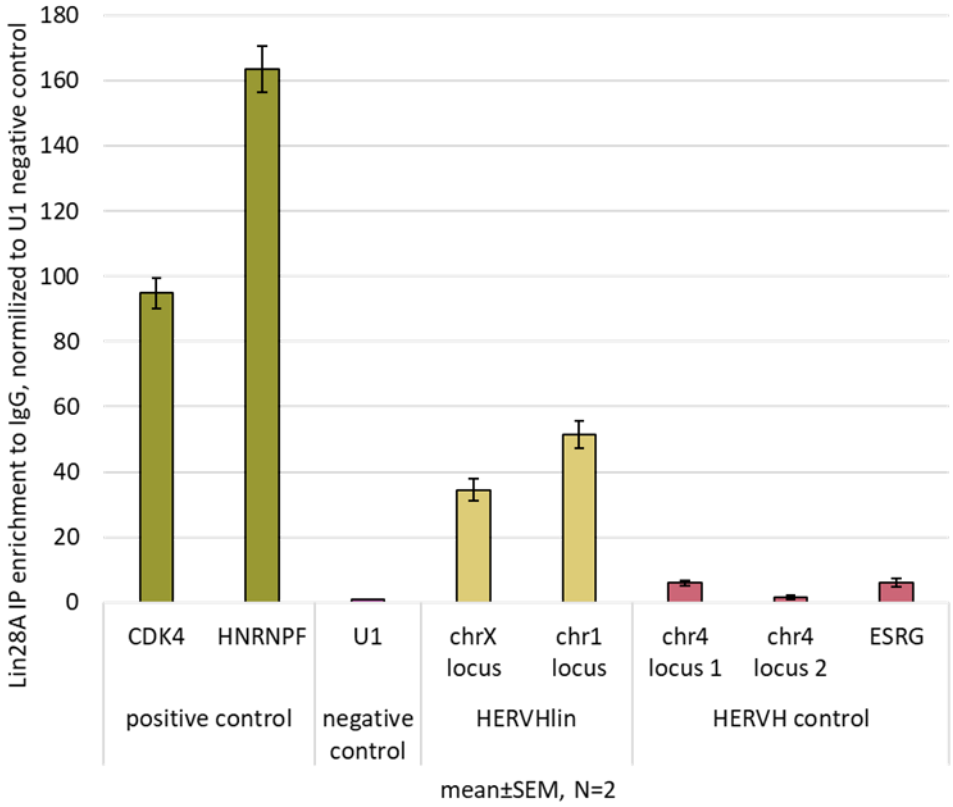


Figure 39. LIN28A predominantly binds HERVHlin transcripts in H9 hESC. LIN28A immunoprecipitation followed by RNA isolation, cDNA synthesis and qPCR validation. CDK4 and HNRNPF – positive controls, U1 – negative control, chrX and chr1 – primers, amplifying HERVHlin loci from X and 1 chromosomes respectively, chr4 locus 1, locus 2 – primers, amplifying HERVHcon loci from 4 chromosome in different locations, *ESRG* – HERVH-derived transcript, not containing *lin* or *con* motif. The mean value from two replicates with standard error of the mean (SEM) is shown. Due to high batch-to-batch variability, based on t-test, the difference between HERVHlin and HERVH control loci is insignificant.

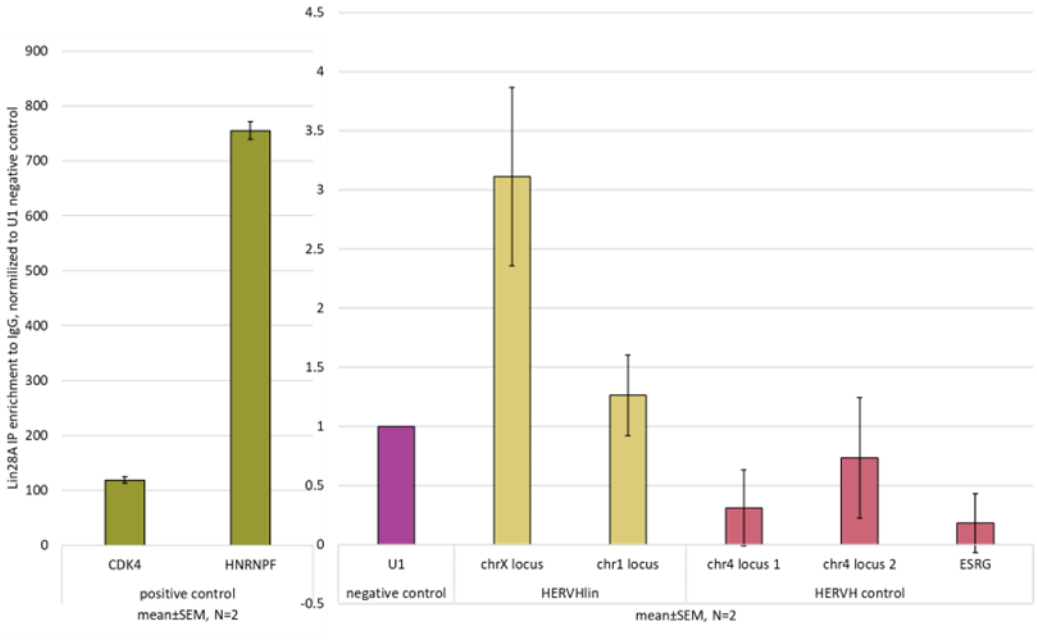


Figure 40. LIN28A predominantly binds HERVHlin transcripts in H1 hESC. All the data is as in Figure 39. Similarly to H9 hESC, due to high batch-to-batch variability, based on t-test the difference between HERVHlin and HERVH control loci is insignificant.

In both H9 and H1 cell lines there is a trend of more efficient LIN28A binding to HERVHlin than HERVHcon, even though the difference is not statistically significant, as assessed via a t-test. After normalization, variation for CDK4 and HNRNPF positive controls between two replicates is more than 4 or 12 times, respectively, which suggests issues with reproducibility of the RIP-qPCR protocol, even operating with the commercial kit. Nevertheless, both CLIP-seq and RIP-qPCR data confirm preferential binding of LIN28A to HERVHlin transcripts.

In the light of the third aim of this study, I hypothesize that HERVHlin binding to LIN28A prevents it from degrading let-7 precursor RNAs [34]. Therefore active let-7 can inhibit L1, as this activity was shown before [38].

3.7. let-7 independent L1 is not affected by HERVH knock-down

Importance of let-7 in the hypothesized mechanism is addressed by an artificially created hyperactive version of L1, L1-ORFeus [234], which has less predicted binding sites for let-7, than a regular L1-RP element. Let-7 binding site or seed regions were analyzed with RNA22 microRNA binding prediction tool [235] on L1-ORFeus sequence (table 5).

element name	Let7 binding coordinates	p-value
L1-RP	745-764	0.31500
L1-RP	2646-2667	0.084900
L1-RP	4254-4274	0.235000
L1-RP *	4592-4611	0.003450
L1-ORFues	2488-2509	0.062700

Table 5. let-7 microRNA binding sites prediction in L1-RP (wild type element) and L1-ORFeus (hyperactive transposon). Coordinates of each binding site are shown as a range of base pairs. P-value reflects significance of predicted sites. * - the experimentally validated let-7 binding site.

The only significant let-7 binding site ($p<0.05$), which was also experimentally validated [38], is located around 4600bp of L1 consensus. The one predicted binding site on L1-ORFeus is not significant, as well as two other sites on L1-RP. Therefore, L1-ORFeus transposition is probably independent of let-7 activity.

If HERVHlin "protects" let-7 through LIN28A binding, then the general HERVH depletion, which also targets HERVHlin loci, decreases the number of active let-7 molecules. In this background, regular L1-RP transposes more (section 3.2.2). But if the L1-ORFeus based reporter is used, its transposition activity should not differ in shHERVH samples when compared with scr controls, due to the absence of let-7 binding sites on L1-ORFeus sequence. A luciferase-based L1-ORFeus reporter (further referred to as ORFeus) [224], active according to the same principles as CAG-L1 reporter (Figure 18) and driven by a CAG promoter, was used in H9 hESC with shscr or shHERVH constructs to deplete HERVH transcripts, including HERVHlin. JM111-Luc reporter served as inactive control. Five days after transfection, cells were collected and the luciferase assay was performed (section 2.1.15). All normalization and technical repetitions were performed similarly to the experiment in section 3.2.2 in four independent replicates. ORFeus reporter alone was transfected in two independent replicates (Figure 41).

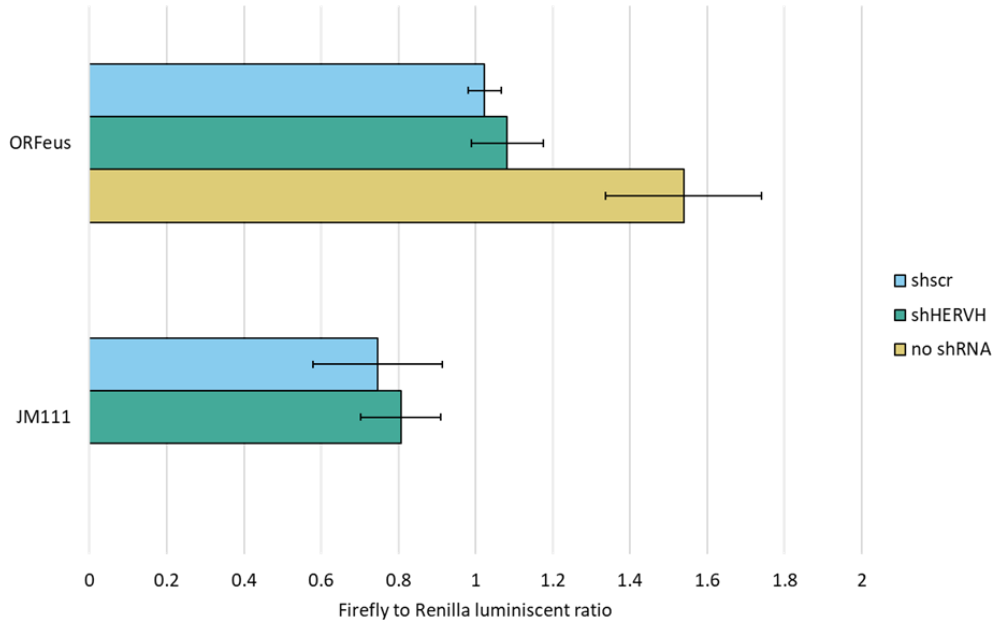


Figure 41. HERVH knock-down does not affect ORFeus transposition in H9 hESC. Hyperactive L1-ORFeus transposition measured with luciferase reporter (ORFeus), shown as a ratio of Firefly to *Renilla* luminescent signal. JM111 is a transposition impaired L1, serves as a negative control. The mean of four independent replicates is shown. Except for the ORFeus-only transfection, measured in two independent replicates. Each replicate is normalized to batch variation, bar represents standard error of the mean (SEM), signal ration does not differ significantly between shscr and shHERVH samples (p=0.7753).

HERVH presence does not affect the transposition activity of L1-ORFeus element, which could serve as an indirect proof of let-7's importance on the HERVHlin control of L1 transposition.

As stated in the third aim of the study, the mechanism of HERVH controlling REs transposition was suggested and partially validated by confirmation of HERVHlin more efficient binding to LIN28A and let-7-independet L1-ORFeus transposition not affected by HERVH depletion.

3.8. An overview of results in agreement with the aims of the study

In this work I have shown that HERVH presence is crucial for the control of young REs activity. According with the first aim of the study, I had shown that general HERVH depletion causes elevated L1 transposition measured by EGFP- and luciferase-based reporters in H9 hESC. The knock-down cells gain *de novo* insertions of L1s, SVAs and mostly Alus after being maintained for several passages, as for the second aim. A part of HERVH, responsible for control of young elements' transposition, has a *lin* motif with two LIN28A binding sites. These HERVHlin loci are present in a similar number in human, chimp and gorilla genomes, half in orangutan and absent from gibbon, rhesus macaque and marmoset. The LIN28A binding sites in HERVHlin sequences are functional and indeed these transcripts are bound to LIN28A more efficiently than other HERVH. HERVHlin sponging of LIN28A is probably crucial to control L1 transposition, because that allows the production of microRNA let-7 and the subsequent L1 activity inhibition. Similar rates of let-7 independent L1 element L1-ORFeus transposition in HERVH depleted and control cells support this hypothesis, which mostly covers the scope of the third aim of the thesis.

4. Discussion

4.1. Young retroelements activity and HERVH

Phylogenetically young REs, which are still active in the human population [96], exhibit their deleterious effect through the activation of inflammation [148, 149] and *de novo* integrations to genes or regulatory sequences, which disrupt the functions. The host has evolved several mechanisms to inhibit the activity of REs. Older REs undergo transcriptional control, and younger are targeted with small RNA and RNA deamination. These protective mechanisms have crosstalk, implying several layers of inhibition for REs loci.

In this work, I show that a part of HERVH family elements has evolved to rewire a conserved protein-miRNA pathway for repression of L1 in hESC. Previously, the contrasting expression was observed in naïve vs primed hESC, where naïve cells were marked by SVAs, HERVK, and to a lesser extent, L1s and HERVH were present in RNA-seq data of primed conventional hESCs [5]. Dr. Trono's research group has reported an evolutionary recent subfamily of SVAs, HERVK with LTR5Hs promoter and younger HERVH elements, driven by LTR7B or LTR7Y, to serve as enhancers in naïve cell cultures and pre-implantation development during EGA [93]. KLF4 and its functional homolog KLF17 were shown to be responsible for opening thousands of genomic loci during EGA, including the aforementioned transposons. These data do not contradict our analysis, where evolutionary older HERVH transcripts are expressed in the opposite manner to younger REs, as we are reporting on events happening later in the span of human preimplantation development. HERVHlin and other HERVH transcripts are expressed profoundly in the pluripotent epiblast [6] (Singh et al, unpublished), whereas SVAs, LTR5Hs with HERVK and younger HERVH are marking earlier developmental stages such as morulae and 8-cell stage. Also, as mentioned in the section 1.3., cultured primed hESCs are derived from later stages of human development than the ones mimicked in forced naïve cultures. That explains higher HERVH expression in conventional cells.

HERVH expression is contrasting to young REs during the reprogramming of human fibroblasts to iPSCs. The waves of HERVH expression during reprogramming were reported in the research of Ohnuki with co-authors, which provided us with the high-throughput data for the downstream analysis [193]. And remarkably, *de novo* REs integrations during reprogramming and further culturing of hESCs and hiPSCs have been shown before [236, 237]. L1s, Alus, and SVAs were reported to mobilize in different cell lines and, for example, four out

of seven L1 integrations were full-length. Nevertheless, the authors do not show the cause of these integrations [236]. We assume, that, at early maturation stages of reprogramming, due to the absence of HERVHlin transcripts, cells suffer from the active transposition of Alus, SVAs, and L1s. Later, when different HERVH subtypes, including HERVHlin, are expressed, young REs activity is inhibited, but the new integrations are maintained in the cells.

The high levels of Alu, SVA, and L1 transcripts in RNA-seq data might be the consequence of REs transposition, as these elements mobilize through an RNA intermediate. How actively should REs transpose to produce enough transcripts, detectable as a significant difference in RNA-seq data, is still an open question. But I don't exclude the possibility of additional expression regulation of young retrotransposons either by HERVH or by an epigenetic factor, controlling both types of elements.

4.2. Challenges and limitations of the research

4.2.1. Phenotype of HERVH depletion

The shRNA sequence used in this research to deplete HERVH has been reported to cause differentiation of hESCs [7]. The different shRNA, applied previously in our research group for HERVH knock-down, also disturbs pluripotency [6]. In both articles, the morphology of pluripotent cells changed to fibroblast-like and the expression of *NANOG* and *OCT4* pluripotency markers was reduced. In my experiments, I do observe only the reduction of *NANOG* expression with no morphological changes either in the transient knock-down (section 3.1.) or stable knock-down (section 3.3.2). This can be explained by the different shRNA sequence in comparison with Wang and co-authors' results, as these shRNAs are predicted to target different subsets of HERVH loci [6]. But the shRNA sequence used here to deplete HERVH is similar to one of the three shRNAs designed by Lu and co-authors [7]. Nevertheless, I cloned the shRNA to a different expression vector and used a different delivery method. Both conditions might have caused the targeting of HERVH transcripts, residing in a specific compartment, for example the cytoplasm, different from the nuclear subset of HERVH depleted in the mentioned research [7]. I used H9 hESCs to perform the knock-down experiments, in contrast to H1 hESCs-based earlier experiments. All these factors might cause the discordant results of my work with the previously published research.

The recent study, performed by Yamanaka's research group questioned the role of *ESRG*, HERVH-derived transcript, which is expressed in hPSCs [125]. In the article, they not

only performed depletion by shRNA, but also a knock-out of HERVH, located in this locus. None of the experiments have shown any effect on pluripotency factors expression, or morphology of hESC colonies. Takahashi and co-authors used a different shRNA from the three used in the publication of Lu and co-authors [7], to compare with the one I applied in this work. However, except for the expression of *NANOG*, Takahashi's and co-authors' research supports my observation of the absence of differentiation phenotype.

Despite the absence of differentiation morphology, the *NANOG* reduction, caused by shHERVH transfection could be the first sign of hESCs differentiation. It is quite crucial, as, during human pluripotent cells differentiation, histone marks are undergoing rearrangement [238]. Many transposons are controlled by epigenetic modifications (see section 1.2.5); therefore the differentiation might cause a shift in control of expression and transposition of young REs. In other types, for example, in neuronal differentiation of rat cells, L1 transposition is activated [189]. Hyslop with co-authors has shown that Nanog depletion could cause differentiation to trophectoderm [219], and trophectoderm was reported to support L1 transposition (Muñoz-Lopez et al., unpublished). Thus, I tested the expression of markers for several differentiation programs and showed similarly low expression between HERVH-depleted and control cells. The effect of HERVH knock-down on L1s, SVAs, and Alus transposition is a direct function of HERVH transcripts and can't be explained by the differentiation of HERVH-depleted cells.

