Human endogenous retroviruses aid embryonic development

Inaugural-Dissertation
to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)
submitted to the Department of Biology, Chemistry and Pharmacy
of Freie Universität Berlin

by

MANVENDRA SINGH

from Jaunpur, INDIA

2017

The dissertation "Human endogenous retroviruses aid embryonic development" was prepared under the direction of Dr. Zsuzsanna Izsvak in the period from October 2012 till October 2017 at the Max-Delbrück-Centrum für Molekulare Medizin in der Helmholtz-Gemeinschaft Berlin.

1st Reviewer	**Dr. Zsuzsanna Izsvak** Max-Delbrück-Centrum für Molekulare Medizin in der Helmholtz-Gemeinschaft Berlin
2nd Reviewer	**Prof. Dr. Udo Heinemann** Freie Universität Berlin

date of defence: 19.03.2018

This thesis is dedicated to every member of my Family back in India
I haven't seen them in last six years
my findings are tribute to the patience we all had in this pursuit

Preface

Transposable or Transposed elements (TrEs) are a productive source of biochemically active non-coding or occasionally coding elements that are tightly regulated in a cell-type specific manner. Many recent studies strengthen the hypothesis that few of these elements are co-opted for the regulation of host genes. Here, we focus on one particular large family of human ERVs, termed Human Endogenous RetroVirus H (HERV-H), which we find as a key regulatory player in the human pluripotent stem cells. I show the co-option of HERV-H has potentially led the evolution and development of human-specific embryogenesis including the pluripotency. I present an updated transcriptomic encyclopedia of early human embryogenesis with their transcriptional flags. I re-define the progression of human embryogenesis at single-cell resolution with their markers. In conjecture with previous definitions, I characterise two unattended cell population in human preimplantation embryos that did or did not commit to any of the known lineages in order to form the stable blastocyst. I show the contrasting pattern of transposon families during the progression of preimplantation embryogenesis inferred from embryonic transposcriptomics. Their cross-talk with host factors could have driven the commitment of host cells to the particular lineages, leading to rapid turnover, compared with non-human species.

My dissertation work shows the probable mechanism of HERV-H modulating the human pluripotency network throughout the embryonic development and primate evolution. My comparative single cell transcriptomic analysis of human vs *Cynomolgus* blastocyst and pluripotent states of *Callithrix, Gorilla, Chimpanzee and Bonobo* provides a high-resolution spectrum of distinct embryonic co-ordinates among primates. Much divergence within primate early development is owed to *de-novo* genes, chimeric transcripts and even the novel human specific genes that were remodelled by HERV-H are central to restructuring the preimplantation development across the primates. With my new classification of early cell types and markers, we can address some practical questions. First, is there a pluripotent cell type in the human blastocyst that would be a good candidate for extraction and stable maintenance of *in vitro* pluripotent cells? I suggest that owing to their homogeneity and self-renewal property, the pluripotent epiblast cells should be kept as a reference frame to validate *in vitro* pluripotent cells. Second, and conversely, we can ask what is the identity of the plethora of recently derived naive or naive-like cell types? Do naive/naive-like cells naturally exist in the human preimplantation embryo?. My study suggests that maintaining homogeneity and properly expressing human specific features might be crucial to both best mimic epiblast and preserve genome stability which is controlled by HERV-H. HERV-H displays human-specific cross-talk with host-factors to maintain the identity of distinct developmental lineages, both *in-vivo* and *in vitro* pluripotent states which sets a reliable foundation for clinical applications.

Acknowledgements

With endeavour, I feel highly obliged & have immense pleasure in expressing my heartfelt gratitude to *Max-Delbrueck Centre for Molecular Medicine* for position and finances that I cherish as an obligation to my hard work and passion for science. I take great pride and privilege to say undoubtedly that as true guide, *Dr. Zsuzsanna Izsvak* for expert planning, peerless guidance and timely carping which allowed me to carry out the desired dissertation work. I am highly thankful to *Prof. Udo Heinemann* for inculcating in me the spirit of sincerity, his availability and critical suggestions led this process as smooth it can get. I can never be enough thankful to *Prof. Laurence D. Hurst* for his expertise technical guidance and critical amendments during the pursuit. Our lab administrator *Beate Valeske* always helped to make my research journey extremely smooth. I am extremely thankful to my colleague cum collaborator, *Julianna Zadora*, for her zeal for science and valuable suggestions. My contemporary colleague *Angellica Garcia Perez*, seeing her working tirelessly was definitely the motivation to keep working on. *Kathy, Sascha and Felix*, the youngest members of our team, endlessly conversations and energy improved the scientist in me everyday. *Eniko, Christine, Atilla, Tamas and Cai* for their for constant encouragement and valuable suggestions throughout the entire course and research work. *Jichang, Suneel, Christine, Guo, Katarina, Himanshu and Rabia* being the wonderful collaborators and colleagues. The entire crew of lab, including the former members including *Ananth, Sanam, Helena, Ana and Vaishnavi* had provided me critical suggestions whenever it was needed. Our lab managers, *Sandra, Martina and Ana* had helped me greatly with technical education.

I have no words to express my feeling for friend/colleague/teammate/ex-flatmate *Vikas Bansal* for emphatic help and intellectual simulations along-with *Andranik,* and my better half *Isabel* for their everlasting support at turbulent times and all those good times we spent together. *Arora, Gyaneshwer, Niraj, Sachin, Vijay, Sagar, Nikhil, Shanti, Arun, Vinesh, Kapil* and many more whose finite capacity for support will always be a beckon of light to lead me in all pursuits of future. I have fond memories of our Beer hour sessions in MDC for its invisible affection showered on me during my short stays. An endless list include mainly *Jette, David, Neel, Sebastian, Joscha, Danny and co., Gul, Bob, Katja, Steffi, Conny, Silvia and Laura*. My current and former cricket team-mates *Rohit, Mohit, Abhi* and entire crew inculcated the spirit of getting up every time I fell. My parents *Mr and Mrs Singh* being the earthly god in my life, deserves much more than what I can weigh in words their silent prayers, aesthetic love & affection, moral support & steel belief in my capabilities have enabled me to make this endeavour possible. An unending support from my brother *Shivendra,* my sister *Deepali*, my home-town friends *Umesh, Hariom, Gaurav, Pallavi, Saumya, Saurabh, Adi, Baba, Vivek, Shail & Ravi* are always cherished.

Statement of Contributions

- High-throughput sequencing platforms were provided by **MDC core sequencing facility**.
- High-throughput sequencing raw datasets were analyzed on **MDC-login1** and **Max-cluster univa grid engine** computational clusters.
- High-throughput sequencing experiments were performed by our colleague **Julianna Zadora and Sandra Niendorf**.
- High-throughput microarray and related experiments were performed by our colleague **JiChang Wang**.
- Inguinity pathway shown in figure 5.11 is drawn by our colleague **Julianna Zadora**.
- Callithrix pluripotent cells were provided by **Gerald G. Schumann** affiliated from Paul-Ehrlich-Institut, Langen, Germany.
- Gorilla pluripotent cells were provided by **Ulrich Martin** affiliated from Center for Regenerative Medicine Hannover Medical School, Hannover, Germany.
- **Jose Garcia Perez**, our collaborator from the University of Edinburgh, UK; gave the *in situ* figure of human embryos 4.1C.
- **Avazeh T. Ghanbarian and Laurence D. Hurst**, our collaborator from University of Bath, UK; constructed the figure 6.1E.
- **Gangcai Xie**, our ex-colleague constructed the figure 6.1A and 6.1D.
- **Tamas Rasko and Atilla Svetnik**, performed the experiment and generated figure 6.1F.
- **Zsuzsanna Izsvak**, my PhD supervisor constructed the figure 8.2F.
- **Vikas Bansal** constructed the figure 2.1 and 2.2, and significant content in the method section.
- **Isabel von Holt, Laura Müller and Silvia Ruzittu** had performed the proof reading.
- **Isabel von Holt and Katja Herzog** performed the summary translations to German language.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Mobile DNA

Mobile DNA is also referred to as Transposable Elements (TrEs), are DNA sequences within the nuclear genome of all living organisms. TrEs have the inherent capability to be transposed into a new genomic location. They are harboured as enormous copy number and represent a huge portion of the eukaryotic genome as a result of subsequent transposition. For instance, more than half of the human genome is represented by TrEs sequences, which exceeds in massive order compared with the copy number of protein-coding genes [1]. TrEs contribute directly to evolution and diseases as they are major source of DNA mutations, recombination and rearrangements [2]. TrEs are being pinned as a central player driving the structural and functional evolution of genes in various organisms [3]. There are several ways TrEs aid biological functions to their respective hosts, such as providing regulatory sequences, coding exons, RNA genes and open reading frames and polyadenylation signals which gradually evolves with a phenomenon called *molecular domestication* also termed as exaptation or co-option [4, 5]. Distinct lineages of species or subspecies acquire concordant and discordant TrEs in their genome that ranges from being consistent to volatile, as been extensively studied in [6]. Briefly, TrEs are key generators of evolution as they can disrupt, activate or inactivate genes, move chromosomal segments, and serve as hot spots for genetic recombination. Mutations caused by TrE's activity are often deleterious so they are gradually kept under tight control via epigenetic modifications during the earlier stages of embryonic development. There is an universal classification scheme of TrE families in eukaryotes, stored in a database called "Repbase". It took around 15 years for "Repbase" to develop, but still not all TrEs of vertebral genome has been annotated [7].

1.1.1 Antiquity of Mobile DNA

TrEs were first understood as ***"jumping genes"***, due to their inherent property of moving from one location (locus) to other within the genome in the form of DNA sequence. DNA sequences which have lost their ability to "jump" during the formative ages of evolutionary periodogram, might be termed as Transposed elements. TrEs were first noticed in maize plants about 67 years ago by a geneticist *Barbara McClintock*. Biologists were initially sceptical of her work and TrEs were called "Junk" DNA sequences for first few decades. Only in the early 1980s, TrEs were recognised as significant entity confronting the stimulus genomes face as per McClintock's *genome shock hypothesis* [8]. *Barbara McClintock* received

an unshared *Nobel Prize* in Physiology or Medicine in 1983 for her classical work challenging the existed perception of what DNA sequences were capable of [9]. Eventually, it became apparent that not all TrEs are ***"junk"*** as they constitute the genome of almost all organisms (both prokaryotes and eukaryotes) typically in large numbers [10]. For example, TrEs make up approximately 50% of the human genome and up to 90% of the maize genome [11]. Now, the exaptation of TrEs in the species specific gene regulation is being acknowledged as few of them evidently contribute to evolution and development of distinct biological processes and moreover, the health and disease of their respective hosts [12].

1.1.2 Theories of Transposable Elements evolution

This is 67 years and counting on from Barbara McClintock's remarkable discovery of mobile DNA. We now understand that genomes are dynamic and erratic, with TrEs being major contributors to their elasticity. We recognise that TrEs are key players in genome evolution and have assisted shaping the function of genes. Nevertheless, TrEs are foremost parasitic DNA, and parasites must be controlled or they could be pathogenic to its host. There is far more junk than treasure in mobilomes. There are two widely used hypothesis for the evolution of TrEs that underwent our understanding for decades.
(i) Controlling elements: *McClintock* along-with *Britten & Davidson* [13, 14] suggested major functions for transposable elements (TEs) in genome regulation. Such a view does not exclude the harmful effects that TrEs may provoke but adds an important functional parameter to the presence of TrEs in the host genome. **(ii) Selfish DNA:** *Doolittle & Sapienza* [15], along with *Orgel & Crick* [16], suggested that TrEs are selfish genetic elements, i.e. , genomic parasites colonising the genomes. Along with the junk DNA premise, these hypotheses do not deny the possible exaptation events but consider their presence due to the selfish nature of TrEs.
Currently, the evolution of TrEs is hypothesised on the based of *Latent hypothesis* that was not discussed much in past two decades. As *Dixie Mager* wrote *"because of their capacity to replicate and spread among species, TrEs constitute the largest repertoire of active and latent gene regulatory sequences"* [17]. This sets a foundation of the present dissertation work.

1.1.3 Classification of Transposable Elements

In general, there are various classes of TrEs and numerous ways to categorise them. The two major classes of TrEs are categorised by the form of intermediates they use during their transposition process. TrEs which are mobile within genomic loci by DNA intermediates, using *transposase* and DNA *polymerase* to catalyse transposition are called DNA transposons (Figure 1.1). However, TrEs which are moved by RNA intermediates using RNA polymerase, *endonucleases* and *reverse transcriptase* which catalyse the retrotransposition are known as retrotransposons. retrotransposition requires *reverse transcription* in order to transpose, are categorised under Class I TrEs, whereas the DNA transposons belong to Class II TrEs (Figure 1.1). Different classes of transposable elements are found in the genomes of different eukaryotic organisms will be discussed in later sections. As it is illustrated in the figure 1.1, TrEs are categorised into two major and several minor classes based on their structure and intermediate products being utilised to complete their life-cycle.

1.1.3.1 Class I Transposable Elements

Class I TrEs are also referred as retrotransposons, duplicate within the genome through RNA intermediates which are reversely transcribed and integrated at a distinct locus [18]. DNA transposons had lost their mobility during early primate evolution at around 37 MYA [19], whereas few families of RNA TrEs are still mobile in primate's genome in a cell-type specific manner (mostly germ cells), so continuing the saga of genome evolution [20]. retrotransposons are subdivided into two major groups, based on the presence or absence of flanking LTRs.

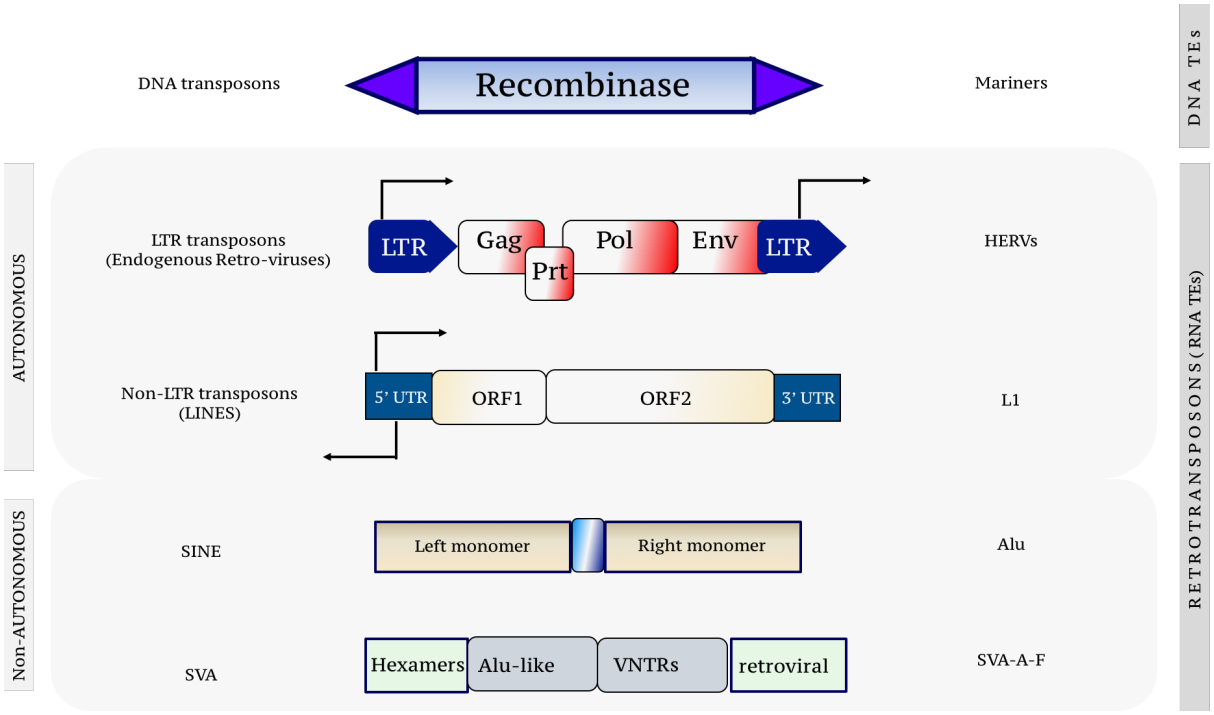


Figure 1.1: This schematic picture illustrates the structure of transposable elements. Additionally, an example of each main class of TrE is depicted with average genome length i. e DNA transposons (1.4 KB), LTRs retrotransposons (7 to 9 KB), and LINE (6 KB) are annotated as autonomous retrotransposons flanked by UTR. Nonautonomous retrotransposons are either classified as SINE or SVA that amplify with the help of LINE-encoded reverse transcriptase via trans-acting functions. Note: This figure is redrawn from [21].

(i) Long Terminal Repeat retrotransposons: Retrotransposons under this subgroup are classified by the presence of flanked LTRs on both ends of internal (ICR) sequence. ICR encodes both *structural and enzymatic proteins*, e.g. Gag codes for structural protein that forms viral like particle, whereas Pol codes enzymatic proteins including protease and reverse transcriptase, lastly Env codes the capsid subunit that is crucial for re-infection. Non-LTR TrEs lack the direct repeats at their terminal sequences (Figure 1.1).

(ii) Non-LTR retrotransposons: They are mostly represented by *L1*, *Alu* and *SVA* elements in human genome. L1 elements comprises around 6 kilobases (kb) thus abbreviated as long interspersed nuclear elements (LINE). In contrast, Alu elements are around 250 bases, thus a short interspersed repeats, or SINE. SVA is a chimeric sequence of Alu, LTRs, tandem repeats and other SINE thus neither fall in LINE nor in SINE category.

1.1.3.2 Class II Transposed Elements

Class II TrEs or DNA transposons are classified as mobile DNA via a single or double-stranded DNA intermediate [18]. Major subclass of Class II TrEs is identified by the presence of terminal inverted repeats (sequentially compliment to each other) on both of their ends. They encode their own protein *Transposase* that mediates the insertion or excision of TrEs within the genome by a *"cut and paste"* mechanism.

1.1.3.2.1 DNA transposons There are two minor sub-classes of DNA transposons that have been found in eukaryotes and are exception to *"cut and paste"* mechanism of transposition. Their transposition mechanism is led by double or single-stranded DNA intermediates. Exceptionally, they could also move via alternate mechanisms compared with those that are moved by *cut and paste* mechanism. Selected examples of those TrEs are following:
(i) MITEs are defined as Miniature inverted repeat transposable elements are non-autonomous transposons as a repercussion of truncated internal sequence from autonomous DNA transposons [22]. MITEs are DNA transposons, containing Terminal Inverted Repeatss flanked by small direct repeats. MITEs exhibit shorter internal part that lacks putative transposase ORF. MITEs transpose through transposase encoded by autonomous DNA TrEs [23].
(ii) Helitrons are the eukaryotic transposable elements, an intrinsic part of the selective eukaryotic genomes. Helitrons follow the displacement by aggressive rolling-circle replication model of transposition (see. *Rolling circle*). [24]; The activity of its reconstructed sequence was recently showed in eucaryotic genome by our group [25].
(iii) Mavericks are likely to replicate via a self-encoded DNA polymerase resembling the DNA viruses mode of transposition [26]. Nevertheless, these elements follow *copy and paste* mechanism of transposition instead of classical *cut and paste* mechanism.
The majority of TrEs from both classes contain flanking direct repeats which are not part of TrEs themselves but they play a role in TE integration. Sequentially, the flanking repeats are left behind as *"footprints"* after the excision of genome. These footprints are utilised as promising tool in order to detect transposition events in a given genome.

1.1.3.3 Autonomous and Non-autonomous

Class I and Class II TrEs could further be classified as autonomous or non-autonomous depending on their self-capability to move around the genome. Autonomous TrEs are those mobile DNA sequences that move on their own, while non-autonomous element's mobility is obligatory on the autonomous ones. As they lack either transposase or reverse transcriptase which is required for their mobility, they acquire these proteins from autonomous ones in order to move.

1.1.4 Impact of DNA Transposons on host genome

Class II TrEs are harboured in the genomes of prokaryotes, protozoans to a wide range of metazoans including insects, worms and in higher ordered organisms such as humans (Table 1.1) [10]. There are at-least 10 sub-classes of DNA transposons and most of them are plugged on an early evolutionary node, thus are diversified and maintained on major branches of the eukaryotic evolutionary tree of life (Table 1.1)

[27]. DNA transposons are extensively studied in the context with genome evolution, adaptive immune system, gene duplication, exon shuffling, allelic diversity, chromosomal rearrangement and epigenetic regulation of gene expression which all have already been extensively reviewed [21]. Additionally, we recently reported that V(D)J recombination that has been evolved with transposition, is prone to evolutionary mutations that might stabilise the transposon-host specific interactions and prepare the DNA transposon for a horizontal transfer from one species to another [28]. Recent advances include the role of DNA transposons in generating introns and exonic splice sites on genomic scales [29]. These findings tag the journey of DNA transposons from ***"junk"*** **to** ***"jewel"*** as TrEs offer positive, negative or even neutral selection (see. *Neutral selection*) to the host not just at individual level but also at population/species level [30]. *In-silico* reconstruction of fish transposon and its capability to be mobile into the eucaryotic genomes as shown by *Zsuzsanna Izsvak* provides enormous applications of Sleeping Beauty transposons for nonviral gene therapy.

1.1.5 Impact of RNA Transposons on its host

Note: The domestication of retrotransposons is reported to be an extremely rare event (approximate number should be known after genome-wide CRISPR screen), but once it is co-opted, its impact is intense and profoundly contribute to the fitness of host.

1.1.5.1 LTR Transposons

Most of the Internal Coding Regions (ICRs) flanked by LTRs are known as Endogenous retroviruses (ERVs), a remnant of ancient retroviral infections. The structure of their gene units are similar with to those of Endogenous retroviruses (XRVs) and so are their properties up to a certain extent. Mammalian/Vertebral genomes are expanded by multiple waves of XRV integrations into germ cells, and consequently propagated through generations in Mendelian fashion [31, 32]. The transformation of XRVs to ERVs in symbiosis with the evolutionary clock by a process known as ***"endogenization"*** of retroviruses. The endogenization events display a classical consequence of *Lateral Gene Transfer* [33, 34]. The polymorphic nature of ERVs leads to genetic diversity within individual and/or population [35]. Polymorphism at an individual level is established through a homologous recombination between two LTRs which leaves ICRs i.e. *provirus* DNA out, whereas LTRs stands alone on locus and established as solo-LTRs [36, 37]. The proviral LTR sequences offer various motif for TFs binding to initiate the transcription (on both strands). For example, TATA-box coupled with a GC/GT-box and AATAAA, the polyadenylation signal at 3' LTR enables them to generate the full-length proviral mRNAs [38]. Since LTRs are frequently integrated in close proximity with protein coding genes [39], so the inherent potential of proviral LTRs for TFs binding could modulate the transcriptional activity of nearby genes [40]. Additionally, proviral sequences and/or LTRs in proximity to cellular gene may modulate their expression by donating splice sites, which leads to the generation of alternative or aberrant transcriptional products [41]. These transcriptional products may get translated and distort the native function of cellular genes resulting into the interference with normal phenotype of hosts [42]. The contribution of ERVs at the level of host fitness have been debated regarding both, the parasitic and symbiotic point of view [43]. Although natural selection would eradicate the deleterious ERVs, still some ERV loci are pathogenic [44] suggesting that the process of neutral selection is still ongoing. Among various diseases in vast diverse species caused by ERVs are multiple sclerosis, schizophrenia, diabetes, systemic lupus erythematosus, seminoma, malignant

melanoma, numerous types of cancer, preeclampsia and azoospermia [45, 46, 47, 48, 49, 50, 51, 52]. The pathogenic molecular processes include antigenic mimicry, immune dysregulation, receptor interference, growth stimulation by cis or trans-activation, loss of physiological functions mediated by retroviral genes and gene loss by recombinations [53, 54, 55, 56, 49, 57, 58, 59].

1.1.5.2 Short Interspersed Nuclear Elements (SINE)

SINE elements have shown tremendous reproductive success during the evolution of mammalian genomes as they are harboured at a very high copy number [1, 60]. In conjecture with selective advantageous nature, SINEs are non-autonomous and dependent on host's resources for its replication, expression and amplification [61]. Alu elements are extensively studied SINE element (4917 pubmed entries with "Alu" in the title) as Alu elements make up more than 10% of human genome indicating their highly prolific nature during the course of evolution [1, 60, 62]. Alu elements have greatly contributed to mammalian genome diversity due to an uneven level of homologous recombination [63]. Various Alu families have been expanded in relatively short time and are responsible for being part of 62% of all new exons in the human genome [64]. Alu elements played a pivotal role in evolution by committing to restructure or recombinations of gene models. Consequently, Alu sequences add a new functionality to the gene by providing a new exon, through the process so called *"exonization"*, which is one of the coherent property of Alu elements [65]. The exonization of Alu elements questions the necessity of the host to harbour such a deleterious mutation. An answer may lie in the mechanisms of splicing as transcripts can have option of being spliced to either exclude or include the "Alu" exons in their mature transcript. [66]. Transcripts that exclude "Alu" exons would be translating as the same "old" peptide whereas, transcripts including "Alu" exons would translate an entire new peptide or variant of the same. Additionally, "Alu" elements potentially monitor the splicing of the gene being "intronic" or "intergenic" in order to produce new classes of transcripts e.g. linear and circular RNAs [67, 68, 69].

1.1.5.3 Long Interspersed Nuclear Elements (LINE)

Among the diverse LINE families, LINE-1 (L1) has been the dominating one with its content in the genome of numerous species; although the copy number of L1 is smaller than Alu elements, its lengthy structure makes ∼17% of the human genome [70, 71]. L1 elements are the only autonomous elements [72] that are currently active in human and non-human primate genomes; SINEs and SVA use L1 machinery to be mobilized [73, 74], so the L1 element has directly or indirectly constructed ∼30% of the human genome sequence [1]. In spite of L1 element's inactivation during their genomic integration due to 5' truncation and mutation accumulations, still there are around 80 active L1 loci per genome [75]. The first active L1 element was discovered when a mutation due to L1 integration distorted the *Factor VIII gene* in a haemophilia patient [76]. Since then, more than 100 human diseases have been linked to the germ-cell integration of L1 elements [77]. Active L1 elements must preserve their intact 6 KB sequence that contains 5'UTR, two ORFs followed by 3'UTR (Figure 1.1) [78]. In addition to huge contributions, recently, a study from **Fred Gage** laboratory has reported the existence of third ORF in human and *Chimpanzee* genomes along-with its significance in neuronal development [79]. Interestingly, 5'UTR has potential to act as a bidirectional promoter (sense and anti-sense transcription) [80] and drives the neighbour genes [81]. ORF1 of L1 makes smaller protein at around 40 KDA with RNA binding activity [82], is required for L1 integration [83]; whereas, ORF2 is a large protein at around 150 KDA. ORF1

does not show any level of similarity with known protein sequences [84], whereas the ORF2 sequence matches with endonuclease and reverse transcriptase activities [85]. Beyond their capability to mobilise themselves, L1 elements generate a spliced transcript that has been implicated in many diseases [86]. For instance, various cancer cell lines are enriched with L1 driven transcripts [87]. L1 mobility has been implicated to cause the activation of oncogenic pathways [88] various human diseases such as colorectal cancer [89], complexity and diversity of human brain [90], neurodegeneration [91], the double strand break repair [92] and human evolution [93]. L1 mobility and L1 mediated normal/abnormal gene expression, are key players to continue the evolutionary saga of human health and diseases which is broadly covered in a recent review [94].

1.1.5.4 SVA elements

Unlike LINE and SINE elements, the existence of SVA (SINE/VNTR/Alu) element is exclusive to primate genomes [95]. At first, SVA elements were classified under SINE and known as "SINE-R" where "R" stood for retroviral origin [96]. Seven years later, SINE-R was re-annotated to SVA as it consists of SINE-R, Variable Numbers of Tandem Repeats (VNTR) and Alu sequences [97]. This forms an interesting structure of SVA elements as repeat of repeats or chimeric sequence, indicating that consensus may have never existed in the ancestral species. SVA elements are transcribed by their own promoters and produce non-coding RNAs [98]. Like other SINEs, SVA is also nonautonomous and require endonucleases and reverse transcriptases encoded by L1 elements to pursue transposition [99]. The alternative transcriptional products may result in the retrotransposition of SVA elements with differential units/lengths that may have contributed to exon-shuffling or exon-trapping inside genome [95, 100]. SVA elements are subdivided into six families, from SVA-A to SVA-F. SVA-F is human-specific and higher transpositionally active than L1 elements in the order of 10 folds [101, 102]. Younger L1 and SVA elements (less than 12.5 million years) are the ones that are still retrotranspositionally active in human genome and currently involved in the evolutionary arm-race with KRAB Zinc finger protein [102, 103]. Despite SVA elements having a canonical polyadenylation signal, AATAAA at its 3'end, its transcription may bypass its own polyA signal. This occasionally show the read-through events and drive the downstream genes [104]. Ontological analysis of genes associated with SVA insertions has been illustrated to be involved in brain function, behavioural features and reproductive functions [105].

1.2 Host's mode of control over Transposable Elements

This decade is witnessing a trendy debate over TrEs, whether they are with us or against us in the journey of evolution, development and fitness. If TrEs are not restrained inside cells, their uncontrolled spread would certainly lead to lethality [127]. Insertional mutagenesis caused by active TrEs may cause adverse consequences for the host system. Even if the TE does not land into the coding sequence of a gene, it can still disturb its expression by altering the conformation or regulatory sequences around it [128]. The disruptive expression of the gene may alter the phenotype of the host and leads to lethal human diseases such as cancer [129], as manifested by the activation of proto-oncogenes during the gene therapy of patients and within the cell populations [130, 131]. The essence of the TrEs restraint has also been illustrated in the early development as TE integrants cripple the retroelements control pathways which leads to embryonic lethality [132, 133]. Majority of human TE sequences are crumbled in our genome

Table 1.1: This table shows that with rare exceptions, transposons are present in every genome studied so far. Transposons are still active today. However, their activity level varies across different species. Somatic transposon activities can be observed in various organisms, including both plants and animals and comprise significant portions of their genomes. Most active TrEs are in context with species under observation. Abbreviations: LINE, long interspersed element; LTR, long terminal repeat; SINE, short interspersed element; TIR, terminal inverted repeat.

Species	Common name	*Significance*	Genome size (Gb)	TE content	Hyperactive TrEs	Hypoactive TE	Reference
Physcomitrella patens	Moss	*model for evolutionary biology*	0.480	~ 52 %	*LTR retrotransposon*	Helitrons	[106]
Selaginella moellendorffii	Lycophyte	*an oldest vascular plant*	0.106	~ 35 %	*LTR retrotransposon*	Basho	[107]
Arabidopsis thaliana	Flowering plant	*genetics & development model*	0.125	~ 10 %	*LTR retrotransposon*	MULEs	[108, 109]
Zea mays	Maize	*Discovery of transposons*	2.3	~ 85 %	*LTR retrotransposon*	Helitrons & TIR	[110, 111, 11]
Oryza sativa	Rice	*staple calorie source in human*	0.420	~ 25 %	*LTR retrotransposon*	*LINE & SINE*	[112, 113, 23]
Hydra magnipapillata	Cnidarian	*regeneration & tissue patterning model*	1.5	~ 60 %	*CR1*	hAT	[114]
Caenorhabditis elegans	Nematode	*neurobiology & development model*	0.097	~ 12 %	Tc1 & Tc3 Mariners	*LTR retrotransposon*	[115, 116]
Drosophila melanogaster	Fruit-fly	*sex-specific development & genetics model*	0.180	~ 10 %	TIR	*LTR, LINE & SINE*	Review [117]
Takifugu rubripes	Putterfish	*compact genome model for comparative genomics*	0.365	~ 03 %	*LINE*	DNA TE	[118]
Anolis carolinensis	Lizard	*the first reptilian genome sequenced*	2.2	~ 20 %	hAT, Tc1 & Helitrons	*LTR retrotransposon*	[119, 120]
Ornithorhynchus anatinus	Platypus	*mammalian & reptilian evolution*	2.3	~ 50 %	*LINE2 & MIR*	***Inactive LINE1***	[121]
Myotis lucifugus	Bat	*show the most aggressive transposon activity*	0.002	~ 25 %	TIRs & Helitrons	*LINEs & SINEs*	[122, 123, 124]
Rattus norvegicus	Rat	*first mammal domesticated for scientific research*	2.7	~ 40 %	*LINE, SINE and LTR*	DNA TrEs	[125, 126]
Mus musculus	Mouse	*genetically manageable mammalian model*	2.5	~ 37 %	*LINE, SINE and LTR*	DNA TrEs	Review [77]
Homo sapiens	Modern human	*We the people*	2.9	~ 52 %	*SVA, LINE, SINE and LTR*	***Inactive DNA TrEs***	Review [17]

so they lie as inactive DNA. retroelements that have emerged recently in primate evolution are the chief target of host factors in order to defend itself against the same (Figure 1.2) [134]. The host mode of defence against intrinsic retroviruses features the innate immune pathways they use for extrinsic infectious retrovirus restriction [135]. The host system arranges the machinery to restrain endogenous retroelements in the multiple layers as following *(i)* transcriptional control via DNA methylation, KRAB-ZNF binding and chromatin status *(ii)* post-transcriptional processing of retroelements using piRNAs and RNA editing machinaries and *(iii)* integration of new retrotransposons via DNA repair subunits. How the multiple layers of the host defence are interconnected is largely unknown. Nevertheless, the existence of multiple layers indicates that even if retroelements are transcriptionally or post-transcriptionally active, there are host factors to check the life cycle of retroelements. The fourth layer of host defence is the cell cycle arrest or undergoing senescence which is reported to be a restraining retrotransposition but is mechanistically unclear [136, 137].

1.2.1 Transcriptional control

The transcriptional competence of TrEs is a consequence of *cis*-regulatory sequences, trans-acting factors and epigenetic layouts acting in combination around a TrEs locus. retrotransposed elements have to be expressed in germ cells and in their embryonic precursors to form a new insertion which passes from one generation to other. TrEs harbour intact 5' LTR and 5' UTR regulatory sequences which appear to be appropriate for transcriptional machineries of early embryos and the germ-line. On the other side, the host system has arranged the barrier of DNA methylation and repressive chromatin states at TrE's promoters to make them inaccessible to transcription factors blocking their transcription.

1.2.1.1 DNA methylation

DNA methylation has been asserted to have evolved as a strive of the host defense to check the TE activity [138]. Cytosines in CpG dinucleotides are converted to 5-methylcytosine by DNA methyltransferase (DNMT) enzymes resulting in the TrE's prominence of methylated cytosines [139, 71]. DNA methylation leads to immediate transcriptional inactivation of the locus by activating the local repressive chromatin state [140]. Additionally, they can also promote permanent inactivation by deaminating C into T. Since TrEs are heavily methylated in the germ-line, the accumulation of TA or TG conversions makes them incompetent to mobilise [141]. There are a handful members of DNA-methyltransferase family in our genome viz. DNMT3A, DNMT3B, DNMT3L and DNMT1 that regulate DNA methylome (Table 1.2). There is a DNMT2 too in the vertebrate genome but is not involved in the regulation of global DNA methylation status like the other members of DNMTs [142]. DNMT3A and DNMT3B execute *de novo* DNA methylation on CpG sites, since they contain a catalytic domain for methyl transferase, whereas DNMT3L is catalytically inactive regulatory factor of DNA methyltransferase [143]. DNMT3L can either promote or inhibit the DNA methylation depending on the context, since it is essential for the function of DNMT3A and DNMT3B [144]. It has been documented that DNMT3L is required for retrotransposon methylation and their silencing in pre-meiotic male germ cells [143]. DNMT1 is the most abundant methyltranferase in adult cells in order to maintain the methylation level [145].