4.2.2. Reporter-based transposition assays

We detected that the most differentially expressed family of retrotransposons, "reacting" on the presence of HERVH transcripts was SVA. Unfortunately, SVA transposition reporters are based on antibiotic selection, which has a high background – a consequence of the single cell selection and further colony maintenance [239, 240]. Additionally, the available SVA trans-mobilization assay reporters were based on neomycin selection, which could interfere with HERVH depletion experiments also utilizing neomycin resistance [141]. The more precise transposition assay types are fluorescence or luminescence-based, where it's possible to select the true positive signal efficiently. These types of transposition assays exist only for the L1 element. Since L1 is the only autonomous element, and its mobilization contributes to non-autonomous REs transposition, the confirmed L1 integrations would indicate the activity of Alus and SVAs. Hence, I performed both transposition assays for L1.

L1-EGFP reporter contained the L1 sequence driven by native 5'-UTR promoter, which is active at the relatively low levels in hESCs. Additionally, the newly integrated L1-EGFP cassette might be 3' truncated or silenced after integration, which would result in the proportion of two out of three false negative results in the assay (Figure 42) [241].

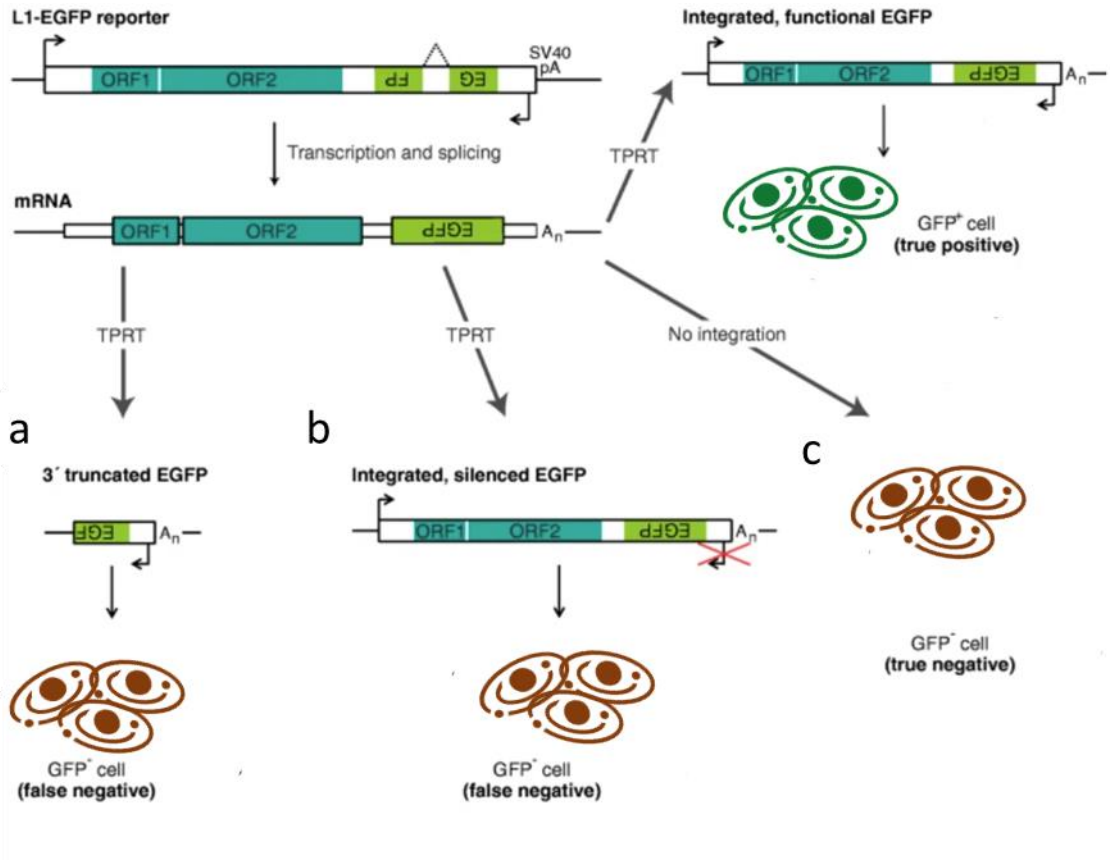


Figure 42. An output of the EGFP-based reporter L1 transposition assay. After *de novo* integration of the L1-EGFP cassette, the false negative cells could be the result of a. 3' truncation of EGFP, b. silencing of the whole cassette after integration. The true negative EGFP cells and true positive signals are the other possible outcomes. TPRT – target-primed reverse-transcription. The image is adapted from [241].

The average percent of EGFP-positive HERVH-depleted cells was around 1% higher than the control, which might be the consequence of the close-to-background L1 transposition or low sensitivity of the reporter. Driven by a stronger EF1α promoter, but generally, the same type of L1-EGFP reporter had been used to decipher the transposition activity in post-reprogramming cultured cells, showing a comparable percent for EGFP-positive cell population (Figure 43) [237].

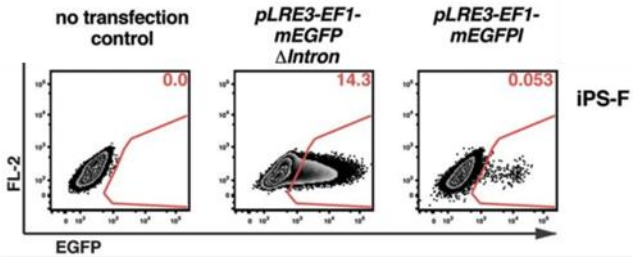


Figure 43. L1-EGFP transposition assay in hiPSC, showing EGFP-positive cells population, as a result of active L1 transposition (adapted from [237]). pLRE3-EF1-mEGFP ΔIntron – a positive control reporter construct, lacking intron in the EGFP sequence, showing the efficiency of the reporter transfection; pLRE3-EF1-mEGFPI – L1 transposition reporter, with an EGFP cassette, interrupted by an intron, EGFP-positive cell appear only after the round of successful L1 transposition; iPS-F – fibroblasts-derived hiPSC; FL-2 – second fluorophore channel, used to sort out autofluorescence.

As I could replicate the previously published transposition frequency in hPSCs [237], the more sensitive luciferase-based reporter was used, to better address the transposition in the HERVH-depleted background [224]. The luciferase-based transposition reporter had the same proportion of false negative samples. But the acquired data could be considered more reliable due to the high intensity of the luminescent signal, the higher sensitivity of the assay, and the strong constitutive CMV promoter driving L1 expression [224].

The main challenge was the duration of the transposition assay. The *Renilla* luciferase luminescence, used to normalize the transposition signal, was detected in parallel with Firefly luciferase five days after transfection when the knockdown reached the desired levels. As the Firefly signal was a product of transposition, which was stable in time after integration, *Renilla* activity was derived from the reporter plasmid, losing the signal after maintenance in cell culture. The previously performed luciferase-based L1 transposition assays did not exceed four days after transfection [38, 224]. Due to the close-to-background luciferases signal, only the more sensitive Dual-Luciferase® reporter assay system (E1980, Promega), recovering decimal higher values, would detect the significant values, unlike the Dual-Glo® luciferase assay system (E2980, Promega).

Considering the low levels of luciferases luminescence, we have decided to avoid reporter-based transposition assays further on and use a high-throughput analysis to detect *de novo* retrotransposons integrations.

4.2.3. Sequencing-based transposition detection

High-throughput integration site analysis is a common technic to address the activity of transposons. The widely used method is retrotransposon capture sequencing (RC-seq) [242]. In this method, Illumina libraries are enriched for fragments containing the 5' and 3' termini of specific mobile elements insertions. RC-seq has been successfully performed for several cell lines, detecting new transposition sites, including the integrations of L1s, Alus, and SVAs [243–245]. The other applied strategy is WGS, followed by computational analysis for discordant reads, performed by the TEBreak algorithm [245]. Due to the short experimental timeframe and less sophisticated sample preparation, we decided to perform WGS to detect *de novo* integrations of young REs in HERVH-depleted cells.

The main challenge of WGS-based analysis for retrotransposon insertions is the low abundance of cells, carrying the new integrations. hESCs grow in colonies and a colony is a clonal cell population, hence, a progeny of one cell. After a transposition of an RE, heterogeneous clones with only one cell carrying an integration will be formed. To be able to detect an integration with WGS, it's important to reach enough cells carrying an integration. For this clonal cell culture has to be maintained for at least 10 passages. Therefore, I generated samples from passages 0 and 10 of control and HERVH-depleted clones. After analyzing with TEBreak and selecting predicted integrations, HERVH knock-down clones had 180 or 83 predicted RE integrations, which corresponds to the previously detected number of integrations in cultured human pluripotent stem cells [236, 237]. To support the results with experimental validation, 12 integrations from each clone, 6 from the high confidence number of mapped discordant reads, and 6 from the low confidence group were analyzed with PCR, confirming one integration. If the false-positive rate of the predicted integrations is continuous, then HERVH-depleted clones had acquired 7 and 15 integrations of mostly Alu elements per 10 generations. It needs to be further addressed if this level of REs transposition could explain the higher number of their transcripts in RNA-seq data, where the HERVH transcripts are not present.

4.3. HERVH functional regions and the novel HERVH subgroups

HERVH sequences have been studied extensively since their discovery in 1984 [29]. The full study of HERVH consensus with descriptions of homology regions to other viruses was performed by Jern and co-authors [100]. *Gag*, *pol*, *pro,* and *env* regions were described, positioning functional sites like PBS, *pre-gag* region, and the former translational start site. The more recent, full-length alignment, including the analysis of functional regions of the element, has been done by Dr. Katzourakis' research group [108]. This work has taken into account the differential expression of HERVH loci in hPSCs, which was earlier analyzed in our research group [6]. Some regions like G2 in *gag* and P6 in *pol* were found to correlate with the expression levels of HERVH in pluripotent cells, implying a possibility for epigenetic marks at these locations (Figure 44). Nevertheless, the *lin* motif has not been addressed in any of the publications and the HERVHlin subfamily of elements did not attract attention until now. The schematic alignment of *gag* and *pol* regions with some important activity features is shown below (Figure 44).

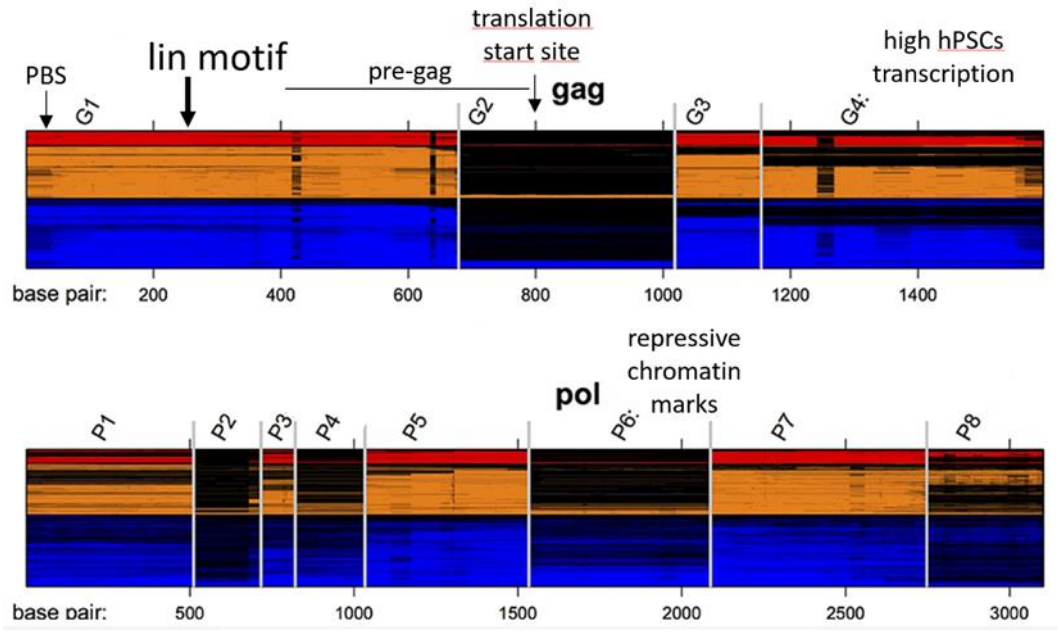


Figure 44. HERVH schematic consensus, adapted from [108] with additional information from [100]. Each horizontal line represents one locus, red – highly active loci, orange – moderately active, and blue – inactive, based on the analysis of [6]. *Gag* and *pol* regions are shown, with *gag* consisting of G1-G4 regions, where the G2 region is often deleted (black). Primer binding site (PBS) is located at around 20bp from the start of gag [100], *lin* motif – 270bp, *pre-gag* region – 400-800bp, followed by the former translation start site, G4 region correlates with high transcription in hPSCs

[108]. *Pol* consists of P1-P8, where P2, P4, P6, and P8 are frequently deleted, P6 correlates with lower transcription, most likely by attracting repressive histone marks [108].

The novelty of this study is not only the discovery of a new functional *lin* motif in the HERVH consensus but also the perception of HERVH loci as subgroups, separated by spatio-temporal expression and sequence. Previously, the research was concentrating on HERVH as a whole family [6, 7, 117], or as a single locus [118, 125].

Since we don't consider a family of paralog genes as one entity, HERVH should be also studied in functional subgroups, and not only as a family. On the other hand, the repetitive nature of the element adds one extra layer of complexity and hinders the research with the same logic as for a single gene. The group-specific expression pattern was adapted from [6] and was developed further, based on single-cell RNA-seq of pre-implantation development and reprogramming. That led to the discovery of the antagonistic HERVH subgroup. Further, the presence of the functional LIN28A binding *lin* motif allowed me to describe HERVHlin loci, based on their expression pattern and sequence-specificity. HERVHlin could integrate into a conservative LIN28A/let-7 pluripotency-specific pathway without disturbing its function and providing an additional layer of control for young REs.

4.4. HERVHlin rewired the LIN28A/let-7 pathway

In this study, I hypothesize that HERVHlin has gained a *lin* motif, tandem LIN28A binding site, which allowed HERVHlin to sponge LIN28A. If LIN28A is inactive, let-7 could mature and control L1 transposition (Figure 45).

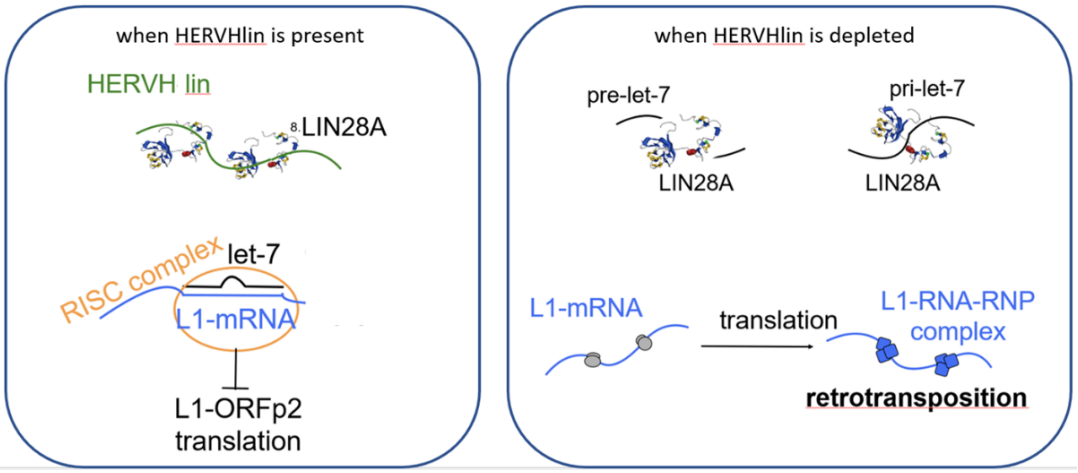


Figure 45. Suggested molecular mechanism behind HERVHlin inhibition of young REs. When HERVHlin (green) transcripts are present (left panel), LIN28A protein (structure image) binds these RNAs and does not control maturation of let-7 miRNA. Active let-7 can inhibit ORF2 of L1 translation by binding L1-mRNA (blue) in RISC complex (yellow). L1 inhibition results in reduced transposition of Alus and SVAs as well. If HERVHlin is depleted (right panel), pre-let-7 and pri-let-7 are degraded by LIN28A. L1-mRNA is translated, L1 forms RNPs (blue) and retrotransposes, allowing integrations of Alus and SVAs.

4.4.1. The canonical LIN28A/let-7 pathway

LIN28A was discovered in nematodes and was shown to be conserved in all bilaterians [246]. The protein is highly expressed in development and downregulated during differentiation, with the exception of erythrocytes-like cell types, as well as cardiac and skeletal muscle tissues [247]. All vertebrates possess two paralogs, *LIN28A* and *LIN28B*. LIN28A can shuffle between cytoplasm and nucleus, while LIN28B is a nuclear protein [248]. *LIN28A* was able to substitute the oncogenic *c-Myc* in reprogramming, and endogenous activation of both *LIN28A* and *LIN28B* was crucial for maximum efficiency of the process [62, 249].

The pluripotency is supported through let-7 inhibition. Let-7 is a microRNA that is normally induced upon the differentiation of stem cells and inhibits stemness-specific factors like HMGA2, RAS, and c-Myc [250–252]. LIN28A and LIN28B could block pri-let-7 processing with Drosha in the nucleus or pre-let-7 maturation through inhibition of Dicer in the cytoplasm [34–37]. To inhibit let-7 maturation, LIN28A binds GGAG or GGAG-like motifs in

pre-let-7's terminal loop [253]. On the other hand, mature let-7 can bind LIN28A mRNA, blocking its translation and representing a bistable switch [254–256].

4.4.2. LIN28A binds mRNAs and HERVH

Based on CLIP-seq analysis, endogenous LIN28A was shown to be binding GGAGA motifs in H9 hESCs, located at unpaired mRNA regions of secondary structures of genes, mostly coding splicing factors [33]. Protein levels of TDP-43, FUS/TLS, TIA-1, and hnRNPF increase in response to the upregulation of LIN28A, which causes widespread changes to alternative splicing. Interestingly, the detected GGAGAU binding motif in HEK cells differed by one last nucleotide to an hESC-specific motif [33]. HERVHlin doesn't have a U base pair at the end of the *lin* motif, which might cause the difference between hESC- and HEK-specific LIN28A binding sites, as HERVHlin has low expression levels in HEK cells, to compare with H9 hESCs (data not shown).