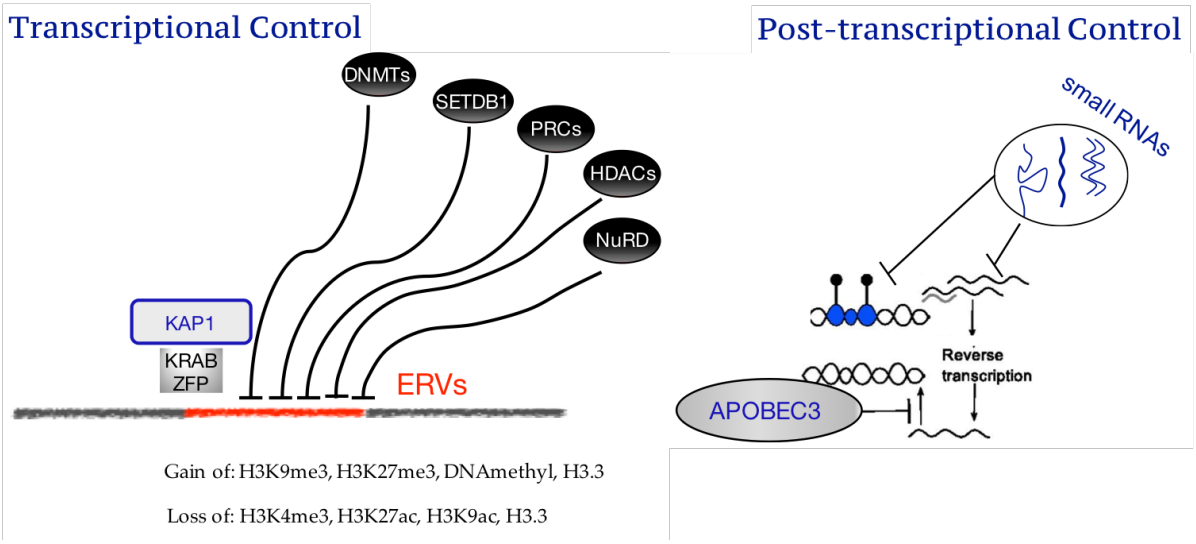


Figure 1.2: The repression of class I and class II ERVs is dependent on SETDB1 histone H3 lysine 9 trimethylase activity (H3K9me3). SETDB1 is believed to be recruited to its ERV targets via KRAB-containing zinc finger proteins and their co-repressor KAP1. DNA methyltransferase enzymes catalyse the DNA substrate converting them to methylated CpG which causes repression of ERVs. The histone deacetylases and NuRD complex remodels the chromatin marked by H3K27Me3 (PRC2 mediated) or H3K9Me3 (SETDB1 mediated) or H3.3 (unknown) are hallmarks of ERV repression. If this barrier of ERV repression is broken, then another layer contains piRNAs or APOBEC group of proteins. If KRAB-ZNF fails to recognise ERV sequence then small RNAs such as piRNAs, which in turn down-regulate their expression via DNA methylation, whereas their retrotransposition would further be blocked by proteins such as APOBEC family members via RNA-editing

It finds and binds CpG sites on DNA with single-stand methylation (hemi-methylated), since the parent-strand underwent replication remains methylated but not the newly synthesised strand [146]. So, the precise function of DNMT1 is to maintain the established *de novo* methylation through mitosis [147]. In order to pursue the function of a genomic guardian as maintaining the methylation status

through cell divisions, DNMT1 requires Ubiquitin like with PHD nnd Ring Finger Domains, to be loaded onto hemimethylated sites [148]. However, there are exceptions to the conventional DNMTs mode of action, e.g. an overexpression of DNMT1 leads to *de-novo* methylation of CpG islands [149], similarly, DNMT3A or DNMT3B may occasionally have a role in the maintenance of methylation, as human cancer cells devoid of DNMT1 have shown the maintenance [150]. Surprisingly, DNMTs were also demonstrated to act as demethylases as mammalian DNMT1, DNMT3a, and DNMT3b can convert 5'methyl cytosine to cytosine [151]. An aberrant expression or mutant version of genes encoding DNA methylation machinery result in renewal of TrEs and a dramatically reduction in viability or fertility [152]. Interestingly, inactivation of DNMT3A and DNMT3B simultaneously in early embryos showed similar phenotype with that of the DNMT1 knockout (Table 1.2) [146]. Since DNMT3A requires DNMT3L as a co-factor to control the expression of retrotransposons in male germ-line, a DNMT3L or DNMT3A mutant enhances the expression of retrotransposons which leads to male sterility (Table 1.2) [143, 153]. Curiously, DNMT3L distinctively emerged around 150 million years ago in eutherians which concurred with a vital expansion of TrEs in mammalian genomes [154, 121]. DNMTs are obligated to several proteins that assist the methylation conversion reaction. Among them, members of the chromatin re-modeller SNF2 family that aid DNMTs accession to the DNA substrate [155]. Deletion of a member from the SNF2 family leads to hypomethylation and enhanced expression of retrotransposons in germ cells and embryonic lineages which results into the post-natal lethality [156, 157]. Hence, the interplay between DNA methylation enzymes and chromatin dynamics resolve the control of transcription of TE and non-TE sequences. Moreover, the role of TP53, a tumour suppressor genes, is emerging in recent literatures to be restraining the mobility of retrotransposons by recruiting DNA methylation over them [158].

1.2.1.2 Chromatin status

In addition to fore-mentioned DNA methylation, the chromatin status determined by histone modifications and several other epigenetic mechanisms is housekeeping regulator of the genes and TrEs expression. The mechanism of histone modifications, chromatin status and its accessibility for other proteins has been extensively studied and reviewed [159]. Methylated DNA and specific histones co-operatives form the condensed heterochromatic states (see. *Heterochromatin* and *Euchromatin*) around TE promoters in order to silent their expression. Mainly, trimethylation on H3K9 in somatic cells and on H3K27 in early development are the hallmarks to induce TrEs silencing [160, 161]. It was shown in the case of mice that trimethylation on H3K9 and H3K20 constructs the heterochromatin environment around the promoters of retrotransposons [162, 163].

Histone modifying enzymes are extremely diverse and redundant a fact that limits our knowledge about their precise significance at a functional level. ERG-associated protein with SET domain gene has been implicated to control TrEs at a broader level (Table 1.2) [166]. Additionally, SUV39H inactivation leads to retrotransposons reactivation at reasonable level [162]. Interestingly, the loss of H4K20 methylation does not affect TE expression in ES cells [176] but combined loss of H3K9 and H4K20 trimethylation reactivates retrotransposons even in somatic cells such as fibroblasts [177]. Finally, the Polycomb Repressor Complex 2 (PRC2) mediated H3K27 trimethylation is potential silencer of TrEs as evident from series of recent reports which show that the loss of PRC subunits upregulates LTRs and ERVs (Table 1.2) [178, 179].

Table 1.2: This table shows the list of host proteins (first column) with their molecular function (second column) which control the transcriptional or post-transcriptional activity of TrEs. Third column displays the mode of cellular action as evidenced in fourth column. Fifth column indicates their impact on organism's phenotype. Finally the last column illustrates the responsive TrEs for a given host factor.

Gene	Molecular Function	*Cellular Role*	Evidence	Consequence	Targetting TrEs	Reference
DNMT1	DNA methyltransferase	*maintainance of methylation*	Knockout mouse	Embryonic lethality	IAP	[164]
DNMT3A	DNA methyltransferase	*de novo methylation*	Knockout mouse	Neonatal lethality	L1, LTR & IAP	[146, 153]
DNMT3B	DNA methyltransferase	*de novo methylation*	Knockout mouse	Male sterlity	L1, LTR & IAP	[165]
DNMT3L	DNA methyltransferase	*de novo methylation*	Knockout mouse	Male sterlity	L1, LTR & IAP	[143]
KAP1	Co-factor KRAB-ZNF	*Transcriptional repressor*	Knockout mouse	Embryonic lethality	L1, LTR & IAP	[133]
LSH	Helicase	*de novo methylation*	Knockout mouse	Embryonic lethality	IAP	[157]
UHRF1	hemimethylated DNA binding	*de novo methylation*	Knockout mouse	Embryonic lethality	IAP	[148]
ESET	Histone methyltransfease	*Histone methylation (H3K9)*	Knockout mouse	*Not Available*	L1, Alu, LTR & IAP	[166]
SUV39	Histone methyltransfease	*Histone methylation (H3K9)*	Knockout mouse	*Not Available*	L1, Alu, LTR & IAP	[162]
PRC1/PRC2	Polycomb Repressor Complex	*Histone methylation (H3K27)*	Knockout mouse	*Not Available*	IAP and MLV	[167]
MIWI2	Rnase H (Piwi family)	*piRNA biogenesis/pathway*	Knockout mouse	Male sterlity	L1 and IAP	[168]
MILI	Rnase H (Piwi family)	*piRNA biogenesis/pathway*	Knockout mouse	Male sterlity	L1 and IAP	[169]
TDRD1	Tudor (binds to MILI)	*piRNA biogenesis/pathway*	Knockout mouse	Male sterlity	L1 and IAP	[170]
TDRD9	Tudor (binds to MIWI2)	*piRNA biogenesis/pathway*	Knockout mouse	Male sterlity	L1 and IAP	[171]
GASZ	Leucin zipper	*piRNA biogenesis/pathway*	Knockout mouse	Male sterlity	L1 and IAP	[172]
ADAR	Adenosine deaminase	*RNA editing*	Reporter assay	*Not Available*	retrotransposons	[173]
APOBEC3	Cytosine deaminase	*RNA editing*	Reporter assay	*Not Available*	L1, LTR, IAP & Alu	[93]
AID	Cytosine deaminase	*RNA editing*	Reporter assay	*Not Available*	retrotransposons	[174]
ERCC	Endonuclease	*Nucleotide excision pathway*	Reporter assay	*Not Available*	L1	[175]

1.2.1.3 KRAB-Zinc Finger Proteins

There are at least 350 and 600 Krueppel-associated box domain containing zinc-finger proteins (KZFPs) encoded by human and mouse genome respectively [103, 180, 181]. KZFPs are some of the most rapidly evolving gene families among tetrapods, whereas, fish and birds contain extremely low numbers of KZFPs (< 50) in their genome [103]. The dynamics of TrEs and KZFPs are extensively studied in pluripotent cells as TrEs are directly targeted by KZFPs, which recruit TRIM28 protein as their co-factor. The TRIM28-SETDB1 complex catalyses H3K9 trimethylation over the regulatory sequences of numerous TrEs. In response, TrEs can acquire spontaneous mutation to escape the KZFP binding and again in counter response, KZFPs have to be rapidly evolved to re-silence them. This evolutionary phenomenon is regarded as ***evolutionary arms race*** between TrEs and host factors [133, 182, 183, 102]. TRIM28, being the co-factor of vast array of KZFPs, serves as a scaffold for a heterochromatin inducing complex enveloping histone methyltransferase, histone deacetylase, nucleosome remodelling, and DNA methyltransferase activities [165]. TRIM28 mediated repression is also reflected on neighbour genes via repressed TrEs which often have an enhancer or promoter activity to drive downstream genes [166, 133, 165, 184, 185, 186].

1.2.2 Post-transcriptional control

1.2.2.1 RNA-editing

The term RNA editing describes the molecular processes by which enzymatic proteins modify the mature mRNA molecules. RNA-editing enzymes can modify various nucleosides of RNA transcripts, mostly by deamination. RNA-editing enzymes are implicated to target against infectious RNA viruses, so these proteins are involved in triggering host innate immune responses [187]. The APOBEC family catalyses the deamination of cytosine residues into uracils, whereas ADAR catalyses adenosine to inosine and has greatly expanded in the primate lineages [178]. They act as a third line of defence against endogenous and exogenous retroelement activity, e. g APOBEC3G reduces the replication of the human immunodeficiency

virus [188]. Primate and vertebrate specific Alu and L1 elements are edited by ADAR which converts adenosine residues into inosines in double-stranded RNAs [189, 190, 191]. APOBEC3A, 3B, 3C and 3F enzymes potentially restrain the different classes of LTR and non-LTR retrotransposons, such as L1, IAP, Alu, human ERVK and MusD elements in human and mouse cells [192, 193, 194].

1.2.2.2 RNA interference

RNA interference through miRNA, piRNA and endogenous siRNA displays another layer of the post-transcriptional regulation of TE restraining [195]. This mechanism proceeds with the generation of small RNA sequences which find and bind their complementary sequence within TE transcripts and induces degradation of TrEs by recruting RISC. The LTR transcripts specific to early embryos and ES cells are transcribed in both sense and antisense orientation [196, 197]. Some of them may form double-stranded transcripts that are putative substrates for DICER and trigger RNA interference response [198]. Another class of small RNAs which has been repeatedly linked to RNA interference based TE repression is piRNA. The piRNAs (PIWI-interacting RNAs) control TrEs in germ cells which are single-stranded RNAs, being little longer than other classes of small RNAs (24-34 nt)) and processed independently. They are loaded onto germ-line specific argonaute proteins also known as Piwi Like RNA-Mediated Gene Silencing proteins, has broad phylogeny than conventional argaonautes [199]. Neverthless, the siRNA/Argonaute complex and the piRNA/PIWI interaction shows a similar mode of action over degradation followed by the recognition of target transcripts. There are four PIWI proteins in human (PIWIL1-PIWIL4), whereas, the mouse genome encodes three PIWI proteins, their loss in the genome causes sterility similarly with the DNMT3L mutant phenotype (Table 1.2) [168, 169]. PIWI mutants loose the methylation over the TrE sequences, indicating that the piRNA pathway not only acts as a post-transcriptional level, but also controls the transcriptional activity of TE elements [200, 201].

1.2.3 Integration restrainment

Lastly, the host has to defend the final stage of the TE life cycle, i. e the integration of the TE cDNA into the genome. The host's mechanism of integration restrainment is largely unknown, however the nucleotide excision repair pathway has been associated with it. DNA repair complex that maintains genomic stability is reduced to enhance L1 retrotransposition into the human cell line, suggesting its role in restraining [202]. However, some DNA-repair enzymes have an contrasting impact on retrotransposition; ATM gene that repairs double strand break facilitates L1 retrotransposition [203], supports that various factors from same pathway could have positive or negative impact on restraining the TE mobility.

1.3 Human Endogenous retroviruses (HERVs)

Endogenous means ***"within the organism"*** or ***"within the cell"*** or even ***"within the genome"***. Hence ***endogenous retroviral*** (ERV) sequence is a ***piece of DNA brought in our genome by retroviruses***. More than half of the human genomic sequence is derived from TrEs (Figure 1.4). The research advancements in the last decade provide substantial evidence that few TrEs are co-opted in service of the human biological system [204, 51, 205]. Co-opted TrEs acquire new physiological functions such as coding/non-coding and regulatory sequences for human genes (Figure 1.7) [206, 207, 208]. Among TrEs, HERVs derived regulatory sequences contribute to transcriptional modulations of host genes as *promoter*, *enhancer*, and

insulator [209, 210, 12, 211, 212]. Every human survives by the maintenance of the delicate equilibrium between HERVs and the host genes.

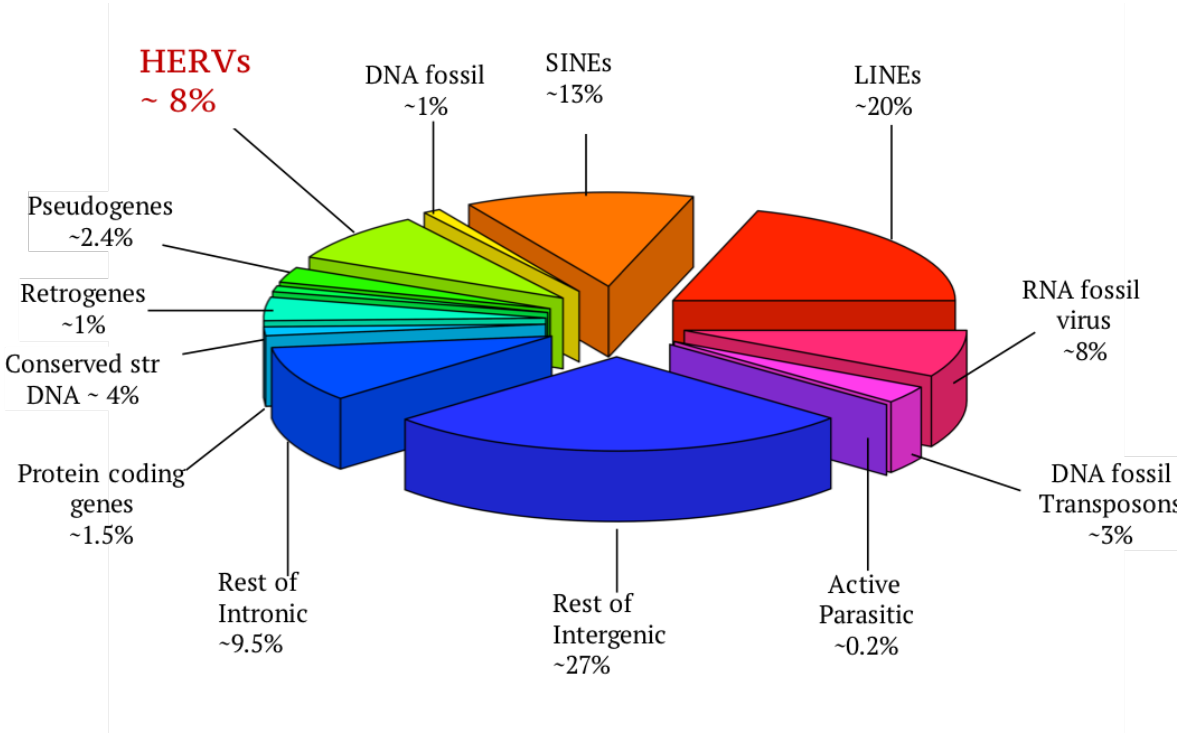


Figure 1.3: More than half of human genome sequences comprises transposable or transposed elements. The human genome content has been modified by multiple waves of retroviral invasions, followed by integration, amplification and lateral expansion reflected as approximately 8% of genomic sequence

They were transmitted from one generation to the other in the *mendelian* fashion and consequently endogenized to serve regulatory functions [213]. Recent advancements explore the substantial evidence that few families of HERVs are co-opted to regulate cell-type specific physiological networks [214]. Amongst all HERVs in the human genome, HERV-H is comparatively conserved and intact in primates, whereas, rest of the intrinsic retroviral families of similar age is highly polymorphic [215]. HERV-H shows limited polymorphism in human population, suggesting that the host system has preserved it throughout its evolutionary journey [216]. This family of HERVs has a Histidine (H) tRNA primer-binding site, hence the name "HERV (H)". Phylogenetic studies plug it in the category of Class I/Type C or gamma-retroviral group (Figure 1.5) [217]. While a low number of HERV-H like elements occur in New World monkeys e.g. *Callithrix*, squirrel monkey, spider monkey etc. The major expansion of this family occurred in the Old World branch (Figure 1.6). Most of the HERV-H locus in human genome has an orthologous sequence (see. *Orthologous*) in Old World monkeys e.g. rhesus, *Cynomolgus*, baboon and apes [218, 219]. Notably, a common partially deleted form, which amplified to several hundred copies in Old World monkeys and which is associated primarily with LTRs, is annotated as LTR7 and LTR7B (originally termed Type I and Type II LTR, [219, 220]. These LTRs were estimated by genomic analysis to be present in no more than 50 copies in two species of New World monkeys [221]. A later expansion in hominoids of approximately 100 elements with a variant LTR (termed Type Ia and annotated as LTR7Y in Repbase, [7, 222] has also been documented [219, 220]. In comparison with LTR7 or LTR7B, LTR7Y is represented by fewer copies in the human genome and has higher promoter activity in reporter assays [219, 220, 223]. Unlike HERV-K-HML2 (the youngest family of HERVs) [224], there is no evidence for

HERV-H integrations after the divergence of human and *Chimpanzee*.
The human genome has been modified to approximately 8% by multiple waves of retroviral invasions, followed by their integration, amplification and lateral expansion, mainly in germ cells (Figure 1.3 and 1.4) [1, 225].

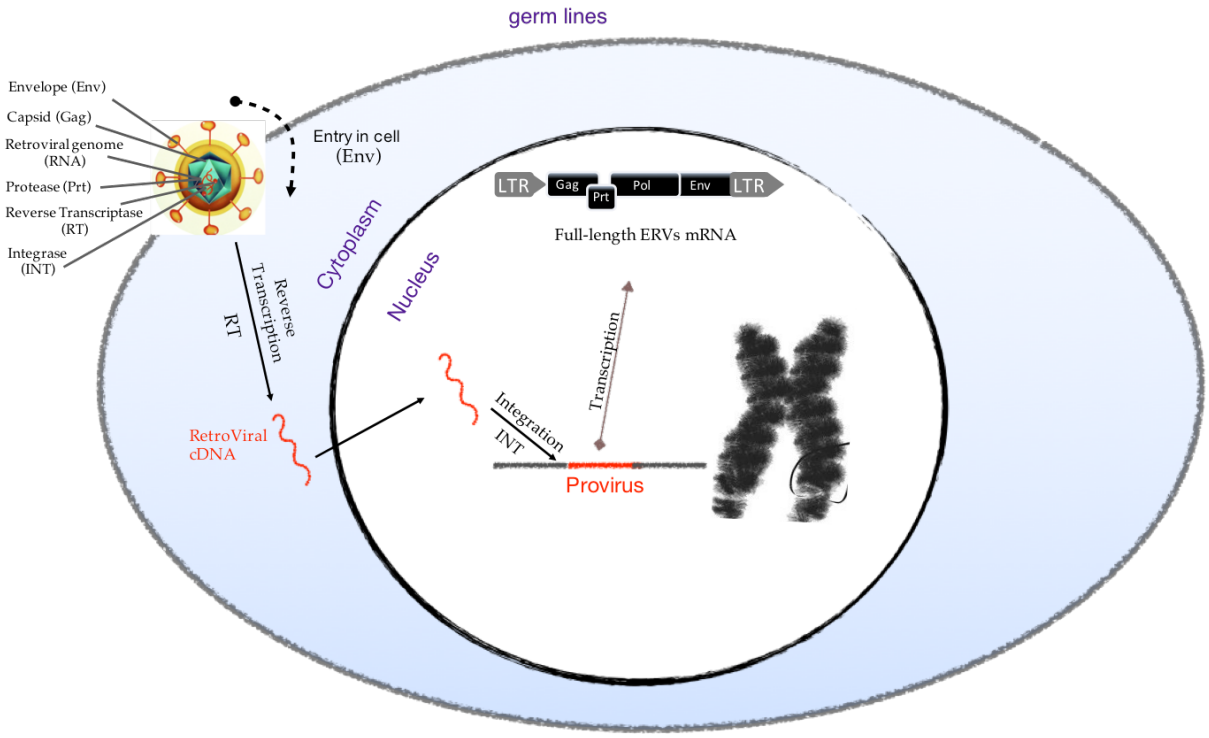


Figure 1.4: Spontaneous germ line virus infection and retroviral insertional mutagenesis in the genome lead to provirus formation. The proviral sequence uses host transcriptional machinery in order to get activated. Once the full-length intact sequences are transcribed followed by translation, they might produce viral particles that are infectious to other cells or they might use their reverse transcriptase to reintegrate elsewhere in the genome

Such studies and the calculated divergence between related elements indicate that the insertional activity of HERV-H was ceased at ∼10 mya. While ∼1000 elements harbour the full complement of retroviral genes, the rest has numerous mutations or deletions with no evidence of replication competence in the human genome [221, 226, 227]. A few copies have an open reading frame (ORF) coding for the immuno-suppressing domain of the envelope, but no evidence of an envelope protein production has ever been reported [226]. Indeed, most of the elements with two LTRs are roughly 5.7 kb in size and share several common deletions [228], suggesting that the partly deleted form of HERV-H became the favoured substrate for retrotransposition using the proteins provided by full-length integrants. HERV-H activity has generated a total of ∼6000 copies (including solitary LTRs), which are fixed in the modern human genome. The number of solitary HERV-H LTRs (∼1000) roughly equals the number of full-length or partly full-length elements. Such a ratio of full-length and solitary LTR is highly unusual among ERV families, which typically have much higher numbers of solitary LTRs, due to the tendency for recombination between 5' and 3' LTRs of proviruses over evolutionary time [229]. The reason for this unusual ratio is unclear, but it is tempting to speculate that maintaining full-length HERV-H elements is selectively advantageous for the host, possibly due to the essence of HERV-H transcripts.

1.3.1 Classification of endogenous retroviruses

ERVs are categorised into three broader classes on the basis of their sequential similarity with XRVs [230]. Class I includes the gammaretroviruses, II the betaretroviruses and III the spumaviruses (Figure 1.5). The conventional term used to categorize HERVs is ***"families"*** which are determined by phylogenetic analysis of their DNA sequences. Each family represent a single invasion followed by their amplification and expansion into the genome of the host [231]. HERV families are further suffixed after the single letter of amino acid on tRNA which is complementary to the PBS of the HERV's sequence. However, this nomenclature does not fit very well as various members of the same family do not harbour the same PBS [227]. HERVs are classified into 30 to 40 families but we will discuss the few of them in the following sections which have distinguished characteristics in human genome (Figure 1.5) [232, 233].

1.3.1.1 HERV-T

HERV-T is a *classical* example of the HERV family in the human genome due to its conventional *structural variation*; representative of those families on whom host had reacted faster to inactivate. Due to their truncated structure they did not proliferate much and exist as ∼60 copies in the human genome. I wrote *'classic'* since the similar host action is reflected over the structure of many other HERVs in the human genome e.g. HERV-P, HERV-S and HERV-K (HML5). Most of these families do not harbour much space in human genome as there are fewer than 60-80 copies indicating the subsequent loss of early transposition. Usually, their *enevelope* gene proliferated mainly by reinfecting human genome rather than retrotransposition [234, 235]. These HERV families had lost their mobility ∼35 Mya. The post-alignment phylogenetic tree of these HERVs gives a *star like* structure, suggesting the inhibition of their proliferation due to the accumulation of subsequent mutations during the course if their evolutionary life [236].

1.3.1.2 HERV-L

HERVL is the most ancient family of HERVs in human genome. HERV-L sequence is also shared by vast variety of the mammalian genome. Coalescence time of HERV-L is dated back to the ancestry of tetrapods [237]. Interestingly, the mouse version MERV-L controls the embryonic development and pluripotency of its host [238, 239, 240]. Strikingly, HERV-L is devoid of the *env* gene in the human genome, indicating that it had lost its re-infection abilities before the mobility, as the higher copy number is consequence of its retrotransposition events [237]. MERV-L is still mobile in the mouse, but HERV-L had lost its mobility in humans ∼30–40 Mya [237].

1.3.1.3 HERV-H

HERV-H is the most abundant intrinsic retrovirus in human genome and reasonably ***"central to present dissertation"***. HERV-H is a primate-specific endogenous retroviral family and was first noticed in human cancer cell on *December,1984* [241]. HERV-H harbours approximately one-third of total HERVs which contain *"pol"* segments [227]. It invaded the primate germ lines ∼30 mya during the bifurcation of old and new world monkey lineages [220, 218]. Unlike HERV-L and HERV-T, there are HERV-H loci in the human genome, which harbour intact full-length sequence with *"envelope"* genes [242]. Their activation performs an immunosuppressive function in human cells [242]. There are various reports demonstrating the mode of HERV-H amplification, mostly through retrotransposition, some through re-infection [221]

and trans-complementation [243]. It is repeatedly reported that HERV-H had lost its mobility ∼ 10 Mya [218].

1.3.1.4 HERV-W

HERV-W is primate-specific family that invaded before the split of Old and New World Monkeys [244]. This family occupies fewer copies in the primate's genome, because its activity perished after around 5 myr of mobility [245]. HERV-W, along-with its sister family HERV-FRD, was the first known HERV to be co-opted as their *envelope* subunits (*syncytin* genes) play pivotal roles in placentation process [51, 205]. Strikingly, HERV-W is also the only HERV family that has been shown to be mobilised, integrated and stabilised by proteins encoded by LINEs elements [246].

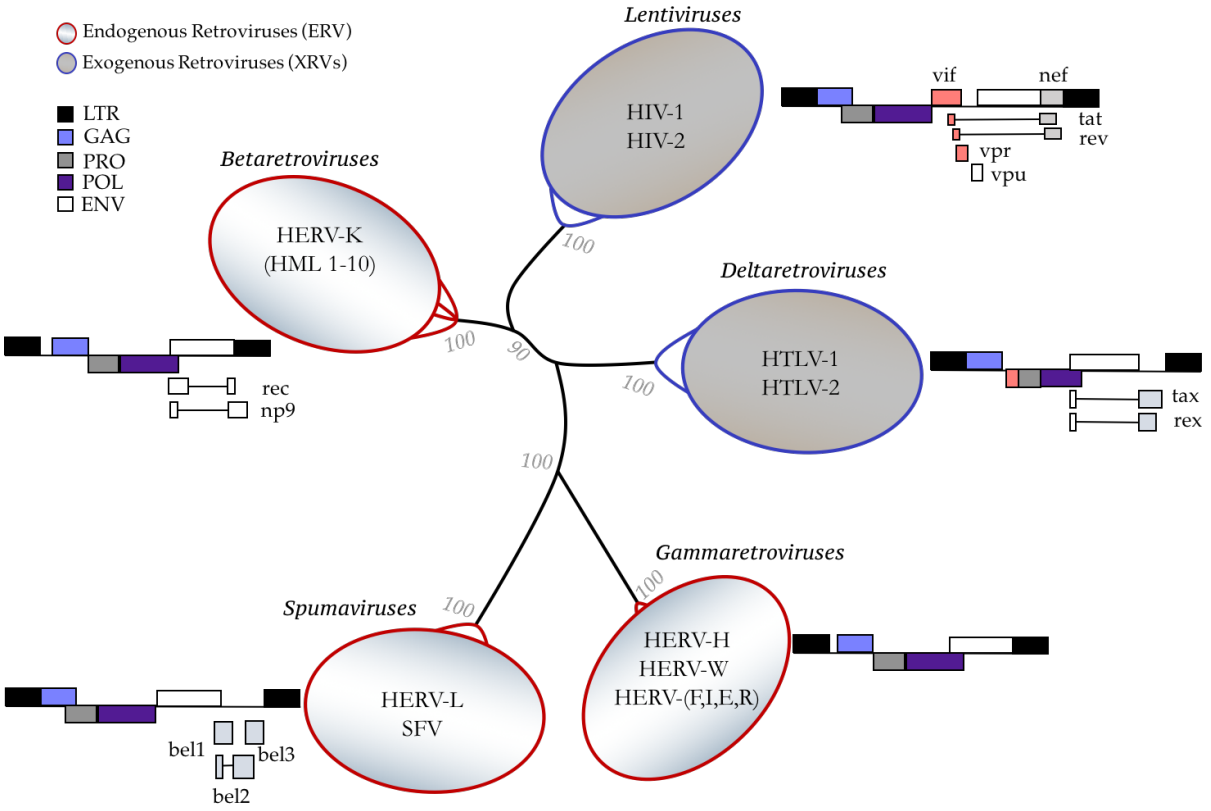


Figure 1.5: The schematic genome organisations for different classes of retroviruses are shown. The drawings are manually made and is not to scale. Note: *This figure is re-drawn from [247].*
Abbreviations HERV = human endogenous retrovirus; HIV = human immunodeficiency virus; HTLV = human T-cell lymphotropic virus; SFV = simian foamy virus; LTR = long terminal repeat, consisting of the U3, R and U5 regions in the integrated provirus, gag = group-specific-antigen, du = dUTPase, pro = protease, pol = polymerase (reverse transcriptase and integrase), env = envelope, bel 1-3 (bel 1 is also known as tas; the bel 2 reading frame overlaps with another one named bet), tax, rex, tat, rev, vpu, vif, nef and vpr encode small additional proteins. In spumavirus, either gag-pro or pro-pol are encoded in the same translational reading frame.

1.3.1.5 HERV-K

HERV-K (HML2) family is the youngest of all HERVs in human genome, being the only family that is still undergoing retrotransposition. It is speculated that HERV-K might still expand in the eventual generations of humans as evidenced by its polymorphic nature in the human population [248, 249, 243, 36, 250]. HERV-K (HML2) is the only family which has all open reading frames intact in our genome, since the host

has not got enough time to induce frameshift indels (see. *Indels*) or premature stop codons to inactivate them [249, 250, 251, 252, 253]. Several laboratories used consensus sequences of various full-length HERV-K loci to reconstruct infectious viruses *in vitro* [254, 255], shedding some light that HERV-K is potential to form viral particles inside the human cells.

1.3.2 HERVs in primate genome evolution

The role of HERVs in primate genome evolution was 'first' shown to cause or intervene with genomic rearrangements or DNA shuffling [256]. Next, it was illustrated that ERVs can alter the function of the beta-globin locus of primates [257, 258]. Most of the HERV families occurred in common ancestors of the primate genome after the divergence of old world and new world monkeys between 30 and 45 mya (Figure 1.6) [259]. Some of the HERV families predated the common ancestors of new and old world monkeys as they are older than 55 myr (Figure 1.6) [44, 260]. Since the host has got enough time to combat the deleterious effect of retroviruses, HERVs are infested by mutations, large deletions and insertions of other repetitive elements in order to be biochemically inactive as full-length viral sequences. An array of accumulated mutations during the course of evolution led the HERVs to loose their mobility in the human genome. As mentioned earlier, HERV-K (HML2) is the exceptional HERV family that has intensified its copy number after the human and *Chimpanzee* split [243, 261]. Although the extremely rare events witnessing the HERVs have been portrayed to be beneficial for their hosts such as in placentation (HERV-W and HERV-FRD) [51, 205], pluripotency (HERV-H and HERV-K) [262], embryonic development (various LTRs) [263] and X-Y Chromosomal dosage compensation [264, 265, 266] but the mechanism of their action is largely unknown. *This dissertation presents an another layer on existing knowledge about what HERVs are capable of.*

1.3.3 *Cis*-regulatory activities of HERVs

As discussed earlier, transcription of HERVs are mostly driven by LTRs that harbour the binding site for array of transcription factors. The diversity of HERVs between and within species in their copy number and structure, enormously impacts the outcome of the *cis-regulatory sequences* they carry. In particular, the new HERV integrant would introduce two copies of LTR in host genome [267]. Additionally, HERVs frequently endure the recombination between their LTRs that removes the coding region and leave the intact solitary LTRs. HERVs sequences make out 8% of the human genome whereas (Figure 1.3) solitary LTRs comprise 90% of total HERVs [1]. *"Exaptation", "Co-option", "Function", "Activity"* and *"Domestication"* are the terms widely used in TrEs research, however TrEs or HERVs gain these terms after undergoing evolutionary processes as illustrated in figure 1.7. Nonetheless, the term *"Co-option", "Exaptation"* [268] or *"Domestication"* [269] is refers to for HERVs when they are strongly selected for a cellular function that contributes to the fitness of the host organism (Figure 1.7). HERVs sequences undergoing co-option is classical example of evolution in the light of *purifying selection*. Evolutionary conservation of thousands of non-coding sequences derived from HERVs shows the strong selective pressure across the primates, while others remain evolutionary neutral [21]. The integration pattern of such repeat families is often close to the genes falling in particular ontological terms [12]. It has been shown for such elements that they are likely to acquire a stronger selection for controlling the transcription of the neighbour gene and and inclined to gain *cis*-regulatory co-option [270].

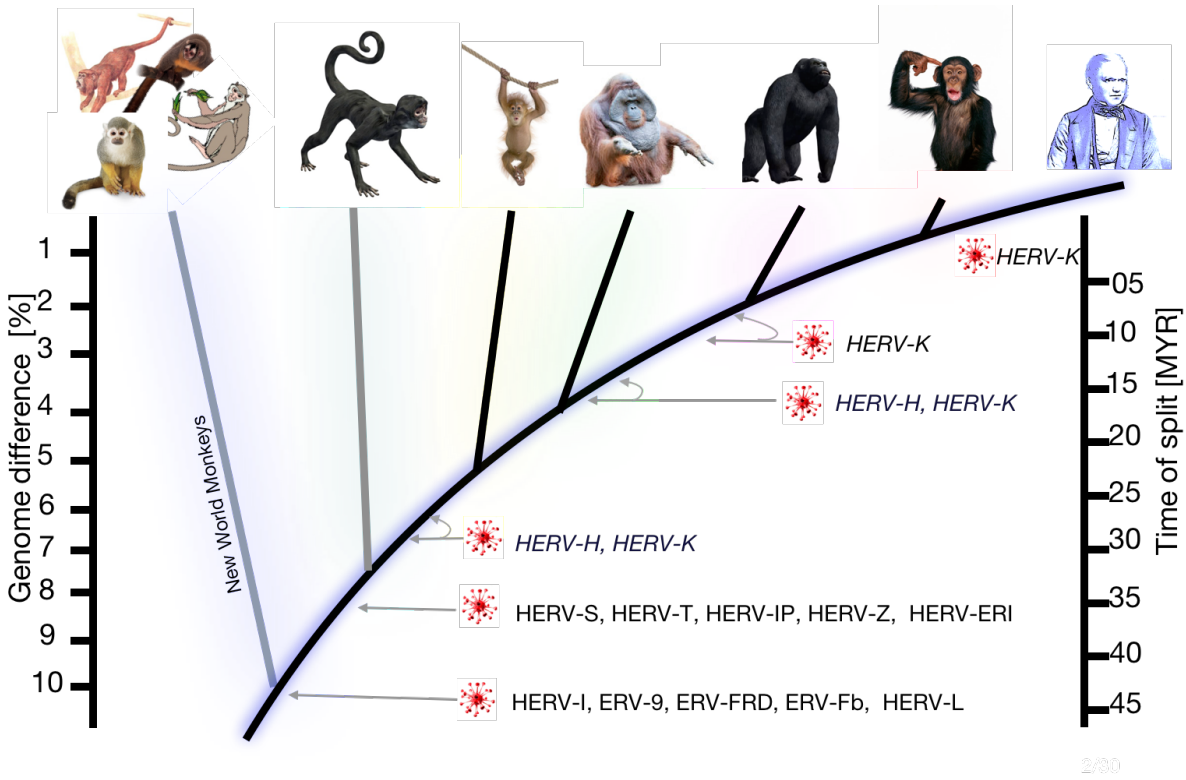


Figure 1.6: Major invasion and expansion of human endogenous retroviruses (HERVs) occurred after the platyrrhine (New World Monkeys) lineage separated from the catarrhines (Old World Monkeys and apes). As the consequence of repeatedly integration, amplification and expansion during evolutionary periodogram of humans, the increment in genome size is directly proportional to waves of integration

Co-option of retroelements has never been shown immediately after their integration into the genome, however, it was unexpectedly found in plants [271]. In addition to provide the *cis*-regulatory DNA elements, TrEs have also been documented to contribute a wealth of non-coding regulatory transcripts, such as microRNAs (miRNAs) and long non-coding RNAs (lncRNAs), which could re-wire the gene expression in *cis* or as *trans-regulatory sequences* [211]. Its abstruse that how evolutionary selective forces such as the *genetic drift* deals with the case that HERV integrants were preferred over original regulatory sequences [272, 273]. Strangely, most of the retroelement-derived orthologous regions across primates are of discrete age & type, unevenly distributed and plugged near TFs and developmental genes [204, 12]. Few of the upstream ERVs, mainly their LTR sequences, are shown to drive their neighbour genes, mainly stabilising the fitness of embryonic development, pluripotency, placentation and innate immunity, as shown in table 1.3. Further examples can be found in the C-GATE (`https://omictools.com/catalogue-of-genes-affected-by-transposable-elements-tool`) [17]. The re-activation of few of these elements elements cause fatal human diseases such as cancer [274].