It is important to note, that most of the HERVH sequences have one additional GGAGA motif closer to the 3' part of predicted transcripts. LIN28A is involved in the regulation of splicing factors [33]. Splicing of transcripts could enhance their expression, probably by showing a cell the importance of these sequences [257, 258]. The original binding to LIN28A might have been an evolutionary feature of previously active transposing HERVH to increase transcription through the number of splicing events.

4.4.3. HERVHlin might sequester LIN28A to condensates

One essential question to ponder on is what might happen after LIN28A binds HERVHlin. HERVHlin sequences additionally differ from HERVHcon by CT-rich partially repetitive regions (section 3.4.3, figure 33). I assume that these regions might play a role in liquid-liquid phase separation, allowing HERVHlin to sequester LIN28A into cytoplasmic droplets, where the protein can no longer function. Previously it was shown that ERVs transcripts tend to form transcriptional condensates in mESC [259]. HERVH in particular was discovered to be crucial for the formation of BRD4 puncta, condensates of activate transcription in a cancer cell line [117].

Therefore, HERVHlin might be a scaffold for LIN28A containing condensates. If a regular LIN28A function is disturbed after droplets formation, that might be inhibitory types of condensates, such as P-bodies, instead of transcriptional activation. Counterintuitively, P-

bodies were shown to generally contain mRNA decay machinery (reviewed in [260]) and stress granules – translation initiation components. But in general, both structures are dynamic and there is a number of proteins which shuffle between one another [261]. Interestingly, LIN28A was detected shuffling between P-bodies and stress granules in hESC but the conditions allowing this shuffle are not yet understood [262]. The factor, that causes LIN28A shuffling between P-bodies and stress granules might be HERVHlin. The pilot experiment to address this idea might be a co-staining of LIN28A with markers of P-bodies and stress granules, compared between wild type and HERVHlin depleted cells.

4.4.4. let-7 might be involved in the control of L1 by HERVHlin

In this work, HERVH depletion was shown to not affect L1-ORFeus transposition (section 3.7., Figure 41). The L1-ORFeus element was created as a hyperactive version of L1 [234], having fewer binding sites for the L1 transposition controlling factor, including the predicted let-7 binding regions (section 3.7., Figure 40). Cells transfected with shRNA against HERVH did not show any increase in the transposition activity of ORFeus reporter, to compare with control cells, unlike a regular L1 reporter transfected cells. That suggests let-7's involvement in the HERVHlin-mediated control of L1 activity.

If HERVHlin inhibits LIN28A activity by sponging it to condensates, mature let-7 miRNA should be detectable in hESC, where HERVHlin is present. Let-7 is believed to be expressed only in differentiating cells. But in one study let-7 was detected specifically in human and not mouse ESC [263]. The authors discovered that let-7 fine-tuned LIN28B and does not affect LIN28A in hESCs. From the other hand, when LIN28A itself was silenced, it did not influence the let-7 level in hESCs. The previously mentioned bistable switch was not active in hESCs, most probably due to LIN28A already being irresponsible for let-7 maturation in wild-type hESC, since the protein is bound by HERVHlin. Consequently, let-7 could control L1 transposition in hESCs.

Let-7 was shown to inhibit L1 transposition in HELA, HEK, and lung cancer cell lines [38]. The miRNA was guiding AGO2 to the human L1 mRNA, and its binding occurred in the L1 coding sequence. ORF2 contained a noncanonical 7-mer let-7 binding site. Mutations in this binding site reduced, but didn't abolish, the effect of let-7 modulation on human L1 mobility. Let-7 was shown to impair the translation of L1 ORF2p without affecting mRNA stability [38].

I assume the same mechanism takes place in hESCs, after HERVHlin sequestered LIN28A, allowing let-7 to function.

This takes me to the question of how can let-7 and LIN28A canonically function in pluripotent cells and during differentiation, considering the novel HERVHlin role? HERVHlin probably resides only in one cell compartment, most likely in the cytoplasm. Therefore, LIN28A located in the nucleus could still regulate let-7 maturation. Additionally, due to HERVHlin established control of L1 transposition being an evolutionary new process, the whole machinery might function at moderate levels, as it's not fully prevalent in the human population yet, allowing some amount of LIN28A to control let-7.

5. Conclusion and outlook

In the first part of this thesis, I had shown that HERVH inhibits the transposition of young REs. HERVH is an endogenous retrovirus, which integrated into the genome of our ancestor 30 MYA [30], mobilization impaired now but is highly expressed in hPSCs [6–8, 32]. In association with the Mediator complex, p300 activator, and OCT4, HERVH transcripts support pluripotency by regulating the expression of neighboring genes [7]. Its promoter, LTR7, can also drive expression of genes, lncRNAs, and chimeric transcripts, crucial for hPSC maintenance [6]. Here I discovered the mechanism of genome protection by HERVH, crucial for pluripotent cells. The phylogenetically young REs like some L1s, SVAs, and Alus are still active in humans, and their mobilization may disturb genome integrity [20–27]. When L1 is active, Alus and SVAs are usually able to mobilize as well [139, 141]. With the two reporter assays, I showed that HERVH depletion causes elevated L1 transposition. We annotated *de novo* integrations of Alus and SVAs in HERVH-depleted hESCs and confirmed the activity of these REs. The novel HERVH function brings an extra layer of complexity to the pluripotency network, supported by HERVH expression.

In the second part of the study, I addressed the mechanism of the HERVH protective function. By analyzing HERVH sequences, I discovered a HERVHlin subgroup. These HERVH loci have a 16bp *lin* motif, which carries two LIN28A binding sites [33]. HERVHlin is younger than the other HERVH elements. HERVHlin is present in humans, chimps, and gorillas, their number is reduced in orangutans and absent from lower primates. By analyzing published CLIP-seq data [33] and performing the RIP-qPCRs, I showed that LIN28A can bind the *lin* motif, and HERVHlin are more frequently bound to LIN28A to compare with other HERVH. LIN28A is known to control the maturation of let-7 miRNA [34–37]. Additionally, let-7 controls L1 transposition in cancer cell lines [38]. I suggest the molecular mechanism of REs control, where HERVHlin sponges LIN28A, and that suppresses LIN28A-mediated degradation of let-7. The mature let-7 then inhibits L1 transposition. When L1 is not active, Alus and SVAs can't mobilize. I performed an experiment, where let-7 independent L1 reporter's transposition rate did not change in HERVH depleted background, which indirectly supports the suggested molecular mechanism.

Crucial steps to further support the molecular mechanism of HERVHlin activity would be, first, to analyze expression and RNA isoforms of HERVHlin in human and primate PSCs, to ensure that most of the HERVHlin sequences are transcribed and *lin* motif is present in the

majority of the transcripts. Second, HERVHlin-specific knock-down or knock-out of the most expressed loci needs to be established, followed by a transposition assay, to confirm the unique role of HERVHlin in comparison with other HERVH. Third, the elegant way to show the specificity of the *lin* motif would be a mutation of LIN28A binding sites on the RNA level with Cas13-based editing. Importantly, let-7 involvement in the process needs to be confirmed with a rescue experiment, applying let-7 mimic in HERVHlin-depleted background and measuring the transposition activity of L1. The general phenotype of HERVHlin depletion, addressed by high-throughput methods, might also help in deciphering the mechanism.

The novelty of this work is additionally reflected in the analysis of HERVH not only as the whole family or a single locus but rather as a subgroup with defined sequence features and spatio-temporal expression pattern. The similar approach to repetitive elements analysis might be useful for future studies.

The most fascinating discovery here, in my opinion, is a new evolutionary event, when a former selfish transposon HERVH embedded in a conservative LIN28A-let-7 pathway. This co-option protected the host from other selfish elements, which were harming the genome through new integrations. Therefore, not only HERVH stayed as an important player in the pluripotency network, but also inhibited the transposition of its competitors.

Bibliography

1. Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, Marshall VS, et al. Embryonic Stem Cell Lines Derived from Human Blastocysts. Science. 1998;282:1145–7.

2. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, et al. Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. Cell. 2007;131:861–72.

3. Yamanaka S. Pluripotent Stem Cell-Based Cell Therapy—Promise and Challenges. Cell Stem Cell. 2020;27:523–31.

4. Theunissen TW, Powell BE, Wang H, Mitalipova M, Faddah DA, Reddy J, et al. Systematic Identification of Culture Conditions for Induction and Maintenance of Naive Human Pluripotency. Cell Stem Cell. 2014;15:471–87.

5. Theunissen TW, Friedli M, He Y, Planet E, O'Neil RC, Markoulaki S, et al. Molecular Criteria for Defining the Naive Human Pluripotent State. Cell Stem Cell. 2016;19:502–15.

6. Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature 2014 516:7531. 2014;516:405–9.

7. Lu X, Sachs F, Ramsay L, Jacques P-É, Göke J, Bourque G, et al. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nature Structural & Molecular Biology 2014 21:4. 2014;21:423–5.

8. Durruthy-Durruthy J, Sebastiano V, Wossidlo M, Cepeda D, Cui J, Grow EJ, et al. The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. Nat Genet. 2016;48:44–52.

9. Kastenberg ZJ, Odorico JS. Alternative sources of pluripotency: science, ethics, and stem cells. Transplantation Reviews. 2008;22:215–22.

10. Hyun I, Wilkerson A, Johnston J. Embryology policy: Revisit the 14-day rule. Nature. 2016;533:169–71.

11. De Trizio E, Brennan CS. The business of human embryonic stem cell research and an international analysis of relevant laws. J Biolaw Bus. 2004;7:14–22.

12. Nichols J, Smith A. Naive and Primed Pluripotent States. Cell Stem Cell. 2009;4:487–92.

13. Messmer T, von Meyenn F, Savino A, Santos F, Mohammed H, Lun ATL, et al. Transcriptional Heterogeneity in Naive and Primed Human Pluripotent Stem Cells at Single-Cell Resolution. Cell Reports. 2019;26:815-824.e4.

14. Nakamura T, Okamoto I, Sasaki K, Yabuta Y, Iwatani C, Tsuchiya H, et al. A developmental coordinate of pluripotency among mice, monkeys and humans. Nature. 2016;537:57–62.

15. Tyser RCV, Mahammadov E, Nakanoh S, Vallier L, Scialdone A, Srinivas S. Single-cell transcriptomic characterization of a gastrulating human embryo. Nature. 2021;600:285–9.

16. Xiang L, Yin Y, Zheng Y, Ma Y, Li Y, Zhao Z, et al. A developmental landscape of 3D-cultured human pre-gastrulation embryos. Nature. 2020;577:537–42.

17. Gafni O, Weinberger L, Mansour AA, Manor YS, Chomsky E, Ben-Yosef D, et al. Derivation of novel human ground state naive pluripotent stem cells. Nature. 2013;504:282–6.

18. Weinberger L, Ayyash M, Novershtern N, Hanna JH. Dynamic stem cell states: naive to primed pluripotency in rodents and humans. Nat Rev Mol Cell Biol. 2016;17:155–69.

19. Ware CB, Nelson AM, Mecham B, Hesson J, Zhou W, Jonlin EC, et al. Derivation of naïve human embryonic stem cells. Proceedings of the National Academy of Sciences. 2014;111:4484–9.

20. Scott AF, Schmeckpeper BJ, Abdelrazik M, Comey CT, O'Hara B, Rossiter JP, et al. Origin of the human L1 elements: Proposed progenitor genes deduced from a consensus DNA sequence. Genomics. 1987;1:113–25.

21. Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD. Molecular archeology of L1 insertions in the human genome. Genome Biology. 2002;3:research0052.1.

22. Quentin Y. Fusion of a free left Alu monomer and a free right Alu monometer at the origin of the Alu family in the primate genomes. Nucleic Acids Research. 1992;20:487–93.

23. Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, et al. SVA Elements: A Hominid-specific Retroposon Family. Journal of Molecular Biology. 2005;354:994–1007.

24. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, et al. Hot L1s account for the bulk of retrotransposition in the human population. Proceedings of the National Academy of Sciences. 2003;100:5280–5.

25. Sassaman DM, Dombroski BA, Moran JV, Kimberland ML, Naas TP, DeBerardinis RJ, et al. Many human L1 elements are capable of retrotransposition. Nat Genet. 1997;16:37–43.

26. Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, et al. LINE-1 Retrotransposition Activity in Human Genomes. Cell. 2010;141:1159–70.

27. Ewing AD, Kazazian HH. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res. 2010;20:1262–70.

28. Izsvák Z, Wang J, Singh M, Mager DL, Hurst LD. Pluripotency and the endogenous retrovirus HERVH: Conflict or serendipity? BioEssays. 2016;38:109–17.

29. Mager DL, Henthorn PS. Identification of a retrovirus-like repetitive element in human DNA. Proceedings of the National Academy of Sciences. 1984;81:7510–4.

30. Magiorkinis G, Blanco-Melo D, Belshaw R. The decline of human endogenous retroviruses: extinction and survival. Retrovirology. 2015;12:8.

31. Carter TA, Singh M, Dumbović G, Chobirko JD, Rinn JL, Feschotte C. Mosaic cis-regulatory evolution drives transcriptional partitioning of HERVH endogenous retrovirus in the human embryo. eLife. 2022;11:e76257.

32. Santoni FA, Guerra J, Luban J. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. Retrovirology. 2012;9:111.

33. Wilbert ML, Huelga SC, Kapeli K, Stark TJ, Liang TY, Chen SX, et al. LIN28 Binds Messenger RNAs at GGAGA Motifs and Regulates Splicing Factor Abundance. Molecular Cell. 2012;48:195–206.

34. Heo I, Joo C, Kim Y-K, Ha M, Yoon M-J, Cho J, et al. TUT4 in Concert with Lin28 Suppresses MicroRNA Biogenesis through Pre-MicroRNA Uridylation. Cell. 2009;138:696–708.

35. Viswanathan SR, Daley GQ, Gregory RI. Selective Blockade of MicroRNA Processing by Lin28. Science. 2008;320:97–100.

36. Rybak A, Fuchs H, Smirnova L, Brandt C, Pohl EE, Nitsch R, et al. A feedback loop comprising lin-28 and let-7 controls pre-let-7 maturation during neural stem-cell commitment. Nat Cell Biol. 2008;10:987–93.

37. Newman MA, Thomson JM, Hammond SM. Lin-28 interaction with the Let-7 precursor loop mediates regulated microRNA processing. RNA. 2008;14:1539–49.

38. Tristán-Ramos P, Rubio-Roldan A, Peris G, Sánchez L, Amador-Cubero S, Viollet S, et al. The tumor suppressor microRNA let-7 inhibits human LINE-1 retrotransposition. Nature Communications 2020 11:1. 2020;11:1–14.

39. Schoenwolf GC, Bleyl SB, Brauer PR, Francis-West PH. Larsen's human embryology. 5th edition. Elsevier Science; 2014.

40. Iwata K, Yumoto K, Sugishima M, Mizoguchi C, Kai Y, Iba Y, et al. Analysis of compaction initiation in human embryos by using time-lapse cinematography. J Assist Reprod Genet. 2014;31:421–6.

41. Rossant J, Tam PPL. Early human embryonic development: Blastocyst formation to gastrulation. Developmental Cell. 2022;57:152–65.

42. Niakan KK, Han J, Pedersen RA, Simon C, Pera RAR. Human pre-implantation embryo development. Development. 2012;139:829–41.

43. Mihajlović AI, Bruce AW. The first cell-fate decision of mouse preimplantation embryo development: integrating cell position and polarity. Open Biology. 7:170210.

44. Boroviak T, Nichols J. Primate embryogenesis predicts the hallmarks of human naïve pluripotency. Development. 2017;144:175–86.

45. Flach G, Johnson MH, Braude PR, Taylor RA, Bolton VN. The transition from maternal to embryonic control in the 2-cell mouse embryo. EMBO J. 1982;1:681–6.

46. Braude P, Bolton V, Moore S. Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. Nature. 1988;332:459–61.

47. Vassena R, Boué S, González-Roca E, Aran B, Auer H, Veiga A, et al. Waves of early transcriptional activation and pluripotency program initiation during human preimplantation development. Development. 2011;138:3699–709.

48. Yan L, Yang M, Guo H, Yang L, Wu J, Li R, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nature Structural & Molecular Biology 2013 20:9. 2013;20:1131–9.

49. Asami M, Lam BYH, Ma MK, Rainbow K, Braun S, VerMilyea MD, et al. Human embryonic genome activation initiates at the one-cell stage. Cell Stem Cell. 2022;29:209-216.e4.

50. Kleinsmith LJ, Pierce GB Jr. Multipotentiality of Single Embryonal Carcinoma Cells. Cancer Research. 1964;24:1544–51.

51. Martin GR. Teratocarcinomas and Mammalian Embryogenesis. Science. 1980;209:768–76.

52. Abbasalizadeh S, Baharvand H. Technological progress and challenges towards cGMP manufacturing of human pluripotent stem cells based therapeutic products for allogeneic and autologous cell therapies. Biotechnology Advances. 2013;31:1600–23.

53. Marino J, Paster J, Benichou G. Allorecognition by T Lymphocytes and Allograft Rejection. Frontiers in Immunology. 2016;7.

54. Draper JS, Pigott C, Thomson JA, Andrews PW. Surface antigens of human embryonic stem cells: changes upon differentiation in culture*. Journal of Anatomy. 2002;200:249–58.

55. Liu X, Li W, Fu X, Xu Y. The Immunogenicity and Immune Tolerance of Pluripotent Stem Cell Derivatives. Frontiers in Immunology. 2017;8.

56. Appleby JB, Bredenoord AL. Should the 14-day rule for embryo research become the 28-day rule? EMBO Molecular Medicine. 2018;10:e9437.

57. Briggs R, King TJ. Transplantation of living nuclei from blastula cells into enucleated frogs' eggs *. Proceedings of the National Academy of Sciences. 1952;38:455–63.

58. Briggs R, King TJ. Factors affecting the transplantability of nuclei of frog embryonic cells. Journal of Experimental Zoology. 1953;122:485–505.

59. Gurdon JB, Elsdale TR, Fischberg M. Sexually Mature Individuals of Xenopus laevis from the Transplantation of Single Somatic Nuclei. Nature. 1958;182:64–5.

60. Wilmut I, Schnieke AE, McWhir J, Kind AJ, Campbell KHS. Viable offspring derived from fetal and adult mammalian cells. Nature. 1997;385:810–3.

61. Takahashi K, Yamanaka S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. Cell. 2006;126:663–76.

62. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, et al. Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells. Science. 2007;318:1917–20.