1.3.4 *Trans*-regulatory activities of HERVs

In addition to contributing to cis-regulatory DNA elements, bidirectional promoters and driving neighbour genes [275], HERVs are also a rich source of non-coding regulatory RNA transcripts such as microRNAs (miRNAs) [211] and long non-coding RNAs (lncRNAs) [209] that can regulate the gene expression either in *cis* or *trans* [211, 209, 263, 276]. This supports the idea that HERVs are creative sources of regulatory networks in host cells referred as *re-wiring the host's network* transcriptionally or post-transcriptionally.

Species	gene product	TrEs	*Cis*-regulation	Function	Evidence	Reference
Human	AIM2	MER41 (Primate)	Interferon-inducible enhancer	inflammatory response	*CRISPR knockout*	[12]
Human	β-globin	ERV9 (Primate)	Erythroid enhancer	developmental switch	*Cre-LoxP knockout*	[278]
Human	Prolactin	MER39 (Primate)	Endometrial-specific promoter	pregnancy	*5' RACE*	[279]
Human	CYP19	LTR7 (Human)	Placenta-specific promoter	placentation	*Reporter assay & 3C*	[280]
Human	cAMP	MER20 (Eutherian)	Placenta-specific regulator	placentation	*ChIP-seq & Reporter assay*	[281]
Human	CSF1R	THE1B (Primate)	Promoter	Embryonic development	*ChIP-seq & RNA-seq*	[274]
Human	HPAT5	LTR8 (Human)	Promoter	Embryonic development	*Reporter assay & RNAi*	[211]
Human	LINC-ROR	LTR7 (Human)	Promoter	Pluripotency	*Reporter assay & RNAi*	[282]

Table 1.3: Chromatin conformation capture (3C); Rapid amplification of cDNA ends (RACE); Chromatin immuno-precipitation followed by sequencing (ChIP-seq); High-throughput sequencing of complementary DNAs (RNA-seq); (*See Reporter assay*)

Notably, transcriptome-wide analyses have revealed HERVs as an abundant source of tissue,specific and/or non-coding RNAs [277, 209, 211].

Mainly, two methods were implemented to map the HERVs derived promoters and non-coding RNAs at a genome-wide scale. First, RNA-seq revealed that chimeric transcripts that initiate within HERV locus constitute a considerable fraction of human transcriptomes, notably during early development and pluripotent states [276, 283, 208]. Second, approaches mapping sites of transcription initiation, such as displays *Pol II* initiation within vast array of HERV locus [284]. Along-with chimeric transcripts formation by providing TSS, HERVs sequences are also enriched in TES contributing to form UTR of protein-coding genes. These events may explain the HERVs provide a convenient mechanism for host genes to evolve new expression patterns and isoforms.

1.3.5 HERVs *re-wire* the host's regulatory networks

The experimental evidences illustrating the role of HERVs in the host gene regulation convincingly persuades that HERVs re-wired networks do exist in a cell-type manner [270]. The exaptation of HERVs can be classified according to the mechanism through which they influence the transcriptome (Figure 1.8). Recently developed reviews focused on how the HERV sequences are transformed as exaptation into enhancers, promoters or as a source of novel non-coding and protein-coding genes (Figure 1.8) [285].

Recent evidences are pointing out that HERVs provide a platform of other types of *cis*-regulatory elements including enhancers, insulators and repressive elements as revealed by ChIP-seq experiments. LTRs are among the substantial fractions of binding sites (5–40%; average 20%) for a given transcription factor and cell type examined across the genomes [163]. Notably, the small number of specific LTRs contribute the majority of binding sites for a given transcription factor compared with the density of these TE families in the genome [103]. The cellular function of diverse HERV-derived lncRNA is described as extremely specific to a particular cell-type. Among the lncRNAs that are initiated from HERV sequence and were shown to be functional are: linc-RoR, is vital for pluripotency and modulates p53 levels as a DNA damage response [286, 287]; HPAT5 is HuERS-P1 (p element like human endogenous retroviral sequences) that modulates early development and reprogramming via let7 miRNA interaction [211]; and BANCR, SAMMSON and UCA1 are the the lncRNAs that have potential to alter the phenotypes of several cancer cells [288, 289, 290]. The diversity of functions is attributed to the mentioned lncRNAs which are all promoted by HERV sequences (Figure 1.8).

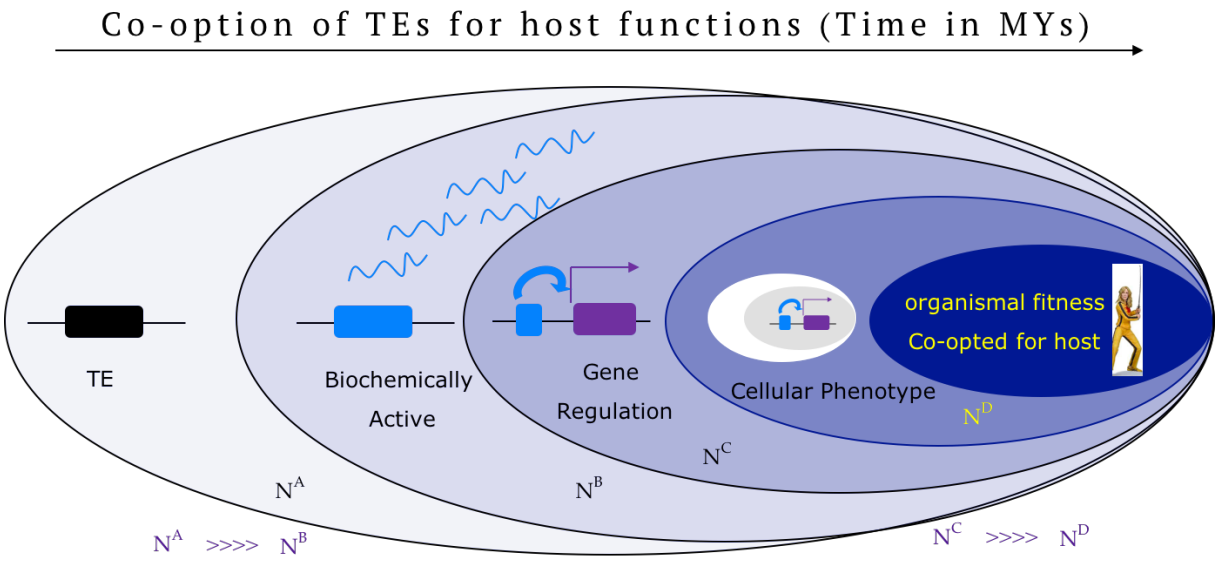


Figure 1.7: A journey of of TrEs from junk sequences to determine whether a TE has been co-opted for host functions. Many TrEs have biochemical hallmarks of regulatory activity on the basis of genome-wide assays. However, additional evidence is required to determine which of these TrEs alter the regulation of host genes and affect the organismal phenotypes and fitness. Note: This figure is redrawn from [12]

1.3.6 HERV-H is the most abundant ERV in the human genome

The first HERV-H copy, a deleted version of the full-length elements, was identified from the human genome in December, 1984 [241]. It was a 5.6 kb repetitive sequence that consisted of the 415bp LTRs, flanking the internal sequence and a histidine tRNA primer binding site that was located directly after the 5'LTR. The regulatory LTR sequences which flank the HERV-H internal sequence (HERV-H,int) are characteristic to HERV-H elements and the HERV-H related LTR subtypes are LTR7, LTR7A, LTR7B, LTR7C and LTR7Y [7], representing different evolutionary age. Other than their regulatory LTRs, HERV-Hs, are further characterised based on their structure as complete, slimmed down, substituted and solo LTR elements [291]. The human genome contains around 50,100 copies of the almost intact forms of HERV-H with a full-length size of 8.7 kb [7, 228]. A few HERV-H copies have an *env* open reading frame which, however, has not been found to produce a protein [292]. Although the almost intact HERV-H copies have the full repertoire of retroviral genes, these carry several mutations or deletions and are not replication competent [32, 7, 228]. The vast majority of the HERV-H integrations in the human genome originate from a common 5.8 kb form with a structure of 5'LTR-gag-pol-3'LTR [292] that carries large deletions in its pol coding region [218] and lacks the *env* coding region. The integrity of an inactive ERV is typically not protected, and their sequences are exposed to various degradation processes such as homologous recombination that generates solitary LTR copies, and various indels, resulting in fragmentation and even in complete *'stochastic loss'*. During the degradation process, the active full-length elements, operating as autonomous retrotransposition machineries, might mobilise the more common and partially deleted non,autonomous forms of ERVs [293, 218]. What is noteworthy about HERV-H compared to other ERV families is the unusually high number of full-length and partially deleted insertions relative to solitary LTRs [285, 263], suggesting that the degradation of HERV-H copies occur at a much slower rate compared to other ERVs [294]. With 1060 copies flanked by LTRs and another 1270 copies of solitary LTR sequences per haploid genome, HERV-H is the HERV family with

the highest copy number in the human genome.

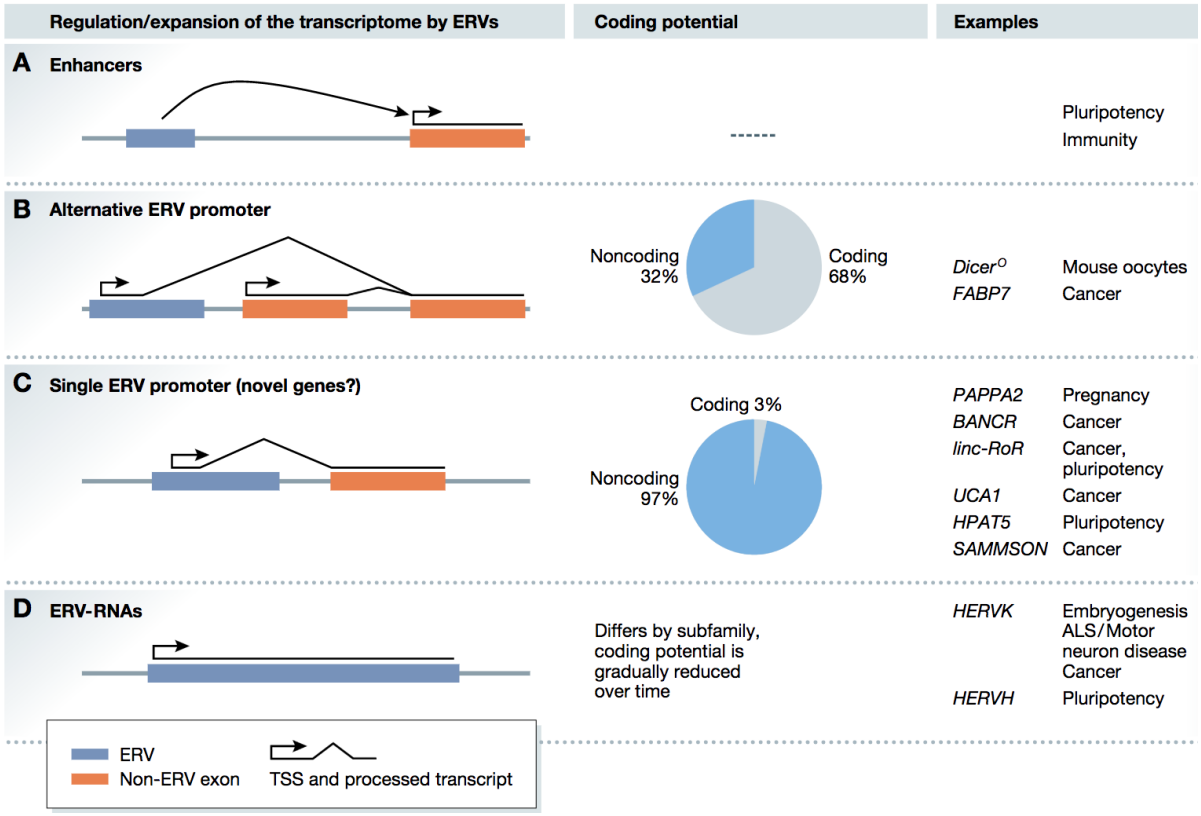


Figure 1.8: Firstly, retrotransposons can be co-opted as enhancers, influencing the expression of nearby genes without activating the retrotransposon itself. Secondly, retrotransposons can act as promoters that initiate transcription at the retroelement. Such elements can increase gene isoform diversity and introduce novel cell-type specificity for existing protein-coding genes. In addition, they may act as their own promoter, driving the expression of retrotransposon-derived RNAs.
A. ERVs can act as enhancers, regulating genes in their proximity.
B. ERVs can act as alternative promoters for protein-coding and noncoding genes.
C. ERVs can provide the only promoter for a gene; such ERV-derived genes are largely noncoding.
D. ERVs can be transcribed over their full length. Transcribed ERVs can generate proteins and peptides, but they can also generate noncoding RNAs.
Note: *This figure is adopted from [285] with the consent of Jonathan Göke.*

1.4 Human early embryogenesis

Our understanding of developmental pathways underlying early embryogenesis owes to the previously studied mammalian embryonic development, mainly the mouse ones. However, there are some major and minor species-specific differences, mainly the gain and loss of genes in transcriptional networks, timing of *embryonic genome activation*, *aneuploidy* and the patterns of epigenetic modifications (Figure 1.9) [295]. This level of disparity may restrict the prognostication of some conclusions about human early embryogenesis. Till date, our knowledge about human early embryoenesis is limited to few *In vitro fertilization* and single cell RNA-seq in order to understand the different aspects of the human pre-implantation development [296, 297, 298]. Recently, the advancement of sequencing technologies enabled us to understand the high resolution of cellular and molecular biology of distinct developmental stages [296]. Currently our understanding about the progression of human embryogensis is described in

figure 1.9, whereas the number of genes participating in major changes in 1.9 are employed from a current study.

1.4.1 Distinct stages of pre-implantation embryonic development

Human embryonic development begins with the loss of the maternal control over the transcriptional machinery during the oocyte to *embryonic transition* that lasts for approximately 3 days. The list of events during the first 3 days circumscribe the fusion of the egg and the sperm, the migration and fusion of the germ cell pronuclei, *genetic and epigenetic reprogramming* and a series of *cleavage divisions* which conclude with the major wave of *embryonic genome activation* (EGA) at the transition of the 4 to 8-cell stages (Figure 1.9) [299, 300]. Some studies also found the paternal transcripts but only at day 2 of embryogenesis [301]. The first transcriptomic study had shown that as much as 1800 mRNAs were dynamic (mainly impeding from oocyte) in the first 3 days of human embryogenesis [302]. The progression of human embryogenesis differs at a considerate level from those in rodents and hamsters, where ZGA is switched-on during the zygote to 2-cell stage [303, 304, 305]. The human EGA corresponds to the mouse ZGA as the first major wave of genome activation after the fertilisation of the egg by the sperm; however, the mechanism of progression is obscure [306].
Following the EGA, the embryo undergoes a successive *compaction* to form a morula that flags the first morphological deviation from radial symmetry (Figure 1.9). From here on, consecutive cell divisions trigger the initiation of the blastocyst formation with the development of fluid-filled cavity, the *inner cell mass* is encircled by *trophectoderm* (Figure 1.9). The initiation of the blastocyst formation at *embryonic day 5* (E5) is followed up by maturation of the blastocyst at *embryonic day 6* (E6) and, finally, the stabilisation of the blastocyst formation at *embryonic day 7* (E7) is achieved just before the blastocyst implants into an uterine wall. The development of the blastocyst is the step-wise process as ICM from E5 further branches out to form *epiblast* [307, 308, 309, 296], the only pluripotent cell population of the human embryos and the *primitive endoderm*, an epithelial layer essential for tissue patterning [310, 309]. So, it takes 7 days of embryonic development before embryos attaches to uterine wall in humans, the process being called *implantation*. This intra-uterine life of embryos which is endured before and after the attachment, termed as pre and post-implantation respectively.

1.4.2 Human specific nature of embryogenesis

The global epigenomic reprogramming during early development is associated with massive transcriptional reactivation of HERVs [165]. HERVs exhibit a developmental stage-specific pattern that might be taken as hallmarks for certain stages of early development [263]. MERVL associated transcripts expressed at 2-cell stage are assumed to be crucial for totipotency in mice [311]. Reactivation of MERVL in mouse induces many 2-cell-specific transcription products. Many of these transcripts are directly promoted by LTRs of MERVL or related MaLR elements, or are chimeric transcripts derived from MERVL/MaLR LTRs. Similarly to mice, TE reactivation occurs in a well-defined waves during human embryogenesis, still involving a distinct set of LTR and HERVs [276, 312, 263]. At morula and blastocyst stages most of the TE-derived transcriptional reads derive from HERV-H [312]. Retroelement's mediated biogenesis of various non-coding transcripts which regulate the gene expression either transcriptionally or post-transcriptionally via microRNAs (miRNAs) [211], long non-coding RNAs (lncRNAs) [209] and circular RNAs (circRNAs) [69].

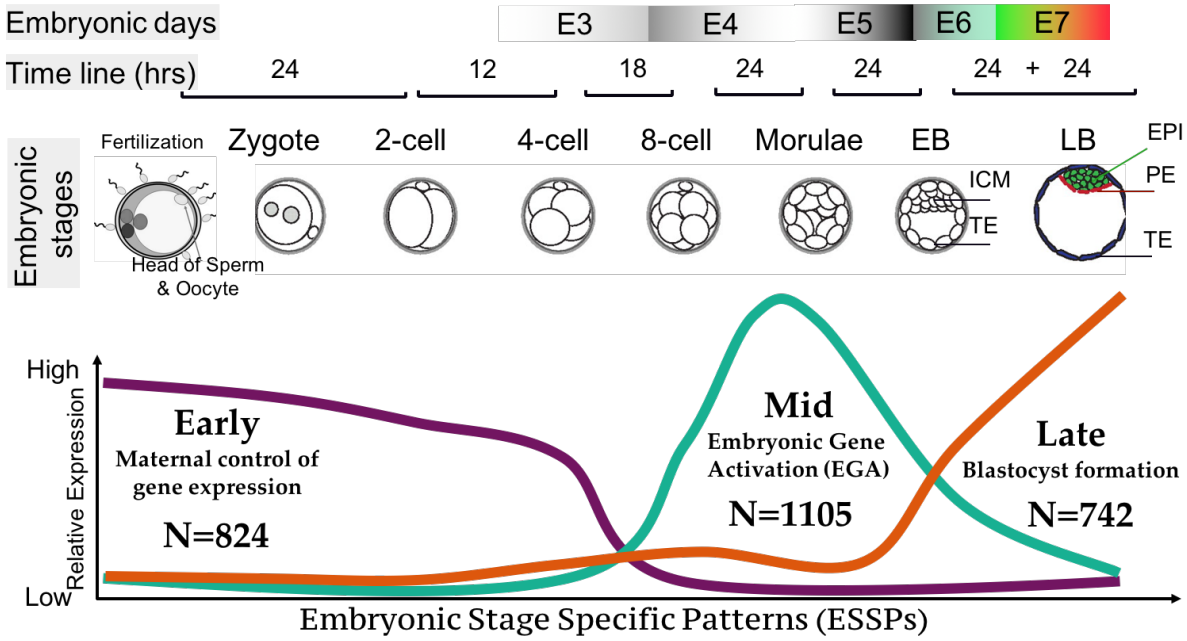


Figure 1.9: Human embryonic development from day (d) 0 to day 7. Following the fertilisation, embryos undergo a series of mitotic cell divisions. Arrowheads in d0 and d1 indicate pro-nuclei. On or around day 4, the embryo compacts, resulting in the formation of a morula or blastomeres that consists of cells in a compact cluster contained within the zona pellucida (the glycoprotein layer that surrounds the embryo). The blastocyst, which forms on day 5, is a fluid-filled structure composed of an *inner cell mass* and *trophectoderm*. On day 6, the blastocyst 'hatches' from the zona pellucida and it is ready to implant into the uterine wall on day 7, where the stabilised blastocyst is formed consisting of *epiblast*, *primitive endoderm* and mature *trophectoderm*
Potential application of predicted embryo developmental potential in *assisted reproductive technology*. Three distinct lines are k-means clustering of single cell transcriptome that show the loss of Maternal control of gene expression at day 3 (8-cell stage). There starts the *embryonic genome activation* (EGA) that is also lost after two days where blastocyst formation starts at day 5 which proceeds to be stabilised at day 7

Early human development is thought to have unique or exceptional properties in at least two regards. First on a gross level, the cell types and developmental trajectories of those cell types are thought to be unusual. In mice, the preimplantation embryogenesis is in step-wise manner and achieved as a consequence of three major waves: first in the zygotes where the maternal control of the gene expression is lost (ZGA). Second in the morula, where TE and the ICM is diverged, and finally in the blastocyst, where Epi and PrE has arisen from ICM [313]. While in human, first wave is embryonic genome activation instead of the zygotic genome activation (discussed in previous section) [295], second wave is segregation of morulae into TE, EPI, and PE simultaneously rather than a step-wise manner [314]. This suggest that an ICM population does not exist as discrete group of the progenitor cells during the initiation of the human blastocyst formation [314]. Similarly, it is in conjecture that multiple cell types develop simultaneously in humans rather than through an ordered series as it has been illustrated in other species (Figure 1.10). [313, 295, 314]. On a different level, there are multiple genes expressed in early human embryos which are either human specific (e.g. ESRG) [312] or primate specific (HPAT5) [211]. Some of these lineages or clade-specific genes are the transcriptional products of *transposed elements* particularly with one endogenous retrovirus, HERV-H, coming in for scrutiny. These two observations suggest a series of questions which will be addressed in present dissertation. An overarching question concerns the relationship between novel gene expression and novel developmental trajectories. However, before addressing any such issue, we must first establish the extent to which development trajectories are in

fact distinguished from that of other species. While it has been a consensus that human preimplantation embryogenesis has unique features compared to the early development of murine or even non,human primates, the fine-tuning of the blastocyst formation, and the definition or lack thereof of the ICM through human early development has been never being fully resolved, even by using single cell data [296]. Previously, around one third of the cells could be unambiguously identified, leading to premature outcomes and stagger speculations regarding the human preimplantation development.

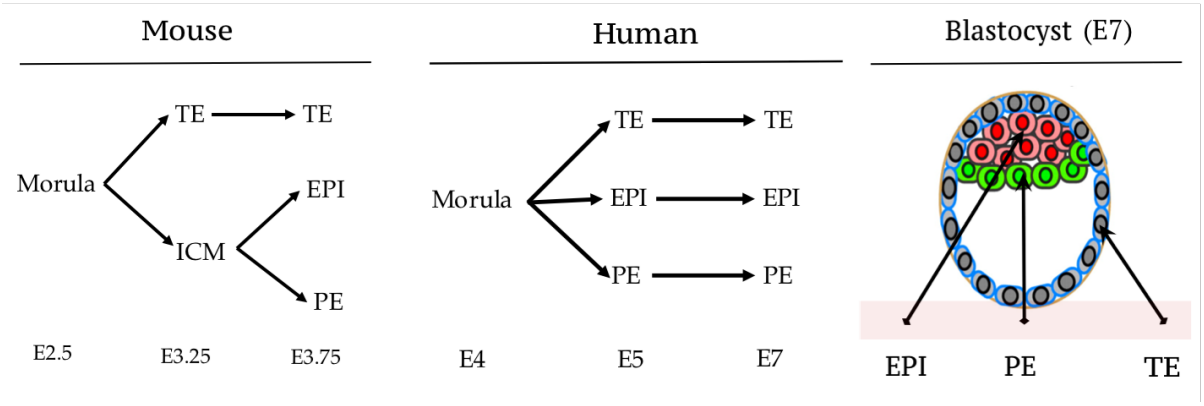


Figure 1.10: Both, mouse and human lineage specification starting from the morula till the formation of the blastocyst. The mouse pre-implantation development is more rapid than that in human, resulting in different timing of events (E represents embryonic day post-fertilization). In the early mouse development, the lineage segregation is stepwise. First, the trophectoderm (TE) and the inner cell mass (ICM) segregate from the morula, later the emergence of the epiblast (EPI) and the primitive endoderm (PE) from the ICM follow. However, humans display a concurrent rather than step-wise lineage segregation where the TE, PE, and EPI emerge simultaneously. The mouse time-course is adapted from recent review [313], human time course is adopted from recent research work [296, 298, 314]

1.5 Human pluripotent states in petri dish

The resolution of the molecular mechanisms underlying the human development occurred when spare human pre-implantation embryos have enabled the derivation of hESCs in the *petri dish*. This led to the establishment of novel tools for human developmental biology and the emergence of hESC-based regenerative medicine. Consequently, these cells in the petridish are maintained to be self-replicating (see. *self-renewal*) and still have potential to form any of three germ layers (see. *pluripotent*) namely, ectoderm, mesoderm and endoderm that would further form the human fetus. The central aspects of the human pre-implantation development might provide not only insights into human developmental biology and common birth defects, but also potential benefits for reproductive health and improvements in regenerative medicine, and the pluripotent cells are the way to do so.

1.5.1 Derived and induced pluripotent states

Since the human embryonic development is progressive in nature, thus it is intricate to catch them *in vivo*. This troubleshoot was duly resolved by deriving cells from different stages of the early embryonic development and maintained indefinitely in an artificially induced self-renewal state *in vitro* [315, 316, 317]. Pluripotent stem cells are annotated based on their origin of the donor cells from which they are

isolated from. that can be isolated from. Embryonic stem cells (ES cells) are isolated from the inner cell mass (ICM) of developing pre-implantation blastocysts of a given organism. [315, 316, 318]. Epiblast stem cells (EpiSCs) are isolated from the mouse post-implantation epiblasts [319, 320]; however there are no such known attempts to derive equivalent cells from human embryos due to ethical reasons. In the case of rodents, even PGC can be converted *in vitro* into pluripotent like cells, known as embryonic germ cells [321, 322]. However, there are no stable and validated embryonic germ cells that has been isolated from primates with higher fidelity, so far, opening an another window to research on human pluripotent states. [323, 324].

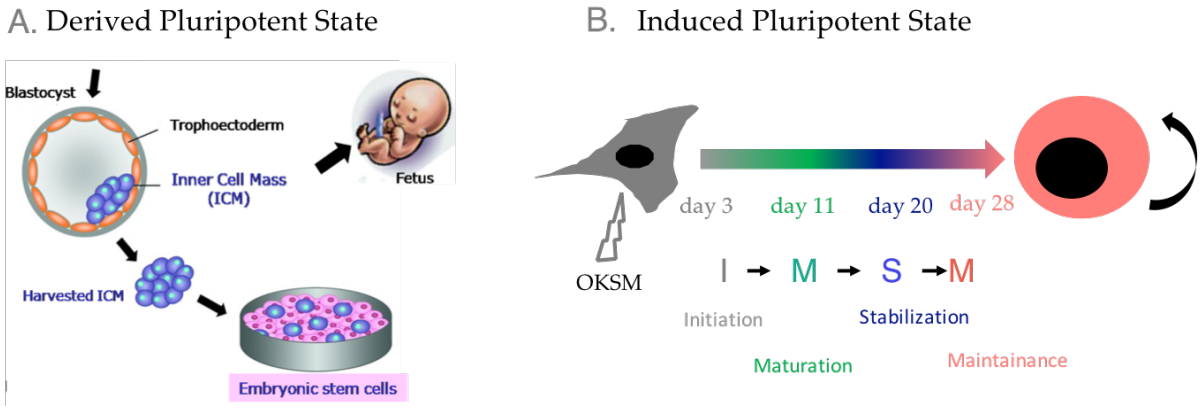


Figure 1.11: This schematic figure illustrates the artificial pluripotency in two dimensions. The left panel displays the derived pluripotency where the cells from early blastocysts are isolated and grown in the petri-dish. Cells tend to loose their native property as they are brought to the petri-dish. Right panel shows the reprogramming of somatic cells to pluripotent state by induction of four transcription factors. Cells tend to undergo multiple stages prior to resemble like pluripotent cells. Initiation of reprogramming costs the loss of somatic genes and activation of early embryonic gene expression. Maturation stage corresponds to the loss of exogenous expression of OSKM (Oct4, Sox2, Klf4 and c-Myc) and activation of endogenous expression of OSKM. Finally, the stabilisation stage is achieved when cells gain self-renewal property that is exclusive to epiblast cells in human embryos [313, 296, 298, 314]

Pluripotency can also be restored in cells of post-implantation embryonic stages, such as EGC [322, 325], SCNT [326, 327] and somatic reprogramming with an induction of transcription factors (iPSCs), which has earned ***Shinya Yamanaka*** a *"Nobel Prize in Physiology and Medicine"*, since it has turned out to be revolutionary tool in regenerative medicine [328]. The outstanding badge of iPSC technology is its simplicity and reprehensibility as few transcription factors turn back the cellular clock. While its efficacy keeps on improving every year, another stunning approach is to skip the iPSCs and reprogramming directly from one somatic cell type to another desired somatic cell type (see. *Transdifferentiation*) [329, 330]. Almost a decade after the discovery of Human iPSCs (hiPSCs), they are being widely used for disease modelling, drug discovery and cell therapy [331]. Several reports indicate its preponderance achievement in the discovery of new drugs from iPSC screens and consequently, several clinical trials using human iPSC-derived products has been initiated [332, 333].
We owe our understanding of pluripotency to the studies mentioned in this section, especially the discovery of iPSCs that has played the role of *"game changer"* in regenerative medicine, *precision medicine* and stem cell biology [334]. These artificially reverted pluripotent cells have been expected to be of higher similarity in the terms of characteristics and molecular functions with the embryonic epiblast. Nevertheless, recent studies have illustrated the crucial verdict that there are significant molecular differences in the spatio-temporal domains between any two pluripotent stem cells analysed [335, 336, 337, 338].

Table 1.4: Summary of cell culture methods for undifferentiated human naive cells. 2i, 3i, 5i and 6i corresponds to the number of inhibitors for human biological pathways i.e. MEK, GSK3β, BRAF, ROCK, SRC and JNK signalling respectively. Fourth column name "Transcriptome R" means the transcriptome-wide resemblance of naive cells

Strategy	Supplements	Morphology	Transcriptome R	Methylation	Clonogenicity	Chimera	Derivation	Ref
KLF4/KLF2 OE	LIF/2i	Dome shaped	*mouse naive ESCs*	NA	Improved	NA	PSCs	*[317]*
small molecules	LIF/3i/4i/5i	Dome shaped	*mouse naive ESCs*	Hypo	Improved	Yes	PSCs, blastocyst	*[357]*
small molecules	LIF/3i/iBMP	Compact raised	*mouse naive ESCs*	NA	Improved	NA	H1-ESCs	*[358]*
NANOG, KLF2 OE	LIF/2i/Gö6983	Dome shaped	*human blastomere*	Hypo	Improved	Yes	H9-ESCs	*[359]*
small molecules	LIF/2i/bFGF	Compact small	*mouse naive ESCs*	NA	Improved	NA	PSCs	*[360]*
Oct4 enhancer	LIF/5i/bFGF/Activin	Dome shaped	*human blastomere*	NA	NA	No	NA	*[361]*
small molecules	LIF/2i/bFGF	Dome shaped	*mouse naive ESCs*	NA	Improved	NA	Blastocyst	*[362]*
small molecules	LIF/3i/bTGF	Compact raised	*mouse naive ESCs*	NA	NA	Yes	PSCs	*[363]*
small molecules	LIF/2i/bFGF/FK	Dome shaped	*mouse naive ESCs*	NA	NA	NA	PSCs	*[364]*
small molecules	LIF/5i/bFGF/Activin	Dome shaped	*mouse naive ESCs*	NA	Improved	NA	Blastocyst	*[356]*
Recombinant protein	NME7AB	Dome shaped	*mouse naive ESCs*	NA	Improved	NA	PSCs	*[365]*
small molecules	LIF/5i/6i/bFGF/Activin	Dome shaped	*human blastomere*	Hypo	NA	NA	PSCs	*[366]*
NANOG, KLF2 OE	LIF/2i/Gö6983	Dome shaped	*human blastomere*	Hypo	NA	NA	Blastocyst	*[367]*

The morphological features of iPSCs and ESCs are similar [339, 340, 341], in spite of significant differences between the transcriptomes and genome-wide methylation patterns among the both of them [342, 343, 344]. Recent updates indicate that most of the variations between pluripotent states are due to the genetic background of donors (i.e. , different donors for iPSCs and ESCs) which has allotted for most of their regulatory differences [345]. A study went further on to show the differences between pluripotent stem cells generated from from one laboratory to another [341]. Unlike epiblast cells (Natural pluripotent cells), all these derived or induced pluripotent cells have the common property of exhibiting self-renewal while being potential to be differentiating to further lineages [307, 295, 309, 308, 346, 296, 314].

The recent revelations of cellular and molecular mechanisms underlying the developmental programs leading to a natural pluripotent state. The inability of artificial pluripotent states to reproduce them is major hitch to the immortality of hiPSCs/hESCs. Repeatedly asked questions while comparing natural and artificial pluripotent states are; does hiPSCs/hESCs have blueprint for setting up the entire body plan as epiblast cells do? The blunt answer is 'no', since the human pluripotent states in dish cannot give rise to germ-line chimeras (see. *xeno-pluripotency*) [347]. It was even indicated that embryonic stem cell therapy could encounter the same problems as organ transplants do [348]. When the first clinical trial proceeded on two AMD patients, their skin cells were reprogrammed followed by differentiating to RPE cells, the implanted RPE sheets generated genetic changes in both the patient's iPS cells and the RPE cells derived from skin cells. Afterwards, the trial was put on hold in order to avoid any complications due to those mutations. Artificial pluripotency, as we see it today, is no longer a perfect system to treat human diseases, as they lack *xeno-pluripotency* [349, 337, 350], the induction of coding-region mutations during reprogramming [351, 352], the lack of immune response [353] and the loss of methylation memory from developmental stages [354, 355, 356].

1.5.2 Naive and Native pluripotent cells

As shown in the previous section, artificial pluripotent states confer some limitations when it comes to *precision medicine*, although there is always room for an improvement. The major pursuit of improvement is to mimic the *in vitro* cells to the *in vivo* cells along-with preserving their potential to form the live embryo. Intriguingly, which cells from preimplantation embryos should be kept as reference frame to compare the *in vitro* cells, is still debatable [335, 357, 362, 366]. The progressing cells from oocyte to the morula in human embryonic development are *totipotent*, since they can give rise to an entire embryo

but have some limitations to be treated as reference frame. The development is not affected even if one or more cells being removed from the 8-cell or morula stage [368] questions the essential role of asymmetric division of the originator cells from oocytes. The zygote is the cell formed post-fertilisation of an egg by the sperm, that has been shown to be differentiation deficient, whereas if the *blastomeres* are dissociated and cultured in petri dish, then less than two cells could form the trophoblast vesicles or microblastocysts [369, 370]. Most importantly, as long as the blastocyst is not formed, the fate of an individual cells is determined by their position within the embryo [371], so it would not be a smart choice to treat these cells as a reference frame for *in vitro* cells. Firstly, the most potent cells were found to be ICM which is flexible and have the capacity to generate the rest of the other cell types [372]. The ICM gives rise to the hypoblast (primitive endoderm), an extraembryonic lineage and the rest of cells develop into the pluripotent epiblast, which has noticeable disparity with the blastomeres and ICM [368, 373, 374]. One of the breakthrough work that sets up the reference frame, has illustrated that mouse embryonic stem cells (mESCs) were derived from ICM but was originated from epiblast [375]. After being contentious for 14 years, this finding was further confirmed and epiblast is established as gold standard for pluripotent states [376].