63. Malik N, Rao MS. A Review of the Methods for Human iPSC Derivation. In: Lakshmipathy U, Vemuri MC, editors. Pluripotent Stem Cells: Methods and Protocols. Totowa, NJ: Humana Press; 2013. p. 23–33.

64. Warren L, Manos PD, Ahfeldt T, Loh Y-H, Li H, Lau F, et al. Highly Efficient Reprogramming to Pluripotency and Directed Differentiation of Human Cells with Synthetic Modified mRNA. Cell Stem Cell. 2010;7:618–30.

65. Anokye-Danso F, Trivedi CM, Juhr D, Gupta M, Cui Z, Tian Y, et al. Highly Efficient miRNA-Mediated Reprogramming of Mouse and Human Somatic Cells to Pluripotency. Cell Stem Cell. 2011;8:376–88.

66. Wiegand C, Banerjee I. Recent advances in the applications of iPSC technology. Current Opinion in Biotechnology. 2019;60:250–8.

67. Hoffmann-La Roche. Phase I/IIa Dose Escalation Safety and Efficacy Study of Human Embryonic Stem Cell-Derived Retinal Pigment Epithelium Cells Transplanted Subretinally in Patients With Advanced Dry-Form Age-Related Macular Degeneration (Geographic Atrophy). Clinical trial registration. clinicaltrials.gov; 2022.

68. Kim JY, Nam Y, Rim YA, Ju JH. Review of the Current Trends in Clinical Trials Involving Induced Pluripotent Stem Cells. Stem Cell Rev and Rep. 2022;18:142–54.

69. Huang C-Y, Liu C-L, Ting C-Y, Chiu Y-T, Cheng Y-C, Nicholson MW, et al. Human iPSC banking: barriers and opportunities. Journal of Biomedical Science. 2019;26:87.

70. Sequiera GL, Srivastava A, Sareen N, Yan W, Alagarsamy KN, Verma E, et al. Development of iPSC-based clinical trial selection platform for patients with ultrarare diseases. Science Advances. 2022;8:eabl4370.

71. Malchenko S, Xie J, de Fatima Bonaldo M, Vanin EF, Bhattacharyya BJ, Belmadani A, et al. Onset of rosette formation during spontaneous neural differentiation of hESC and hiPSC colonies. Gene. 2014;534:400–7.

72. Kikuchi T, Morizane A, Doi D, Magotani H, Onoe H, Hayashi T, et al. Human iPS cell-derived dopaminergic neurons function in a primate Parkinson's disease model. Nature. 2017;548:592–6.

73. Sougawa N, Miyagawa S, Fukushima S, Kawamura A, Yokoyama J, Ito E, et al. Immunologic targeting of CD30 eliminates tumourigenic human pluripotent stem cells, allowing safer clinical application of hiPSC-based cell therapy. Sci Rep. 2018;8:3726.

74. Okita K, Ichisaka T, Yamanaka S. Generation of germline-competent induced pluripotent stem cells. Nature. 2007;448:313–7.

75. Amps K, Andrews PW, Anyfantis G, Armstrong L, Avery S, Baharvand H, et al. Screening ethnically diverse human embryonic stem cells identifies a chromosome 20 minimal amplicon conferring growth advantage. Nat Biotechnol. 2011;29:1132–44.

76. Merkle FT, Ghosh S, Kamitaki N, Mitchell J, Avior Y, Mello C, et al. Human pluripotent stem cells recurrently acquire and expand dominant negative P53 mutations. Nature. 2017;545:229–33.

77. Mandai M, Watanabe A, Kurimoto Y, Hirami Y, Morinaga C, Daimon T, et al. Autologous Induced Stem-Cell–Derived Retinal Cells for Macular Degeneration. New England Journal of Medicine. 2017;376:1038–46.

78. Sugita S, Mandai M, Hirami Y, Takagi S, Maeda T, Fujihara M, et al. HLA-Matched Allogeneic iPS Cells-Derived RPE Transplantation for Macular Degeneration. Journal of Clinical Medicine. 2020;9:2217.

79. Zhao T, Zhang Z-N, Rong Z, Xu Y. Immunogenicity of induced pluripotent stem cells. Nature. 2011;474:212–5.

80. Deuse T, Hu X, Agbor-Enoh S, Koch M, Spitzer MH, Gravina A, et al. De novo mutations in mitochondrial DNA of iPSCs produce immunogenic neoepitopes in mice and humans. Nat Biotechnol. 2019;37:1137–44.

81. Araki R, Uda M, Hoki Y, Sunayama M, Nakamura M, Ando S, et al. Negligible immunogenicity of terminally differentiated cells derived from induced pluripotent or embryonic stem cells. Nature. 2013;494:100–4.

82. Guha P, Morgan JW, Mostoslavsky G, Rodrigues NP, Boyd AS. Lack of Immune Response to Differentiated Cells Derived from Syngeneic Induced Pluripotent Stem Cells. Cell Stem Cell. 2013;12:407–12.

83. Osafune K, Caron L, Borowiak M, Martinez RJ, Fitz-Gerald CS, Sato Y, et al. Marked differences in differentiation propensity among human embryonic stem cell lines. Nat Biotechnol. 2008;26:313–5.

84. Kattman SJ, Witty AD, Gagliardi M, Dubois NC, Niapour M, Hotta A, et al. Stage-Specific Optimization of Activin/Nodal and BMP Signaling Promotes Cardiac Differentiation of Mouse and Human Pluripotent Stem Cell Lines. Cell Stem Cell. 2011;8:228–40.

85. Choi J, Lee S, Mallard W, Clement K, Tagliazucchi GM, Lim H, et al. A comparison of genetically matched cell lines reveals the equivalence of human iPSCs and ESCs. Nat Biotechnol. 2015;33:1173–81.

86. Kajiwara M, Aoi T, Okita K, Takahashi R, Inoue H, Takayama N, et al. Donor-dependent variations in hepatic differentiation from human-induced pluripotent stem cells. Proceedings of the National Academy of Sciences. 2012;109:12538–43.

87. Takashima Y, Guo G, Loos R, Nichols J, Ficz G, Krueger F, et al. Resetting Transcription Factor Control Circuitry toward Ground-State Pluripotency in Human. Cell. 2014;158:1254–69.

88. Yu H, Chen M, Hu Y, Ou S, Yu X, Liang S, et al. Dynamic reprogramming of H3K9me3 at hominoid-specific retrotransposons during human preimplantation development. Cell Stem Cell. 2022;29:1031-1050.e12.

89. Ohinata Y, Ohta H, Shigeta M, Yamanaka K, Wakayama T, Saitou M. A Signaling Principle for the Specification of the Germ Cell Lineage in Mice. Cell. 2009;137:571–84.

90. Smith A. Formative pluripotency: the executive phase in a developmental continuum. Development. 2017;144:365–73.

91. Kinoshita M, Barber M, Mansfield W, Cui Y, Spindlow D, Stirparo GG, et al. Capture of Mouse and Human Stem Cells with Features of Formative Pluripotency. Cell Stem Cell. 2021;28:453-471.e8.

92. Yu L, Wei Y, Sun H-X, Mahdi AK, Pinzon Arteaga CA, Sakurai M, et al. Derivation of Intermediate Pluripotent Stem Cells Amenable to Primordial Germ Cell Specification. Cell Stem Cell. 2021;28:550-567.e12.

93. Pontis J, Planet E, Offner S, Turelli P, Duc J, Coudray A, et al. Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. Cell Stem Cell. 2019;24:724-735.e5.

94. McClintock B. Controlling Elements and the Gene. Cold Spring Harb Symp Quant Biol. 1956;21:197–216.

95. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.

96. Hancks DC, Kazazian HH. Roles for retrotransposon insertions in human disease. Mobile DNA. 2016;7:9.

97. Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, et al. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature. 2000;403:785–9.

98. Mager DL, Freeman JD. Human endogenous retroviruslike genome with type C pol sequences and gag sequences related to human T-cell lymphotropic viruses. Journal of Virology. 1987;61:4060–6.

99. Fraser C, Humphries RK, Mager DL. Chromosomal distribution of the RTVL-H family of human endogenous retrovirus-like sequences. Genomics. 1988;2:280–7.

100. Jern P, Sperber GO, Ahlsén G, Blomberg J. Sequence Variability, Gene Structure, and Expression of Full-Length Human Endogenous Retrovirus H. Journal of Virology. 2005;79:6325–37.

101. Goodchild NL, Wilkinson DA, Mager DL. Recent Evolutionary Expansion of a Subfamily of RTVL-H Human Endogenous Retrovirus-like Elements. Virology. 1993;196:778–88.

102. Mager DL, Freeman JD. HERV-H Endogenous Retroviruses: Presence in the New World Branch but Amplification in the Old World Primate Lineage. Virology. 1995;213:395–404.

103. Wilkinson DA, Goodchild NL, Saxton TM, Wood S, Mager DL. Evidence for a functional subclass of the RTVL-H family of human endogenous retrovirus-like sequences. Journal of Virology. 1993;67:2981–9.

104. Jern P, Sperber GO, Blomberg J. Definition and variation of human endogenous retrovirus H. Virology. 2004;327:93–110.

105. Nelson DT, Goodchild NL, Mager DL. Gain of Sp1 Sites and Loss of Repressor Sequences Associated with a Young, Transcriptionally Active Subset of HERV-H Endogenous Long Terminal Repeats. Virology. 1996;220:213–8.

106. Sjøttem E, Anderssen S, Johansen T. The promoter activity of long terminal repeats of the HERV-H family of human retrovirus-like elements is critically dependent on Sp1 family proteins interacting with a GC/GT box located immediately 3’ to the TATA box. Journal of Virology. 1996;70:188–98.

107. O’Connor L, Gilmour J, Bonifer C. The Role of the Ubiquitously Expressed Transcription Factor Sp1 in Tissue-specific Transcriptional Regulation and in Disease. Yale J Biol Med. 2016;89:513–25.

108. Gemmell P, Hein J, Katzourakis A. The Exaptation of HERV-H: Evolutionary Analyses Reveal the Genomic Features of Highly Transcribed Elements. Frontiers in Immunology. 2019;10.

109. Ramsay RG, Gonda TJ. MYB function in normal and cancer cells. Nat Rev Cancer. 2008;8:523–34.

110. Göke J, Lu X, Chan Y-S, Ng H-H, Ly L-H, Sachs F, et al. Dynamic Transcription of Distinct Classes of Endogenous Retroviral Elements Marks Specific Populations of Early Human Embryonic Cells. Cell Stem Cell. 2015;16:135–41.

111. Scherer WF, Syverton JT, Gey GO. STUDIES ON THE PROPAGATION IN VITRO OF POLIOMYELITIS VIRUSES. J Exp Med. 1953;97:695–710.

112. Graham FL, Smiley J, Russell WC, Nairn RY 1977. Characteristics of a Human Cell Line Transformed by DNA from Human Adenovirus Type 5. Journal of General Virology. 36:59–72.

113. Wilkinson DA, Freeman JD, Goodchild NL, Kelleher CA, Mager DL. Autonomous expression of RTVL-H endogenous retroviruslike elements in human cells. Journal of Virology. 1990;64:2157–67.

114. Feuchter A, Mager D. Functional heterogeneity of a large family of human LTR-like promoters and enhancers. Nucleic Acids Research. 1990;18:1261–70.

115. Ramsay L, Marchetto MC, Caron M, Chen S-H, Busche S, Kwan T, et al. Conserved expression of transposon-derived non-coding transcripts in primate stem cells. BMC Genomics. 2017;18:214.

116. Hsieh F-K, Ji F, Damle M, Sadreyev RI, Kingston RE. HERVH-derived lncRNAs negatively regulate chromatin targeting and remodeling mediated by CHD7. Life Science Alliance. 2022;5:e202101127–e202101127.

117. Yu C, Lei X, Chen F, Mao S, Lv L, Liu H, et al. ARID1A loss derepresses a group of human endogenous retrovirus-H loci to modulate BRD4-dependent transcription. Nat Commun. 2022;13:3501.

118. Zhang Y, Li T, Preissl S, Amaral ML, Grinstein JD, Farah EN, et al. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. Nature Genetics 2019 51:9. 2019;51:1380–8.

119. Goodchild NL, Wilkinson DA, Mager DL. A human endogenous long terminal repeat provides a polyadenylation signal to a novel, alternatively spliced transcript in normal placenta. Gene. 1992;121:287–94.

120. Feuchter-Murthy AE, Freeman JD, Mager DL. Splicing of a human endogenous retrovirus to a novel phospholipase A2 related gene. Nucleic Acids Research. 1993;21:135–43.

121. Kowalski PE, Freeman JD, Nelson DT, Mager DL. Genomic Structure and Evolution of a Novel Gene (PLA2L) with Duplicated Phospholipase A2-like Domains. Genomics. 1997;39:38–46.

122. Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biology. 2012;13:R107.

123. Loewer S, Cabili MN, Guttman M, Loh Y-H, Thomas K, Park IH, et al. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Nat Genet. 2010;42:1113–7.

124. Wang Y, Xu Z, Jiang J, Xu C, Kang J, Xiao L, et al. Endogenous miRNA Sponge lincRNA-RoR Regulates Oct4, Nanog, and Sox2 in Human Embryonic Stem Cell Self-Renewal. Developmental Cell. 2013;25:69–80.

125. Takahashi K, Nakamura M, Okubo C, Kliesmete Z, Ohnuki M, Narita M, et al. The pluripotent stem cell-specific transcript ESRG is dispensable for human pluripotency. PLOS Genetics. 2021;17:e1009587.

126. Dombroski BA, Mathias SL, Nanthakumar E, Scott AF, Kazazian HH. Isolation of an Active Human Transposable Element. Science. 1991;254:1805–8.

127. Kolosha VO, Martin SL. High-affinity, Non-sequence-specific RNA Binding by the Open Reading Frame 1 (ORF1) Protein from Long Interspersed Nuclear Element 1 (LINE-1) *. Journal of Biological Chemistry. 2003;278:8112–7.

128. Cook PR, Jones CE, Furano AV. Phosphorylation of ORF1p is required for L1 retrotransposition. Proceedings of the National Academy of Sciences. 2015;112:4298–303.

129. Feng Q, Moran JV, Kazazian HH, Boeke JD. Human L1 Retrotransposon Encodes a Conserved Endonuclease Required for Retrotransposition. Cell. 1996;87:905–16.

130. Mathias SL, Scott AF, Kazazian HH, Boeke JD, Gabriel A. Reverse Transcriptase Encoded by a Human Transposable Element. Science. 1991;254:1808–10.

131. Swergold GD. Identification, characterization, and cell specificity of a human LINE-1 promoter. Molecular and Cellular Biology. 1990;10:6718–29.

132. Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH. High Frequency Retrotransposition in Cultured Mammalian Cells. Cell. 1996;87:917–27.

133. Hohjoh H, Singer MF. Ribonuclease and high salt sensitivity of the ribonucleoprotein complex formed by the human LINE-1 retrotransposon11Edited By D. E. Draper. Journal of Molecular Biology. 1997;271:7–12.

134. Gilbert N, Lutz S, Morrish TA, Moran JV. Multiple Fates of L1 Retrotransposition Intermediates in Cultured Human Cells. Molecular and Cellular Biology. 2005;25:7780–95.

135. Ostertag EM, Kazazian HH. Twin Priming: A Proposed Mechanism for the Creation of Inversions in L1 Retrotransposition. Genome Res. 2001;11:2059–65.

136. Ullu E, Tschudi C. Alu sequences are processed 7SL RNA genes. Nature. 1984;312:171–2.

137. Ahl V, Keller H, Schmidt S, Weichenrieder O. Retrotransposition and Crystal Structure of an Alu RNP in the Ribosome-Stalling Conformation. Molecular Cell. 2015;60:715–27.

138. Deininger P. Alu elements: know the SINEs. Genome Biology. 2011;12:236.

139. Dewannieux M, Esnault C, Heidmann T. LINE-mediated retrotransposition of marked Alu sequences. Nat Genet. 2003;35:41–8.

140. Hancks DC, Kazazian HH. Active human retrotransposons: variation and disease. Current Opinion in Genetics & Development. 2012;22:191–203.

141. Raiz J, Damert A, Chira S, Held U, Klawitter S, Hamdorf M, et al. The non-autonomous retrotransposon SVA is trans -mobilized by the human LINE-1 protein machinery. Nucleic Acids Research. 2012;40:1666–83.

142. Bodak M, Yu J, Ciaudo C. Regulation of LINE-1 in mammals. Biomolecular Concepts. 2014;5:409–28.

143. Kobayashi K, Nakahori Y, Miyake M, Matsumura K, Kondo-Iida E, Nomura Y, et al. An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. Nature. 1998;394:388–92.

144. Taniguchi-Ikeda M, Kobayashi K, Kanagawa M, Yu C, Mori K, Oda T, et al. Pathogenic exon-trapping by SVA retrotransposon and rescue in Fukuyama muscular dystrophy. Nature. 2011;478:127–31.

145. Lee J, Han K, Meyer TJ, Kim H-S, Batzer MA. Chromosomal Inversions between Human and Chimpanzee Lineages Caused by Retrotransposons. PLOS ONE. 2008;3:e4047.

146. Boissinot S, Davis J, Entezam A, Petrov D, Furano AV. Fitness cost of LINE-1 (L1) activity in humans. Proceedings of the National Academy of Sciences. 2006;103:9590–4.

147. Xing J, Wang H, Belancio VP, Cordaux R, Deininger PL, Batzer MA. Emergence of primate genes by retrotransposon-mediated sequence transduction. Proceedings of the National Academy of Sciences. 2006;103:17608–13.

148. De Cecco M, Ito T, Petrashen AP, Elias AE, Skvir NJ, Criscione SW, et al. L1 drives IFN in senescent cells and promotes age-associated inflammation. Nature. 2019;566:73–8.

149. Simon M, Meter MV, Ablaeva J, Ke Z, Gonzalez RS, Taguchi T, et al. LINE1 Derepression in Aged Wild-Type and SIRT6-Deficient Mice Drives Inflammation. Cell Metabolism. 2019;29:871-885.e5.

150. Deniz Ö, Frost JM, Branco MR. Regulation of transposable elements by DNA modifications. Nat Rev Genet. 2019;20:417–31.