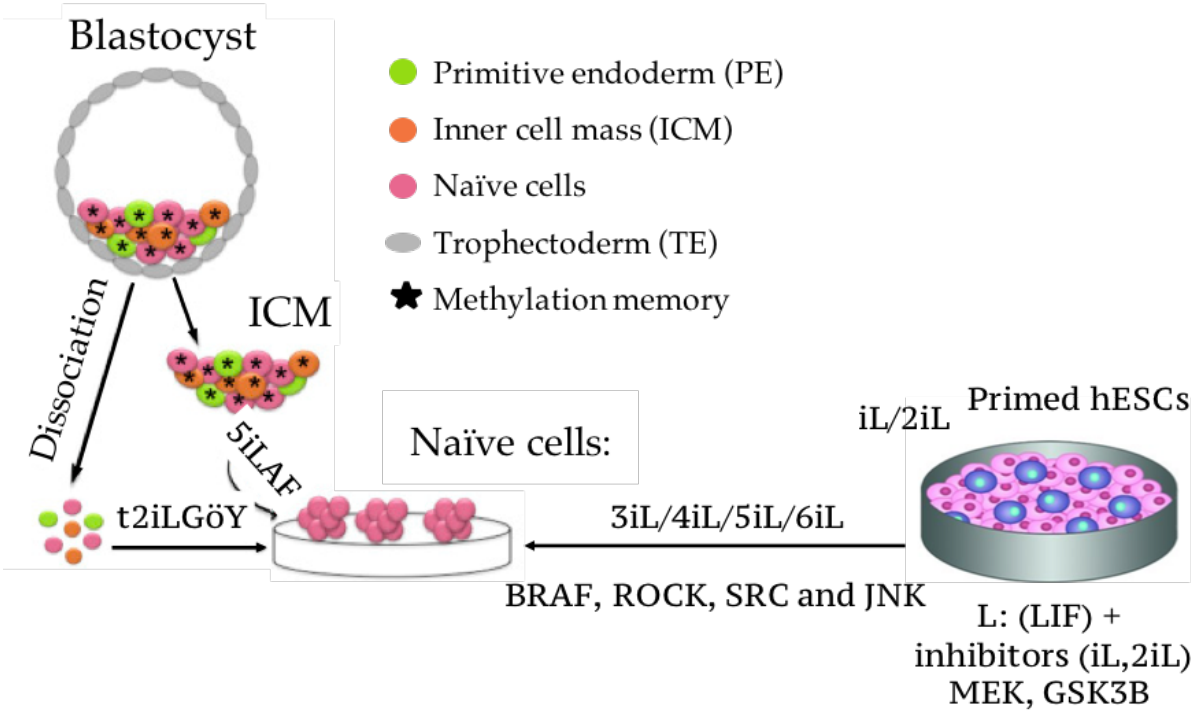


Figure 1.12: Derivation of Human Pluripotent Cells with Naive Properties: Human naive pluripotent cells can be derived from individual ICM cells of pre-implantation embryos in the t2iLGöY condition, or from ICM or primed PSCs in the 5iLAF condition. The naive cells reverted from primed hESCs or derived from ICMs in 5iLAF, characterized by the loss of SSEA4 expression, exhibit a methylation landscape with little similarities to that of the blastocyst as well as loss of imprinting. Note: This figure is adopted from [356]

So, here we make an argument that the epiblast is a ***Native pluripotent state*** (Natural niche of pluripotency) that should be kept as reference frame to compare it to any *in vitro* cell line that is being claimed as ***Naive pluripotent cells***. The generation of naive cells is an attempt to diminish the gap between natural and artificial pluripotency and a contest to apprehend the missing ICM like human pluripotency in petri dish; their advantages and disadvantages have already been heavily reviewed

[336, 337, 377, 378, 379]. There has been several claims from various research groups that the human naive state is successfully derived encompassing hESCs or hiPSCs in the last five years (Table 1.4) [380, 317, 357, 367, 359, 361, 366]. Nevertheless, attempts to generate preimplantation chimeras using claimed naive cells have shown either low efficacy or no chimeraes at all, on the top of this, the yield was too low to perform pluripotency assays (Table 1.4) [357, 361, 359, 381].

1.5.3 Evolution of pluripotency in primates

The blastocysts comprise the pluripotent epiblast lineage which potentially gives rise to rest of the somatic and the germ cell lineages [382, 383]. The duration of the preimplantation development i.e. the time interval between fertilisation and implantation of the embryo into the uterus, is highly dynamic cellular and molecular process. The progression of preimplantation embryogenesis lasts for 7 days in human, 9 days in *Cynomolgus* and 4 days in mouse, indicating the higher level of inter-species evolutionary divergence [384, 385, 313, 296]. The old-world monkeys share a high degree of similarity in physiology and genome sequence (92.5% to 97.5%) with the human genome [386, 387]. It is crucial to understand the differences between the human and non-human primate's pluripotent states, since the latter is considered to be a reliable clinical and physiological model for human biology and medical treatments [388, 389, 390, 391]. While the preimplantation development in all mammals is phenotypically similar, nevertheless, the series of recent findings [385, 309] suggest the remarkable differences in the gene expression profiles. In consistence with an hourglass model for gene expression conservation [392], the dynamics of early embryos seem to have evolved in similar fashion. The recently materialised cases indicate that species-specific genes have re-wired the transcriptional network leading to diverged evolutionary lineages [393]. The challenges in tractability of transposable elements from the transcriptome repertoire, masks the ravelling of numerous biological networks at greater extent. Hence, there is an essence to detect the entire stock of candidate genes using the availability of revolutionary high-throughput sequencing technologies. This might allow us to dissect the genome and transcriptome data at bulk and at single cell resolution for numerous organisms [394, 395, 396].

In particular, an advanced transcriptome sequencing coupled with machine learning methodologies could hasten the discovery of the unforeseen origin of new genes and their contribution to evolutionary alterations. Indeed, the primate-specific endogenous retrovirus, HERV-H, was shown to have high functionally important expression levels, at least in some naive-like cells, positively mimicking the inner cell mass/pre-implantation epiblast (ICM/EPI) [282, 294, 212, 211]. As nearly half of total number of full-length HERV-H loci is primate specific and rest half of it could even be human-specific [397, 209], indicating the species-specific mode of action of HERV-H in terms of the evolution of new genes or as a regulatory sequences within primates. Species-specific genes of various types have been illustrated to have contributed to the evolution of cellular, physiological, morphological, behavioural, and reproductive traits [398, 399, 400]. Previously, cross-species orthologous genes were analysed and developmental co-ordinates of epiblast was found to be distant among human, *Cynomolgus* and mouse [385].

1.6 Aims and significance of the Thesis

To what extent the inescapable expansion of genomes by HERVs has stirred the evolution of the primate gene regulation or early development is still an enigma. As *Cedric Feschotte* concludes as ***"TEs have***

indeed provided a rich source of non-coding sequence material fuelling regulatory innovation during vertebrate evolution". These young transcriptional networks remodelled by intrinsic retroviruses throw the series of basic biological questions e.g. *Why was there a requirement of a network involved in an essential biological system as pluripotency has to be re-wired*, *Is it a consequence of positive or negative selection to hosts?*. If the cell-type specificity of HERVs causes the deviation then *How did the host deal with unequal "hour-glass" mode of evolution?*. The present dissertation work leverages the availability of vast variety of genome-scale data sets. We perform the strategic analysis to immaculate the molecular interactions of hosts with HERVs. In order to understand the involvement of HERVs, we carefully chose biological process where HERVs could be best studied, such as ***1. human early embryogenesis 2. primate pluripotency 3. human diseases***. *The impact of the diffusing cis-regulatory elements of HERVs on the genome regulation and affecting cellular function is the central theme of the present dissertation.*
An overarching question concerns the relationship between novel gene expression and novel developmental trajectories. However, before addressing any such issue, we must first establish the extent to which development trajectories are in fact distinguished from that of other species. While it has been a consensus that human preimplantation embryogenesis has unique features compared to the early development of murine or even non-human primates, the fine-tuning of the blastocyst formation, and the definition or lack thereof of the ICM through human early development has been never being fully resolved, even by using single cell data [296]. Previously, around one third of the cells could be unambiguously identified, leading to premature outcomes and stagger speculations regarding the human preimplantation development. I sought to solve the mechanisms underlying the progression of early human development. Early human development is unusual in that an inner cell mass (ICM) is not evidenced and the key cell types develop simultaneously, not in sequence.
The recent developments revealed a primate-specific element that orchestrates the transcriptome-wide regulation of biological functions. Selection can therefore act on selfish genetic elements to generate novel gene networks. *We ask whether the re-wiring of transcription networks is driving the phenotypical evolution and biology between and within given species*. Although the *cis* and *trans* activity of HERVs might not alone infer that it is exaptated to perform proper function in genome [401, 277]. To date, the majority of blastocyst cells, however, resist classification. We aim to strategically survey the single cell transcriptomics of early embryogenesis that might provide us the blue-print of human body formation. Moreover, the strategical analysis could reveals that whether the human development is or is not that unusual as previously thought. Moreover, the similarity and disparity of human development with murine and other primates might fix the heterogeneity between natural and artificial pluripotent states. Nevertheless, high resolution of the distinctive features would set up the reliable foundation for regenerative medicine. Importantly, our recently reported articles regarding HERV-H-driven regulatory network, we intend to provide the degree of pluripotency that is primate specific and even specific to us humans. My study with stringent design would aid to understand the significance of intrinsic retroviruses in embryonic development and diseases. Comparison of HERV-H-marked populations with all the available native, formative and artificial pluripotent cells would be optimal to resolve the fidelity of generated pluripotent states and self-renewing populations. Lastly, we would dissect the transcriptomes and epigenomes of many forced naive cultures to calculate their degree of heterogeneity and resemblance with conventional stages of human early development. Finally, to compare the cross-talk of transcriptome with transposcriptome would set up the resource for future studies.

Chapter 2

Methods

The advent of the high-throughput sequencing (HTS) technology has greatly accelerated the research in life sciences. Due to its low cost and high throughput, HTS is being commonly used by laboratories to carry out research at broader scale. Moreover, it is now possible to sequence the entire human genome, transcriptome, regulatome and proteome in a day. Besides whole genome sequencing, HTS has other applications like the identification of genome-wide protein-DNA interactions and quantification of the gene expression. In general, HTS is used to determine the sequence of millions of DNA fragments in parallel, and these fragments can be generated using various methods. cDNA sequencing (RNA-seq) is now being used widely for uncovering multiple facets of transcriptome to facilitate the biological applications. However, the large-scale data analyses associated with RNA-seq harbours challenges especially in the case of repetitive elements. In this study, we performed our analysis in a strategic way to achieve the inferences with higher fidelity, including data preprocessing, differential gene expression analysis, alternative splicing analysis, variants detection and allele-specific expression, pathway analysis, co-expression network analysis, and applications combining various experimental procedures beyond the achievements that have been made. We also leveraged the availability of Single cell sequencing (SCS), that has emerged as a powerful new set of technologies for studying rare cells and delineating complex populations. Over the past 5 years, SCS methods for DNA and RNA have had a broad impact on many diverse fields of biology, including microbiology, neurobiology, development, tissue mosaicism, immunology and cancer research. We, hereby employed it for cross-platform, cross-species analysis of primate embryogenesis datasets.

2.1 Microarray data analysis

Expression data was processed from bead-level expression intensity values preprocessed from Illumina's software in the form of ".txt" or ".bab" files carrying 47,324 probesets targeted by HumanHT-12 v4 Expression BeadChips. Green intensities were extracted after adjusting non-positive values by BeadArray's (http://bioconductor.org/Rpackage) built-in functions. Further, to the BeadArray output data, we fetched significance level of normalized expression values corresponding to probe ID using lumi R's (http://bioconductor.org/Rpackage) variance-stabilising transformation (VST) to deal with sample replicates and robust spline normalization (RSN), for normalization, of which (p-value < 0.05) were further transformed onto log2 scale of and IDs were annotated from illuminaHumanv4.db of Bioconductor annotation data package. The expression values of multiple probes for one gene were

assigned by their median, resulting in 20394 unique genes for GFP samples. Microarray data for Human naive and primed, Human ES and blastocyst, Mouse naive and primed were fetched that are available at the National Center for Biotechnology Information Gene Expression Omnibus database under the series accession number *GSE46872, GSE29397 and GSE15603* respectively. In present study, fold change of differential expression between samples on log2 scale, were analyzed by using linear and Bayesian model algorithms from limma http://bioconductor.org/Rpackage and pairwise differential expression between samples from various datasets were performed by the correction of batch effect arising from two different platforms was by normalizing (*quantile*) each data set to a sample of the same genotype and merging data sets for downstream analysis. Heatmaps shown for differential expression among GFP and Knock-Down samples were drawn for genes, showing significantly highest standard deviations, on their Z-score (relative enrichment). Priory, matrix was hierarchically clustered (Spearman correlation and distances between observations were calculated using euclidian distances and average linkage). We explored the online tool *Gorilla* http://cbl-gorilla.cs.technion.ac.il/, to check for biological processes functional enrichment of differentially expressed genes where entire gene list was used as background. A false discovery rate,corrected P,value threshold was set at 0.05.

2.1.1 Retrieving significant variable genes

In the next step, genes showing significant variation (we applied Poisson's distribution model on the data, genes lying out of curve were considered as genes showing higher variance in the matrix of samples) were retrieved (which always resulted in total of > 2000 genes). Furthermore, we removed genes, which were showing more variation within the group than between the groups. Finally, the heatmap and clusters were constructed as defined using R packages built for them. To find a significant correlation gene networks we applied Weighted Gene Co-expression Network Analysis (WGCNA), which is used to identify clusters (modules) of highly correlated genes. We employed this algorithm to find the set of genes co-expressed in our total merged sample's transcriptome. Total microarray datasets provide higher level of variability thus giving stronger power to this study to generate a significantly enriched co-expression networks and moreover, to check whether a candidate gene is part of them. WGCNA gave us significant co-expression networks as various modules with numerous genes into it. We then used candidate gene as a probe to find out the networks associated with it from WGCNA output. The metamodel derived from Spearman's correlation threshold of positive (60%) and negative (55%) gave the probable significant networks (Note: threhold values are defined by us). We kept this threshold to extract all genes, which were correlated at provided correlation thresholds with candidate gene from extracted networks. Finally, pairwise ranked correlation matrix of identified genes was generated.

2.1.2 Cross-platform analysis

Quantile normalized datasets from the two microarrays described above were used in the analysis. In order to generate a matrix of expression level for unique genes in each sample, two datasets from different platforms were merged by their unique gene names in total samples. The batch effect arising from two different platforms was corrected by normalizing surrogate variances from *"Combat"* package from R Bioconductor. The corrected batch effect was confirmed by Principal Component Analysis (PCA). The resulting expression matrix was subjected to hierarchical clustering (Spearman rank correlation, average linkage). P-value threshold for the correlation test of matrix was kept up to 0.01 where the

hierarchically clustered dendrogram displays gene expression patterns relative to the mean of all the replicated samples included in this analysis. In order to examine statistical reliability of clustering we performed bootstrapping (1000 replicates) on unbiased hierarchically clustered dendrogram of distance matrix using Ward method.

2.1.3 Cross-species analysis

Cross-species gene expression analysis was performed on Human, viz. Illumina HumanHT,12 v4 (expression beadchip containing 47,324 probes, present study) and Affymetrix HuGene 1.0 ST microarrays (containing 33,252 probes, GSE46872) and on Mouse i.e. Agilent 4x44K array platform (containing 45,018 probes, GSE15603) microarray expression sets. Human–mouse orthologous genes were downloaded by online tool (biomart) from Ensemble (http://www.ensembl.org/biomart/martview/) containing 18,657 pairs of orthologous genes, out of these 9,583 genes were mapped by probes of both Human and mouse array platforms explored in present study which were implemented for further analysis. Expression value of each gene was determined by median of all probes targeting to it. As mentioned above, the batch effect was corrected; correction was further confirmed by Principal Component Analysis. Next, these independent datasets were merged in one for further analysis. Each gene value was further assigned as their relative abundance value which is the expression value of gene in each sample divided by mean of expression values of corresponding gene across the samples within same species. The resulting expression matrix was subjected to hierarchical clustering (Spearman correlation, average linkage), p-value threshold for cor. test for matrix was kept upto 0.01 and so outliers are not shown in colored matrix but hierarchically clustered dendrogram displays all the samples included in this particular analysis. Comparison of global expression profile of Blastocyst (ICM), hESC (*GSE29397*) and GFP samples (present study) represented gene wise (19,103 genes possessing common probes between two platforms) which were subjected to hierarchical clustering (Pearson correlation, centroid linkage, k=3) whereas, samples are represented in the order of euclidean distance were clustered using Spearman correlation and centroid linkage. Differentially expressed gene,list between GFP (High), GFP(+) and GFP (-) samples (FDR < 0.05) were intersected to cross,platform, pair wise comparison of rescaled expression values of genes assigned as their row wise Z,score (expression value subtracted by mean of its row values and divided by its standard deviation). heatmap showing comparison with Blastocyst and ES. neighbour genes were fetched using bedtools falling in the window of 50 kb from HERV-H genomic co,ordinates, fold changes between naive and Primed were calculated independently, keeping thresholds for Human and mouse samples in the same way as mentioned above, datasets were intersected by gene names and heatmaps were drawn on their calculated Z-scores.

2.1.4 Pathway analysis of dysregulated genes

Pathway analysis of deregulated genes Canonical pathways and biological function of the differentially expressed genes identified in microarray data sets were investigated using QIAGEN's Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity). The Ingenuity Knowledge Base (www.ingenuity.com) was used as reference gene sets (pathways). Overrepresentation of biological pathways was assessed by Fisher's exact Test and corrected for multiple testing by the Benjamini-Hochberg procedure. The ratio (overlap) is calculated as a number of genes from the dataset that map to the pathway divided by the number of total genes included into the pathway.

2.2 Alignment of high-throughput sequencing data to a Reference Sequence

High-throughput sequencing(HTS) technology is rapidly evolving and revolutionising the research in the life sciences. HTS generates millions of short sequences (reads) that is followed up by computational analysis to fetch out the information from the data. Usually, the first step in high-throughput sequencing data analysis is the alignment (mapping) of the generated reads to a reference sequence [402]. The mapping process is complicated by several factors, including sequencing errors, genetic variations, short-read length, multi-mapped reads, and the huge amount of reads to be mapped [402]. Therefore, during the past decade, numerous software tools have been developed to accomplish this task (e.g. for DNA mappers: SOAP [403], MAQ [404], Bowtie [405], BWA [406], SHRiMP [407], RazerS [408], mrFAST [409]; and for RNA mappers: TopHat [410], SpliceMap [411], MapSplice [412], SOAPsplice [413], STAR [414]). The most widely used DNA mappers are Bowtie (10,517 citations) and BWA (12,269 citations), and the most popular RNA mapper is TopHat (6,844 citations) and STAR (4,692 citations). Due to sequencing errors and genetic variations, we may not find exact matches for all the reads. Therefore, BWA and Bowtie uses the modified matching algorithms i.e. backtracking algorithm for BWA and quality-aware backtracking algorithm for Bowtie. BWA searches for matches between the read and the corresponding genomic position within a certain defined distance whereas Bowtie uses a quality threshold. It is important to note that these software tools are being updated regularly and new versions could have modified algorithms. For example, Burrows-Wheeler Aligner's Smith-Waterman Alignment (BWA-SW) can align long reads up to 1 Mb against a reference genome and Bowtie 2 was mainly designed to map reads longer than 50 bps and supports gapped alignments. TopHat (or the latest version TopHat2) is a fast splice junction mapper for RNA-seq reads. In general, TopHat is a pipeline to map RNA-seq reads to transcriptome and/or genome using Bowtie (or Bowtie 2) and then analyzes the mapping results to identify splice junctions between exons. The steps involved in the TopHat2 pipeline are depicted in Figure 2.1.

2.2.1 Quantification of transcripts

A major application of RNA-seq technology is the quantification of transcripts and differential expression of given transcripts in a group of samples. We employed simplest but highly reliable mode of quantifying the transcripts from transcriptome repertoire. We quantify the expression using featureCounts that aggregates the raw count over the length of transcript. I provided gene transfer format (GTF) file containing the genome coordinates of exons and genes, and often discarded reads mapping over multi-positions. Raw read counts alone are not sufficient to compare expression levels among samples, as these values are affected by factors such as transcript length, total number of reads, and sequencing biases. The measure RPKM (reads per kilobase of exon model per million reads) is a within-sample normalization method that will remove the feature-length and library-size effects. This measure and its subsequent derivatives FPKM (fragments per kilobase of exon model per million mapped reads), a within-sample normalized transcript expression measure analogous to RPKs, and TPM (transcripts per million) are the most frequently reported RNA-seq gene expression values. We used TPM each time when compared the change of expression between the samples instead of FPKM, whereas FPKM was utilized to obtain expression level within the sample. Correcting for gene length is not necessary when comparing changes

in gene expression within the same gene across samples, but it is necessary for correctly ranking gene expression levels within the sample to account for the fact that longer genes accumulate more reads. Furthermore, programs such as Cufflinks that estimate gene length from the data can find significant differences in gene length between samples that cannot be ignored.

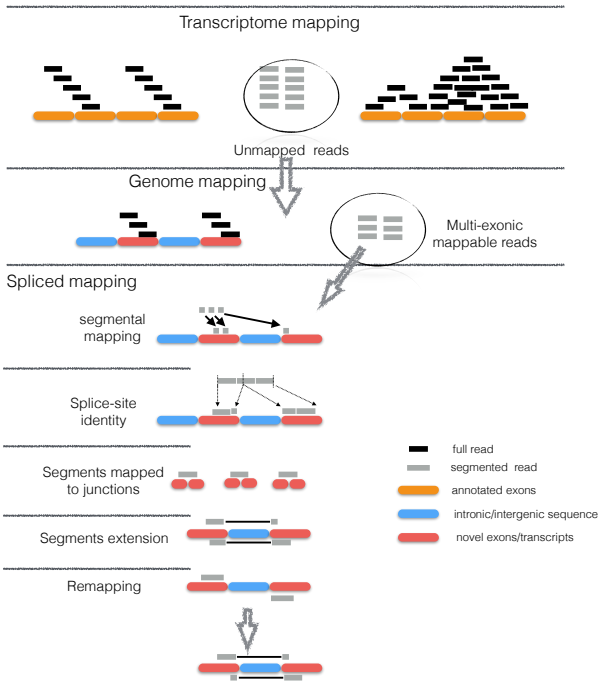


Figure 2.1: Illustration of steps involved in mapping RNA-seq reads using TopHat2. TopHat2 pipeline uses Bowtie (or Bowtie 2) to align the reads to reference transcriptome and unmapped reads are then aligned to the reference genome. Figure is taken from the dissertation [415] of my colleague Vikas bansal with his consent

2.3 ChIP-seqencing data analysis

Chromatin immuno-precipitation (ChIP) followed by high-throughput DNA sequenc- ing allows genome-wide identification of protein-DNA interactions such as transcription factor bindings, transcriptional co-factors binding, RNA polymerases binding and chemical modifications of histone proteins. In general, the first step in the ChIP-seq analysis is read mapping as described in previous section, followed by the peak calling step, which aims to identify the genome-wide binding sites of a protein. In this study, ChIP-seq data of histone marks and transcription factor has been used.

2.3.1 Identification of genome-wide binding events

After mapping, the most common step is the peak calling to identify the genome-wide binding sites of a protein. In the past, various tools have been developed to find peaks from the ChIP-seq data [416]. In general, binding sites or peaks are the regions of a genome where sequence reads are significantly enriched as compared to the control. Before defining a region as a peak, distinct steps are carried out such as read shifting and background estimation. During the mapping step, ChIP-seq reads can align to either the sense or anti-sense strand and therefore, location of mapped reads form two peaks. Next, the reads are shifted towards the centre to determine the most likely location involved in protein binding. The shift

parameter is determined by the fragment size generated in the ChIP-seq library preparation. Interestingly, Model-based analysis of ChIP-seq (MACS) can empirically model the shift size of ChIP-seq reads without any prior knowledge [417].

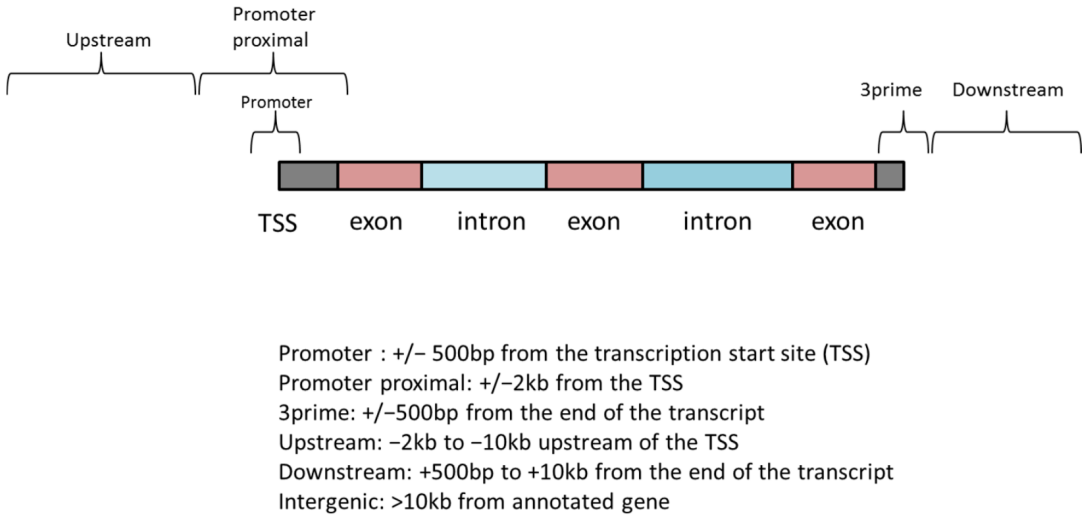


Figure 2.2: Illustration of gene model, its Transcription start site and transcription end site of genes and repetitive elements, moreover, the genomic rage around it that was considered to be annotated. Figure is taken from the dissertation [415] of my colleague Vikas bansal with his consent

2.3.2 Annotation of ChIP-seq Peaks

One of the important step after peak calling is to summarize the location of the peaks in the genome. This step can be used as a validation criterion for certain chromatin- associated modifications and proteins. For example, using prior knowledge, we can confirm if a particular sequence-specific transcription factor preferentially binds near transcription start site (TSS). If a peak overlaps two regions, it was annotated for both of them. More often, researchers are interested in associating peaks with the genes. Peaks were assigned to the genes or repetitive elements only if they are located within 10 kb upstream of the TSS or in the transcribed region (Figure).

2.3.3 Discovery of Sequence Binding Motifs

Sequence-specific transcription factors preferentially binds to a short DNA sequence, which is expected to be enriched in ChIP-seq peaks. Therefore, when a motif of the protein is already known, this step can be used as a proof of principle of a successful experiment. Moreover, if the motif is not known, discovery of a centrally located motif can lead to the identification of DNA-binding motifs of other proteins that bind in complex, which highlights the mechanism of transcriptional regulation. A commercially available database ”TRANSFAC” provides the experimentally- proven binding sites of eukaryotic transcription factors or the freely available database is "JASPAR". These motifs are stored as position weight matrices (PWMs), also called as position-specific weight matrix (PSWM) or position-specific scoring matrix (PSSM), and is the most common way to represent the motifs. In the past, various tools have been introduced to identify over-represented motifs in the ChIP-seq peaks. For example, MATCH tool uses the matrix library collected in TRANSFAC database and calculates two score values: the matrix similarity

score (MSS) and the core similarity score (CSS). These two scores measure the quality of a match between the sequence and the matrix, which ranges from 0.0 to 1.0, where 1.0 denotes an exact match.

2.4 RNA-sequencing data generation

Total RNA was isolated from *Callithrix* jacchus [418] and *Gorilla* PSCs [419] using Trizol lysis reagent and Direct-zolTM RNA MiniPrep kit including DNase I on-column digestion (Zymo Research) according to the manufacturer‘s protocol. Concentration of RNA was quantified on NanoDrop Spectrophotometer ND-1000 and the quality of RNA was analysed using Aglient RNA 6000 Nano Kit on Aglient 2100 Bioanalyzer machine. Library for RNA sequencing was prepared from 550 ng of RNA using Illumina TruSeq Stranded mRNA LT Set A kit (cat. no. RS-122-2101), according to TruSeq Stranded mRNA Sample Prep LS Protocol. Samples were indexed with sample-specific indices which allow for the pooling and sequencing of all libraries in two pools of five samples. Expression profiling of transcriptomes by high throughput sequencing were performed on BIMSB Genomics Platform of Max-Delbrueck Center for Molecular Medicine (Berlin, Germany) on Illumina HiSeq2000 sequencing platform. The clustering of the index-coded samples was performed on a cBot Cluster Generation System using PE TruSeq Cluster Kit v3-cBot-HS (Illumina) according to the manufacturer's instructions. After cluster generation, sequencing was performed on an Illumina HiSeq 2000 platform as 100 bp first strand specific paired-end reads.

2.5 RNA-sequencing data analysis

Sample-specific barcoded sequencing reads were de-multiplexed from multiplexed flow cells. The resulting BCL files were converted to FASTQ format files using CASAVA 1.8.2. The quality of the raw sequence reads was determined with the FastQC. Reads with quality score < 30 were removed. We also truncated 2nt from the end of sequencing reads, since their average quality score was not same as the rest of nucleotides. This resulted in at least 70 million reads per sample. Next, reads were mapped over the reference genome (Human hg19/GRCh37) and transcriptome model (hg19.refseq. gtf), downloaded from USCS tables (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/) using TopHat v2.0.8, samtools 0.1.17.0 and Bowite 2.0.5.0 applying parameters as: "tophat2 -p 8 –r 150–mate-std-dev 140 –library-type fr-firststrand". On average 75% of total reads were uniquely mapped on the annotated gene models, approximately 10% of total reads are uniquely mapped on repeated fraction of genome (data not shown). Transcript assembly for each individual sample was conducted using Cufflinks v2.08 measured as FPKM. For calculation of differentially expressed genes (DEGs) we calculated Counts Per Million (CPM) using featureCounts and algorithms from "DESeq2" which performed quantization and statistical inference of systematic changes between conditions, as compared to within-condition variability. The package "DESeq2" provides methods to test for differential expression by use of negative binomial generalized linear models; the estimates of dispersion and logarithmic fold changes incorporate data-driven prior distributions. In addition, for the outlier's samples, DEGs were calculated using a single-replicate model. The read counts were calculated with featureCounts from subread package (http://subread.sourceforge.net/), FPKM was calculated using bamutils (http://ngsutils.org/modules/bamutils/count/). Next, Random Variable1 (Var1) = n. l. x formula was used, where x (Random Variable 2) is the expression level of this gene, n is reflecting the sequencing depth and l is the gene length. The two random variables were used in the published model of

GFOLD algorithm, which calculate the normalization constant and variance to extract fold changes from unreplicated RNA-seq data.

2.5.1 RNA-seq from Non-human primates

Reads were mapped to their respective reference genomes i.e. human (hg19), *Chimpanzee*(PanTro4), *Gorilla* (gorGor4), *Callithrix* (calJac3) and mouse (mm10) using STAR [414] with our defined settings *–alignIntronMin 20 –alignIntronMax 1000000 –chimSegmentMin 15 –chimJunctionOverhangMin 15 –outFilterMultimapNmax 20*
followed by constructing STAR genome/transcriptome indices providing their respective RefSeq gtf annotations. As per STAR default we permitted at most two mismatches. We obtained uniquely mapped read counts using featureCounts [420] at gene level with refSeq annotations. Gene expression levels were calculated at Transcript per million (TPM) from counts over the entire gene (defined as any transcript locating between TSS and TrEs). This we did using our in-house R script (available on request).

2.5.2 Visualization of reads

Single cell mapped reads merged for each stage in bam format were obtained using *samtools* [406] for human embryonic development. It was converted into bedGraph format using bedtools [421] to visualize through IGV over refseq genes (hg19). Conservation track was visualized through UCSC genome browser under net/chain alignment of given Non-Human Primates (NHPs) and, later on, merged beneath IGV tracks.

2.6 Single cell RNAseq data processing

We reanalyzed single cell sequencing data from previously published studies of human [297, 298, 296] and *Cynomolgus* [385] embryogenesis. For cross-platform single-cell RNAseq data, counts were merged on gene names. Variation due to batch effects was adjusted using COMBAT [422, 423] from R package "sva" [422, 423]. The resulting dataframe was converted into log2 TPM scale (as mentioned before).

We chose samples expressing more than 5000 genes. We subsequently selected for further analysis those genes that express (Log2 TPM > 2) in at least in 5 of these samples. This resulted in 1285 single cells carrying expression levels of 15,501 genes for human samples. For mouse we recovered 259 single cells carrying expression levels of 15181 genes. We separated cells by applying to the most variable genes dimension reduction methods, notably principal component analysis (PCA) and t-stochastic neighbour embedding (*t-SNE*). To define the set of discriminating genes for PCA for any given species, we calculated the z-score for each gene in each sample in the dataframe of all genes across all the samples (i.e. for human 15501 genes in 1285 samples). Each gene was then represented by an across sample vector of z-scores. The resulting z vectors for all the genes were clustered by Pearson's correlation method. We took any cluster for further analysis that carried more than 20 genes. For each resulting cluster, we considered the vector for each gene in turn and calculated log(Variance/mean) of the vector. We then determined the mean of this value across all the genes within the cluster. Those clusters showing a mean(log(Variance/mean)) > 1 were considered as most variable clusters. All the genes (for humans n=1681) in these clusters were considered as most variable genes (MVG). Our murine analysis confirms previously identified clusters, so is not presented here. The above PCA analysis clearly resolves human E3 and E4. However, E5

onwards appears as an unresolved cloud. In order to resolve this cloud, we first ran t-SNE on single cell data for stages E4 and E5 together and then E5 independently. As this resolved the stages we were interested in, we then applied t-SNE to the full dataset, enabling full resolution of discrete stages in early human development. We used *"Seurat"* http://satijalab.org/seurat/ [395] and *"SCDE"* http://hms-dbmi.github.io/scde/ [424] packages from R. Seurat was used to obtain most variable genes, markers for given clusters, most significant principle components, t-SNE analysis and visualizations. DEGs between single cell clusters were calculated using SCDE algorithms [424]. A gene qualified as a marker of a given cluster if it fulfilled two criteria: the gene must be over-expressed in that particular cluster (average fold difference >2 compared to the rest of clusters) and must also be expressed (Log2TPM > 2) in at least 70% of cells in that particular cluster. To estimate the expression level for repetitive elements per locus, only the data from [297] was employed this sample uniquely having long reads. The above analyses were applied to single cell data. To compare such data with bulk RNAseq datasets from naive and primed stem cells we adopted a different approach. Rather than merging entire transcriptome datasets, instead we identified the most discriminating genes from single cell analysis and extracted their expression values from the bulk data. The set of most discriminating genes was defined as previously described (see PCA section above). PCA was performed on these genes' expression values (n=1681). I show that single-cell ESC samples were clustered with bulk RNAseq samples from comparable cells, suggesting an absence of distortion of signal owing to sequencing protocol. Prior to any analysis, bulk data was normalized using COMBAT [423].

2.6.1 ESRG analysis and self-renewal regulatory network

We created a dataframe of single-cell data for stages E5, E6 and E7, carrying expression values (TPM) of all Human MVGs. We then computed pairwise Pearson's correlation for all MVG. We then selected only those genes that show strong correlation or anti-correlation with ESRG (threshold rho > |0.70|), as shown in heatmaps (Figure 5.7. A network was constructed on genes showing the highest level of ranked correlation among each other, with rho > 0.70, using "igraph" http://igraph.org/r/ package from R. Arrows show the linking (links based on a preset level of preferential attachment (Barabasi-Albert model)) between genes. Their direction is manually set under the criterion that "from" genes appear first in human pre-Epi, "to" genes appearing next. The size of a circle represents the number of instances a gene is upstream (nodes) of its paired partners (edges). Genes in the network are markers of human EPI and colors are assigned as to their expression, or lack thereof, in mouse or non-human primate's embryogenesis.