151. Jönsson ME, Ludvik Brattås P, Gustafsson C, Petri R, Yudovich D, Pircs K, et al. Activation of neuronal genes via LINE-1 elements upon global DNA demethylation in human neural progenitors. Nat Commun. 2019;10:3182.

152. Castro-Diaz N, Ecco G, Coluccio A, Kapopoulou A, Yazdanpanah B, Friedli M, et al. Evolutionally dynamic L1 regulation in embryonic stem cells. Genes Dev. 2014;28:1397–409.

153. Quenneville S, Verde G, Corsinotti A, Kapopoulou A, Jakobsson J, Offner S, et al. In Embryonic Stem Cells, ZFP57/KAP1 Recognize a Methylated Hexanucleotide to Affect Chromatin and DNA Methylation of Imprinting Control Regions. Molecular Cell. 2011;44:361–72.

154. Robbez-Masson L, Tie CHC, Conde L, Tunbak H, Husovsky C, Tchasovnikarova IA, et al. The HUSH complex cooperates with TRIM28 to repress young retrotransposons and new genes. Genome Res. 2018;28:836–45.

155. Matsui T, Leung D, Miyashita H, Maksakova IA, Miyachi H, Kimura H, et al. Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. Nature. 2010;464:927–31.

156. Schultz DC, Friedman JR, Rauscher FJ. Targeting histone deacetylase complexes via KRAB-zinc finger proteins: the PHD and bromodomains of KAP-1 form a cooperative unit that recruits a novel isoform of the Mi-2α subunit of NuRD. Genes Dev. 2001;15:428–43.

157. Sripathy SP, Stevens J, Schultz DC. The KAP1 Corepressor Functions To Coordinate the Assembly of De Novo HP1-Demarcated Microenvironments of Heterochromatin Required for KRAB Zinc Finger Protein-Mediated Transcriptional Repression. Molecular and Cellular Biology. 2006;26:8623–38.

158. Ohtani H, Liu M, Zhou W, Liang G, Jones PA. Switching roles for DNA and histone methylation depend on evolutionary ages of human endogenous retroviruses. Genome Res. 2018;28:1147–57.

159. Russell SJ, LaMarre J. Transposons and the PIWI pathway: genome defense in gametes and embryos. Reproduction. 2018;156:R111–24.

160. Carmell MA, Girard A, van de Kant HJG, Bourc'his D, Bestor TH, de Rooij DG, et al. MIWI2 Is Essential for Spermatogenesis and Repression of Transposons in the Mouse Male Germline. Developmental Cell. 2007;12:503–14.

161. De Fazio S, Bartonicek N, Di Giacomo M, Abreu-Goodger C, Sankar A, Funaya C, et al. The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements. Nature. 2011;480:259–63.

162. Roovers EF, Rosenkranz D, Mahdipour M, Han C-T, He N, Chuva de Sousa Lopes SM, et al. Piwi Proteins and piRNAs in Mammalian Oocytes and Early Embryos. Cell Reports. 2015;10:2069–82.

163. Marchetto MCN, Narvaiza I, Denli AM, Benner C, Lazzarini TA, Nathanson JL, et al. Differential L1 regulation in pluripotent stem cells of humans and apes. Nature. 2013;503:525–9.

164. Sasaki T, Shiohama A, Minoshima S, Shimizu N. Identification of eight members of the Argonaute family in the human genome☆. Genomics. 2003;82:323–30.

165. Orecchini E, Frassinelli L, Galardi S, Ciafrè SA, Michienzi A. Post-transcriptional regulation of LINE-1 retrotransposition by AID/APOBEC and ADAR deaminases. Chromosome Res. 2018;26:45–59.

166. Bogerd HP, Wiegand HL, Hulme AE, Garcia-Perez JL, O'Shea KS, Moran JV, et al. Cellular inhibitors of long interspersed element 1 and Alu retrotransposition. Proceedings of the National Academy of Sciences. 2006;103:8780–5.

167. Chen H, Lilley CE, Yu Q, Lee DV, Chou J, Narvaiza I, et al. APOBEC3A Is a Potent Inhibitor of Adeno-Associated Virus and Retrotransposons. Current Biology. 2006;16:480–5.

168. Feng Y, Goubran MH, Follack TB, Chelico L. Deamination-independent restriction of LINE-1 retrotransposition by APOBEC3H. Sci Rep. 2017;7:10881.

169. Horn AV, Klawitter S, Held U, Berger A, Jaguva Vasudevan AA, Bock A, et al. Human LINE-1 restriction by APOBEC3C is deaminase independent and mediated by an ORF1p interaction that affects LINE reverse transcriptase activity. Nucleic Acids Research. 2014;42:396–416.

170. Ikeda T, Abd El Galil KH, Tokunaga K, Maeda K, Sata T, Sakaguchi N, et al. Intrinsic restriction activity by apolipoprotein B mRNA editing enzyme APOBEC1 against the mobility of autonomous retrotransposons. Nucleic Acids Research. 2011;39:5538–54.

171. Kinomoto M, Kanno T, Shimura M, Ishizaka Y, Kojima A, Kurata T, et al. All APOBEC3 family proteins differentially inhibit LINE-1 retrotransposition. Nucleic Acids Research. 2007;35:2955–64.

172. Koyama T, Arias JF, Iwabu Y, Yokoyama M, Fujita H, Sato H, et al. APOBEC3G Oligomerization Is Associated with the Inhibition of Both Alu and LINE-1 Retrotransposition. PLOS ONE. 2013;8:e84228.

173. Liang W, Xu J, Yuan W, Song X, Zhang J, Wei W, et al. APOBEC3DE Inhibits LINE-1 Retrotransposition by Interacting with ORF1p and Influencing LINE Reverse Transcriptase Activity. PLOS ONE. 2016;11:e0157220.

174. Muckenfuss H, Hamdorf M, Held U, Perković M, Löwer J, Cichutek K, et al. APOBEC3 Proteins Inhibit Human LINE-1 Retrotransposition*. Journal of Biological Chemistry. 2006;281:22161–72.

175. Niewiadomska AM, Tian C, Tan L, Wang T, Sarkis PTN, Yu X-F. Differential Inhibition of Long Interspersed Element 1 by APOBEC3 Does Not Correlate with High-Molecular-Mass-Complex Formation or P-Body Association. Journal of Virology. 2007;81:9577–83.

176. Richardson SR, Narvaiza I, Planegger RA, Weitzman MD, Moran JV. APOBEC3A deaminates transiently exposed single-strand DNA during LINE-1 retrotransposition. eLife. 2014;3:e02008.

177. Stenglein MD, Harris RS. APOBEC3B and APOBEC3F Inhibit L1 Retrotransposition by a DNA Deamination-independent Mechanism *. Journal of Biological Chemistry. 2006;281:16837–41.

178. Athanasiadis A, Rich A, Maas S. Widespread A-to-I RNA Editing of Alu-Containing mRNAs in the Human Transcriptome. PLoS Biol. 2004;2:e391.

179. Bazak L, Levanon EY, Eisenberg E. Genome-wide analysis of Alu editability. Nucleic Acids Research. 2014;42:6876–84.

180. Blow M, Futreal PA, Wooster R, Stratton MR. A survey of RNA editing in human brain. Genome Res. 2004;14:2379–87.

181. Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallegger M, Shemesh R, et al. Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat Biotechnol. 2004;22:1001–5.

182. Platt RN, Vandewege MW, Ray DA. Mammalian transposable elements and their impacts on genome evolution. Chromosome Res. 2018;26:25–43.

183. Lowe CB, Bejerano G, Haussler D. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proceedings of the National Academy of Sciences. 2007;104:8005–10.

184. Samuelson LC, Wiebauer K, Snow CM, Meisler MH. Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution. Molecular and Cellular Biology. 1990;10:2513–20.

185. Ting CN, Rosenberg MP, Snow CM, Samuelson LC, Meisler MH. Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. Genes Dev. 1992;6:1457–65.

186. Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature. 2011;479:534–7.

187. Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, et al. L1 retrotransposition in human neural progenitor cells. Nature. 2009;460:1127–31.

188. Evrony GD, Cai X, Lee E, Hills LB, Elhosary PC, Lehmann HS, et al. Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain. Cell. 2012;151:483–96.

189. Muotri AR, Chu VT, Marchetto MCN, Deng W, Moran JV, Gage FH. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature. 2005;435:903–10.

190. Upton KR, Gerhardt DJ, Jesuadian JS, Richardson SR, Sánchez-Luque FJ, Bodea GO, et al. Ubiquitous L1 Mosaicism in Hippocampal Neurons. Cell. 2015;161:228–39.

191. Bedrosian TA, Quayle C, Novaresi N, Gage FredH. Early life experience drives structural variation of neural genomes in mice. Science. 2018;359:1395–9.

192. Petropoulos S, Edsgärd D, Reinius B, Deng Q, Panula SP, Codeluppi S, et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. Cell. 2016;165:1012–26.

193. Ohnuki M, Tanabe K, Sutou K, Teramoto I, Sawamura Y, Narita M, et al. Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. Proceedings of the National Academy of Sciences. 2014;111:12426–31.

194. Lin AJ, Slack NL, Ahmad A, George CX, Samuel CE, Safinya CR. Three-Dimensional Imaging of Lipid Gene-Carriers: Membrane Charge Density Controls Universal Transfection Behavior in Lamellar Cationic Liposome-DNA Complexes. Biophysical Journal. 2003;84:3307–16.

195. Uemura M, Zheng Q, Koh CM, Nelson WG, Yegnasubramanian S, De Marzo AM. Overexpression of ribosomal RNA in prostate cancer is common but not linked to rDNA promoter hypomethylation. Oncogene. 2012;31:1254–63.

196. Outram SV, Varas A, Pepicelli CV, Crompton T. Hedgehog Signaling Regulates Differentiation from Double-Negative to Double-Positive Thymocyte. Immunity. 2000;13:187–97.

197. Wu B, Cao X, Liang X, Zhang X, Zhang W, Sun G, et al. Epigenetic Regulation of Elf5 Is Associated with Epithelial-Mesenchymal Transition in Urothelial Cancer. PLoS One. 2015;10:e0117510.

198. Saryu Malhotra S, Suman P, Kumar Gupta S. Alpha or beta human chorionic gonadotropin knockdown decrease BeWo cell fusion by down-regulating PKA and CREB activation. Sci Rep. 2015;5:11210.

199. Huang H-W, Chen C-Y, Huang Y-H, Yeh C-T, Wang C-S, Chang C-C, et al. CMAHP promotes metastasis by reducing ubiquitination of Snail and inducing angiogenesis via GM-CSF overexpression in gastric cancer. Oncogene. 2022;41:159–72.

200. Livak KJ, Schmittgen TD. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. Methods. 2001;25:402–8.

201. Russell DW, Sambrook J. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Cold Spring Harbor, NY; 2001.

202. Elyanow R, Wu H-T, Raphael BJ. Identifying structural variants using linked-read sequencing data. Bioinformatics. 2018;34:353–60.

203. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.

204. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 2009;10:R25.

205. Smit A, Hubley R, Green P. RepeatMasker. 2013.

206. McLeay RC, Bailey TL. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics. 2010;11:165.

207. Butash KA, Natarajan P, Young A, Fox DK. Reexamination of the Effect of Endotoxin on Cell Proliferation and Transfection Efficiency. BioTechniques. 2000;29:610–9.

208. Bertero A, Madrigal P, Galli A, Hubner NC, Moreno I, Burks D, et al. Activin/Nodal signaling and NANOG orchestrate human embryonic stem cell fate decisions by controlling the H3K4me3 chromatin mark. Genes Dev. 2015;29:702–17.

209. Babaie Y, Herwig R, Greber B, Brink TC, Wruck W, Groth D, et al. Analysis of Oct4-Dependent Transcriptional Networks Regulating Self-Renewal and Pluripotency in Human Embryonic Stem Cells. Stem Cells. 2007;25:500–10.

210. Kuroda T, Tada M, Kubota H, Kimura H, Hatano S, Suemori H, et al. Octamer and Sox Elements Are Required for Transcriptional cis Regulation of Nanog Gene Expression. Molecular and Cellular Biology. 2005;25:2475–85.

211. Rodda DJ, Chew J-L, Lim L-H, Loh Y-H, Wang B, Ng H-H, et al. Transcriptional Regulation of Nanog by OCT4 and SOX2 *. Journal of Biological Chemistry. 2005;280:24731–7.

212. Wang Z, Oron E, Nelson B, Razis S, Ivanova N. Distinct Lineage Specification Roles for NANOG, OCT4, and SOX2 in Human Embryonic Stem Cells. Cell Stem Cell. 2012;10:440–54.

213. Zhang X, Huang CT, Chen J, Pankratz MT, Xi J, Li J, et al. Pax6 Is a Human Neuroectoderm Cell Fate Determinant. Cell Stem Cell. 2010;7:90–100.

214. Pevny LH, Sockanathan S, Placzek M, Lovell-Badge R. A role for SOX1 in neural determination. Development. 1998;125:1967–78.

215. Richter A, Valdimarsdottir L, Hrafnkelsdottir HE, Runarsson JF, Omarsdottir AR, Oostwaard DW, et al. BMP4 Promotes EMT and Mesodermal Commitment in Human Embryonic Stem Cells via SLUG and MSX2. Stem Cells. 2014;32:636–48.

216. Zhang P, Li J, Tan Z, Wang C, Liu T, Chen L, et al. Short-term BMP-4 treatment initiates mesoderm induction in human embryonic stem cells. Blood. 2008;111:1933–41.

217. Calero-Nieto FJ, Joshi A, Bonadies N, Kinston S, Chan W-I, Gudgin E, et al. HOX-mediated LMO2 expression in embryonic mesoderm is recapitulated in acute leukaemias. Oncogene. 2013;32:5471–80.

218. Gering M, Yamada Y, Rabbitts TH, Patient RK. Lmo2 and Scl/Tal1 convert non-axial mesoderm into haemangioblasts which differentiate into endothelial cells in the absence of Gata1. Development. 2003;130:6187–99.

219. Hyslop L, Stojkovic M, Armstrong L, Walter T, Stojkovic P, Przyborski S, et al. Downregulation of NANOG Induces Differentiation of Human Embryonic Stem Cells to Extraembryonic Lineages. Stem Cells. 2005;23:1035–43.

220. Zhou Z, Gong M, Pande A, Lisewski U, Röpke T, Purfürst B, et al. A missense KCNQ1 Mutation Impairs Insulin Secretion in Neonatal Diabetes. 2021;:2021.08.24.457485.

221. Ostertag EM, Goodier JL, Zhang Y, Kazazian HH. SVA Elements Are Nonautonomous Retrotransposons that Cause Disease in Humans. The American Journal of Human Genetics. 2003;73:1444–51.

222. Ostertag EM, Luning Prak ET, DeBerardinis RJ, Moran JV, Kazazian Jr HH. Determination of L1 retrotransposition kinetics in cultured cells. Nucleic Acids Research. 2000;28:1418–23.

223. Zamai L, Bareggi R, Santavenere E, Vitale M. Subtraction of autofluorescent dead cells from the lymphocyte flow cytometric binding assay. Cytometry. 1993;14:951–4.

224. Xie Y, Rosser JM, Thompson TL, Boeke JD, An W. Characterization of L1 retrotransposition with high-throughput dual-luciferase assays. Nucleic Acids Research. 2011;39:e16.

225. Alexopoulou AN, Couchman JR, Whiteford JR. The CMV early enhancer/chicken β actin (CAG) promoter can be used to drive transgene expression during the differentiation of murine embryonic stem cells into vascular progenitors. BMC Cell Biol. 2008;9:2.

226. Ding S, Wu X, Li G, Han M, Zhuang Y, Xu T. Efficient Transposition of the piggyBac (PB) Transposon in Mammalian Cells and Mice. Cell. 2005;122:473–83.

227. Wilson MH, Coates CJ, George AL. PiggyBac Transposon-mediated Gene Transfer in Human Cells. Molecular Therapy. 2007;15:139–45.

228. Colbére-Garapin F, Horodniceanu F, Kourilsky P, Garapin A-C. A new dominant hybrid selective marker for higher eukaryotic cells. Journal of Molecular Biology. 1981;150:1–14.

229. Pieroni L, Fipaldini C, Monciotti A, Cimini D, Sgura A, Fattori E, et al. Targeted Integration of Adeno-Associated Virus-Derived Plasmids in Transfected Human Cells. Virology. 1998;249:249–59.

230. Haqqi TM, Sarkar G, David CS, Sommer SS. Specific amplification with PCR of a refractory segment of genomic DNA. Nucleic Acids Research. 1988;16:11844.

231. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

232. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: A Sequence Logo Generator. Genome Res. 2004;14:1188–90.

233. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res. 2002;12:996–1006.

234. Han JS, Boeke JD. A highly active synthetic mammalian retrotransposon. Nature. 2004;429:314–8.

235. Miranda KC, Huynh T, Tay Y, Ang Y-S, Tam W-L, Thomson AM, et al. A Pattern-Based Method for the Identification of MicroRNA Binding Sites and Their Corresponding Heteroduplexes. Cell. 2006;126:1203–17.

236. Klawitter S, Fuchs NV, Upton KR, Muñoz-Lopez M, Shukla R, Wang J, et al. Reprogramming triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells. Nat Commun. 2016;7:10286.

237. Wissing S, Muñoz-Lopez M, Macia A, Yang Z, Montano M, Collins W, et al. Reprogramming somatic cells into iPS cells activates LINE-1 retroelement mobility. Human Molecular Genetics. 2012;21:208–18.

238. Bhanu NV, Sidoli S, Garcia BA. Histone modification profiling reveals differential signatures associated with human embryonic stem cell self-renewal and differentiation. Proteomics. 2016;16:448–58.

239. Hancks DC, Goodier JL, Mandal PK, Cheung LE, Kazazian HH Jr. Retrotransposition of marked SVA elements by human L1s in cultured cells. Human Molecular Genetics. 2011;20:3386–400.

240. Savage AL, Bubb VJ, Breen G, Quinn JP. Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns. BMC Evolutionary Biology. 2013;13:101.

241. Faulkner GJ, Billon V. L1 retrotransposition in the soma: a field jumping ahead. Mobile DNA. 2018;9:22.

242. Sanchez-Luque FJ, Richardson SR, Faulkner GJ. Retrotransposon Capture Sequencing (RC-Seq): A Targeted, High-Throughput Approach to Resolve Somatic L1 Retrotransposition in Humans. In: Garcia-Pérez JL, editor. Transposons and Retrotransposons: Methods and Protocols. New York, NY: Springer; 2016. p. 47–77.

243. Carreira PE, Ewing AD, Li G, Schauer SN, Upton KR, Fagg AC, et al. Evidence for L1-associated DNA rearrangements and negligible L1 retrotransposition in glioblastoma multiforme. Mobile DNA. 2016;7:21.

244. Schauer SN, Carreira PE, Shukla R, Gerhardt DJ, Gerdes P, Sanchez-Luque FJ, et al. L1 retrotransposition is a common feature of mammalian hepatocarcinogenesis. Genome Res. 2018;28:639–53.

245. Sanchez-Luque FJ, Kempen M-JHC, Gerdes P, Vargas-Landin DB, Richardson SR, Troskie R-L, et al. LINE-1 Evasion of Epigenetic Repression in Humans. Molecular Cell. 2019;75:590-604.e12.

246. Moss EG, Lee RC, Ambros V. The Cold Shock Domain Protein LIN-28 Controls Developmental Timing in C. elegans and Is Regulated by the lin-4 RNA. Cell. 1997;88:637–46.

247. Tsialikas J, Romer-Seibert J. LIN28: roles and regulation in development and beyond. Development. 2015;142:2397–404.

248. Wu K, Ahmad T, Eri R. LIN28A: A multifunctional versatile molecule with future therapeutic potential. World J Biol Chem. 2022;13:35–46.

249. Zhang J, Ratanasirintrawoot S, Chandrasekaran S, Wu Z, Ficarro SB, Yu C, et al. LIN28 Regulates Stem Cell Metabolism and Conversion to Primed Pluripotency. Cell Stem Cell. 2016;19:66–80.

250. Lee YS, Dutta A. The tumor suppressor microRNA let-7 represses the HMGA2 oncogene. Genes Dev. 2007;21:1025–30.

251. Johnson SM, Grosshans H, Shingara J, Byrom M, Jarvis R, Cheng A, et al. RAS Is Regulated by the let-7 MicroRNA Family. Cell. 2005;120:635–47.

252. Shah YM, Morimura K, Yang Q, Tanabe T, Takagi M, Gonzalez FJ. Peroxisome Proliferator-Activated Receptor α Regulates a MicroRNA-Mediated Signaling Cascade Responsible for Hepatocellular Proliferation. Mol Cell Biol. 2007;27:4238–47.

253. Lee H, Han S, Kwon CS, Lee D. Biogenesis and regulation of the let-7 miRNAs and their functional implications. Protein Cell. 2016;7:100–13.

254. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, et al. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature. 2000;403:901–6.

255. Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, et al. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature. 2000;408:86–9.

256. Shyh-Chang N, Daley GQ. Lin28: Primal Regulator of Growth and Metabolism in Stem Cells. Cell Stem Cell. 2013;12:395–406.

257. Mordstein C, Savisaar R, Young RS, Bazile J, Talmane L, Luft J, et al. Codon Usage and Splicing Jointly Influence mRNA Localization. Cell Systems. 2020;10:351-362.e8.

258. Nott A, Le Hir H, Moore MJ. Splicing enhances translation in mammalian cells: an additional function of the exon junction complex. Genes Dev. 2004;18:210–22.

259. Asimi V, Sampath Kumar A, Niskanen H, Riemenschneider C, Hetzel S, Naderi J, et al. Hijacking of transcriptional condensates by endogenous retroviruses. Nat Genet. 2022;54:1238–47.

260. Anderson P, Kedersha N. RNA granules. Journal of Cell Biology. 2006;172:803–8.

261. Buchan JR, Parker R. Eukaryotic Stress Granules: The Ins and Outs of Translation. Molecular Cell. 2009;36:932–41.

262. Balzer E, Moss EG. Localization of the Developmental Timing Regulator Lin28 to mRNP Complexes, P-bodies and Stress Granules. RNA Biology. 2007;4:16–25.

263. Rahkonen N, Stubb A, Malonzo M, Edelman S, Emani MR, Närvä E, et al. Mature Let-7 miRNAs fine tune expression of LIN28B in pluripotent human embryonic stem cells. Stem Cell Research. 2016;17:498–503.

List of publications

1. **Kondrashkina AM**, Antonets KS, Galkin AP, Nizhnikov AA. Prion-like determinant [NSI+] decreases expression of the SUP45 gene in Saccharomyces cerevisiae. [Article in Russian]. Molecular Biology (Mosk). 2014 Sep-Oct;48(5):790-6.

2. Nizhnikov AA, Magomedova ZM, Rubel AA, **Kondrashkina AM**, Inge-Vechtomov SG, Galkin AP. [NSI+] determinant has a pleiotropic phenotypic manifestation that is modulated by SUP35, SUP45, and VTS1 genes. Current Genetics. 2012 Feb;58(1):35-47.

3. Ponomartsev SV, Sinenko SA, Skvortsova EV, Liskovykh MA, Voropaev IN, Savina MM, Kuzmin AA, Kuzmina EY, **Kondrashkina AM**, Larionov V, Kouprina N, Tomilin AN. Human AlphoidtetO Artificial Chromosome as a Gene Therapy Vector for the Developing Hemophilia A Model in Mice. Cells. 2020 Apr 3;9(4). pii: E879.

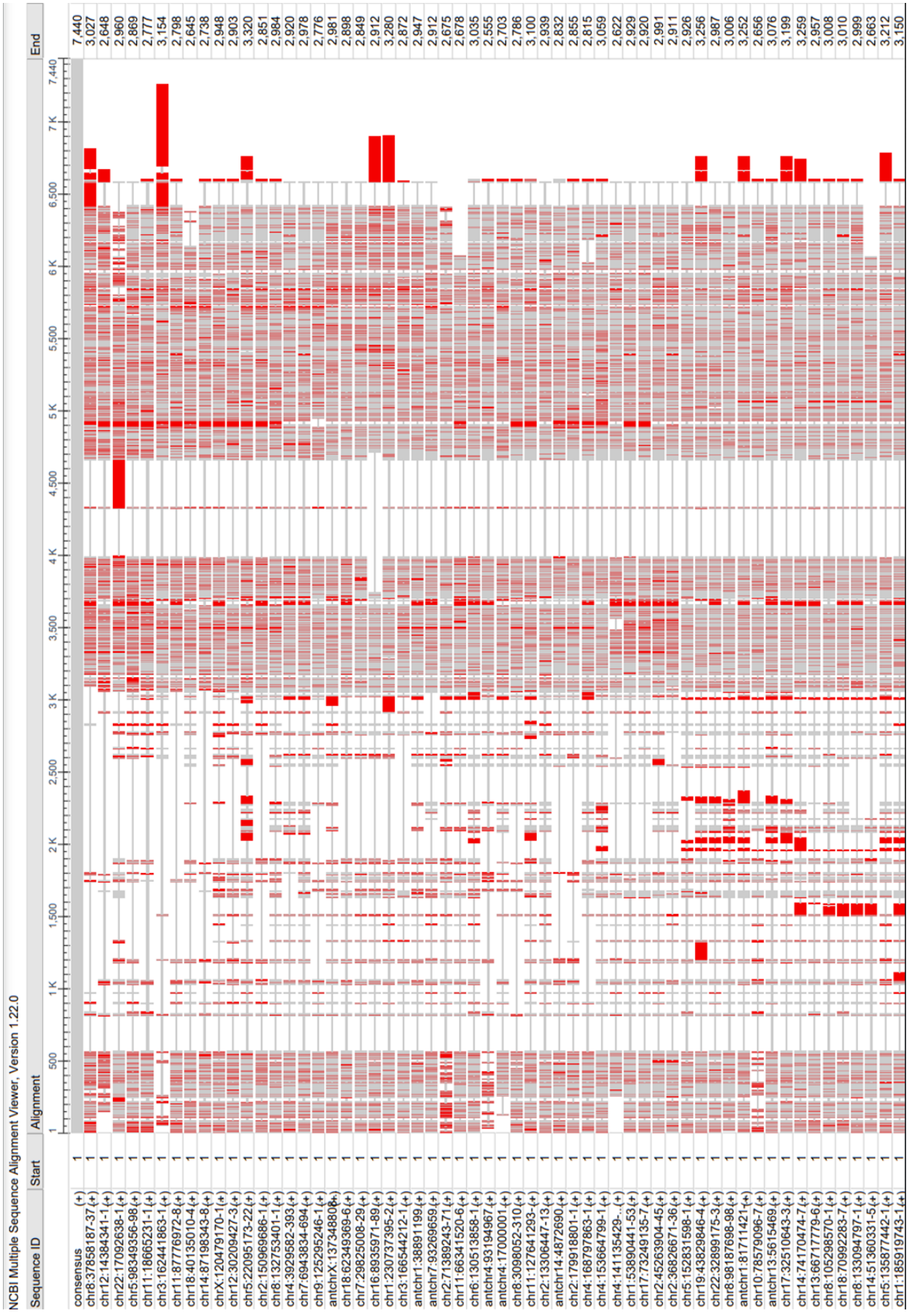
4. Interactions of [NSI+] determinant with SUP35 and VTS1 genes in Saccharomyces cerevisiae. [Article in Russian]. Nizhnikov AA, **Kondrashkina AM**, Galkin AP. Genetika. 2013 Oct;49(10):1155-64.

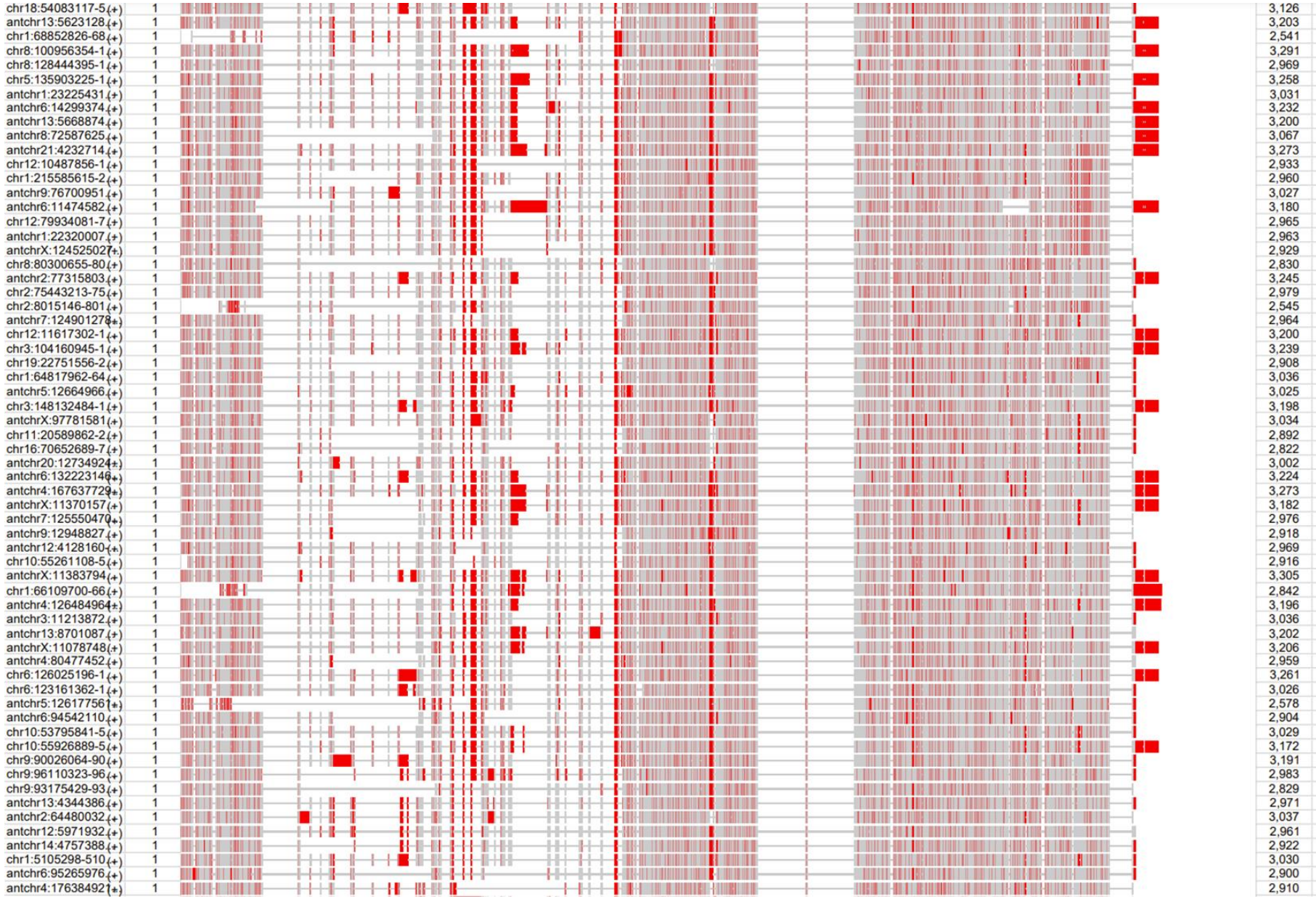
5. Overexpression of genes encoding asparagine-glutamine rich transcriptional factors causes nonsense suppression in Saccharomyces cerevisiae. [Article in Russian]. Nizhnikov AA, **Kondrashkina AM**, Antonets KS, Galkin AP. Ecological Genetics. 2013 Jan;11(1):49-58.

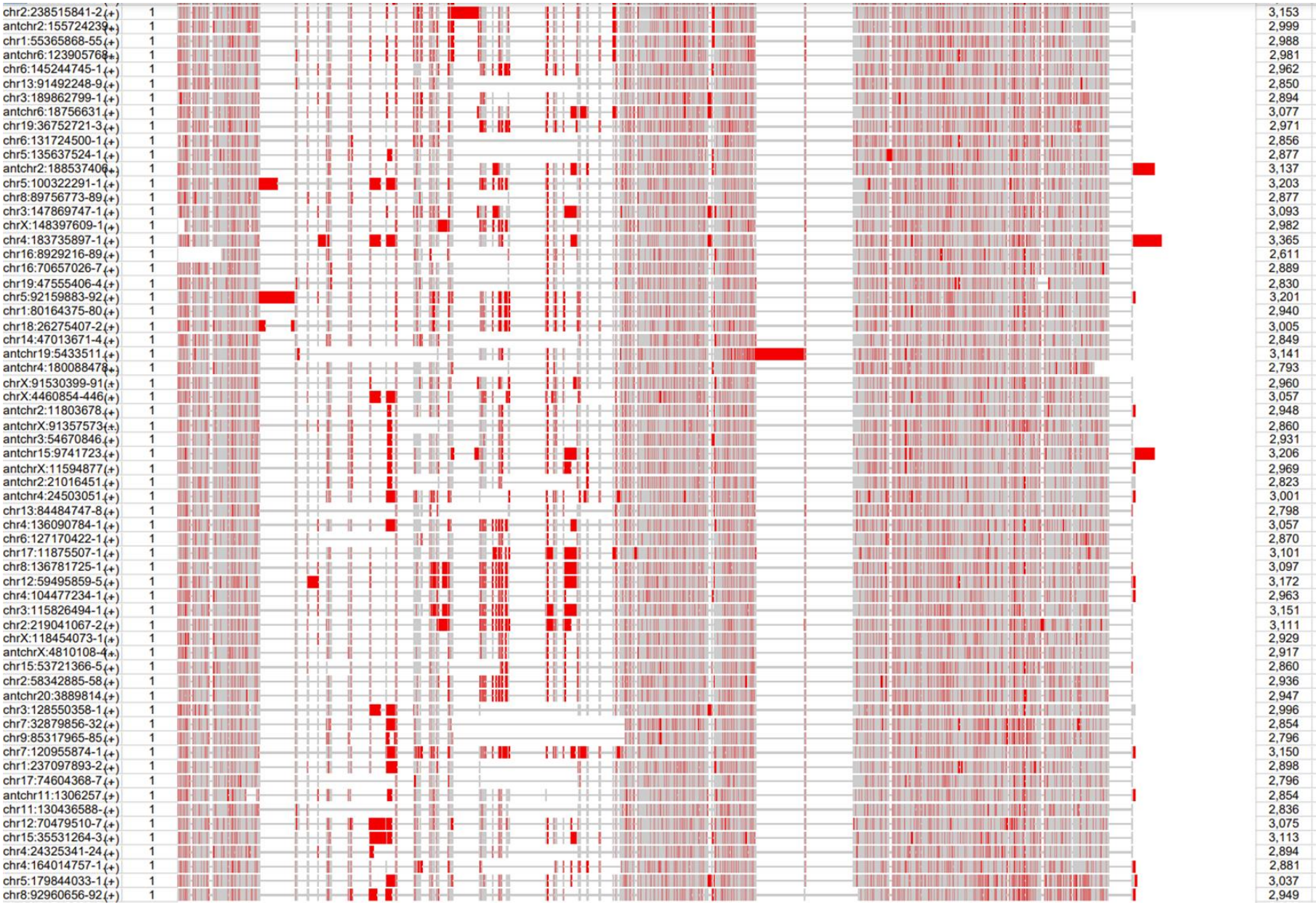
Supplementary I. Predicted Alu integrations