2.6.2 Cross-species analyses

Genes that are differentially expressed (DEGs) between species were obtained by cross-species mapping of RNAseq reads. Reads mappable on both genomes to be compared were further mapped on human genome (hg19) using STAR. Cross-species read counts, FPKM and effective fold change was calculated using GFOLD [425] on obtained replicated and unreplicated datasets. We mapped human and non-human iPSC RNAseq reads against the human reference genome and gene models to determine the expression level of human genes and repeat elements in NHPs.

2.7 Detection of chimeric transcripts from RNAseq data

In order to determine chimeric transcripts, we first aligned the reads using universal aligner "STAR" using the parameters written above that can discover canonical and non-canonical splice and chimeric (fusion) sites. We kept only the junctions that were identified with a minimum of 6 uniquely mapped reads. Any novel genes with resemblance to mitochondrial genes were excluded from all analysis. Either donor site or acceptor site mapping to the mitochondrial genome was considered grounds for exclusion. To exclude chimeras derived from repeat elements, we identified those novel transcripts that had at least 6 consecutive bps from known repeat elements (repeats specified in hg19 rmsk. gtf).

2.8 Reduced Represented BiSulphite sequencing (RRBS) analysis

We sought to indentify DNA methylation marks in the vicinity of repetitive elements and genes including imprinted ones. We employed RRBS and WGBS methylation data from single cell analysis of early human development [426] and Naive/primed cells [361, 356]. We considered only those methylated regions that were supported by 10 reads or more. To analyze methylation around the promoter of any given co-ordinate, we converted methylation sites into bed format and intersected with 5 KB extension upstream and downstream from TSS genomic coordinates. We also employed catalogue of imprinted genes `http://igc.otago.ac.nz/home.html` in the case of the analysis of imprinted genes. Methylation level was defined as the ratio of methylated sequencing reads versus all of the reads on the given locus. We calculated the average across-promoter ratio for all imprinted genes during human embryonic development. Lines in the methylation plot show a polynomial regression at 6^{th} degree fitted between distance from TSS and average methylation fraction. All of the statistics were performed on R `https://www.r-project.org/`.

Chapter 3

Single-cell transcriptomic blueprint of human early embryos

While it has been a consensus that human preimplantation embryogenesis has, unique features compared to the early development of murine or even other primates [385], the fine-tuning of the blastocyst formation, and the definition (or lack thereof) of the ICM through human early development has been never being fully resolved. Indeed, even using single cell data [297, 298, 296], only around one third of the cells could be unambiguously classified. Owing to the fast progression of the developmental process, suddenly generating a large number of cells with high heterogeneity, the most challenging task is to catch the distinct lineages during blastocyst formation. To address this, and to enable characterization of all stages of early human development, we use a strategy of clustering cells instead of genes. We employ a recently developed dimension-reduction clustering methodology on single cell transcriptomics [395]. Importantly, this technique focuses on defining local resemblances and is ideally suited both to single cell data and to resolving transitory stages and otherwise unresolvable clusters. . By resolving the distinct groups we aimed at re-defining the lineages and establishing markers of respective lineages.

3.1 Updating the transcriptomic atlas of human early embryos

To identify markers of distinct stages of human preimplantation embryogenesis, we used available single cell transcriptome data [297, 296]. In contrast to previous analyses that identified differentially expressed genes, we developed a strategy of clustering whole cell trancriptomes. First, we identified 1597 genes exhibiting high variability across single cells and thus potentially useful for defining cell types. Next, we performed principle component analysis (PCA) to reduce the dimensionality of the data, and identified nine significant PCs using a previously described jackstraw method [395]. We used these PC loadings as inputs to t-distributed stochastic neighbour embedding (t-SNE) [395] for visualisation. This approach allowed us to distinguish 10 clusters that we annotate on the basis of previously reported markers (Figure 3.1B and C). In E3-E4, it is relatively straightforward to identify the clusters such as oocytes, zygote, 2, 4, 8 cells stage (E3) and even the more heterogeneous morula (E4) (Figure 3.1A). Though the homogenous 8-cell stage shows maximum level of cell to cell variation (heterogeneity) (Figure 3.1A) is marked by known factor LEUTX (Figure 3.1G).

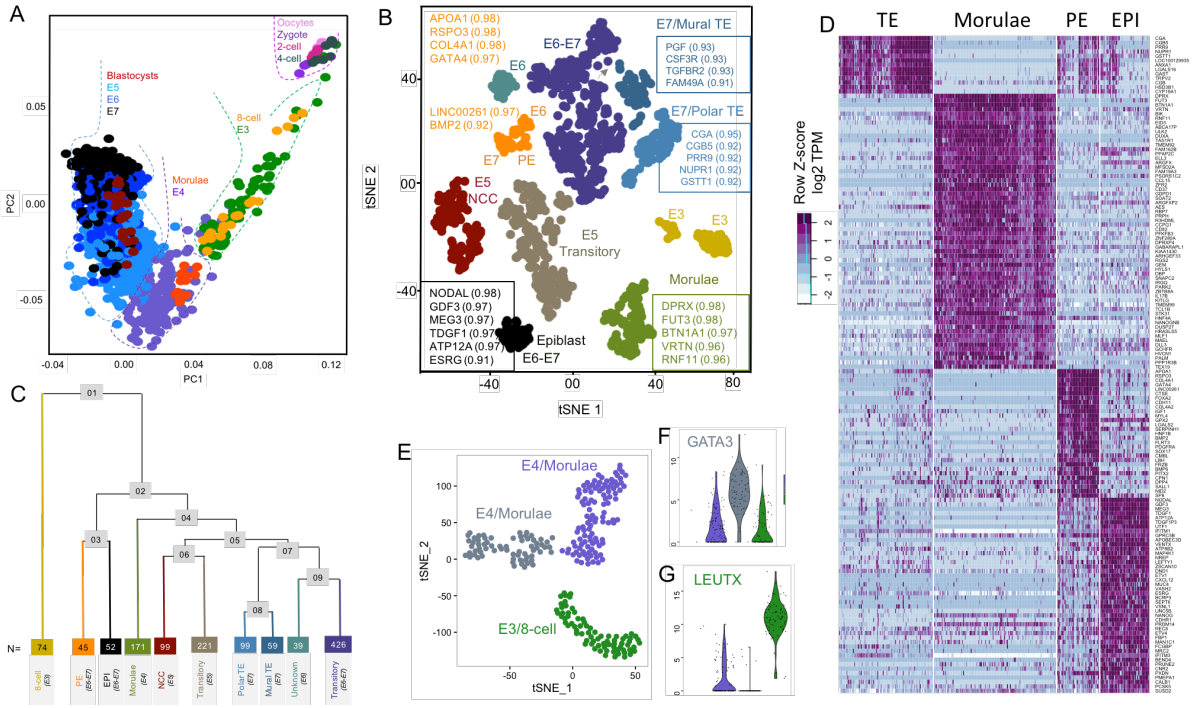


Figure 3.1:

A. Tracing of human embryonic development progression from zygote to blastocyst. Principal component analysis (PCA) of cross-platform 1385 single-cell preimplantation transcriptome using 1583 most variable genes (MVGs). Developmental stages defined as in [297, 296].

B. Two-dimensional t-SNE analysis from single-cell preimplantation transcriptomes using 1651 MVGs re-defines cell populations from distinct lineages as 8-cell at E3, Morulae at E4, Non-committed cells (NCC) and transitory cells at E5, pluripotent epiblast (EPI) at E6-E7, primitive endoderm (PE) at E6-E7, Mural and Polar trophoectoderm (TE) at E7. At E5, cells presenting none of the known lineage markers were assigned as non-committed (darkred) and cluster forming rest of cells express multiple markers so we annotated as transitory cells. Up to 6 most discriminatory genes are listed in box next to each cluster; numbers in bracket refers to AUC value. Colors indicate unbiased classification via graph based clustering where each dot is single cell.

C. Re-ordered phylogenetic tree of clusters shown in previous figure. The tree is constructed using *"BuildClustertree"* built-in function of the *R package 'Seurat'*. Nodes are shown in grey boxes and numbers are in the order of their position on tree. Colour code as in previous figure. Numbers in colored boxes denote the number of cells in each representative cluster. Note: From the 1285 single cells, the independently clustering 16 single cells were added to the transitory category from E6 and E7 (originally 410 cells).

D. Heatmap displaying the scaled expression (log TPM values) of discriminative gene sets (AUC cutoff > 0.90) defining cell populations of morula (n=171), EPI (n=52), PE (n=45) and Polar TE (n=99) reported in previous figures. Heatmap color scheme is based on z- score distributio n from -2 (light blue) to 2 (purple). See Table 5 for detailed list of discriminative markers of this study and their ranking comparison with published ones [297, 298, 296]. Note: In the previous study, EPI, PE and TE markers were not defined. Instead, the authors reported upregulated genes of previously defined developmental stages at defined embryonic days (E5, E6 and E7). Genes identified by our analysis (see Method) are marked with asterisks ($*$).

E. t-SNE clustering analysis results of E3 and E4 cells (n = 245) generated using the R software package 'Seurat' using first 2 PCs (previous figures) as input loadings. Single cells were color-labeled in 3 group wise manner as we obtained 3 distinct clusters viz. one homogenous cluster from E3 whereas two clusters from E4. Since E3 represents 8-cell stage and E4 as Morulae, so we kept their names as it is. Each dot represents an individual cell.

F. Violin plot visualize the density and distribution of gene expression (log TPM value) of LEUTX that was upregulated in 8-cells compared with Morulae cells. Each dot represents an individual cell

G. Violin plot visualize the density and distribution of gene expression (log TPM value) of a preimplantation transcription factor GATA3 that was upregulated in a group of Morulae cells compared with another group of Morulae cells and 8-cells. Each dot represents an individual cell

Contrastingly, Morulae population being homogenous group, could be clustered into two subgroups;

one marked by GATA3 that would probably form the lineages (Figure 3.1E and 3.1F). Human blastocyst formation initiates at E5, progresses at E6 and stabilises at E7 prior to implantation. Our analysis reveals previously unidentified clusters after morula (Figure 3.1A and 3.1B).
Our strategy distinctly identifies EPI and PE, as well as TE (polar and mural) clusters in E5, E6 and in E7 stages, respectively [area under curve (AUC) $>$ 0.90]. These clusters homogenously express a specific set of marker genes (Figure 3.1B and C). A remaining cluster from E6 and one from E6-E7 is yet to be defined since they express markers heterogeneously(Figure 3.1B and 3.1C). These clusters homogenously express their specific, previously reported markers [area under curve (AUC) $>$ 0.90] and some newly discovered markers by present study (Figure 3.1D and table)

3.2 Identification of novel non-committed cells in E5 blastocysts

Not all the clusters are so easy to classify as EPI, PE and the early cell types. Given the heterogeneity between cells of the same type, might some also be transitory types? At E5, we observed two types of cells, either expressing (even multiple) lineage markers or those that fail to express any markers of known blastocyst lineages (EPI-PE-TE). Cells expressing multiple markers at E5 do not segregate on the t-SNE plot using unbiased approaches of clustering E3-E7 cells, and we defined them as transitory cells. By contrast, cells expressing none of the known markers form a clearly segregated cell population on the t-SNE plot that we name non-committed cells (NCCs) (Figure 3.1B and 3.1C). This clearly definable, but novel cell type we analyse in more detail below.

3.3 Resolving the identity of cells segregating from morula unmasks the human inner cell mass (ICM)

The presence or absence of the ICM is possibly the most contentious of issues in early human development [346, 296]. If it exists, we would expect it to be resolved as a type segregating from morula. In order to further resolve the broad spectrum of transcriptomes of cells segregating from morula, we restricted our analysis to E5 only, and removed cells with low quality transcriptomes (expressing (log2 TPM $>$ 1) less than 5000 genes). The transcriptomes of the remaining 300 cells were subjected to a similar strategy that we used for dissecting E3-E7. This approach resulted in six significant PCs, and we enlisted the top 30 genes contributing to their respective eigen vectors (Figure 3.2A). Loading the above PCs as input, we observe three distinct transcriptome clusters on t-SNE (Figure 3.1B) (that could also be distinguished on the first two principal components) using the expression dynamics of Most Variable Genes (MVGs) (Figure 3.3A). Altogether, we identified 235 genes (AUC $>$ 0.80) (Fig. 1C) that we used to characterize the obtained clusters at E5.
Expression of DLX3, a known marker of a precursor population of TE defines the first cluster (n=86) as pre-TE, while the second cluster (n=97) homogeneously expresses BIK, a key inducer of the apoptotic process. The third cluster (n=71) co-expresses known EPI (e.g. NANOG) and PE (e.g. BMP2) markers (Figure 3.3C, all genes star marked would also mark EPI in E6-7 whereas hash marked genes would mark PE at E6-E7, other would lose their specificity with ICM), defining the ICM (Figure 3.3A). Thus, our strategy enabled us to resolve distinct cell populations segregating from morula, and unmask the human ICM. Based on their ranking in the corresponding clusters, we propose the top markers of human ICM,

pre-TE and NCC populations as IL6R, TMPRSS2 and BIK, respectively at E5 (Figure 3.3A and (Figure 3.2C-H)).

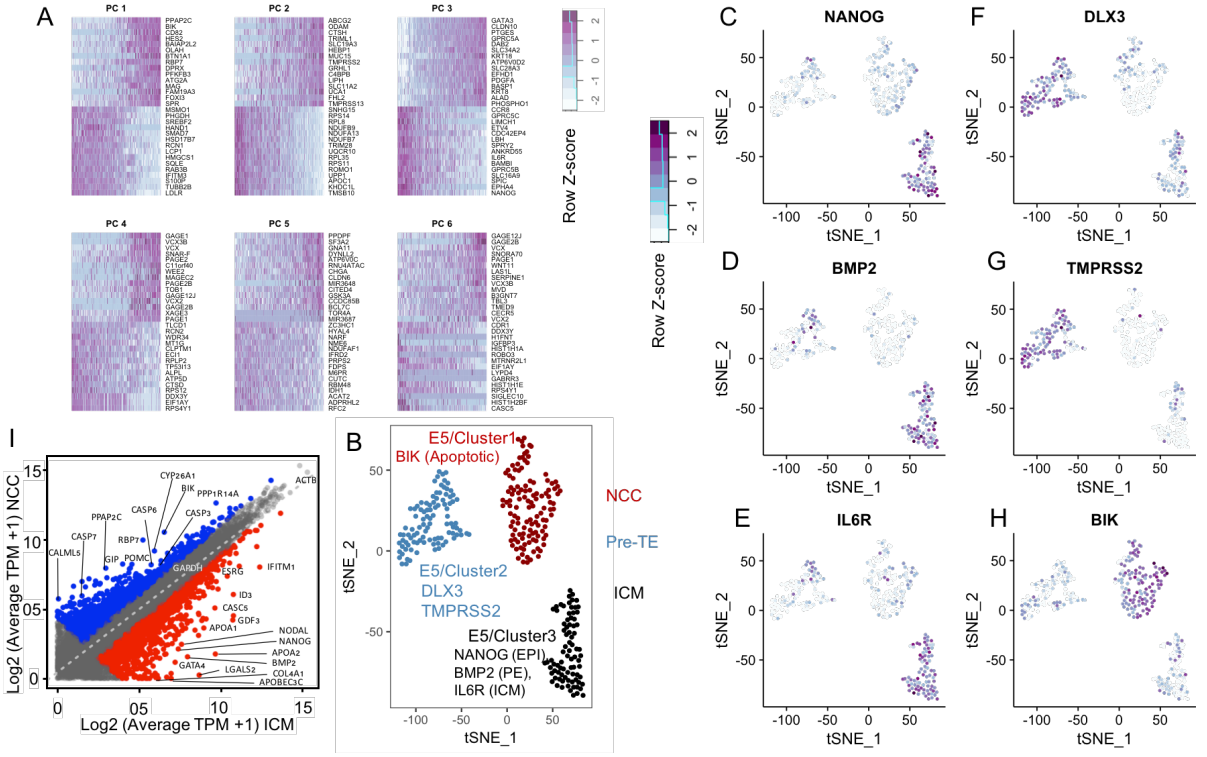


Figure 3.2:
A. Heat maps showing scaled expression (log TPM values) of discriminative gene sets per PC (Principle Component) in 300 E5 cells visualized on the first 6 most significant components. These 6 PCs were used as input loading for further t-SNE analysis of E5 cells. This figure explores the correlated and anti-correlated gene sets among E5 cells.
B. t-SNE clustering of E5 cells (n = 300) using the first 6 PCs (Figure S1A) as input loadings reveals 3 distinct clusters. Each dot represents an individual cell.
C-H. The 3 clusters were also distinguished by their strong expression of known markers ICM (IL6R), EPI (NANOG), PE (BMP2), Pre-TE (DLX3 and TMPRSS2). The unattended cluster is flagged by BIK (BCL2-Interacting Killer/Apoptosis-Inducing NBK) as illustrated in the feature plot where each dot represents an individual cell and color indicating the expression of mentioned marker genes.
I. Scatterplot displays the comparison of averaged gene expression of NCC and ICM cells pooled together in pairwise manner. Red and blue dots are genes whose expression is enriched in either ICM or NCCs. Annotated genes are: unchanged housekeeping genes (GAPDH and ACTB), genes enriched in either NCCs (blue) or in ICM (red).

3.4 Human preimplantation embryogenesis is a sequential process segregating from morula

Based on the most extensive study done so far of single-cell transcriptomics from human preimplantation embryos, it was suggested that the human morula segregates to EPI, PE and TE simultaneously around E5 [296]. This dynamic would be a surprising deviation from the step-wise lineage specification dynamics of mouse or macaque [313, 385]. To evaluate the 'simultaneous' vs 'sequential' models, we employed scaled expression of our cluster-specific markers, DLX3 (Pre-TE), BMP2 (PE), NANOG (EPI), IL6R (ICM) and BIK (NCC), and determined co-expression patterns at single cell resolution (Figure 3.3D and

3.3E).

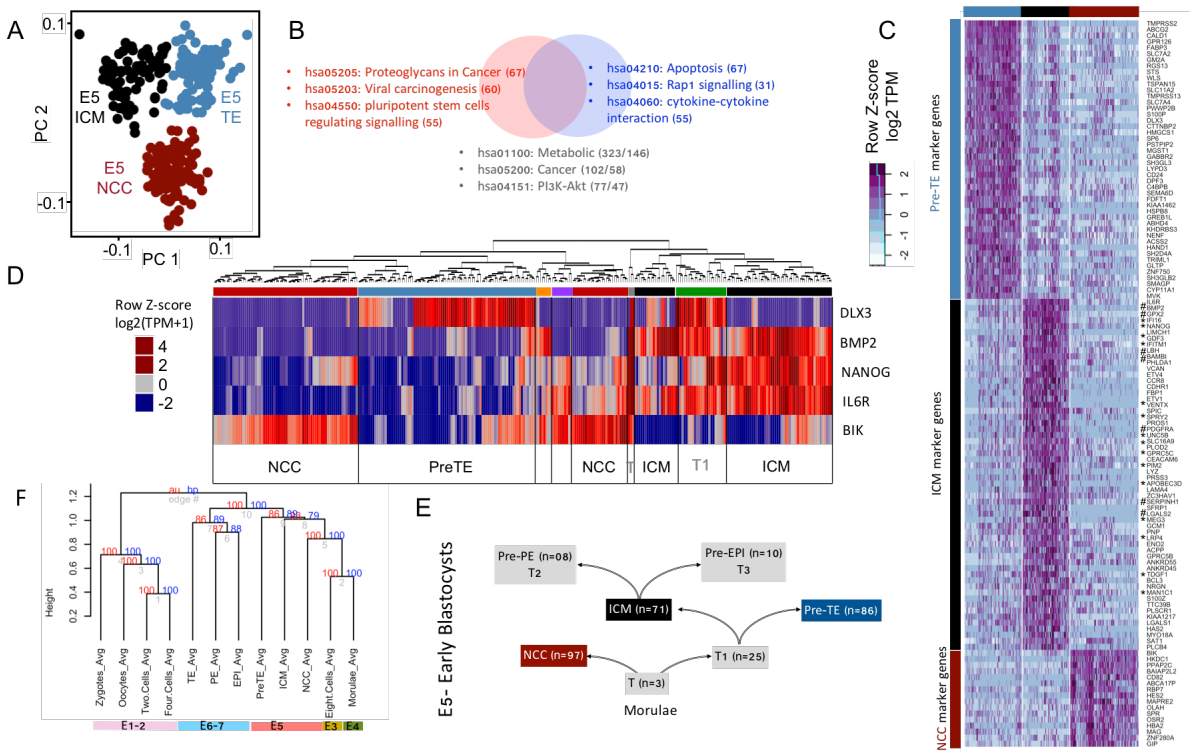


Figure 3.3: ***A***. PCA of E5 cells (n=300) by most discriminating genes (n=526) among three obtained clusters represented on PC1 and PC2.
B. Venn diagram shows the top 3 KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways enriched either in ICM (red) or in NCC (blue) and shared between them (merged, grey). Numbers in bracket indicate the number of genes involved in the respective pathways. Black brackets show ICM numbers slashed by NCC ones.
C. Heatmap visualization of scaled expression [log TPM (transcripts per million)] values of discriminative set of genes for each cluster defined in figure A (AUC cut-off >0.90). Color scheme is based on z-score distribution from -2.5 (light blue) to 2.5 (purple). Left margin color bars highlight gene sets specific to the respective clusters in figure A and top margin color bars define the same for cells. ICM specific genes are marked by (∗) or (#) are also expressed at E6-E7 in EPI or PE, respectively.
D. Heatmap of the row-wise scaled expression (log TPM) levels of selected marker genes for Pre-TE (DLX3), EPI (NANOG), PE (BMP2), ICM (IL6R) and NCC (BIK). Colour bars under the dendrogram were set manually showing the clusters of distinct cells expressing differential combination of markers. Transitory cells (T) are co-expressing multiple lineage markers.
E. Schematic of stepwise initiation of blastocyst formation based on the heatmap in previous figure. Bracketed numbers indicate the number of cells showing the characteristics of the various cells types.
F. Dendrogram via bootstrapping (1000 replicates) based hierarchical clustering using ranked correlation and complete linkage method on averaged expression from the distinct cell populations of human preimplantation embryogenesis (the transcriptomes of all the distinct populations are pooled together). Height of dendrograms represents the euclidian distance of dissimilarity matrix, numbers in red and blue indicate "au" and "bp" values from the bootstrapping.

Out of 300 we detected 3 cells expressing all the four markers, indicating that these cells could be the precursors of early blastocysts. Another subset of cells expressed the markers of the three layers of the blastocyst (n=25), featuring a transitory/precursor state (T1) prior the segregation to ICM (n=71) and pre-TE (n=86). ICM and either PE or EPI markers were enriched exclusively in the transitory cell population of T2 (n=8) and T3 (n=10), respectively. We hypothesized that these ICM-derived cells are the Pre-EPI and Pre-PE cells that would commit to PE and EPI at E6-7 stages (Figure 3.3D and 3.3E).

Our analysis supports the model that the human early embryogenesis is a sequential process, proceeding through well-defined steps. As we have shown in previous figures, we not only define the distinct stages during the formation of human blastocyst but also their progenitors along-with transcriptional flags marking their respective stages (Figure 3.3D and 3.3E).

3.5 The non-committed cells of the human blastocyst are exposed to programmed cell death

With the trajectories and cell types clarified it is instructive to ask what markers define each cell type. Of particular interest in this context are the newly defined NCCs. The top marker of NCCs, BIK is an apoptosis-inducing factor, suggesting that this cell population that we have defined as non-committed might be cells that are being selectively purged (Figure 3.2B, 3.2H, 3.3D-F). To address whether this might be the case, we averaged the expression for individual genes across the groups and performed pairwise analysis, enabling identification of the genes that are differentially regulated in committed (ICM, EPI and PE) vs non-committed cells (NCCs).
Our KEGG pathway mapping for differentially regulated genes revealed that ICM genes were enriched in Pluripotency regulating signalling, whereas NCC genes are mapped to the Apoptosis pathway (Figure 3.3B), consistent with the involvement of BIK. Besides NANOG and BMP2, the ICM expresses a unique combination of genes that have the potential to affect the cytokine or chemokine signalling (e.g. , IL6R, CCR and ETV1/4)(Figure 3.3B and 3.3C). Curiously, in ICM we also detected genes implicated in host defence against retroviruses and retrotransposons e.g. APOBECs, IFITM1 (Figure 3.3B and 3.3C). In NCCs, in addition to BIK, we observed the differential upregulation of several other genes associated with programmed cell death (e.g. BAK1, various caspases or MAPK3, etc.) (Figure 3.2I and 3.3C), indicating that these cells are likely subjected to apoptosis. Our assumption is further supported by the observation that NCCs are not detectable after E5, thus excluded from the developmental process (Figure 3.3F, 3.1C and 3.1B).

To identify the origin of NCCs, we performed a transcriptome-wide clustering with 1,000 bootstraps. Since, the cells are in continuous progression exhibiting a dynamic transcriptome in the otherwise clearly definable clusters, we averaged their expression for the clustering (Figure 3.3F. The analysis detects NCCs as an alternative population of cells commit either to ICM or Pre-TE originating from morula (Figure 3.3F). We also compared the expression of apoptotic marker genes in EPI or PE populations compared with NCC ones, expectedly they were significantly upregulated in NCCs, supports that these populations do not commit to any further lineages (Figure 4.1E). Thus, cells that fail to express any lineage markers are filtered out following cell fate decision at E5 by apoptosis, a mechanism that possibly serves as a quality control measure to eradicate certain (e.g. damaged) cells from development.

Chapter 4

Retro-element's guide to human early embryos

By resolving the distinct groups we aimed at re-defining the lineages and establishing markers of respective lineages. To characterize the cells, we ought to add an extra layer of analysis based on presence in the transcriptome of transposable element (TE) fractions. Embryonic gene activation (EGA) is accompanied with a transition of DNA methylation [426], and also affects the expression of TrEs. Indeed, TE families of different phylogenetic age (both transpositionally active and inactive) are de-repressed, and get transcriptionally activated in characteristic patterns [134, 263, 285, 276, 312]. We ought to determine their expression profiling in the freshly resolved clusters. The regulation of TrEs is highly precise and specific to the given cell-type. We expect a contrasting pattern whereby young and old TrEs are expressed in different cell types: possibly mutagenic young TrEs are expressed in NCCs while older TrEs, most notably HERV-H and HERV-H-associated transcripts, could define many cell types with a developmental future. These cells include, but are not restricted to pluripotent stem cells, where their high expression level was previously shown to be functionally relevant [209, 211].

4.1 Cells upregulating potentially mutagenic transposable elements are exposed to apoptosis

Why might the NCC be expressing apoptopic factors? One possibility is that they might be damaged by the activity of mutagenic transposable elements (TEs). To further characterize the freshly resolved clusters, we examined the expression of their TrEs. In human, the mutagenic, young ($<$ 7MY) elements include certain transpositionally active TrEs, such as Line1, SVA and Alu. The majority of the Young elements in the human genome are activated following EGA and peaking in morula [263, 427]. To monitor the dynamics of TrEs following morula, in the blastocyst, we averaged the expression (Log2 CPM+1) of each particular TE family and compared their expression in NCCs against the other blastocyst cells at E5. We detected transcriptional upregulation of TrEs in both NCCs and ICM. However, while the activated families in the ICM are phylogenetically Old ($>$ 7MY) and transpositionally inactive, the upregulated TrEs in NCCs are Young, and potentially mutagenic. The old TrEs upregulated in the ICM are ancient human endogenous retroviruses (HERVs), dominantly represented by their full-length transcripts: LTR2B-HERVL18, LTR41B-HERVE-a, LTR17-HERV17, LTR10-HERVI, MER48-HERV-H48, and

LTR7-HERV-H in ascending order of average expression (Figure 4.1A). The list of activated young TrEs in NCCs includes LTR5-Hs, HERV-K13-int and the reportedly mutagenic SVA, Alu and Line1 [77] (Figure 4.1A). The upregulation of the old LTR7/HERV-H stays high and strikingly antagonistic to young (e.g. SVA) transcription in either EPI or PE, where the expression of SVA is low (Figure 4.1B). Importantly, Line1 open reading frame 1 (ORF1) is clearly detectable, and marks a subpopulation of the blastocyst in situ, indicating the presence of translated Line ORF1 in these cells. It is not a great leap to suppose that expression of mutagenic elements could result in genome instability that could induce the apoptotic process in NCCs (Figure 4.1C).
The relatively large number of transitory cells, and a built-in potential quality control appear as features of human preimplantation embryogenesis in E5. In addition, domesticated retroviral sequences (HERV-H-derived) have been implicated in human/primate preimplantation embryogenesis [212, 211]. Specifically, in human pluripotent stem cells (PSCs), the LTR7/HERV-H promoter/enhancer drives transcription of numerous regulatory transcripts that modulate pluripotency [312, 209, 211, 212]. Indeed, in ICM we detect the upregulation of neighbour genes (at least 10KB downstream of HERV-H locus) in the proximity of expressed HERV-H loci (Wilcoxon test, p-value < 0.0001), including HERV-H derived gene products (e.g. ESRG, ABHD12B) (Figure 4.1D). While many such effects may reflect innocent bystander effects, the function of a number of HERV-H-derived transcripts (e.g. LincROR, ESRG) has been confirmed to affect pluripotency [286, 428].

4.1.1 Knocking down of HERV-H in pluripotent stem cells results in upregulation of Young TrEs

The mutual exclusion of Young and Old TrEs could be owing to some third part switches. Alternatively, might it be that activity of one suppresses the activity of the other? If so then the activity of TrEs may be involved not just in pluripotency and self-renewal, but also in the maintenance of genome stability. HERV-H-derived products promote, maintain pluripotency and inhibit differentiation. Knocking down (KD) HERV-H in PSCs can model certain aspects of this developmental stage, when cells discontinue to self-renew and commit to differentiate [212, 209]. Does this transitory stage associate with turmoil of TE activities? Our analysis focusing on TE expression in the transcriptome of KD-HERV-H-h1 cells [209] reveals that a repression of HERV-H expression leads to robust upregulation of Young TrEs (Figure 4.2A). This would suggest that the future viability of cells with a potential developmental fate is dependent on HERV-H somehow suppressing the activity of potentially mutagenic Young TrEs.

4.2 HERV-H is repeatedly co-opted during early embryogenesis

While HERV-H has been previously identified as having a key role in pluripotency in humans to date, HERV-H has been treated as one entity. Are then all HERV-H loci expressed at the same time or are there discrete classes of HERV-H expressed in given cell types? LTR7 drives only a fraction of HERV-H genomic loci ($\sim$350/1285) in pluripotent stem cells [263, 285], however HERV-H is transcribed from variants of LTR7 in earlier stages of embryogenesis [263, 312]. Are the time-separated variants different? Our inability to define discrete HERV-H subtypes has in no small part been because we are not able to map the short sequence reads to a particular genomic loci. Using the dataset of [297] enables us to calculate individual locus expression for HERV-H and SVA elements at certain level. This strategy

uncovers that HERV-H expression is not restricted to the blastocyst, but also detectable in earlier stages of embryogenesis. Similarly to the blastocyst,, both HERV-H and SVA expression appears to follow characteristic, but contrasting pattern during the entire pre-implantation development (Figure 4.2B).

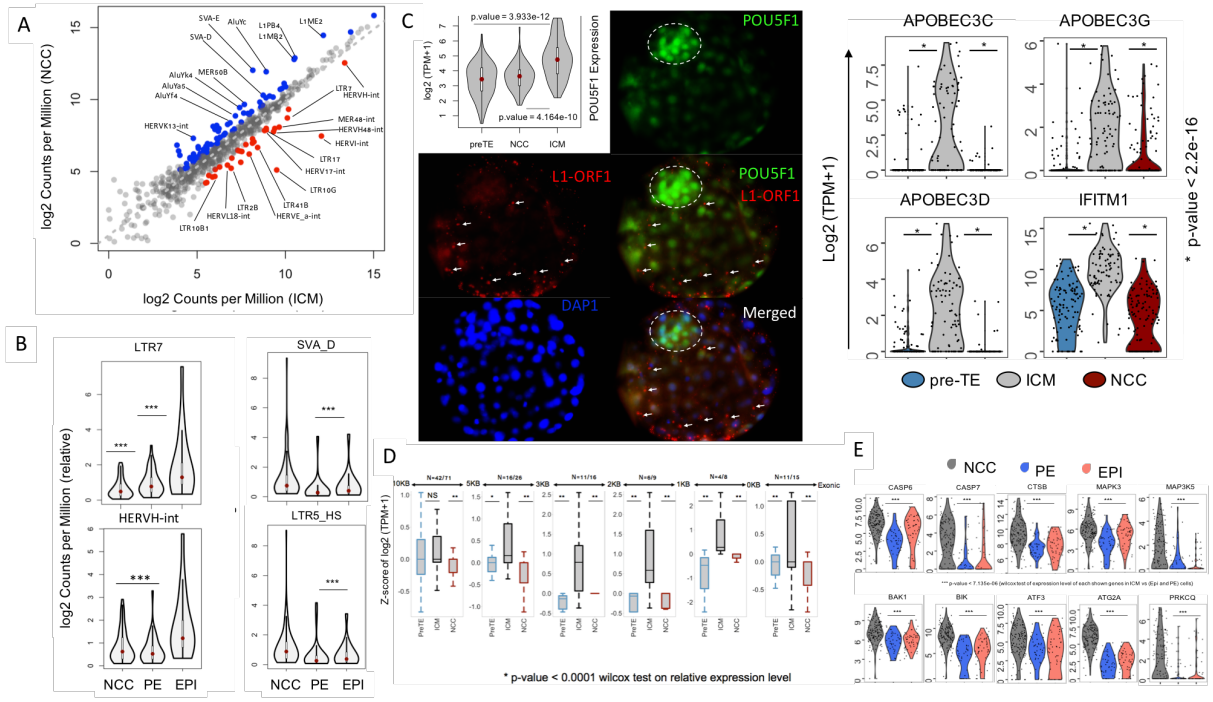


Figure 4.1:
A. Scatterplot shows the comparison of normalized mean expression in CPM (Counts Per Million) of various TE families between averaged pool of ICM (x-axis) and NCC (y-axis) cells. Read counts are normalized to total mapable reads per TE family. Note: Uniquely mapped reads were considered as one alignment per read. We considered one alignment to multimapping reads only if they were mapped to multiple loci, but exclusively in a TE family. Every point corresponds to a TE family. TE families enriched in ICM (red) vs NCC (blue).
B. Violin plots visualize the density and expressional dynamics of the retroelements, SVA, LTR5-HS (Young) vs LTR7/HERV-H (Old) the in non-committed (NCC) vs committed cell populations (PE and EPI) of the human blastocyst. Young retroelements enriched in NCCs vs committed cells (p-value < 0.000072), LTR7 and HERV-H-int enriched in EPI vs PE and NCCs (p-value < 0.00037).
C. Co-staining graph shows the immuno-florescence of ORF1 (is a trimer peptide which is required for L1 retrotransposition) from LINE1 elements (red dots) that is required for retrotransposition of L1, Alu and SVA elements in human genome, POU5F1 (Oct4, enriched in ICM as green dots) and DAPI (stains nucleus chromatin in blue color).
Right panel shows the enrichment of host defensive genes against retrotransposons in ICM cells.
D. Boxplots show the distribution of expression (log2 TPM relative to mean) of genes in multiple windows with respect to distances (0-25KB) from the left boundaries of HERV-H loci on the same strand. Genes used for the analysis are expressed (Log TPM > 1) in at least 1% of cell populations at E5 (pre-TE, ICM and NCC). Boxplots show relative gene expression where HERV-H is part of introns (N=122) or exons (N=11) lower left and top right, respectively. The 11 genes are the following: C6orf99, ESRG, IL34, LINC00263, LINC-ROR, HHLA1, MOK, TLR8-AS1, TPTEP1, UCA1 and ZNF177).
E. Violin plots visualize the density and distribution of gene expression (log TPM values) of selected genes that are upregulated in NCC vs EPI or PE. ICM and Pret-TE (Figure 1C and 1D) and this figure shows the comparison vs E6-E7 (EPI and PE) cells, and are involved in regulating apoptotic pathways (KEGG: hsa04210, Gene Ontology GO:008219, GO:0012501 and GO:0006915) (Wilcox test, p-value < 0.00000071).