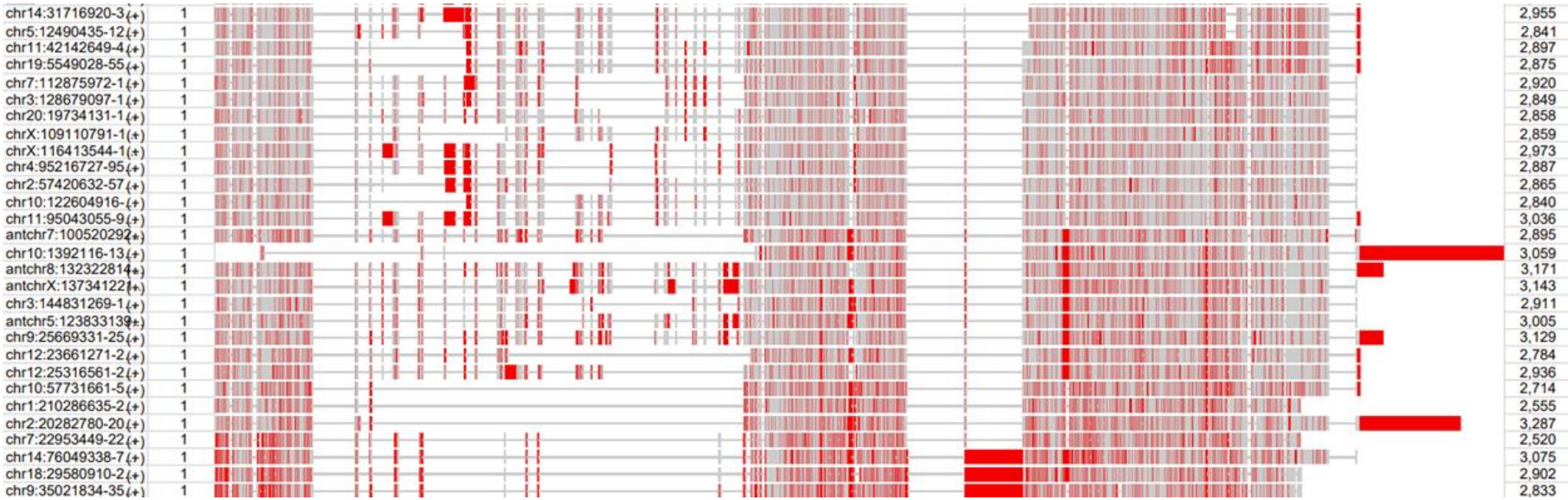
sample	qantiles remapped discordant reads	ID	Chromosome	Left_Extreme	Right_Extreme	X5_Prime_End	X3_Prime_End	Superfamily	Subfamily	TE_Align_Start	TE_Align_End	Orient_5p	Orient_3p	Split_reads_5prime	Split_reads_3prime	Remapped_Discordant
		630514d5-4680-444b-b1e5-8318da33750*	1	36944157	36944362	36944250	36944272	ALU	AluYb9	10	301	NA	-	2	2	5
		39116a3e-e071-4ee3-9c59-0d6f96d30c37	7	139067734	139067942	139067857	139067851	ALU	AluYb9	10	234	-	-	3	1	8
		4da958be-7cd6-4d61-a2c0-0b9387215237	4	71722670	71722896	71722793	71722789	ALU	AluYa5	0	286	+	NA	2	2	10
	0-25%	29e355c6-b2fb-43ec-b7c6-2a716e3e4f0d	16	24216694	24216867	24216744	24216727	ALU	AluYb9	6	229	-	NA	1	3	5
		05364082-7579-4d76-9eaf-301904b233f4*	12	8067281	8067514	8067410	8067426	ALU	AluYb9	0	300	-	NA	1	3	61
		10ef1146-4dcc-4b7e-85f3-0faac7c6de31	19	4899293	4899485	4899401	4899380	ALU	AluYb9	0	298	+	NA	1	3	50
		776ef181-4b14-44eb-ad3e-b2f82e0f02d6*	15	72444984	72445178	72445059	72445078	ALU	AluYb9	0	306	+	+	3	1	32
		49c1b582-7c5b-4949-a3d2-30d69d60031e	12	7838588	7838791	7838673	7838662	ALU	AluYa5	0	294	+	+	2	2	60
clone 1		76f45b03-5378-4a0c-ba64-0796392426e2*	19	30423440	30423681	30423543	30423563	ALU	AluYb9	0	313	-	-	3	1	52
HERVHkd	75-100%	dd84f645-7012-4501-96e0-cf7a860e9e6d	5	175248545	175248751	175248670	175248689	ALU	AluYa5	0	302	-	-	1	3	49
		681c4984-e51a-4e64-a09f-5f2fbdec3f78	19	46139245	46139423	46139387	46139377	ALU	AluYb9	4	204	-	+	2	2	4
		cd47b46a-6fb6-475e-8627-4f7a53762132	13	26109503	26109829	26109695	26109706	ALU	AluYb9	0	246	+	NA	3	1	6
	0-25%	7e4f627d-f321-4db9-9446-b1a248b869a5	9	130884074	130884226	130884190	130884206	ALU	AluYb9	0	309	+	NA	1	3	8
		bd2c13a7-f8e9-4bc4-8e03-ab88125d3257	12	32340950	32341180	32341051	32341039	ALU	AluYb8	0	312	-	-	1	3	58
		b0cf30b7-8977-4d69-8488-a2ae6f8a6210*	12	20090575	20090814	20090677	20090696	ALU	AluYb9	0	303	NA	+	4	1	43
		645ba40d-73d9-4d2a-9b48-66ed2f659dbc*	3	73077736	73077961	73077841	73077849	ALU	AluYb9	0	306	-	NA	3	1	29
clone 3		8203ea94-19ac-4390-9c98-baed1acf6820	17	57814040	57814288	57814161	57814146	ALU	AluYb9	0	321	+	NA	2	2	49
HERVH kd	75-100%	30012f02-27c9-4617-b2ad-ec84da58ccb9	19	10743960	10744222	10744106	10744089	ALU	AluYa5	0	301	-	-	2	3	47
* only not efficient primers possible																

Supplementary II. Full HERVHant alignment to HERVH control

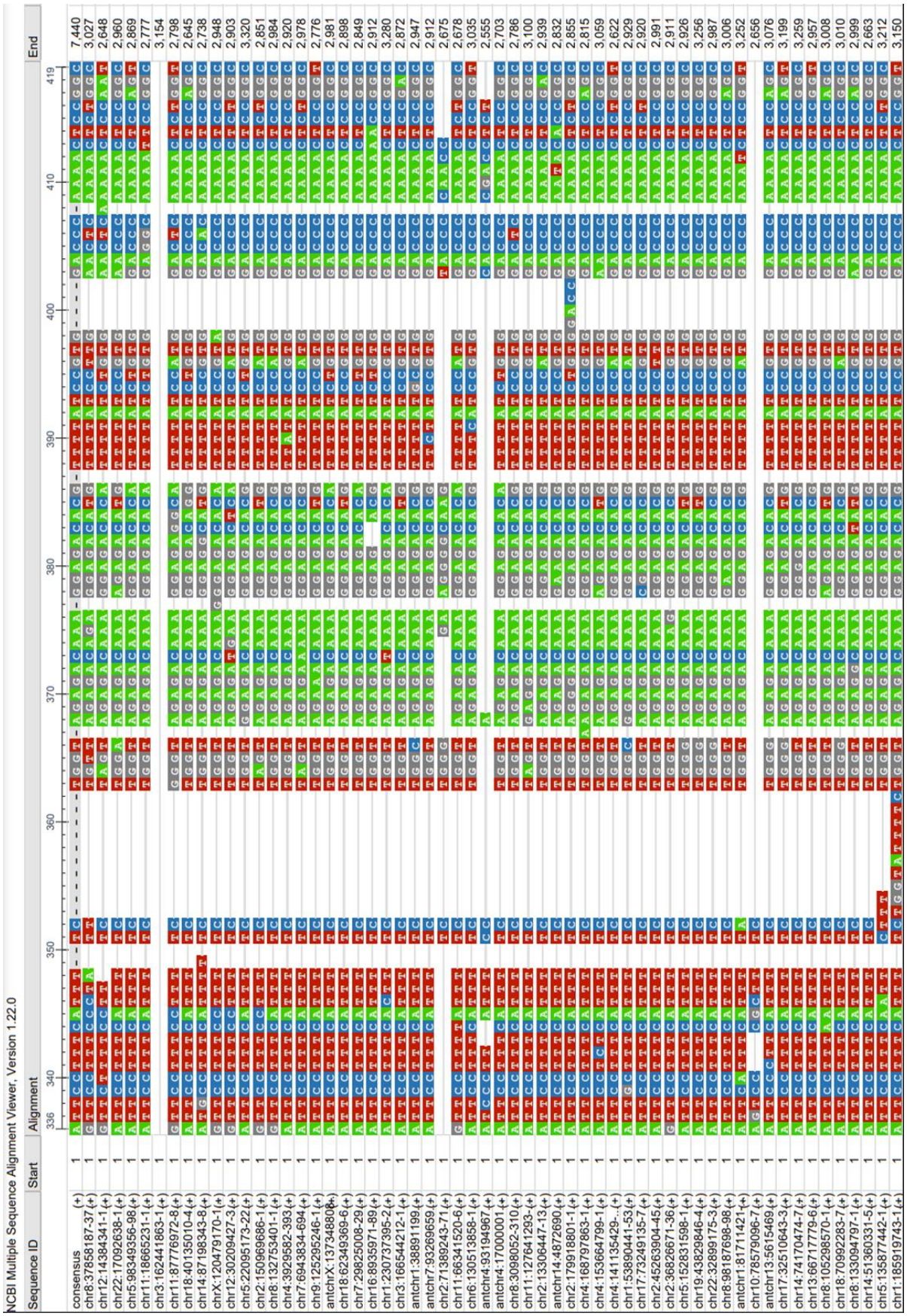


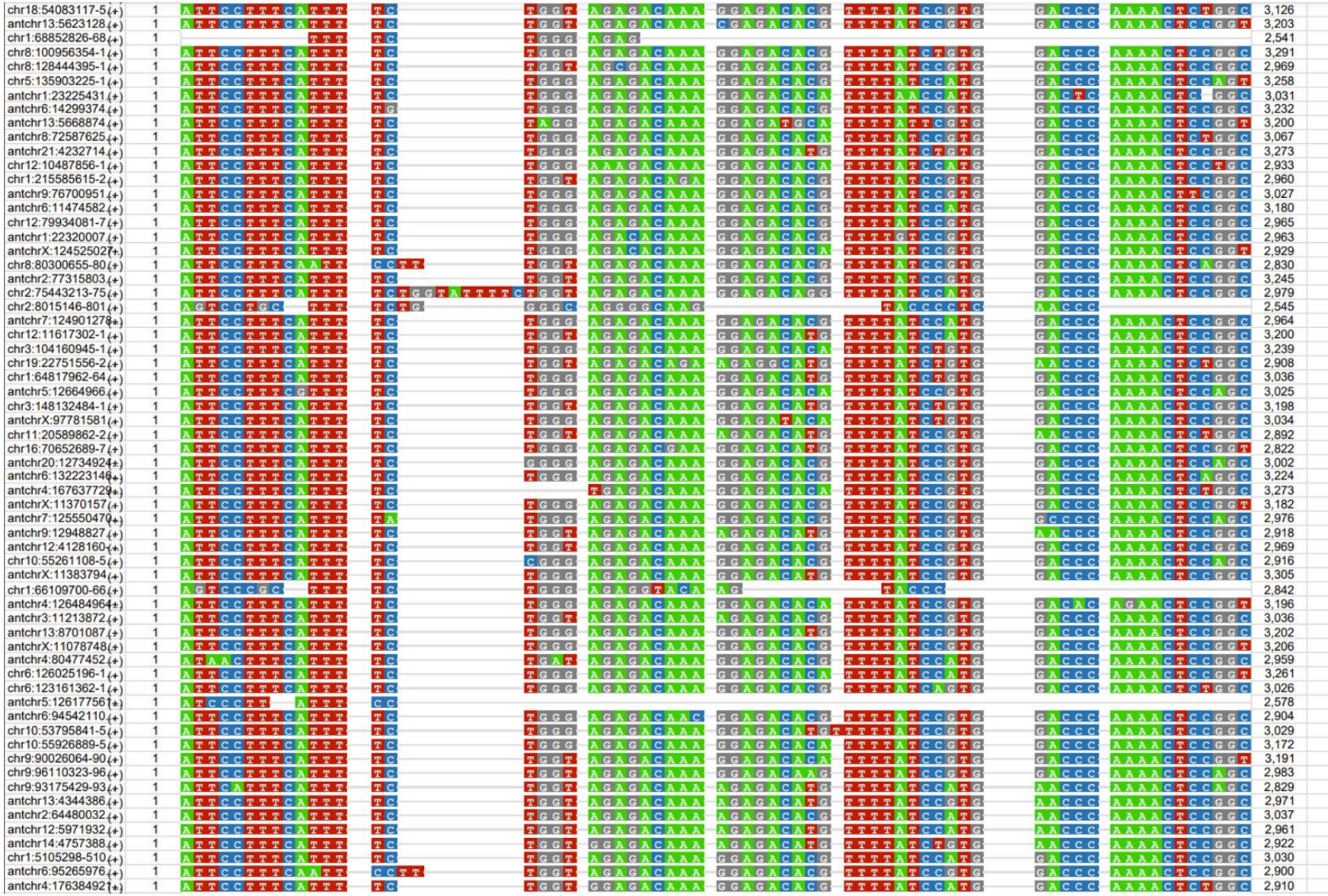

chr18:54083117-5 (+) 1 3,126
antchr13:5623128 (+) 1 3,203
chr1:68852826-68 (+) 1 2,541
chr8:100956354-1 (+) 1 3,291
chr8:128444395-1 (+) 1 2,969
chr5:135903225-1 (+) 1 3,258
antchr1:23225431 (+) 1 3,031
antchr6:14299374 (+) 1 3,232
antchr13:5668874 (+) 1 3,200
antchr8:72587625 (+) 1 3,067
antchr21:4232714 (+) 1 3,273
chr12:10487856-1 (+) 1 2,933
chr1:215585615-2 (+) 1 2,960
antchr9:76700951 (+) 1 3,027
antchr6:11474582 (+) 1 3,180
chr12:79934081-7 (+) 1 2,965
antchr1:22320007 (+) 1 2,963
antchrX:12452502 (+) 1 2,929
chr8:80300655-80 (+) 1 2,830
antchr2:77315803 (+) 1 3,245
chr2:75443213-75 (+) 1 2,979
chr2:8015146-801 (+) 1 2,545
antchr7:124901278 (+) 1 2,964
chr12:11617302-1 (+) 1 3,200
chr3:104160945-1 (+) 1 3,239
chr19:22751556-2 (+) 1 2,908
chr1:64817962-64 (+) 1 3,036
antchr5:12664966 (+) 1 3,025
chr3:148132484-1 (+) 1 3,198
antchrX:97781581 (+) 1 3,034
chr11:20589862-2 (+) 1 2,892
chr16:70652689-7 (+) 1 2,822
antchr20:1273492 (+) 1 3,002
antchr6:132223146 (+) 1 3,224
antchr4:167637729 (+) 1 3,273
antchrX:11370157 (+) 1 3,182
antchr7:125550470 (+) 1 2,976
antchr9:12948827 (+) 1 2,918
antchr12:4128160 (+) 1 2,969
chr10:55261108-5 (+) 1 2,916
antchrX:11383794 (+) 1 3,305
chr1:66109700-66 (+) 1 2,842
antchr4:126484964 (+) 1 3,196
antchr3:11213872 (+) 1 3,036
antchr13:8701087 (+) 1 3,202
antchrX:11078748 (+) 1 3,206
antchr4:80477452 (+) 1 2,959
chr6:126025196-1 (+) 1 3,261
chr6:123161362-1 (+) 1 3,026
antchr5:126177561 (+) 1 2,578
antchr6:94542110 (+) 1 2,904
chr10:53795841-5 (+) 1 3,029
chr10:55926889-5 (+) 1 3,172
chr9:90026064-90 (+) 1 3,191
chr9:96110323-96 (+) 1 2,983
chr9:93175429-93 (+) 1 2,829
antchr13:4344386 (+) 1 2,971
antchr2:64480032 (+) 1 3,037
antchr12:5971932 (+) 1 2,961
antchr14:4757388 (+) 1 2,922
chr1:5105298-510 (+) 1 3,030
antchr6:95265976 (+) 1 2,900
antchr4:17638492 (+) 1 2,910

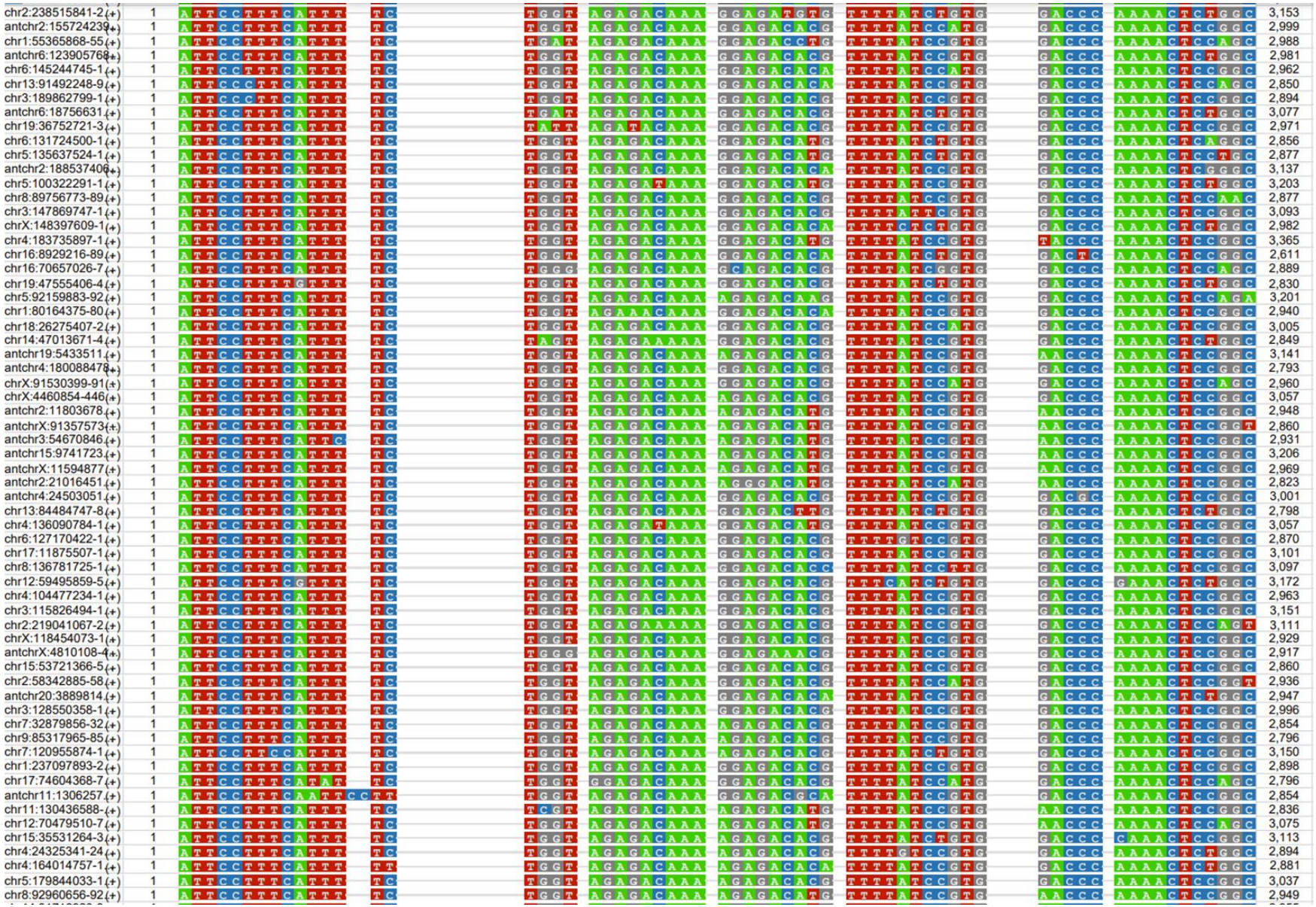

chr2:238515841-2 (+) 1 3,153
antchr2:155724239 (+) 1 2,999
chr1:55365868-55 (+) 1 2,988
antchr6:123905768 (+) 1 2,981
chr6:145244745-1 (+) 1 2,962
chr13:91492248-9 (+) 1 2,850
chr3:189862799-1 (+) 1 2,894
antchr6:18756631 (+) 1 3,077
chr19:36752721-3 (+) 1 2,971
chr6:131724500-1 (+) 1 2,856
chr5:135637524-1 (+) 1 2,877
antchr2:188537406 (+) 1 3,137
chr5:100322291-1 (+) 1 3,203
chr8:89756773-89 (+) 1 2,877
chr3:147869747-1 (+) 1 3,093
chrX:148397609-1 (+) 1 2,982
chr4:183735897-1 (+) 1 3,365
chr16:8929216-89 (+) 1 2,611
chr16:70657026-7 (+) 1 2,889
chr19:47555406-4 (+) 1 2,830
chr5:92159883-92 (+) 1 3,201
chr1:80164375-80 (+) 1 2,940
chr18:26275407-2 (+) 1 3,005
chr14:47013671-4 (+) 1 2,849
antchr19:5433511 (+) 1 3,141
antchr4:180088478 (+) 1 2,793
chrX:91530399-91 (+) 1 2,960
chrX:4460854-446 (+) 1 3,057
antchr2:11803678 (+) 1 2,948
antchrX:91357573 (+) 1 2,860
antchr3:54670846 (+) 1 2,931
antchr15:9741723 (+) 1 3,206
antchrX:11594877 (+) 1 2,969
antchr2:21016451 (+) 1 2,823
antchr4:24503051 (+) 1 3,001
chr13:84484747-8 (+) 1 2,798
chr4:136090784-1 (+) 1 3,057
chr6:127170422-1 (+) 1 2,870
chr17:11875507-1 (+) 1 3,101
chr8:136781725-1 (+) 1 3,097
chr12:59495859-5 (+) 1 3,172
chr4:104477234-1 (+) 1 2,963
chr3:115826494-1 (+) 1 3,151
chr2:219041067-2 (+) 1 3,111
chrX:118454073-1 (+) 1 2,929
antchrX:4810108-4 (+) 1 2,917
chr15:53721366-5 (+) 1 2,860
chr2:58342885-58 (+) 1 2,936
antchr20:3889814 (+) 1 2,947
chr3:128550358-1 (+) 1 2,996
chr7:32879856-32 (+) 1 2,854
chr9:85317965-85 (+) 1 2,796
chr7:120955874-1 (+) 1 3,150
chr1:237097893-2 (+) 1 2,898
chr17:74604368-7 (+) 1 2,796
antchr11:1306257 (+) 1 2,854
chr11:130436588-1 (+) 1 2,836
chr12:70479510-7 (+) 1 3,075
chr15:35531264-3 (+) 1 3,113
chr4:24325341-24 (+) 1 2,894
chr4:164014757-1 (+) 1 2,881
chr5:179844033-1 (+) 1 3,037
chr8:92960656-92 (+) 1 2,949

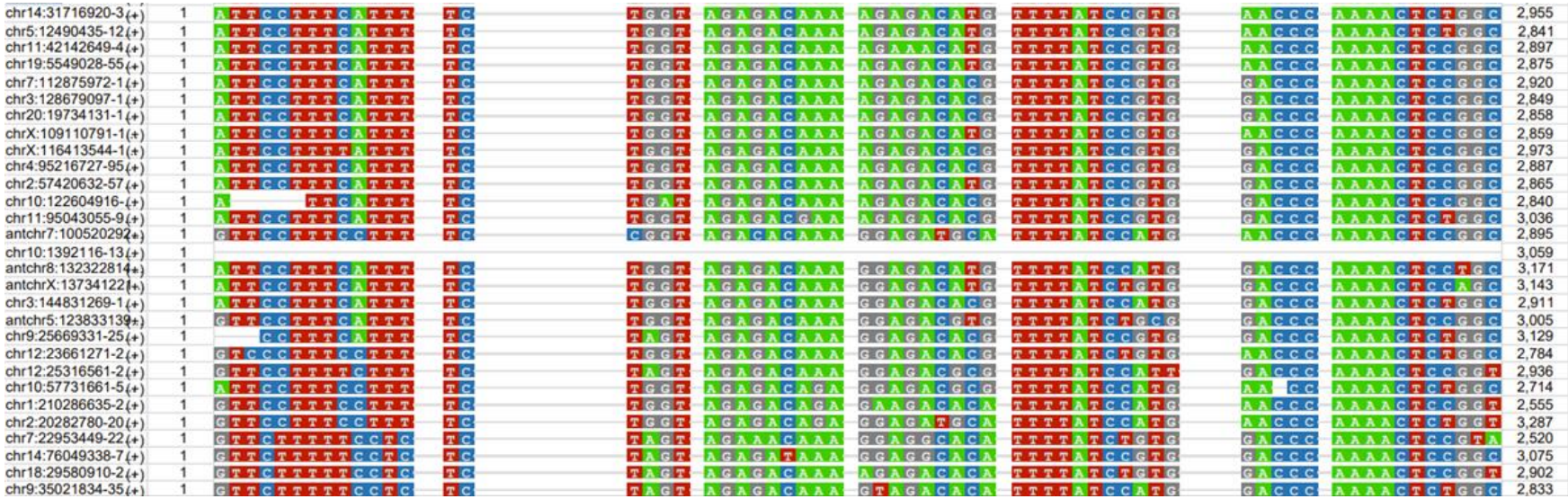

chr14:31716920-3 (+) 1 2,955
chr5:12490435-12 (+) 1 2,841
chr11:42142649-4 (+) 1 2,897
chr19:5549028-55 (+) 1 2,875
chr7:112875972-1 (+) 1 2,920
chr3:128679097-1 (+) 1 2,849
chr20:19734131-1 (+) 1 2,858
chrX:109110791-1 (+) 1 2,859
chrX:116413544-1 (+) 1 2,973
chr4:95216727-95 (+) 1 2,887
chr2:57420632-57 (+) 1 2,865
chr10:122604916- (+) 1 2,840
chr11:95043055-9 (+) 1 3,036
antchr7:100520292 (+) 1 2,895
chr10:1392116-13 (+) 1 3,059
antchr8:13232281 (+) 1 3,171
antchrX:13734122 (+) 1 3,143
chr3:144831269-1 (+) 1 2,911
antchr5:123833139 (+) 1 3,005
chr9:25669331-25 (+) 1 3,129
chr12:23661271-2 (+) 1 2,784
chr12:25316561-2 (+) 1 2,936
chr10:57731661-5 (+) 1 2,714
chr1:210286635-2 (+) 1 2,555
chr2:20282780-20 (+) 1 3,287
chr7:22953449-22 (+) 1 2,520
chr14:76049338-7 (+) 1 3,075
chr18:29580910-2 (+) 1 2,902
chr9:35021834-35 (+) 1 2,833

Supplementary III. Alignment of the *lin* motif region in HERVHlin and HERVHcon




chr18:54083117-5(+) 1 3,126
antchr13:5623128(+) 1 3,203
chr1:68852826-68(+) 1 2,541
chr8:100956354-1(+) 1 3,291
chr8:128444395-1(+) 1 2,969
chr5:135903225-1(+) 1 3,258
antchr1:23225431(+) 1 3,031
antchr6:14299374(+) 1 3,232
antchr13:5668874(+) 1 3,200
antchr8:72587625(+) 1 3,067
antchr21:4232714(+) 1 3,273
chr12:10487856-1(+) 1 2,933
chr1:215585615-2(+) 1 2,960
antchr9:76700951(+) 1 3,027
antchr6:11474582(+) 1 3,180
chr12:79934081-7(+) 1 2,965
antchr1:22320007(+) 1 2,963
antchrX:12452502(+) 1 2,929
chr8:80300655-80(+) 1 2,830
antchr2:77315803(+) 1 3,245
chr2:75443213-75(+) 1 2,979
chr2:8015146-801(+) 1 2,545
antchr7:124901278(+) 1 2,964
chr12:11617302-1(+) 1 3,200
chr3:104160945-1(+) 1 2,239
chr19:22751556-2(+) 1 2,908
chr1:64817962-64(+) 1 3,036
antchr5:12664966(+) 1 3,025
chr3:148132484-1(+) 1 3,198
antchrX:97781581(+) 1 3,034
chr11:20589862-2(+) 1 2,892
chr16:70652689-7(+) 1 2,822
antchr20:12734924(+) 1 3,002
antchr6:132223146(+) 1 3,224
antchr4:167637729(+) 1 3,273
antchrX:11370157(+) 1 3,182
antchr7:125550470(+) 1 2,976
antchr9:12948827(+) 1 2,918
antchr12:4128160(+) 1 2,969
chr10:55261108-5(+) 1 2,916
antchrX:11383794(+) 1 3,305
chr1:66109700-66(+) 1 2,842
antchr4:126484964(+) 1 3,196
antchr3:11213872(+) 1 3,036
antchr13:8701087(+) 1 3,202
antchrX:11078748(+) 1 3,206
antchr4:80477452(+) 1 2,959
chr6:126025196-1(+) 1 3,261
chr6:123161362-1(+) 1 3,026
antchr5:126177561(+) 1 2,578
antchr6:94542110(+) 1 2,904
chr10:53795841-5(+) 1 3,029
chr10:55926889-5(+) 1 3,172
chr9:90026064-90(+) 1 3,191
chr9:96110323-96(+) 1 2,983
chr9:93175429-93(+) 1 2,829
antchr13:4344386(+) 1 2,971
antchr2:64480032(+) 1 3,037
antchr12:5971932(+) 1 2,961
antchr14:4757388(+) 1 2,922
chr1:5105298-510(+) 1 3,030
antchr6:95265976(+) 1 2,900
antchr4:17638492(+) 1 2,910


chr2:238515841-2 (+) 1 3,153
antchr2:155724239 (+) 1 2,999
chr1:55365868-55 (+) 1 2,988
antchr6:123905768 (+) 1 2,981
chr6:145244745-1 (+) 1 2,962
chr13:91492248-9 (+) 1 2,850
chr3:189862799-1 (+) 1 2,894
antchr6:18756631 (+) 1 3,077
chr19:36752721-3 (+) 1 2,971
chr6:131724500-1 (+) 1 2,856
chr5:135637524-1 (+) 1 2,877
antchr2:188537406 (+) 1 3,137
chr5:100322291-1 (+) 1 3,203
chr8:89756773-89 (+) 1 2,877
chr3:147869747-1 (+) 1 3,093
chrX:148397609-1 (+) 1 2,982
chr4:183735897-1 (+) 1 3,365
chr16:8929216-89 (+) 1 2,611
chr16:70657026-7 (+) 1 2,889
chr19:47555406-4 (+) 1 2,830
chr5:92159883-92 (+) 1 3,201
chr1:80164375-80 (+) 1 2,940
chr18:26275407-2 (+) 1 3,005
chr14:47013671-4 (+) 1 2,849
antchr19:5433511 (+) 1 3,141
antchr4:180088478 (+) 1 2,793
chrX:91530399-91 (+) 1 2,960
chrX:4460854-446 (+) 1 3,057
antchr2:11803678 (+) 1 2,948
antchrX:91357573 (+) 1 2,860
antchr3:54670846 (+) 1 2,931
antchr15:9741723 (+) 1 3,206
antchrX:11594877 (+) 1 2,969
antchr2:21016451 (+) 1 2,823
antchr4:24503051 (+) 1 3,001
chr13:84484747-8 (+) 1 2,798
chr4:136090784-1 (+) 1 3,057
chr6:127170422-1 (+) 1 2,870
chr17:11875507-1 (+) 1 3,101
chr8:136781725-1 (+) 1 3,097
chr12:59495859-5 (+) 1 3,172
chr4:104477234-1 (+) 1 2,963
chr3:115826494-1 (+) 1 3,151
chr2:219041067-2 (+) 1 3,111
chrX:118454073-1 (+) 1 2,929
antchrX:4810108-4 (+) 1 2,917
chr15:53721366-5 (+) 1 2,860
chr2:58342885-58 (+) 1 2,936
antchr20:3889814 (+) 1 2,947
chr3:128550358-1 (+) 1 2,996
chr7:32879856-32 (+) 1 2,854
chr9:85317965-85 (+) 1 2,796
chr7:120955874-1 (+) 1 3,150
chr1:237097893-2 (+) 1 2,898
chr17:74604368-7 (+) 1 2,796
antchr11:1306257 (+) 1 2,854
chr11:130436588- (+) 1 2,836
chr12:70479510-7 (+) 1 3,075
chr15:35531264-3 (+) 1 3,113
chr4:24325341-24 (+) 1 2,894
chr4:164014757-1 (+) 1 2,881
chr5:179844033-1 (+) 1 3,037
chr8:92960656-92 (+) 1 2,949


chr14:31716920-3(+) 1 2,955
chr5:12490435-12(+) 1 2,841
chr11:42142649-4(+) 1 2,897
chr19:5549028-55(+) 1 2,875
chr7:112875972-1(+) 1 2,920
chr3:128679097-1(+) 1 2,849
chr20:19734131-1(+) 1 2,858
chrX:109110791-1(+) 1 2,859
chrX:116413544-1(+) 1 2,973
chr4:95216727-95(+) 1 2,887
chr2:57420632-57(+) 1 2,865
chr10:122604916-(+) 1 2,840
chr11:95043055-9(+) 1 3,036
antchr7:100520292(+) 1 2,895
chr10:1392116-13(+) 1 3,059
antchr8:132322814(+) 1 3,171
antchrX:13734122(+) 1 3,143
chr3:144831269-1(+) 1 2,911
antchr5:123833139(+) 1 3,005
chr9:25669331-25(+) 1 3,129
chr12:23661271-2(+) 1 2,784
chr12:25316561-2(+) 1 2,936
chr10:57731661-5(+) 1 2,714
chr1:210286635-2(+) 1 2,555
chr2:20282780-20(+) 1 3,287
chr7:22953449-22(+) 1 2,520
chr14:76049338-7(+) 1 3,075
chr18:29580910-2(+) 1 2,902
chr9:35021834-35(+) 1 2,833

Supplementary IV. Nomenclature and abbreviations

Human genes' names are written in capital letters, italic, e.g., *NANOG*

Human proteins' names are written in capital letters, e.g., NANOG

Mouse genes' names are written with the first capital letter, italic, e.g., *Nanog*

Mouse proteins' names are written with the first capital letter, e.g., Nanog

Long non-coding RNAs' names and chimeric transcripts are written in capital letters, italic, e.g., *ESRG*

Retroelements' names are written in capital letters, e.g., HERVH

microRNAs are named as miR abbreviation, followed by a number, e.g., miR302

Promoters' names are written in capital letters, e.g., CAG or LTR7

HERVH internal regions' names are written in lowercase, italic, e.g., *gag*

Plasmid reporters' names are written in capital letters, e.g., L1-EGFP

(c)DNA – (complementary) deoxyribonucleic acid

(E)GFP – (enhanced) green fluorescent protein

(h)iPSC – (human) induced pluripotent cells

(m)EpiSC – (mouse) epiblast-like stem cell

(q)PCR – (quantitative) polymerase chain reaction

(RNA) Pol II – (RNA) polymerase II

AME – analysis of motif enrichment

bFGF – beta fibroblasts growth factor

Bp – base pair

CAG promoter – CMV immediate enhancer/β-actin promoter

CLIP-seq – cross-linking immunoprecipitation-high-throughput sequencing

CNV – copy number variation

EGA – embryonic genome activation

Epi – epiblast

ERV – endogenous retrovirus

FACS – fluorescence-activated cell sorting

FPKM – fragments per kilobase of exon per million mapped fragments

FSC – formative stem cell

FSC – forward scatter

GMP – good manufacturing practice

HEK – human embryonic kidney

HERVH(K/W) – human endogenous retrovirus H (K/W)

HERVHant – HERVH antagonistic

HERVHint (HERVH-int) – HERVH internal region

HERVHcon – HERVH control

HERVHlin – HERVH with *lin* motif

hESC – human embryonic stem cell

HIV – human immunodeficiency virus

hPSC – human pluripotent stem cell

ICM – inner cell mass

Kb – kilobase

L1 (LINE1) – long interspersed nuclear element

L1Hs – L1 human-specific

LIF – leukemia inhibitory factor

lincRNA – long intergenic non-coding RNA

LncRNA – long non-coding RNA

LTR – long terminal repeat

mESC – mouse embryonic stem cell

MHC – major histocompatibility complex

miRNA – microRNA

MYA – million years ago

ORF – open reading frame

PBS – primer binding site

PE – primitive endoderm

PGC – primordial germ cell

piRNA – Piwi-interacting RNA

RE – retroelement

RIP – RNA Immunoprecipitation

RISC – RNA-induced silencing complex

RNA – ribonucleic acid

RNA-seq – RNA high-throughput sequencing

RNP – ribonuclear particle

SEM – standard error of the mean

shRNA – short hairpin RNA

SNV – single nucleotide variation

SSC – side scatter

SVA – SINE-VNTR-Alus

TAD – transcription-associated domain

TE – trophectoderm

TEs – transposable elements

TPM – transcript per million

tRNA – transfer RNA

UTR – untranslated region

WGS – whole genome sequencing

ZNF – zinc finger