In order to decipher the pattern of HERV-H expression during development, we determine the clusters. We redefine ICM, EPI, PE and TE cells according to our previously established criteria, and include cells exposed to *in vitro* culturing (ESCs) in our analysis. We calculate FPKM for the individual loci,

keeping the threshold of log2 FPKM > 1 to catalogue robust expression only. We observe overlapping, but distinguishable subsets of active HERV-Hs in the oocyte (n=39) and in the early embryo until 4-cell stage (n=87), respectively (Figure 4.2C). Furthermore, a different group of HERV-Hs gets activated during the embryonic gene activation (EGA) phase (n=43) and in EPI/ICM (n=31). We can also identify a specific group of genomic loci getting activated when cells of the blastocyst are exposed to *in vitro* culturing (e.g. ESCs) (Figure 4.2C).

Do the HERV-H loci, similarly to the blastocyst, drive their own expression, and affect neighbouring gene expression in other stages? Indeed, we find a number of examples when HERV-H is taking over transcriptional regulation of host genes in earlier stages of preimplantation development (Figure 4.2D). For example, FUT3, a top marker of morula is being remodelled by HERV-H, driven by a younger variant LTR7B. Similarly, OR1N1, TFPI and NLRP12 marking different stage of early embryogenesis are also remodelled by HERV-H (Figure 3B), suggesting that HERV-H might have been repeatedly co-opted during early embryogenesis. Notably, these remodelling events are relatively recent during the evolution of primates and appear to be specific to humans (Figure 4.2E).

4.3 HERV-H expression also has a characteristic pattern during the somatic reprogramming process

While in the above analysis we can identify time-stratified HERV-H expression, is there any reason to suppose that this is in any manner controlled and deterministic? One way to address this is to ask whether the forward series seen in embryonic development is recapitulated in reverse during somatic reprogramming, as somatic reprogramming mimics certain aspects a developmental process in a 'reverse' mode [429]. We use K-means clustering of human RNAseq data collected during somatic reprogramming [208]. The analysis identifies a list of genes that define distinct stages of the progression of reprogramming from 'initiation' via 'maturation', 'stabilisation' and finally 'maintenance' of pluripotent stem cells (Figure 4.3A). We observe a HERV-H expression cluster (C1) being generally active during the reprogramming process (Figure 4.3B). C1 (n=28) includes HERV-H-derived transcripts reported to promote the reprogramming process e.g. ESRG, ABHD12B, LINC-ROR, HPAT2/3 (Figure 4.3B).

In addition to C1, we identify two additional clusters of expressed HERV-H loci, C2 (n=159) and C3 (n=145) exhibiting a sharp change of expression at the border of 'maturation/stabilisation'. Thus, besides promoting reprogramming, we also identify subsets of HERV-H loci whose expression overlaps with the condition when pluripotency and self-renewal get stabilised (Figure 4.3B). Genes neighbouring C2 and C3 loci with phase specific expression display similar dynamics of expression to their respective HERV-H loci, and are potentially controlled by HERV-H (Figure 4.3C and 4.3D).

To identify a potential overlap between embryogenesis and somatic reprogramming regarding HERV-H expression, we intersect the clusters of HERV-H from embryonic development [297] and somatic reprogramming. Importantly, we find a high overlap (82%) of expressed HERV-H loci between *in vitro* cultured PSCs derived from either the blastocyst [297] or reprogrammed from somatic cells [208].

The self-renewal network, shared between EPI and iPSCs, is getting activated from the stabilisation stage of somatic reprogramming (Figure 4.3A and 4.3B). We still observe a partially similar set of expressed HERV-Hs (and their corresponding neighbours) between EGA and maturation stages (65%) (Figure 4.3E), however there is almost no shared expression of HERV-H loci between early embryonic development and

the stages of somatic reprogramming. These results confirm that the expression of certain HERV-Hs is likely deterministic and that these variants thus can be efficient markers of discrete stages.

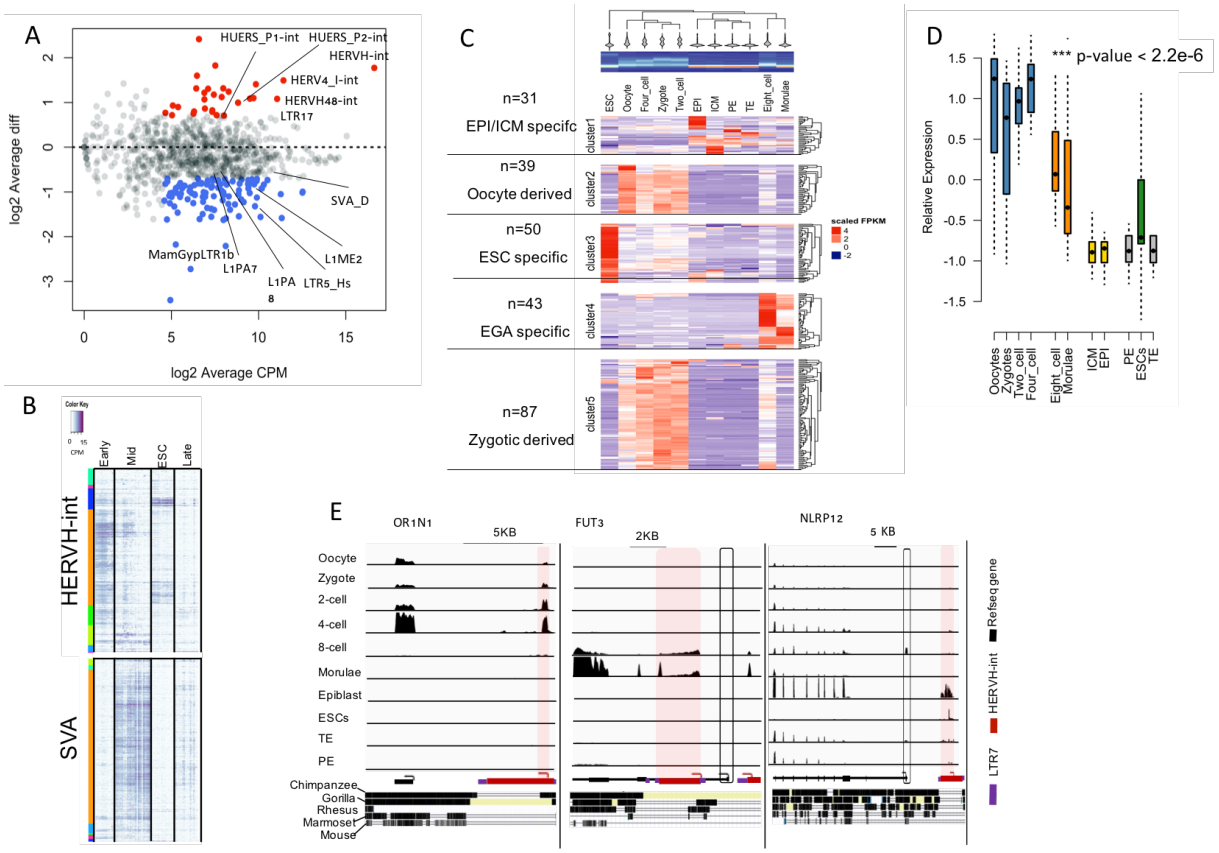


Figure 4.2:
A. MA plot displaying the comparison of average difference of normalized expression (CPM) of TE families between HERV-H-KD (n=2) vs GFP-KD (n=2) in H1-hESCs [209] at y-axis and average expression of TE families in HERV-H-KD and GFP-KD (n=4) in H1-hESCs combined at x-axis. Dots represent TE families downregulated (red) or upregulated (blue) (Log2 Average CPM $>$ 5 and average difference $>$ 2) in HERV-H-KD H1ESCs. .
B. Heatmap visualization of expression of significantly expressed (CPM $>$ 1 in any of the single cells) genomic loci of HERV-H-int and SVA(A-F) during preimplantation embryogenesis (early, oocyte/4-cell; mid, 8-cell/morula, late blastocyst stages and ESCs (passages 0 and 10 [297]) Note complementary expression pattern. The row side colours are indicating pearson's correlation clusters.
C. Heatmap shows the five distinct clusters of highly expressed (log2 FPKM $>$ 2) full-length HERV-H loci (n=250) in 124 single cells of human early embryos and hESCs [297] averaged into eleven cell populations (shown on top of the heatmap). Dendrogram shows the clustering of scaled FPKM values by euclidian distance and spearman's correlation (above cell type labels).
D. Boxplot represents the distribution of early embryo (oocyte, zygote, 2-cell and 4-cell) specific active HERV-H (n=126) from previous neighbour (upto 10 KB downstream) genes expression (log TPM > 1). Note: Relative Expression (log2 TPM – mean(log2 row-wise TPM) dynamics of genes neighbour ($<$ 10 KB) to HERV-H (cluster 2 and cluster 5 shown in previous) during distinct stages of human preimplantation embryonic development. These genes are upregulated in early embryos compared with mid and late preimplantation embryos.
E. HERV-H-enforced gene expression marks distinct stages of early development. Integrative Genome Visualization (IGV) of uniquely mapped reads over a specific gene and the proximal full-length HERV-H locus. Arrows show the annotated (black) and HERV-H-enforced (red) transcription start sites (TSSs). Transcription skipped at annotated (empty box) and HERV-H-enforced TSSs (shaded box) are shown. Lowest panels show phylogenetic conservation status, the presence (thick line) and the absence (narrow line) of the human sequence compared to the *Chimpanzee*, *Gorilla*, Rhesus, *Callithrix* and Mouse assemblies. In humans, OR1N1 is expressed until Embryonic Genome Activation (EGA), controlled by an upstream full-length HERV-H copy. The HERV-H-enforced FUT3 transcript is expressed in 8-cell/morula stages. LTR7B/HERV-H provides TSS and contributes to the chimeric FUT3 transcript in humans only. An upstream HERV-H locus trans-activates NLRP12 in epiblasts.

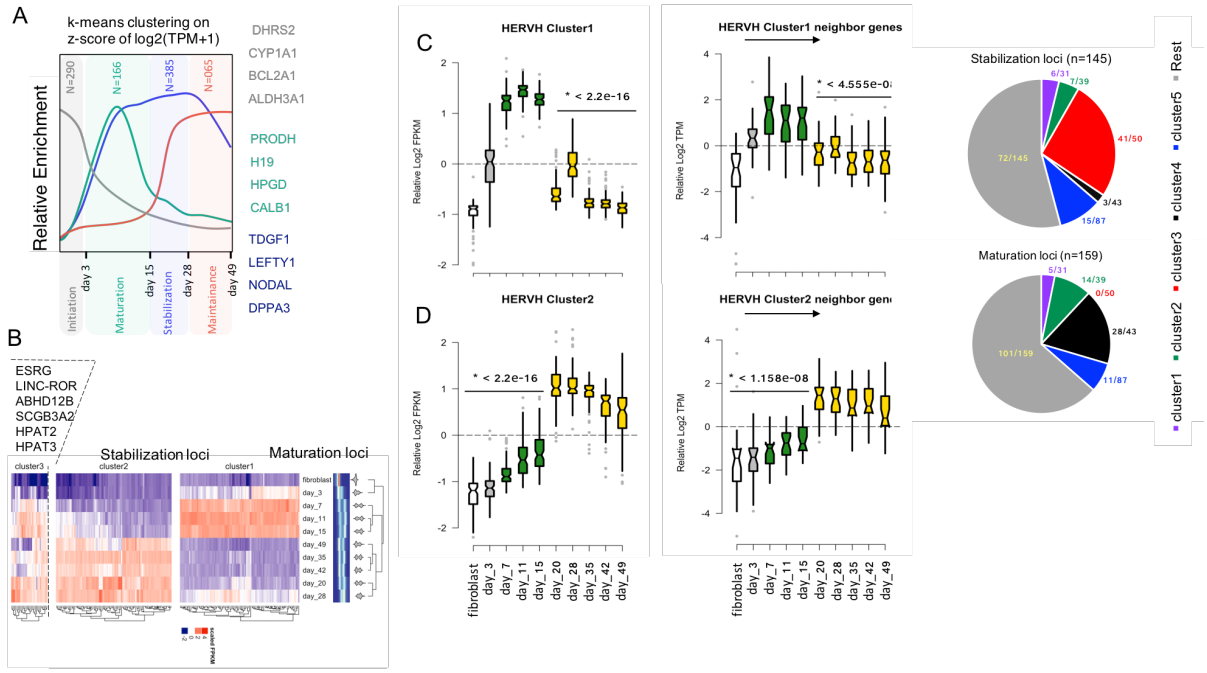


Figure 4.3:
A. Line plot represents connecting relative gene expression (log2 TPM) to the mean value of data frame during the reprogramming process of human fibroblast to pluripotency. 'Initiation', day 3, grey; 'Maturation', days 7, 11, 15, green; 'stabilisation', days 20, 28, blue; 'Maintenance', days 35, 42, 49, orange. Lines on the plot are connecting medians of relative expression levels of the four-major k-means clustered genes from each sample during the reprogramming process. Examples of stage specific genes during somatic reprogramming are shown next to the plot.
B. Heatmap shows the three distinct clusters of highly expressed (same criteria as in Figure 3C) full-length HERV-H loci (n=340) during the reprogramming process from fibroblast to hiPSCs. Cluster 1 (n=159) gain its expression during the maturation (day7-day15) stage. Cluster2 (n=145) is expressed during the stabilisation stage of reprogramming (day20-day28) and during passaging (day35-day49). HERV-H loci (n=36) of Cluster3 are expressed throughout the reprogramming process.
C. Notched boxplots show the expression (log2 FPKM) dynamic of 159 HERV-H loci (left panel) and their neighbour genes (up to 10 KB downstream) (right panel) specific to 'maturation' phase of reprogramming.
D. As on D, but 145 loci specific to 'stabilisation' phase of reprogramming.
E. Pie chart illustrates the overlapping HERV-H loci between reprogramming and early development. Clusters 1-5, as shown on Figure 4.2C

4.4 Both ICM and EPI are pluripotent, but only EPI has self-renewal potential

While the above analyses define the anatomy of early human embryogenesis and reveal markers for the different cell types, we can in addition ask more applied questions. We can ask both whether there is a cell type that would be a good candidate for a self-renewing pluripotent cell population that could be stably maintained in the lab? We can ask also about the relationship between lab maintained cell types and the new classification of true embryonic cell types. To answer the first question, we further dissected the transcriptional differences between EPI and the newly identified ICM. Using the top 1217 MVGs, we observed distinct clusters of ICM (n=75) and EPI (n=53) on PCA (Figure 4.4A). Curiously, pluripotency-specific markers (e.g. NANOG, KLF4, OCT4, etc.) do not exhibit differential gene expression between EPI and ICM (p-value insignificant) (Figure 4.4C), arguing against EPI being the

only pluripotent cell population in the preimplantation embryos. [298, 296]. By contrast, we find 22 and 9 genes, whose expression distinguishes ICM from EPI (AUC > 0.90), including BMP2, FOXR1, CLDN19 and NODAL LEFTY2 etc. , respectively (Figure 4.4B). Consistently with the association of HERV-H with pluripotency, we observe several HERV-H-derived gene products expressed in both ICM and EPI. However, while SCGB3A2 shows equal expression in both cell types, LINC-ROR, HHLA1, LINC00263, ABHD12B or ESRG are expressed more abundantly in EPI (Figure 4.4E), suggesting a possible functional diversification of HERV-H-enforced transcripts in pluripotency. Furthermore, while TFPI2 is exclusively expressed in ICM, the expression of its HERV-H-remodelled paralogue, TFPI has been shifted to both pluripotent cell types (Figure 4.4G). In addition to HERV-H-mediated remodelling events, gene duplication, and functional modification of the duplicated copy can be also observed in EPI vs ICM (e.g. NANOG vs NANOGNB) (Figure 4.4H). The EPI cluster can be also characterised by its low cell-to-cell variation. Indeed, EPI cells from both E6 and E7 cluster together, suggesting that these cells stably maintain their transcriptome in the blastocyst (Figure 3.1B). Notably, the low cell-to-cell variation is thought to be a property of cultured cells referred as self-renewal potential. This assumption is supported by NODAL, GDF3 and LEFTY1, implicated in triggering the self-renewal cascade in human [295, 309] being top-markers of EPI (Rank1, 2 and 10) in our analysis (Figure 3.1A, 3.1D and table). Expression profiling of a set of self-renewing genes suggests that the self-renewal potential might be a property of EPI, but ICM (Figure 4.4D), defining EPI as a strong candidate for *in vitro* work.

4.5 Transcription profiles of transposable elements characteristically mark early stages, and peak in morula

Next we ask whether DNA methylation based control might help to further define features of the human morula-blastocyst boundary, featured with a special set of embryonic genes activated (EGA). The global DNA methylation drops at zygotic state and reaches its minimum at the blastocyst stage [426], and the process is not yet completed in morula (shown around ERVs, Figure 4.5A), orchestrated by enzymes controlling DNA methylation. In contrast to mice, where expression of their dynamic robustly alters during the 2 cell-stage, human cells exhibit a sharp change at the 4 to 8 cell boundary [299](Figure 4.5A-B). In 8 cell/morula stage, levels of DNMT1/UHRF1, performing hemi-methylation are declining, TET1/TET2, playing a role in active DNA demethylation are rising, while DNMT3A and 3B, implicated in *de-novo* methylation and marking repetitive/transposable elements have a characteristic drop (Figure 4.5D). This renders morula an epigenetically transitory stage.

A further peculiarity of early embryogenesis (morula and before) is the generation of chimeric transcripts in which a transcript derives from two physically independent genomic loci. This species-specific process is highly active during early embryogenesis, gradually decaying to an approximately steady state number after morula (Figure 4.5C). The transition of DNA methylation that runs through early development also affects the expression of TrEs. TE families of different phylogenetic age are de-repressed, and get transcriptionally activated in characteristic waves during early embryogenesis [134, 184, 312]. While the phenomenon is observed in both mouse and human, the waves are species-specific, due to the unique repertoire of TrEs in each species [311]. Intriguingly, PC analysis of 968 most variably expressed distinct TE loci, mostly attributed to LTR12B, SVA, L1-PA and HERV-H (Figure 4.6A-B), reveals three major waves of activation during early, mid and late preimplantation embryogenesis (Figure 4.6A-B). We

also subjected a dataset of TE expression in cells derived from both developmental stages and *in vitro* culturing to PCA, an unsupervised clustering and pairwise correlation analysis. Ranked correlation matrix visualized the three waves of TE expression during development, but placed cultured cells separately (Figure 4.6C).

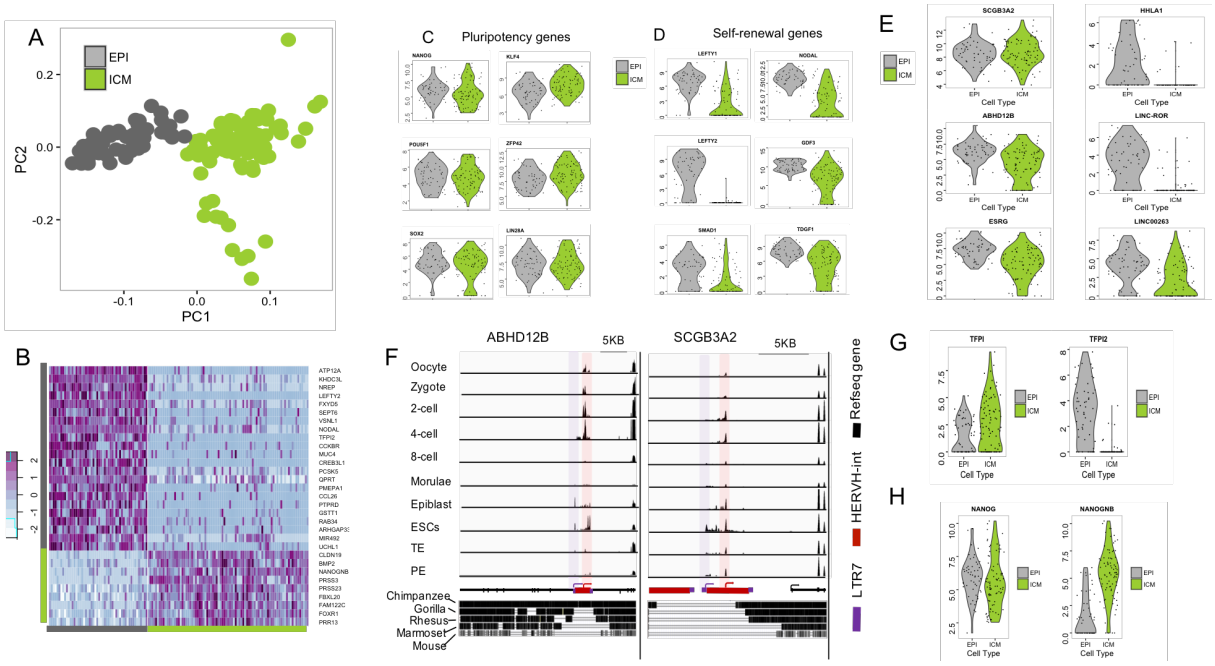


Figure 4.4:
A. PCA biplot showing the analysis of ICM (n = 75) and EPI (n = 52) cells using the R software package, Seurat. PC1 versus PC2 demonstrates the splitting process of ICM to EPI based on transcriptional proximity between the mentioned lineages; each dot represents an individual cell; colored legend for each subset is shown on the top of the plot.
B. Heatmap showing scaled expression (log TPM values) of discriminative gene sets defining EPI and ICM (AUC cutoff > 0.90). Color scheme is based on z-score distribution, from –2.5 (blue) to 2.5 (purple).
C-E, G-H. Violin plots illustrate expression distribution of candidate genes (left) associated with pluripotency, (p-value > 0.23); (middle) self-renewal (p-value < 0.00005); (right) HERV-H-remodelled genes (p-values on plots). (grey for EPI; green for ICM); TFPI and TFPI2 dynamics is in figure "G" whereas, NANOG and NANOGNB is in figure H showng dynamics of paralogous genes between plutripotent states of blastocyst.
F. IGV of uniquely mapped reads over the chosen genes (that has been shown to forming chimeric transcripts in our previous published work ref) and closest full-length HERV-H loci indicates the usage of multiple TSS arising from HERV-H at distinct stages. Both genes lose their annotated TSS and proximal exons to form HERV-H chimera. The chimeric ABHD12B transcript is expressed from zygote to EPI, but expression pauses in 8-cell/morula, Exons of ABHD12B upstream of HERV-H/LTR7 are skipped. While ABHD12B HERV-H appears to be intact in *Chimpanzee*, it has several deletions compared to the human version (not shown). SCGB3A2, implicated in pluripotency, exhibit partially overlapping expression patterns, usage of distinct human-specific HERV-H TSS and loss of annotated TSS and proximal exons.

4.6 HERV-H is exceptional in breaking the old-early/young-late rule

TE invasion of a genome is typically considered to be deleterious to the host. If so we expected to see a difference in the profile of young TrEs, yet to be fully suppressed, and older ones. To address this, we evaluated transcription patterns of Young (< 7 MYA) and Old (> 7 MYA) TE families during human early development.

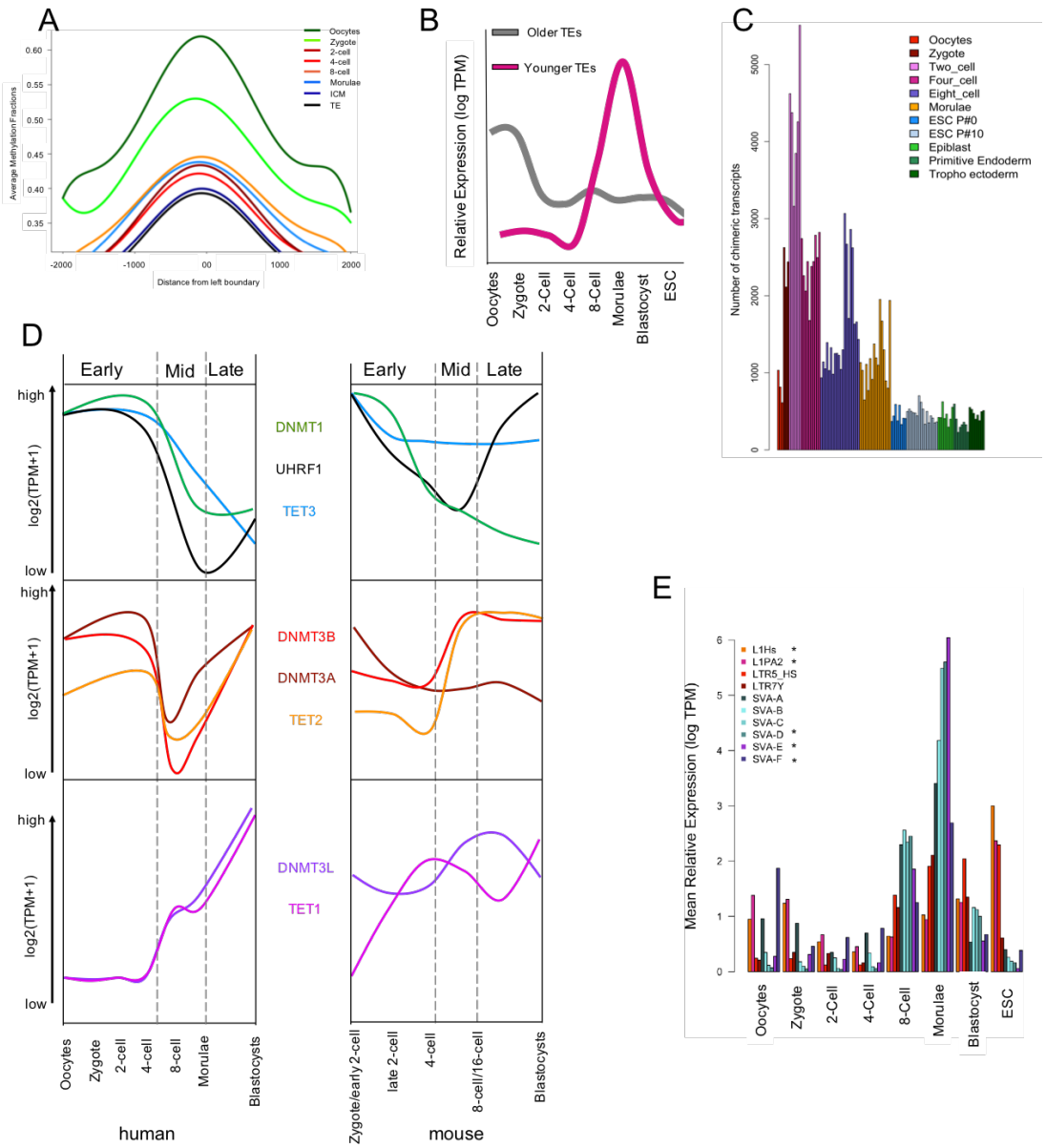


Figure 4.5:
A. DNA methylation pattern around TSSs of human placental imprinted genes from RRBS analysis during human early embryogenesis. Arrows indicate changes between oocyte to 8 cell stage and morula to ICM/Trophoectoderm. Lines in the plot are 6th degree polynomial regression values of average methylation with distance to left boundary (TSS) of Young TrEs. Lines are smoothed by spline function (spar=0.45).
B. Expressional dynamics of Old (>7MY) and Young (<7MY) retroelements `http://www.girinst.org/repbase/` during human pre-implantation-stages. Young TE expression peaks in 8-cell/morula. LTR7Y is older (> 7 MYA) but its younger with respect to LTR7 (< 30 MYA). TrEs marked with $*$ were demonstrated to retrotranspose in the human genome [77].
C. Generation of chimeric transcripts is tuned down after morula (excluding any chimera arising from Transposed Elements). Bar plot showing the number of chimeric transcripts detected from single cells of human preimplantation stages. Only data from [297] was considered since other studies had shorter reads in order to determine chimeric transcripts.
D. Dynamic of selected enzymes affecting genomic DNA methylation. Lines on the plot are connecting medians of single cell expression levels (log2 TPM) from the depicted stages of development. Early includes oocytes, zygote, 2-cell, 4-cell stages in human, while zygote, early 2-cell, late 2-cell, 4-cell in mouse. Mid represents 8-cell, 16-cell/morula stages, and Late denotes blastocysts in both human and mouse. Lines are smoothed by spline function (spar=0.45).
E. Expressional dynamics of Old (> 7MY) and Young (<7MY) retroelements during human pre-implantation-stages (elaborated version if Figure B).

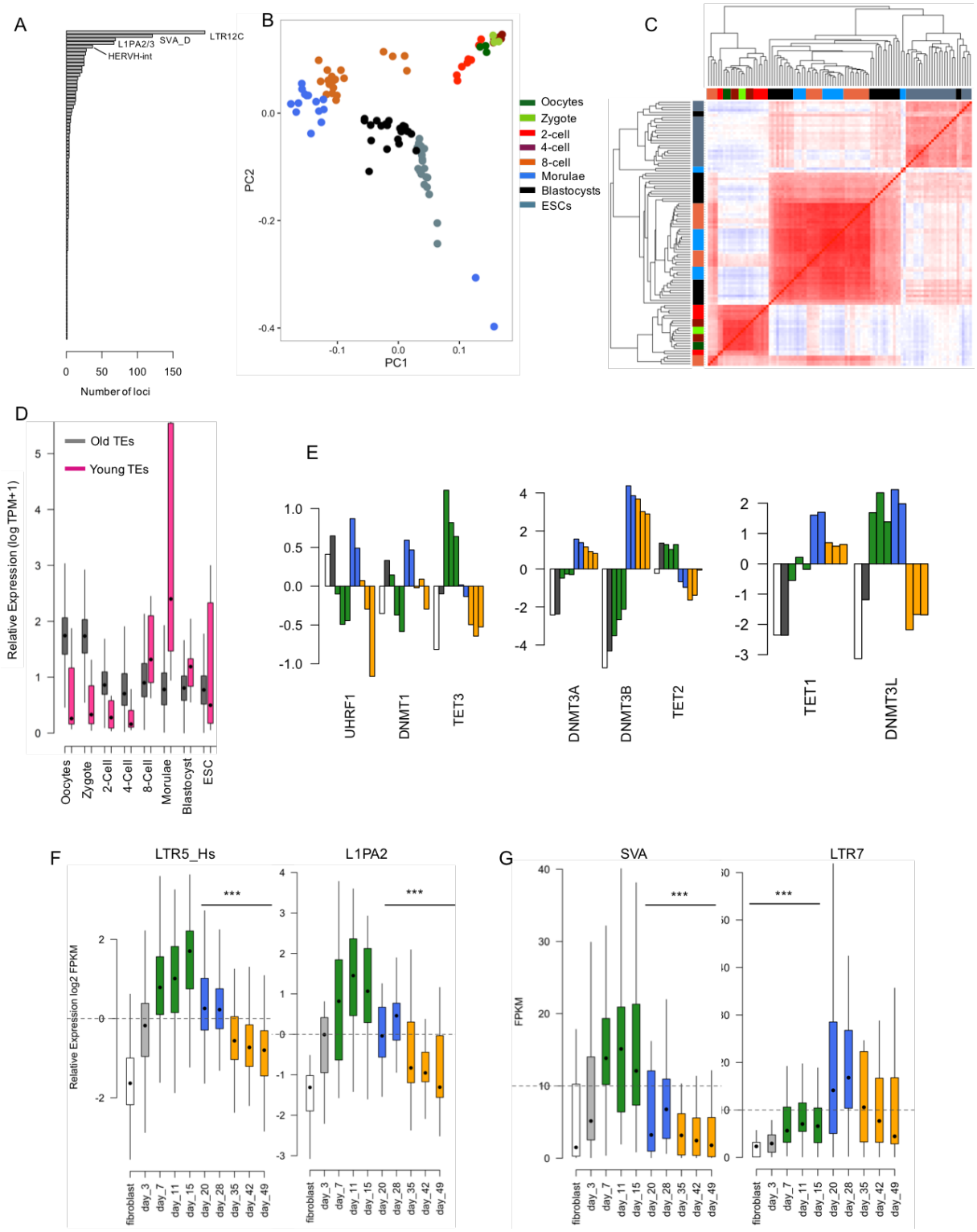


Figure 4.6:

A. Number of independent TE loci. Most variable 968 TrEs expressed during human early embryogenesis. Bars are showing the number of independent loci from each TE families containing the 968 most variable TE loci.

B. Transposable elements mark distinct human developmental stages of embryogenesis. PCA of the most variable 968 TrEs shows that human early embryogenesis can be clustered distinctly on the two most significant eigen vectors.

C. Unsupervised clustering and ranked correlation analysis. Ranked correlation matrix visualization of single cells from human preimplantation human development and cultured PSCs shows the dynamic expression of the most variable TrEs (n=968). ***D***. Expressional dynamics of Old (> 7MY) and Young (<7MY) retroelements during human pre-implantation-stages (See also Figure 4.5B).

E. Expression dynamic of enzymes affecting DNA methylation during somatic reprogramming. Relative expression to the row means. (Color codes as on Figure 4.3A).

F-G. Expressional dynamics relative to mean of the Young retroelements, LTR5-HS, L1PA2 and SVA vs expression dynamics of Old LTR7/HERV-H during the reprogramming process of human fibroblast to pluripotent stem cells.

Old elements are transpositionally inactivated, while some of the Young TrEs are still actively transposing or have been inactivated recently [74]. DNA transposon-derived transcripts are relatively abundant in zygotes and 2-cells stage, but their levels, together with other Old TrEs, gradually decline as development proceeds (Figure 4.5B and 4.6D). This we presume reflects decay of remnant RNAs expressed in oocytes as in humans gene expression isn't reactivated until the 4 to 8-cell stage [299]. Curiously, the transcriptional activation of the Young elements, including L1 (L1-Hs) and SVA (SVA-D, E and F), capable of retrotransposition [74, 77], is substantial from 8-cell stage, peaking at morula with a contrasting dynamic to DNMT3A and 3B, but declining in blastocyst (Figure 4.5E). The high expression of these Young potentially mutagenic transposing elements appears to be a signature also for non-committed vs committed (PE and EPI) cell populations (Figure 4.5E). "Old" LTR7-HERV-H, expressed in committed cells and peaking at EPI is the one exception to the Old/Young difference, and curiously opposing the expression pattern of Young TrEs (e.g. SVA) (Figure 4.5E and 4.6D). SVA elements were found to be activated in HERV-H-KD ESCs supports the exclusive expression pattern of these two distinct families of TrEs (Figure 4.2A-B).

TE expression supports the view that somatic cell reprogramming mimics certain aspects of development in 'reverse' and the exceptionalism of HERV-H Does the old-early/young-late rule transfer to the reprogramming process where development is done "in reverse". If it does, is HERV-H still an exception? Comparable to normal development, the somatic reprogramming process includes genomic demethylation and fluctuations in the expression of DNA methylation genes (Figure 4.6E), resulting in reactivation of TrEs. By the end of the process, the cells acquire many features of pluripotency that also includes silencing of TrEs [426]. We detect massive activation of Young TrEs in the 'maturation' stage (Figure 4.6F). Again, LTR7/HERV-Hs have a dynamic opposite, the expression values and the number of expressed HERV-H loci plateau in the 'stabilisation' stage of induced pluripotency, when the expression from Young elements declines (Figure 4.6G). Maturation phase is non-committed for iPSCs, and here too we see the signature of Early embryo's TE dynamics. Activation of LTR5-HS, L1PAs, SVAs in human embryos, Pluripotency is achieved in EPI that is marked by higher expression of HERV-H loci, during reprogramming pluripotency is achieved at day 28, where we notice similar dynamics of HERV-H (Figure 4.6G-H).

Chapter 5

Human-specific nature of early embryos

The blastocyst comprises the pluripotent epiblast lineage that programs the blueprint of body plan and potentially gives rise to the rest of somatic and germ cell lineages [382, 383]. The most potent cells were claimed to be ICM that could be flexible enough to generate the rest of other cell types [372]. The ICM gives rise to hypoblast (primitive endoderm), an extraembryonic lineage and the rest of cells develop into pluripotent epiblast, that has noticeable disparity with blastomeres and ICM [368, 373, 374]. Embryonic stem cells (ESCs) were derived from ICM, but they were originated from epiblast [375, 376]. So, the epiblast is source of pluripotent stem cells (PSCs) *in vitro*, and *in vivo*, that has played the role of *"game changer"* in regenerative medicine, *precision medicine* and stem cell biology along-with induced pluripotent cells (iPSCs) [334]. Epiblast is the "Native pluripotent state" that should be kept as reference frame to compare any *in vitro* cell line that is being claimed as "Naive pluripotent cell". Generation of naive cells is an attempt to diminish the gap between natural and artificial pluripotency and a contest to apprehend the missing ICM like human pluripotency in petri dish. In last five years, there has been numerous claims of derivating human naive state encompassing hESCs or hiPSCs [380, 317, 357, 367, 359, 361, 366]. The *pros and cons* of forced Naive cells has recently been frequenctly reviewed as a cellular habitat for developmental panacea [336, 337, 377, 378, 379]. Among 12 functional frameworks to evaluate human naive pluripotency, five of them are primate-specific [430]. The prerequisite for the determination of efficacy of naive cells would be on the basis of similarity and disparity between human and non-human primate's epiblast. For example, the modulation of pluripotency marker, NANOG is dependent on NODAL/TGFβ in human PSCs [431, 298] but not in new-world monkeys [432]. These observations indicate that validation of human naive cells, where rodent models are insufficient, the species-specific mechanisms underlying human epiblast formation might be the beckon of light.

Is early embryogenesis in humans in any manner distinct from that of primates? Prior analysis suggested that on the broadest scale it was exceptional. First, it was suggested that the trajectories of cell types in early human development are different from those seen in other mammals, including primates. This model suggested that in human the cell type, known as inner cell mass (ICM) might not even exist. An alternative second model poses that in broad scale cell types and trajectories are not evolving but the underlying genetic architecture is. Here we examine the two models.

Our comparative single cell high-resolution analysis of human vs *Cynomolgus* blastocyst reveals that

human development is not that unusual as previously thought. In fact, human development is basically classical, preserving all the major cell types. Instead, we find that much divergence is owing to rewired regulation of genes, remodelled genes and even novel, human specific genes. I show that primate specific, domesticated endogenous retrovirus HERV-H has played a significant role during primate evolution as HERV-H remodelled, chimeric and *de-novo* genes are central to restructuring the blastocyst, primarily the pluripotent epiblast development. Our analysis using primate iPSCs as models of the pluripotent epiblast suggest rapid lineage specification occurred following HERV-H invasion in new world monkeys. I show that the major gene expression gain (19) and loss (29) events in regulating pluripotency between human and new world monkeys (Callithrix) are due to the HERV-H-governed gene expression. This involvement of HERV-H underpins the great majority of the gene-level differences in genetic architecture between otherwise similar cell types across species.

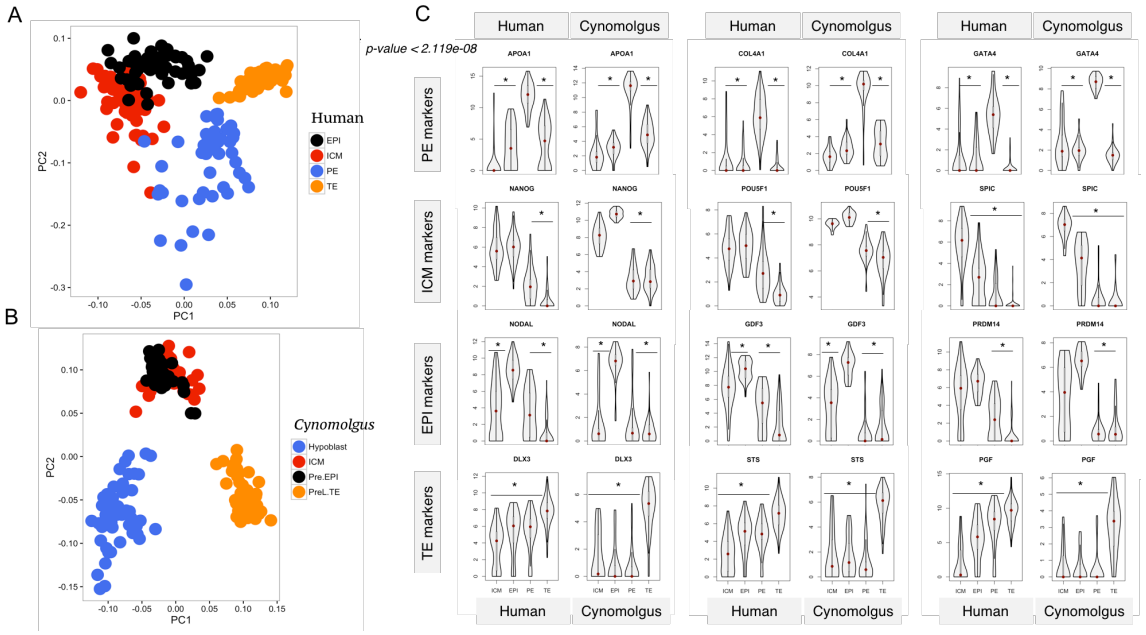
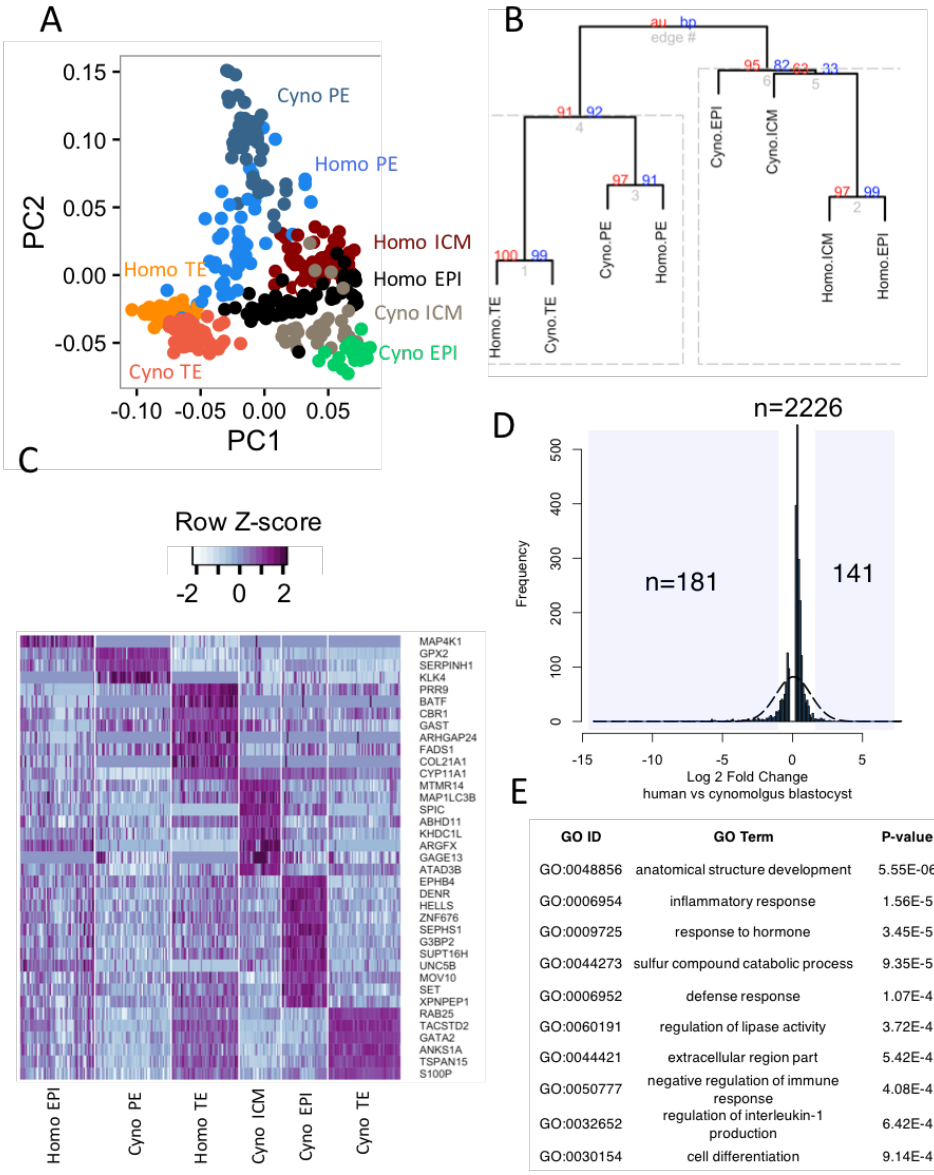


Figure 5.1:
A. PCA of the distinct lineages of human blastocysts by most variable genes among the shown groups (228 cells, 1055 genes). Note: We considered ICM, EPI, PE and TE cells and only those genes were taken into account that were annotated in refseq gene track of both human and *Cynomolgus* species. Every dot represents single cell. Colors are flag for distinct cell types in human blastocyst. **B**. PCA of the distinct lineages of *Cynomolgus* blastocysts by most variable genes among the shown groups (170 cells, 1237 genes). Note: We considered ICM, EPI, PE and TE cells and only those genes were taken into account that were annotated in refseq gene track of both human and *Cynomolgus* species. Every dot represents single cell. Colors are flag for distinct cell types in human blastocyst. **C**. Multiple violin plots visualize the density and distribution of selected gene expression that are conserved markers of distinct lineages across the vertebrates blastocysts [433]. Plot shows the similar pattern of detected key markers for analyzed stages e. g NANOG, POU5F1 marking pluripotent states (ICM and EPI), SPIC marking ICM (primate-specific), EPI is marked by NODAL, GDF3 and PRDM14, PE is APOA1, GATA4 and COL4A1 enriched and finally DLX3, STS and PGF marking TE in both species.

5.1 The pluripotent epiblast displays the most diverged transcriptome in the blastocysts of *Cynomolgus* and human

To unravel divergence in preimplantation embryogenesis in primates, we started with reanalysing single cell transcriptomes of the blastocysts from human and *Cynomolgus* [385, 296].



GO ID	GO Term	P-value
GO:0048856	anatomical structure development	5.55E-06
GO:0006954	inflammatory response	1.56E-5
GO:0009725	response to hormone	3.45E-5
GO:0044273	sulfur compound catabolic process	9.35E-5
GO:0006952	defense response	1.07E-4
GO:0060191	regulation of lipase activity	3.72E-4
GO:0044421	extracellular region part	5.42E-4
GO:0050777	negative regulation of immune response	4.08E-4
GO:0032652	regulation of interleukin-1 production	6.42E-4
GO:0030154	cell differentiation	9.14E-4

Figure 5.2:
A. PCA visualization (PC1 and PC2) on the cross-species normalized scaled genes (commonly annotated in human and *Cynomolgus* refseq gene track format aka gtf) expression (Log2 TPM) estimates in human and *Cynomolgus* blastocyst single cells aka. ICM, EPI, PE and TE. The 1055 genes that showed the most variation (methods) across the merged cross-species data-sets were selected for input loading. Every dot represents a single cell and color code for respective cells are penned next to dots with same color.
B. Dendrogram via bootstrapping (1000 replicates) based hierarchical clustering using ranked correlation and complete linkage method on averaged expression from the cell populations (mentioned in Figure 1A) transcriptome pooled together. Height of dendrograms represent the euclidian distance of dissimilarity matrix, numbers in red and blue indicate au and bp values from bootstrapping
C. Heatmap showing scaled expression (Log2 TPM values) discriminating gene or gene sets defining each lineages and unique to their respective species with AUC cut-toff > 0.90. Note: ABHD12B and SCGB3A2 is commonly annotated but expression is human specific genes but not shown on this heatmap because their expression is shared but unique to human ICM and EPI both, whereas heatmap displays only unique cell type markers.
D. Barplot showing the distribution of differentially regulated genes between human and *Cynomolgus* blastocyst as a whole from 11, 053 orthologous genes detected to be expressed in any 5 cells. There are 181 down-regulated, 141 up-regulated genes in human blastocysts, 2, 226 genes were not significant (p-value < 0.05 and log2 Fold change $>$ 1) and rest of genes did not show any differential expression
E. Table displays the gene ontology enriched in differentially expressed genes keeping entire orthologous gene-sets in background. Analysis was performed using *Gorilla* tool.

In contrast to the human study [296], the lineage specific cells were extracted prior to sequencing in *Cynomolgus* [385], thus we take comparable cell populations of ICM, EPI, PE and TE. We classified their transcriptomes on PCA that revealed the similar pattern of distinct cell types in both *Cynomolgus* and human (Figure 5.1A-B). For comparison, we only use the genes that are annotated in both species. Using the top transcription markers that are expressed in both species, we could identify EPI (e.g. NODAL, GDF3, PRDM14) PE/hypoblast (e.g. APOA1, GATA4 and COL4A1) and TE (e.g. DLX3, STS and PGF). This strategy also identified the ICM unambiguously marked by NANOG, POU5F1 and SPIC (primate-specific) in both species (Figure 5.1C). These conserved markers exhibit similar transcriptional dynamics in blastocyst lineages (Figure 5.1C). Interestingly hypoblast aka PE exhibited a scattered distribution on first two principal components suggesting the higher level of heterogeneity compared with rest of three lineages in both species (Figure 5.1A-B). To compose the comparable data for cross-species analysis, we calculate the scaled expression of Homo- *Cynomolgus* common 16,222 genes and merge the data in a single pool. Applying quality control thresholds, we end up with 11,043 genes for further analysis. PCA plotting these merged cross-species data kept the PE and TE lineages together, regardless of their phylogenetic divergence (Figure 5.2A-B and 5.4 A).

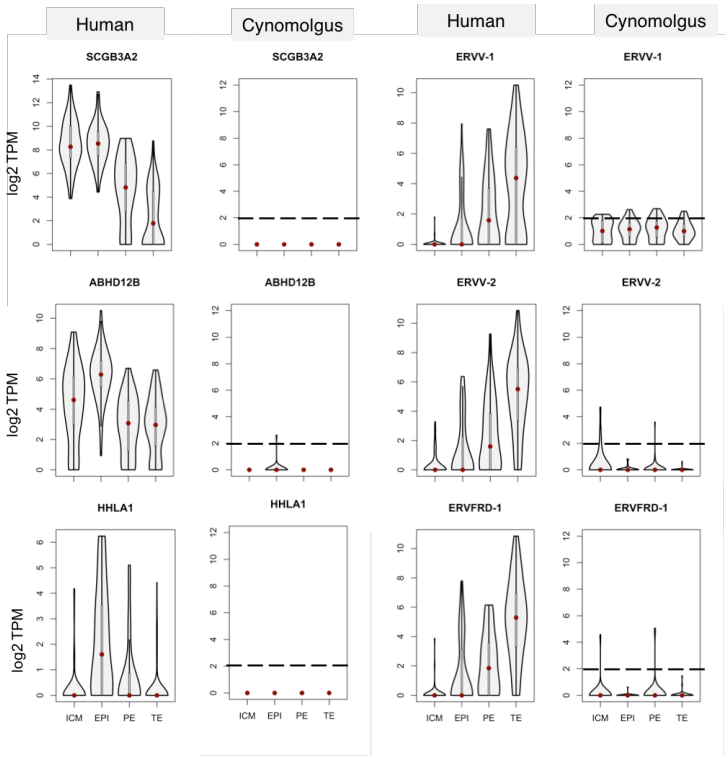


Figure 5.3:
Violin plots visualize the density and distribution of gene expression (log TPM values) of selected orthologous genes that were expressed exclusively in blastocyst cells of human but not in *Cynomolgus* as shown in comparative human and *Cynomolgus* cell type independently. Left panel shows the genes that are remodeled by HERV-H sequence marking specifically the pluripotent populations in human-specific way.
Right panel shows the violin plot visualization of density and distribution of gene expression (log TPM values) of selected orthologous genes that were expressed exclusively in blastocyst cells of human but not in *Cynomolgus* as shown in comparative human and *Cynomolgus* cell type independently. Left panel shows the envelope genes from endogenous retroviral family HERVV-1, HERVV-2 and HERV-FRD.

PCA has plotted the cross-species lineages close to each other instead of intra-species lineages, displaying the fidelity of our merged cross-species data analysis. Upon loading the top four most

significant principal components (PCs) to segregate the cells on two t-SNE dimensions, we found the compact clusters of cross-species lineages, also evolutionary divergent behaviour of EPI followed by ICM lineages between human and *Cynomolgus* (Figure 5.2A and 5.5A). Indeed, 1K unbiased hierarchical clustering using correlation method shows the significant level of similarity between cross-species TE and PE cells (Figure 5.2B). In contrast, the Homo ICM displays transcriptome-wide similarity rather to *Cynomolgus* EPI than ICM (Figure 5.2A-B and 5.4 A), perhaps following a similar trend of directionality also observed between *Callithrix* ICM and *Cynomolgus* EPI [385]. Upon loading the first four most significant PCs to segregate the cells on two t-SNE dimensions, we see again the compact clusters of PE and TE, and the divergent behaviour of ICMs, but mostly in the EPI populations between Homo and *Cynomolgus* (Figure 5.4A).

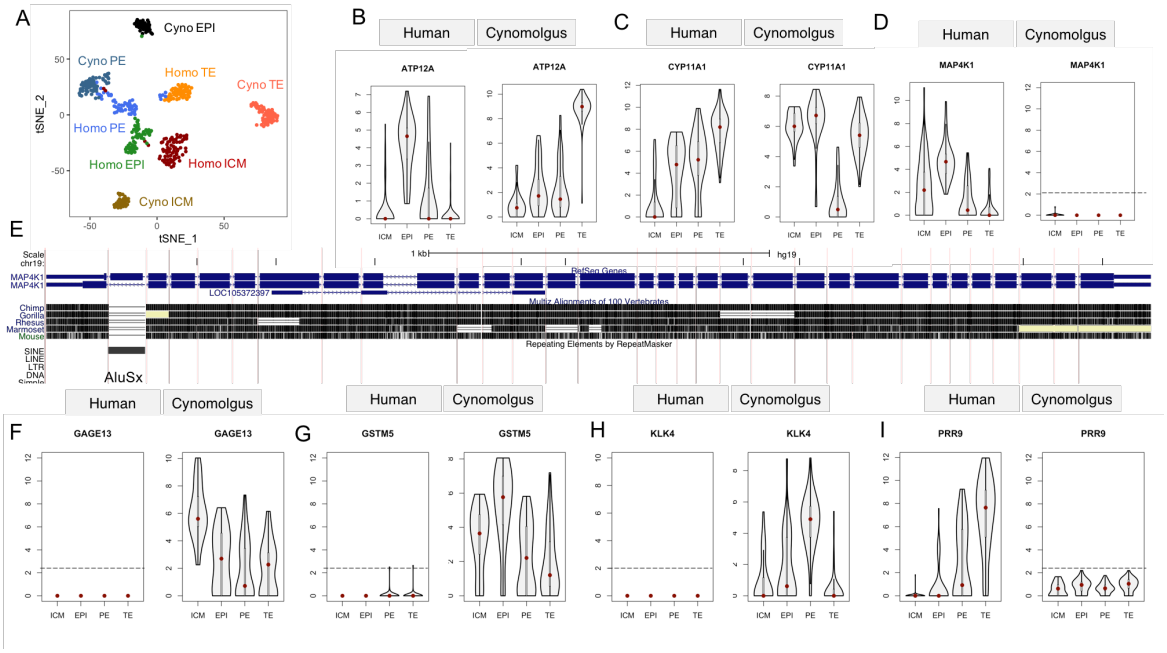


Figure 5.4:
A. t-SNE plot of human- *Cynomolgus* merged datasets (398 cells, 12605 genes) mentioned in Fig S1A and S1B using 948 most variable genes contributing to first 4 PCs. Every single dot is single cell and text next to clusters are the cross-species clustered lineages resolved on t-SNE plot. Note: Here we see 6 single human PE cells clustered with the rest of human TE cells, that were not resolved in our analysis. Similarly, 3 of Cyno PE cells clustered with rest of Cyno EPI cells, but that doesn't affect the inferences we draw in this study.
B. Violin plots showing expression distribution of ATP12A gene that is highly expressed in EPI lineage of human blastocysts but in TE lineage of *Cynomolgus* blastocyst.
C. Violin plots showing expression distribution of CYP11A1 gene that is highly expressed in TE lineage of human blastocysts but in the case of *Cynomolgus* blastocyst, it is enriched in EPI/ICM at higher level followed by TE.
D. Violin plots showing expression distribution of MAP4K1 gene that is annotated in both species but expressed in exclusively in EPI lineage of human blastocysts
E. UCSC snapshot of coding region from MAP4K1 gene illustrating the exonization of human-specific AluSx element as it forms the last exon of mentioned gene.
F. Violin plots showing expression distribution of GAGE13 gene that is annotated in both species but expressed in exclusively in *Cynomolgus* blastocysts at higher level in ICM lineage
G. Violin plots showing expression distribution of GSTM5 gene that is annotated in both species but expressed in exclusively in *Cynomolgus* blastocysts at higher level in EPI lineage
H. Violin plots showing expression distribution of KLK4 gene that is annotated in both species but expressed in exclusively in *Cynomolgus* blastocysts at higher level in PE lineage.
I. Violin plots showing expression distribution of PRR9 gene that is annotated in both species but expressed in exclusively in human blastocysts at higher level in TE lineage

5.2 Cross-species shifts of gene expression between blastocyst lineages

We also find several examples of cross-species shifts of gene expression between the distinct cell types of the blastocyst. For example, ATP12A gene that marks *Cynomolgus* TE is expressed in human EPI (Figure 5.4B), whereas CYP11A1 expressed in ICM and EPI in *Cynomolgus*, marks human TE (Figure 5.4C). We also identify several orthologous genes that mark a certain blastocyst lineage in one, but not even expressed in the blastocyst of the other species, suggesting that their expression has been evolved in a species-specific manner to a particular developmental stage. We observe GAGE13, GSTM5, KLK4 expression in the distinct lineages of *Cynomolgus* whereas PRR9, MAP4K1 are detectable exclusively in the human blastocyst (Figure 5.4D-I). While, there is no obvious explanation to the shift in expression of each cases, we find an evidence of recent exonization event of an AluSx insertion specifically in human the MAP4K1, suggesting this remodelled gene has been recruited to the human EPI transcriptome quite recently (not present in chimp) (Figure 5.5E).

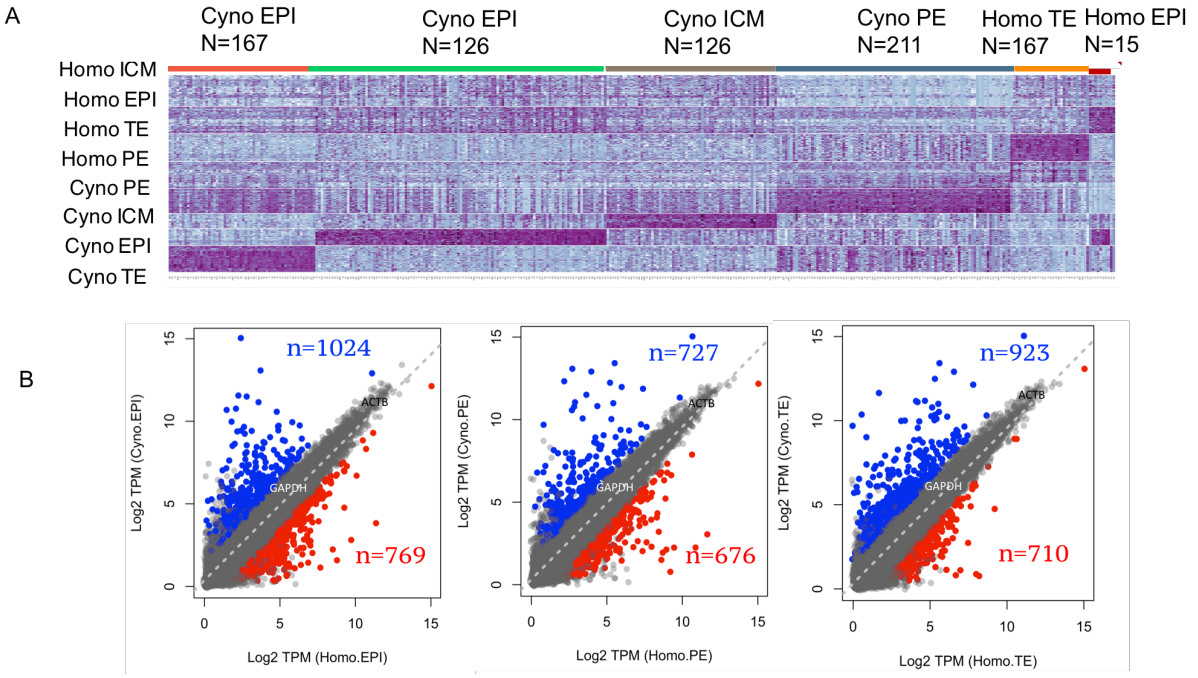


Figure 5.5:
A. Heatmap displaying z-score of cross-species expression level across the different stages of human and *Cynomolgus* blastocysts viz. ICM, EPI, PE and TE. heatmap enlists all upregulated genes obtained while comparing any given lineage against the rest of all 7 lineages. This plot highlights the species-specific expression of genes marking distinct lineages of blastocysts as illustrated in Fig S1. Note: Human ICM and PE had none of expressed genes that could be used as their marker (AUC > 0.75).
B. Scatterplots show the pairwise comparison of cross-species normalized transcriptomes of contemporary blastocyst layers between human and *Cynomolgus* EPI.
C. Scatterplots show the pairwise comparison of cross-species normalized transcriptomes of contemporary blastocyst layers between human and *Cynomolgus* PE.
D. Scatterplots show the pairwise comparison of cross-species normalized transcriptomes of contemporary blastocyst layers between human and *Cynomolgus* TE.

To see which cellular mechanisms are primarily diverged between *Cynomolgus* and Homo, we signature approximately 300 gene expression changes upon comparing entire cross-species blastocyst's

single cells in a pairwise manner (Figure 5.2D). The observed differentially regulated genes (DEGs) are categorized under anatomical structure, inflammation response, defence response and immune related ontologies (Figure 5.2E). In order to identify the genes that have been positively selected to mark a particular blastocyst lineage, we fetch the genes with higher level of expression [area under curve (AUC) > 0.90] in any given cluster of cross-species cells against the rest of the 7 clusters (Figure 5.2C). These 37 genes can be considered as species-specific lineage markers of the blastocyst (Figure 5.2C).

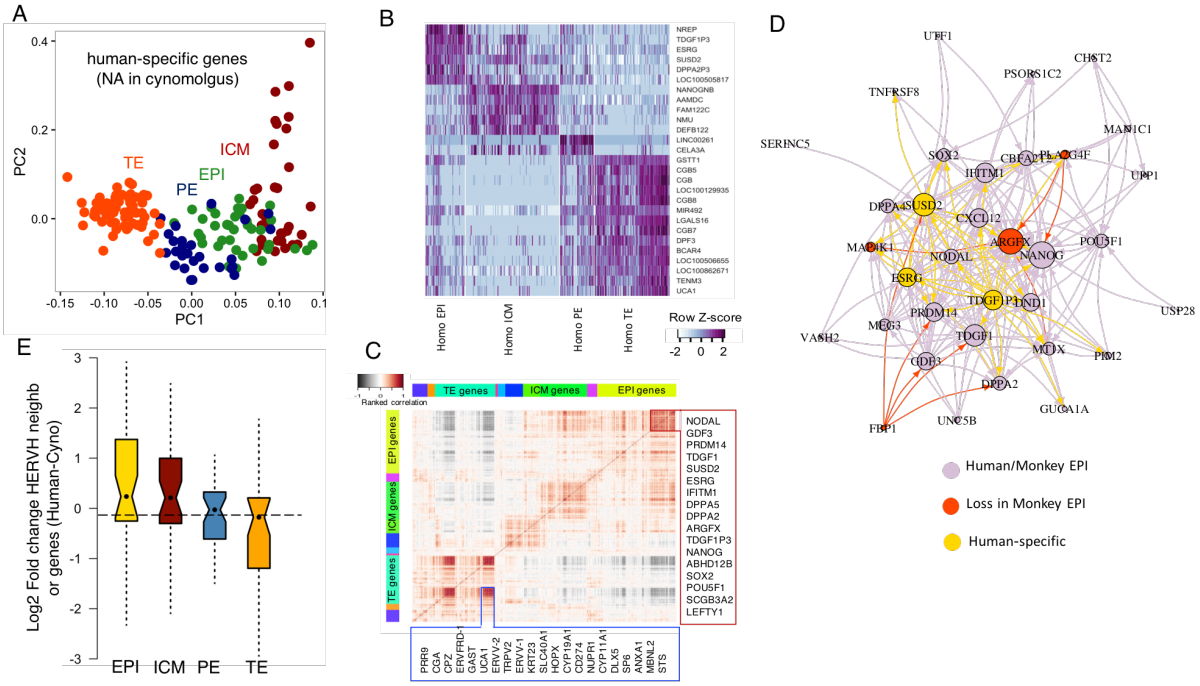


Figure 5.6:
A. PCA visualization (PC1 and PC2) of the normalized scaled genes (exclusively annotated in human refseq gtf) expression (Log2 TPM) estimates in human blastocyst single cells of ICM, EPI, PE and TE. The 526 genes that showed the most variation (methods) across the merged cross-species datasets were selected for input loading. Every dot represents a single cell and color code for respective cells are penned next to dots with same color.
B. Heatmap showing scaled expression (Log2 TPM values) discriminating human-specific gene sets (annotated in human but not in *Cynomolgus* refseq gene track format) defining each lineages of human blastocysts with AUC cuttoff > 0.90. Note: ABHD12B and SCGB3A2 is commonly annotated but expression is human specific genes but not shown on this heatmap because their expression is shared but unique to human ICM and EPI.
C. Heatmaps of correlation matrix visualizing the pair-wise correlation of most variable genes in single cells of human blastocysts lineages (1076 genes out of total genes in human refseq) whose dynamic expression is enough to segregate distinct blastocyst lineages on first two principal components (Fig S1A). K-means clustering provided three major clusters carrying genes marking EPI, ICM and TE (from right to left). Red box next to heatmap contains a cluster of EPI markers that are tightly correlated where ESRG is pinned in whereas, blue box contains a cluster of trophectoderm markers where UCA1 is pinned in.
D. Differential transcriptional network regulating self-renewal between monkey and human by analysing single cell transcriptomes. Only pairs having a strong correlation with ESRG are considered. For each pair, nodes and edges are decided on their expressional dynamics in EPI. Size of the nodes is proportional to the number of components the gene is paired with in the network. Colors denote species specificity: expressed in human EPI only but annotated in both human and monkey (red), in both human and monkey EPI (light pink), human-speciifc genes in network (gold). Note: genes whose expression is shifted from monkey ICM to human EPI could not be found in this network.
E. Notched boxplot represent the distribution of average difference (at log2 scale) of LTR7-HERV-H neighbour (upto 10 KB downstream) genes expression (cell populations pooled together, scaled and averaged). Note: Only cross-species genes (commonly annotated in human and monkey refseq gtf) were taken for this action where ICM and EPI showed considerable level of upregulation of HERV-H neighbour genes expression suggests the recent co-option.

Finally, to compare cross-species expression divergence in any given lineage, we combine the distinct cell types from each species and calculate their pairwise differential expression. As expected, this analysis identified the highest numbers of DEGs in ICM and EPI (Figure 5.5B) between the two species. Curiously, among the top species-specific lineage markers, we notice several genes that were remodelled by HERV-H (e.g. ABHD12B, HHLA1 and SCGB3A2, expressed in human ICM and EPI). Furthermore, HERV-H-enforced genes could be involved in rewiring the transcriptome of the blastocyst, as we observe the HERV-H neighbour genes to be up- and downregulated in EPI/ICM and TE respectively, while PE neighbours were not affected in our cross-species analysis (Figure 5.6E). In human, envelop derived genes of another HERVs, HERV-V (e.g. HERV-V1, HERV-V2) and HERV-FRD1) are expressed specifically in TE (Figure 5.3 right panel). Curiously, while these HERV remodelled genes are conserved in both human and *Cynomolgus* genome, their expression pattern is not, and they mark species-specific lineages, and might be even the potential drivers of the cross-species lineages segregation.

5.3 HERV-H-derived transcripts might define human-specific features of the blastocyst

Do we also identify genes that are not annotated in *Cynomolgus*, and might be even human specific? To answer, we used 3,023 human Refseq genes that were not annotated in *Cynomolgus* Refseq gene track format, but expressed in the human blastocyst. To see if the transcriptions of these genes are stage specific, we performed a PCA on their expression estimates. We find that the expression dynamics of these genes clearly segregate the distinct blastocyst lineages on the first two principal components (Figure 5.6A). This enabled us to discover those genes whose expression could be putative marker of human-specific ICM, EPI and TE (Figure 5.6B). Intriguingly, among these genes, we find UCA1 (TE), LOC100505817 (ICM) and ESRG (EPI) REF that are almost completely comprised by a full-length HERV-H sequence, flanked by different LTR7 variants (e.g. LTR7C, LTR7Y and LTR7, respectively) (Figure 5.7A-B). We then took ESRG and UCA1 marking EPI and TE, respectively and asked how was it incorporated in the human trancriptome.

5.3.1 The HERV-H derived ESRG is integrated into the regulatory circuitry of self-renewal in human pluripotency

Next, we examine a case to decipher how a HERV-H remodeled gene incorporate into the human transcriptome. Of the above, we choose a human specific *de-novo* gene ESRG that is expressed in EPI [312] Due to the low cell-to-cell variation in EPI, it was possible to investigate the gene co-expression dynamics by calculating pairwise weight correlation network analysis (WGCNA) (Figure 5.6D and 5.7C), and to observe significant pairwise ranked correlations on scaled data. This approach allowed us to identify genes whose expression is either strongly anti- or co-paired with ESRG within human blastocyst (Figure 5.6C-D and 5.7D-E). Our analysis reveals that genes whose expression is most significantly correlated and anti-correlated with ESRG mark EPI and PE respectively, suggesting that the presence or the lack of ESRG expression contributes to determine lineage identity in the human blastocyst (Figure 5.6D and 5.7D-E). Consistent with a role of ESRG in self-renewal and pluripotency, as indicated by knockdown data 48, 67, some of the genes tightly co-regulated with ESRG include NODAL, GDF3, TDGF1 and PRDM14 (Figure 5.6D-E and 5.7E), associated with regulating self-renewal in human [434, 435]. Thus, in addition

to its function in promoting pluripotency [312, 212], ESRG appears to be incorporated into, and predicted to modulate, the regulatory circuitry of self-renewal in human pluripotency (Figure 5.6D). Importantly, ESRG expression can be used as a marker of naive-like cells in a heterogenous cell population of human embryos [212, 312].

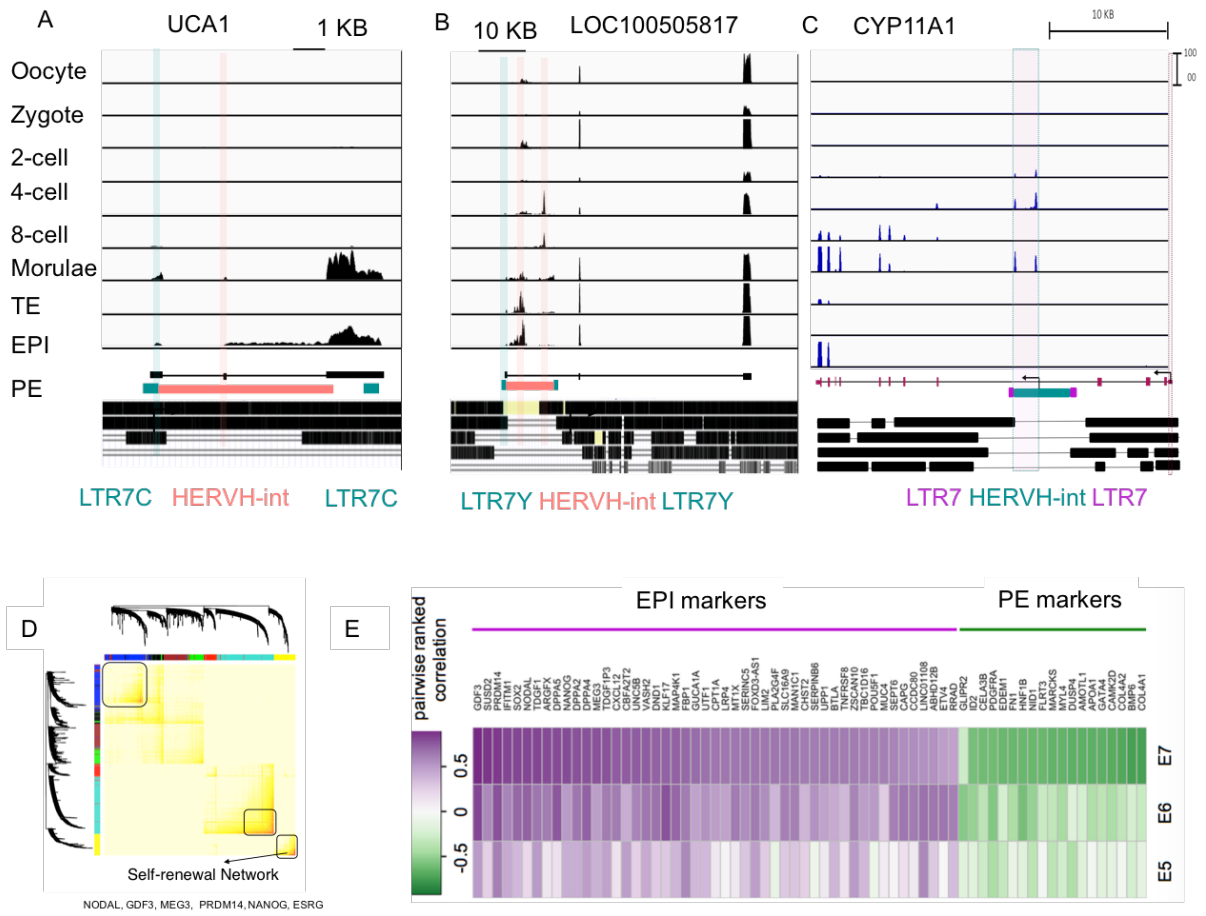


Figure 5.7:
A. Integrative Genome Visualization (IGV) of uniquely mapped reads over the specific genes harbouring full-length non LTR7 HERV-H (HERV-H driven by LTR7C and LTR7Y) inside shows the HERV-H dependent expression of various human-specific genes marking distinct stages of blastocyst. Black arrows show the annotated transcription start sites (TSSs) for shown genes viz. UCA1 and LOC100505817. Shaded areas on the plot shows exonization of HERV-H. Lowest panels show conservation status, the presence (thick line) and the absence (narrow line) of the human sequence compared to the *Chimpanzee*, *Gorilla*, Rhesus, *Callithrix* and Mouse assemblies. UCA1 gene expression marks TE cells driven by LTR7C. UCA1 gene model is also present in *Chimpanzee* and *Gorilla* assemblies.
B. The HERV-H-enforced LOC100505817 transcript is expressed in 8-cell/morula stages using unique TSS compared with rest of human embryonic lineages. LTR7B-HERV-H provides TSS and is part of the matured LOC100505817 transcript in humans only.
C. Integrative Genome Visualization (IGV) of uniquely mapped reads over the specific genes harbouring full-length human-specific LTR7 HERV-H shows the HERV-H dependent expression of CYP11A1 marking TE stage of blastocyst. Black arrows show the annotated transcription start sites (TSSs). Shaded areas on the plot shows exonization of HERV-H. Lowest panels show conservation status, the presence (thick line) and the absence (gap) of the human sequence compared to the *Chimpanzee*, *Gorilla*, Rhesus, *Callithrix* and Mouse assemblies. CYP11A1 gene expression marks TE driven by LTR7. CYP11A1 gene model is conserved across vertebrates.
D. Expression status of ESRG defines cell fate of primitive endoderm and pluripotent epiblast. Expression heat map of top hits of gene pairs harbouring strong (> 0.70) ranked correlation (purple) or (< -0.70) anti-correlation (green) with ESRG during human embryonic preimplantation development (E5-E7). The expression matrix shows the genes that are upregulated at the given stages.

5.3.2 UCA1 expression correlates with genes involved in preparing the embryo for implantation

Correlation network studies reveal that UCA1 belongs to an expression network of TE specific genes. This group of genes include STS, PGL, CGA, DLX5, the ERV-FRD envelope-derived ERVFRD-1 (also known as syncitin 2), and ERVV-1/2 (ERVV-type envelope), but also the HERV-H-rewired CYP11A1 (Figure 5.6C). While UCA1 expression has been observed in various cancers [436, 290, 437] or preeclampsia [438], when expressed in TE, its role, similarly to ERVFRD-1 might be involved with preparing the embryo for implantation. Curiously, quite a few members of this group are either *de-novo* human specific genes or their regulation and have been shifted from postimplantation stages to TE during evolution. Besides UCA1, deregulation of several genes of this group has been associated with various human diseases e. g DLX5, CYP11A1, CYP19A1 are upregulated in Preeclampsia [439].

5.4 Robust divergence of pluripotency following the split of old and new world monkeys

To further decipher the potential role of ERVs during the evolution of preimplantation embryogenesis, we concentrate on pluripotent stem cells (PSCs). As models of the pluripotent epiblast (EPI), we use induced pluripotent stem cells (iPSCs), established from human and various non-human primates (NHPs). To determine the differentially expressed genomic loci between the human and NHPs transcriptomes, we include male iPSCs from human, *Chimpanzee*, *Bonobo* [93] and our own *Gorilla* data [419]. To address the role of HERV-H-driven transcription affecting pluripotency during primate evolution, we additionally generate RNASeq data from comparable *Callithrix* [418], where HERV-H is not present [312] as a control (Figure 5.8A-D). We also extract HERV-H-governed genes defined as those differentially regulated in the knockdown cells (HERV-H-KD) compared to control human ESC-h1 [209].

Our cross-species mapping demonstrates how dramatically the expression of human EPI markers (e.g. including LEFTY1/2, NODAL) changes between human and *Callithrix* PSCs (Figure 5.10A), supporting the dramatic restructuring of the pluripotency network after the split of New World Monkeys (NWM) and Old World Monkeys (OWM). The divergence of PSCs transcriptomes is also high among OWMs (Figure 5.8B). Compared to human PSCs, we observe 2340, 375, 172 and 81 differentially expressed genes (DEGs) in *Callithrix*, *Gorilla*, *Chimpanzee* and *Bonobo* PSCs, respectively, whereas only 82 genes were shared between them (Figure 5.8B). The number of unique DEGs is also proportional to the total number of DEGs, and the degree of transcriptome diversity agrees with the predicted evolutionary path as inferred from clustering fold change values of all observed DEGs (Figure 5.8C-D). The most contrasting transcriptional pattern is observed between the pluripotent cells of NWM and OWMs, nevertheless, the higher transcriptomic divergence of PSCs among OWM is also observed (Figure 5.8B-C).

5.5 The transcriptome divergences of primate pluripotent stem cells are mainly due to HERV-H expression

As HERV-H has been implicated in modulating the regulation network of pluripotency in human [209], we determine the expression of human HERV-H loci by mapping reads from the comparative species

against the human genome, and calculating the level of relative transcription at each locus.

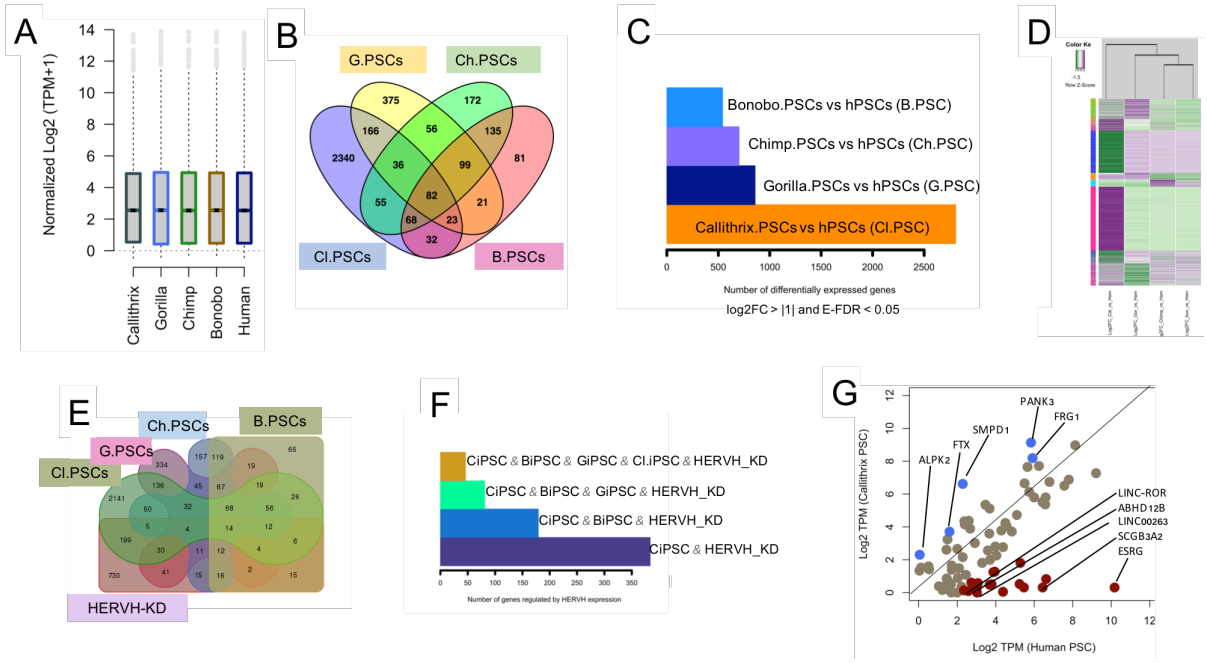


Figure 5.8:
A. Boxplots show the normalized global expression estimates of cross-species mapped RNA-seq data-sets from primate's pluripotent stem cells analysed in this study.
B. Venn diagram displays diverged transcriptomes of pluripotent stem cells during primate evolution. The numbers in the Venn diagram denote significantly differentially regulated genes (FDR < 0.05 and fold-change > 2 or FDR < 0.05 and fold-change < -2). NHPs (*Bonobo*, *Chimpanzee*, *Gorilla* and *Callithrix*) PSCs. RNAseq data was compared with that of human iPSCs considering cross-species reads mapped onto both genomes with the expression of gene was calculated on human genome. Final expression is calculate using the human model.
C. Barplot showing number of differentially expressed genes as in inferred from reads mappable to both genomes under comparison.
D. Heat map showing level of differential expression of cross mapped genes between human and non-human primate iPSCs. Any gene that was detected as differentially expressed in any of the given comparisons was included in the analysis.
E. Diverged transcriptomes of pluripotent stem cells during primate evolution due to HERV-H-mediated regulation. As in Figure S5B, but with addition of genes differentially expressed in HERV-H-KD in H1-ESCs compared with GFP-KD in H1-ESCs.
F. Barplot showing number of differentially expressed genes under the control of HERV-H regulation. Barplot as in Figure S4C showing differentially expressed genes in the annotated comparison.
G. Scatter plot displays the dynamics expression (Log2 TPM) of genes physically downstream to HERV-H genomic sites in *Callithrix* (New world Monkey; doesn't have HERV-H) and human PSCs.

Differences between the NWM and OWM are also reflected in gene loss/gain events. Remarkably, the major gain (19) and loss (29) events in regulating pluripotency between human and *Callithrix* are due to the HERV-H-governed gene expression (Figure 5.10D and 5.8G), underpinning the centrality of HERV-H to early human development. Among those genes whose expression has been toned down, we identified NR2F2, whose repression was reported to enhance iPSC reprogramming in human [440] (Figure 5.10D). Curiously, PRODH is also among the HERV-H-controlled gained genes, suggesting that PRODH is under a dual HERV-governed regulation e.g. LTR5/HERV-K and LTR7B in brain [441] (Figure 5.10D), respectively.

To root back when the co-option of HERV-H initiated, we employ the gene expression profile of the HERV-H knockdown as a surrogate of the ancestral – before HERV-H – expression profile. Calculating

all observed DEGs in any comparison including the ones by HERV-H-KD results in around 2,000 genes (FDR < 0.05) (Figure 5.10B). Hierarchical clustering applying ranked-correlation on their fold-change values reflects the evolution of primate transcriptome, and pushes human HERV-H-KD (ESC-h1) between *Gorilla* and *Callithrix* (Figure 5.10B), suggesting the domestication of HERV-H predates the human-Gorilla common ancestor. The analysis of DEGs across nonhuman primate PSCs uncovers an association between HERV-H-regulation and the gradual evolution of pluripotency (Figure 5.8E-F and 5.9).

In order to decipher the transcriptional gain and loss of existing HERV-H loci between human and NHPs, we scale the expression of orthologous loci between human-gorilla and human-chimp. The orthologous HERV-H loci [397] exhibit a species-specific pattern (Figure 5.9), indicating that HERV-H expression has dynamically changed during primate evolution. Notably, the number of expressed HERV-H loci appears to be dramatically different even between PSCs of human and apes (Figure 5.9), suggesting that the HERV-H-driven regulatory network of pluripotency has reached its human shape quite recently. We observe heavy loss of HERV-H expression in human PSCs (hPSCs) compared with non-hPSCs, whereas, few orthologous loci gained expression in hPSCs (Figure 5.9A-B). Nevertheless, the overall expression of HERV-H neighbour genes in hPSCs is significantly higher compared to their nonhuman primate counterparts (Figure 5.10E and 5.9G). Furthermore, HERV-H affected neighbour genes are expressed at higher level or exclusively in hPSCs compared with non-hPSCs (Figure 5.10E), suggesting that robust HERV-H control over neighbour genes occurred quite recently, in conjunction with the evolution of human pluripotency. Curiously, upon comparing the orthologous transposable element (TrE) loci between primate species, we notice, by contrast to HERV-H, a heavy loss of overall TrE expression in human pluripotent state (Figure 5.10G and 5.9C-D).

5.5.1 Fine-tuning the pluripotent stem cell function between *Chimpanzee* and human

A number of HERV-K loci are also expressed differentially in primates (Figure 5.10C right panel). Unlike HERV-H, several orthologous HERV-K loci are also polymorphic in the human population [442], raising the possibility that HERV-K might still function as an active retrovirus, and its domestication process is not complete.

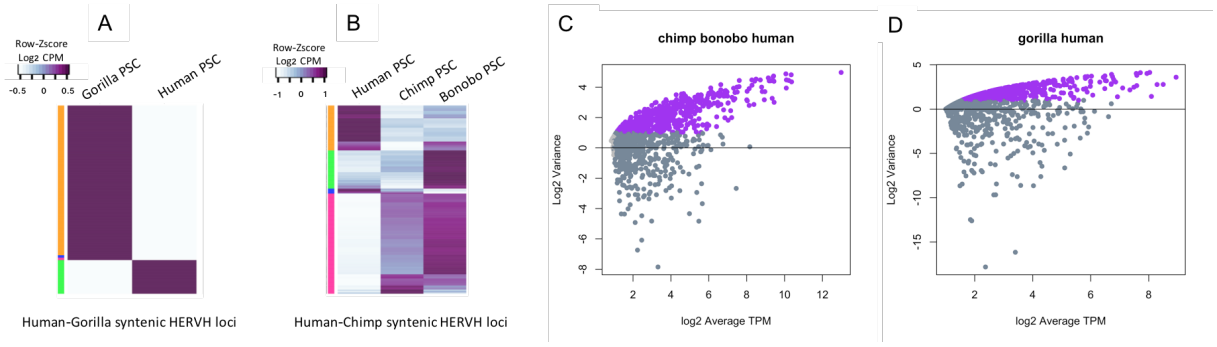


Figure 5.9:
A. Heatmap displays the loss and gain of orthologous HERV-H loci between human and *Gorilla* PSCs (see method) **B**. Heatmap displays the loss and gain of orthologous HERV-H loci between *Bonobo*, *Chimpanzee* and human PSCs. **C**. Differential behaviour of TrEs shown upon plotting Variance against their expression. Plots display loss of orthologous TE expression in human PSCs compared with non-human PSCs (*Chimpanzee* and *Gorilla* in this case) **D**. Differential behaviour of TrEs shown upon plotting Variance against their expression. Plots display loss of orthologous TE expression in human PSCs compared with non-human PSCs (*Chimpanzee* and *Gorilla* in this case)

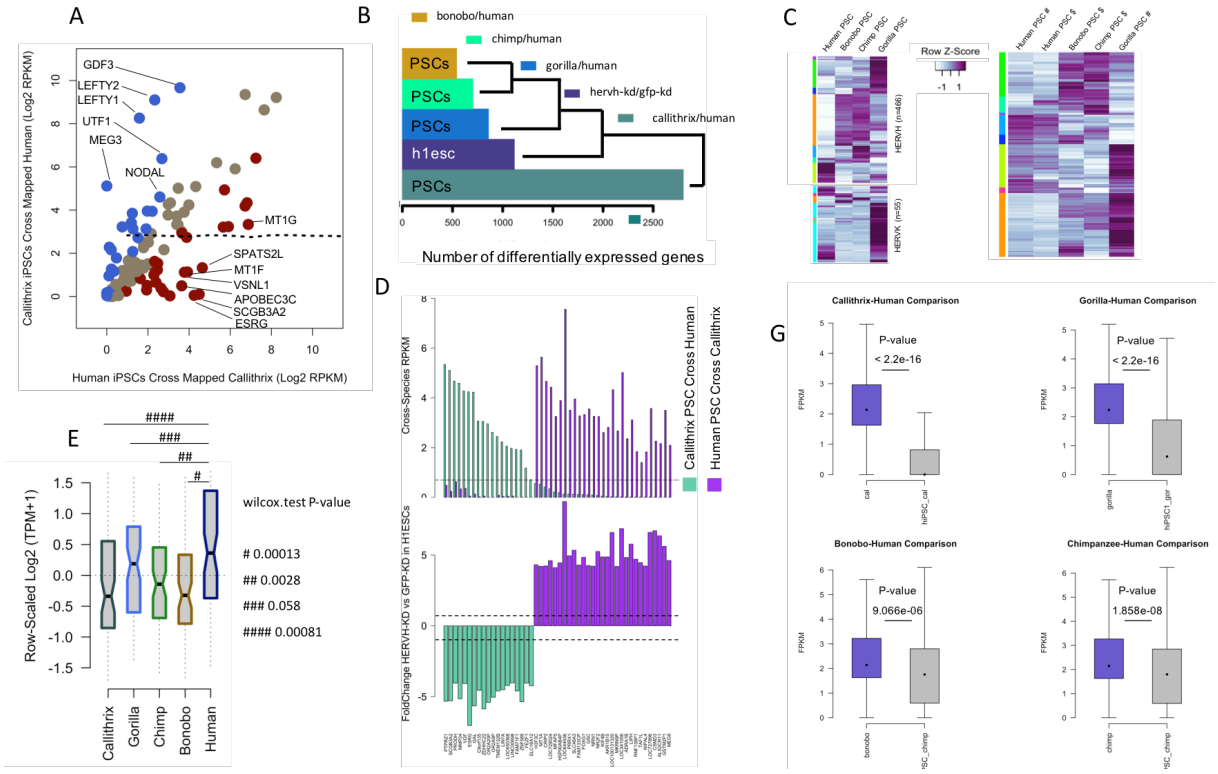


Figure 5.10:

A. Scatterplot shows the differential expression of human pluripotent EPI genes (n=308, AUC cut off $>$ 80%) in human and *Callithrix* PSCs obtained as RPKM on human genome from reads mapped on both genomes. Blue dots represent the cross-species genes that has lost its expression in human PSCs and darkred ones are those that gained its expression in human PSCs. Note: Since we were analyzing lesser set of genes so we also considered those that contained zero mappable reads in either of analyzed species e. g ESRG

B. Barplots combined with dendrograms display the comparison of genes in NHPs PSCs, hPSCs and human ESCs controlled by HERV-H transcription. Barplots show the number of significant DEGs (FDR$<$0.01 and fold change$>$2 or $<$ -2) of gene lists obtained from *Callithrix* PSCs (n=1) vs human iPSCs (HPSCs) (n=4) (CaPSC), *Gorilla* PSCs (n=2) vs HPSCs (n=4) (GPSC), *Chimpanzee* iPSCs (n=4) vs HPSCs (n=4) (CPSC), *Bonobo* iPSCs (n=4) vs HPSCs (n=4) (BPSCs), HERV-H-KD vs GFP-KD (n=2) in H1-hESCs (HERV-H-KD). In case of two replicates, we had selected only those genes which were differentially regulated in both replicates in a similar fashion. Sorted according to HERV-H-KD. Dendrogram is calculated by "ranked correlation" and Euclidian distance method.

C. Heat map of HERV-H and HERV-K dynamic expression (Log FPKM) in primates. Z-scores are calculated by using reads mapped on the human genome from Human, *Bonobo*, *Chimpanzee* and *Gorilla* iPSCs. The numbers represent expressed genomic loci. Note that more genomic copies of HERV-K is active in *Gorilla* compared with the rest of Primate PSCs.

D. Notched boxplot represent the distribution of gene expression (at log2 scale) of LTR7-HERV-H neighbour (upto 10 KB downstream) from cross-species mappable reads in analyzed primate PSCs. Hash (#) bars show the pairwise calculation of p-value between human and NHP PSCs.

E. Boxplots show the pairwise distribution of entire TrEs expression (log2 FPKM) from reads mappable to both species genome calculated against hg19 version of human genome. We only considered those TE locus for this analysis that was expressed (Log2 FPKM $>$ 1). P-value was calculated by wilcoxon test.

F. Combined barplots show the gain and loss of genes expression between human and *Callithrix* (no HERV-H is present) PSCs due to HERV-H regulation. Upper and lower panels show cross-species expression values (RPKM) and fold-change values in KD-HERV-H vs KD-GFP in H1-ESCs (control), respectively. Green bars in upper panel are genes expressed significantly in human (see methods), but not in *Callithrix* PSCs (reads are mapped on human genome), lower panel shows that same set of genes are heavily down-regulated when HERV-H is depleted using by RNAi (HERV-H-KD vs GFP-KD in H1-ESCs). The opposite scenario is shown in purple. (FPKM $<$ 1 was considered as loss).

G. Heatmap displaying the expression (Log2 FPKM) of all Refseq ZNF genes (FPKM $>$ 1) from cross-species uniquely mapped reads on human hg19. Note: ZNFs are fast evolving and sequence variable genes within family so its challenging to be mapped even within species, differential expression on heatmap could also mean the alteration of sequence of ZNFs during evolution of genomes.

Nevertheless, the certain HERV-K-driven transcripts might have been co-opted in human and regulates pluripotency. For example, *rec*, implicated to play a domesticated role in a viral restriction pathway is capable of upregulating the interferon-induced viral restriction factor IFITM1 [276] (Figure 5.11), the top-ranked EPI marker [298]. Similarly to the innate immune response, fine-tuning of the oxidative phosphorylation in PSCs has occurred between *Chimpanzee*-human (Figure 5.11).

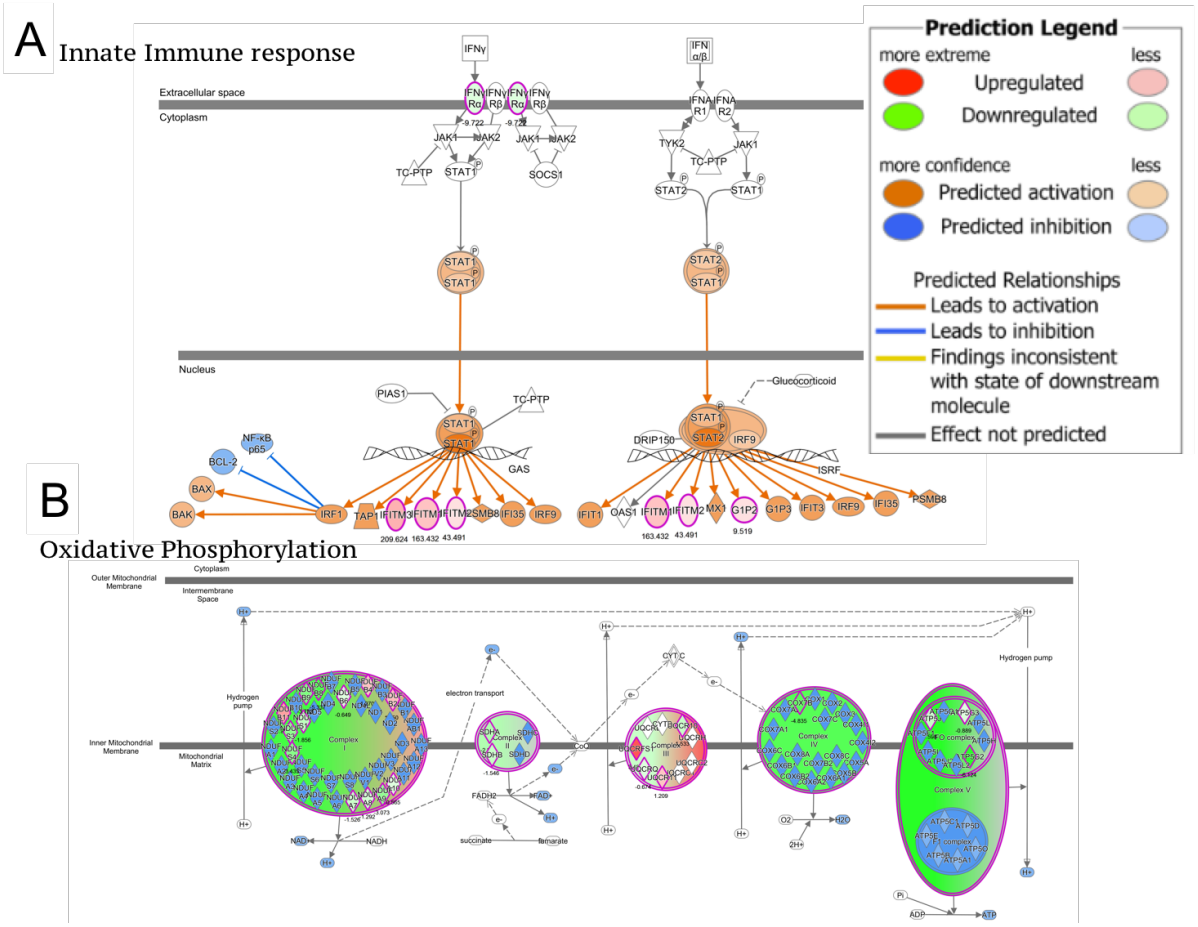


Figure 5.11:
Schematic shows the ingenuity pathway analysis of differentially expressed genes between human and *Chimpanzee* iPSCs.
A. innate immune response and interferon pathway. Red circles upregulated, while green circles show down-regulated genes in human.
B. Oxidative phosphorylation.

Chapter 6

Transcriptional regulation of HERV-H

HERVs are silenced during early embryogenesis by a combination of various repressing pathways, including the DNA methylation, however some HERVs are demethylated and expressed. We sought to understand both the underlying regulatory mechanisms of ERV expression and the downstream consequences, what might be called 'cross-talk'. To decipher the 'cross-talk', we performed high throughput data analyses to survey direct and indirect associations of various HERV families with known host encoded regulators. We also applied in silico approaches to perform cross-species data analysis in order to resolve the evolution of cross-talk. The analysis revealed specific transcripts of different classes of HERVs expression does not associate with the level of methylation, suggests the existence of alternate mechanism of HERVs regulation by host factors. Interestingly, we see the positive selection of a subset of HERV expression, since the genes which are physically close to HERVs, are regulated by HERVs and mark the specific stages of development [263]. Particularly, we investigated the cross talk between primate specific ERVs transcripts and host-factors, involved in maintaining the identity of primate embryonic stem cells. Our analyses reveal differential cross regulation between subsets of HERVs and host factors in the inner cell mass (ICM) and human embryonic stem cells. We propose that certain host encoded factors, including those that carry Krueppel boxes, have interplay with ERVs during primate stem-cell evolution and re-wired the human specific transcriptional network.

While many genes are involved in the maintenance of pluripotency in iPSCs and ESCs, evidence from mouse and human suggests that expression owing to binding of transcription factors to TrEs plays a role [210]. Indeed, TrEs have re-wired a different set of genes into the regulatory network of ESCs in humans and mice [210]. While in mice the ERV family (M)ERVL is implicated in regulating the transitional state between toti and pluripotency [311], less is known about the role of human ERVs (HERVs). Reasoning that primate-specific retrotransposons might be involved in the specification of pluripotency and the naive state in particular, we surveyed RNAseq data of human pluripotent stem cells (hPSCs) to identify highly transcribed families of mobile elements.

6.1 HERV-H is the most enriched TE in hPSCs

Compared to fibroblasts, HERV-Hs were the most upregulated HERV in the hPSC transcriptomes (Figure 6.1A), consistent with previous reports that HERV-H is a hESC marker [294]. A higher level of transcription was associated with elements containing consensus LTR7 rather than the diverged variants

LTR7C or 7Y. HERV-K, W and L (Figure 6.1A) were the next most abundantly expressed, respectively, confirmed via qRT-PCR using primer-set specific for conserved HERV-H loci [212]. Analysis of uniquely aligned reads reveals that around 300 out of the 1225 full-length HERV-H genomic copies are transcribed in hPSCs (Figure 6.1D).

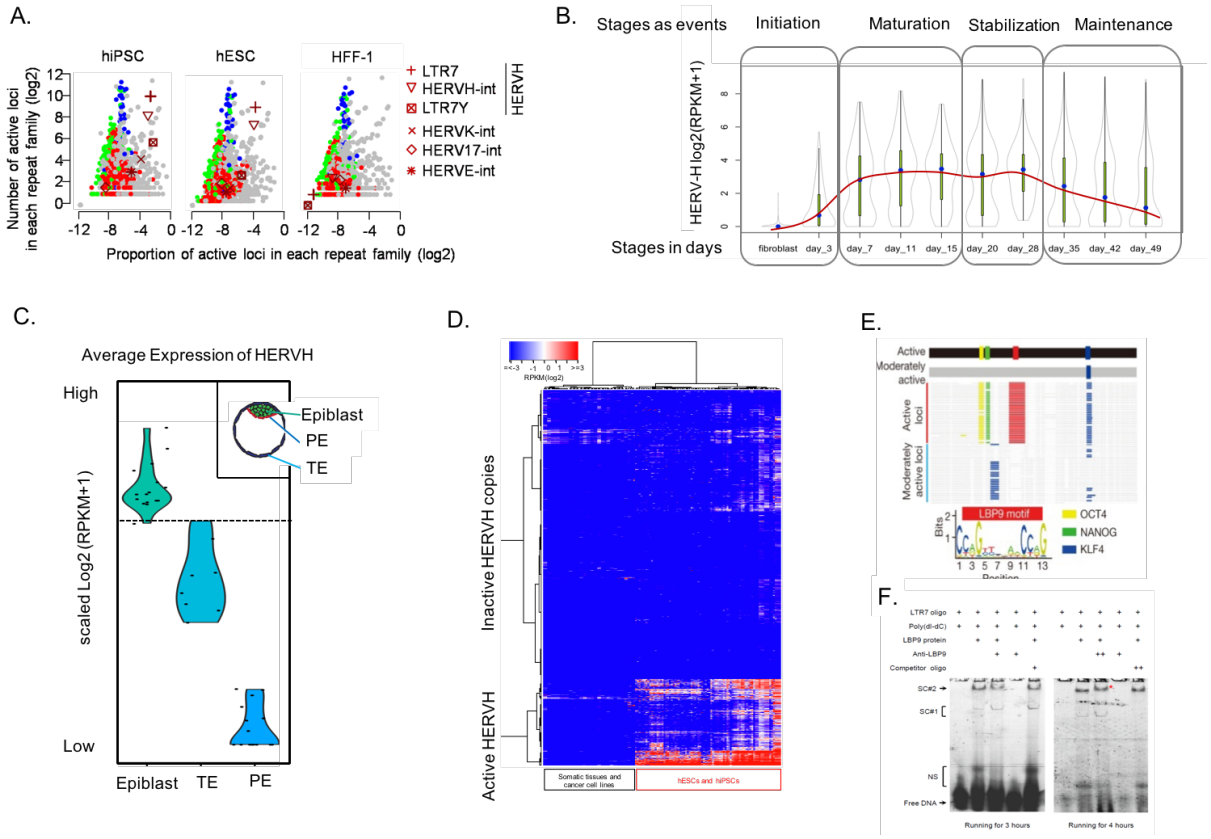


Figure 6.1:
A. Expression of various transposable elements in hiPSCs, hESC-H1 and human fibroblast HFF-1. Colours indicate different classes of transposable elements (red, LTR; green, long interspersed nuclear elements (LINE); blue, short interspersed nuclear elements (SINE); grey, other repeat elements)
B. Expression patterns of the HERV-H family during reprogramming. Data are shown as HERV-H reads per kilobase of exon per million mapped reads (RPKM) in during reprogramming on day 0 (Fibroblast), dat 3 (Initiation of reprogramming), day 7-15 (maturation of reprograming), day 20-28 (stabilisation of reprogrammed cells) and day 28-49 (passaging or maintenance of reprogrammed cells)
C. Violin plot show the distribution of normalised expression (FPKM) of activated HERV-H loci in three distinct layers of human blastocyst viz. Epiblast, primitive endoderm (PE) and trophectoderm (TE). Every dot represents a single cell
D. Expression profile of HERV-H in 43 normal somatic tissues, 8 cancer cell lines and 55 hESC (H1, H6 and H9) and 26 hiPSC samples, including our hiPSC [443] line. The rows represent the transcription from 1,225 full-length HERV-H loci.
E. Conserved binding sites of OCT4, NANOG, LBP9 and KLF4 are shown in active LTR7s versus moderately active versions of LTR7C/Y. The Jaspar consensus sequence of the LBP9 motif is shown.
F. EMSA confirms the binding of LBP9 to LTR7 sequence *in vitro*.

This last observation prompted us to examine the transcriptome-wide HERV-H activity during reprogramming. RNA-seq analysis showed that HERV-H expression picks up the after 3 days of OSKM-mediated reprogramming and displays stabilised expression throughout the process (Figure 6.1B). However, once the somatic cells are reprogrammed to iPSCs, they are subjected to multiple passaging in order to maintain them in petri-dish that slightly looses HERV-H expression (Figure 6.1B). To confirm if

HERV-H is marker of global human pluripotency, we checked it's expression in blastocyst cells (*niche of natural pluripotent state*) using single-cell transcriptomics [297]. HERV-H displayed highest expression in pluripotent epiblast cells compared with TE and PE cells which for extra-embryonic lineages (Figure 6.1C). To address how specific HERV-H transcription is to pluripotent cell types we compared RNAseq datasets of human hPSCs and several differentiated cells and multiple tissues for high levels of HERV-H expression. RNASeq data for 43 normal somatic, 8 cancer cell lines and tissues as well 55 hESCs (H1, H6 and H9), 26 hiPSC samples, including our hiPCS line were analysed (Figure 6.1D). In agreement with our hiPSC data, HERV-H transcription was the highest in hESCs. The vast majority of the transcribed loci is identical between hiPSCs and hESCs (Figure 6.1D). In contrast to hPSCs, transcription levels HERV-Hs are significantly lower in differentiated and cancer cell lines (Figure 6.1D). Our results suggest that HERV-H expression marks pluripotent cell populations of all kind (natural, artificial and formative) in the midst of heterogeneous cells.

6.2 Co-evolution of hESC-specific TFs and HERV-H

In addition to being a marker for pluripotency, might HERV-H also affect pluripotency, and if so how? HERVs may promote transcription by providing transcription factor (TF) binding sites [444]. We interrogated hESCs ChIP-Seq data [445] for DNA binding proteins commonly reported in pluripotency. This identified the core pluripotent factors NANOG and OCT4 (Figure 6.1E). A candidate KLF4 binding site was also identified within the LTR of HERV-H (Figure 6.1E).

We additionally asked which TF motifs are significantly enriched across four in silico independent tests. Only one, LTR-binding protein 9 (LBP9), also known as TFCP2L1 in mice, was significant across all analyses (Figure 6.1E). To confirm specific DNA binding of LBP9 to LTR7, we performed EMSA and ChIP-qPCR (not shown). Binding of LBP9 to LTR7 was significantly enriched relative to controls in both assays (Figure 6.1F). LBP9 specific binding was detected upstream of NANOG (Figure 6.1E). These data-sets reveal that embryonic stem cell specific transcription factors such as OCT4, NANOG KLF4 and LBP9 drive the expression of HERV-H elements in hPSCs. Thus, in contrast to mice, the key pluripotent TFs, such as OCT4, NANOG, KLF4, LBP9 are clustered in humans with the primate-specific HERV-H (Figure 6.1E).

As we already have shown (on the basis of theoretical and experimental evidences) that the genomic expansion of HERV-H over time was likely facilitated by the presence of core key pluripotent TF binding sites in the LTR7. The pluripotency factors NANOG, OCT4, KLF4 and LBP9/TFCP2L1 bind to LTR7 of HERV-H and drive transcription of HERV-H-derived transcripts (Figure 6.1E) [263, 211]. Curiously, upon the depletion of either the level of LBP9 or HERV-H in H9-ESCs by RNAi mechanisms, their transcriptomes display similar pattern (Figure 6.2D) in sharp contrast with that of H9-ESCs. This suggests the functional co-evolution of a transcription factor LBP9 and HERV-H to re-wire the primate-specific transcriptional networks.

In addition to LBP9, we have investigated the binding characteristics of the transcription factor, KLF4 found to be driving HERV-H expression (Figure 6.2A). Data mining of ChIP-seq and RNA-seq data-sets in hiPSCs lines [208] reveal the significant pattern of KLF4 occupancy over the subsets of HERV-H loci. Firstly, HERV-H loci was partitioned into two subsets, (i) active HERV-H loci which has FPKM > 2 (n=245) and (ii) inactive HERV-H loci which did not show any expression (n=782).

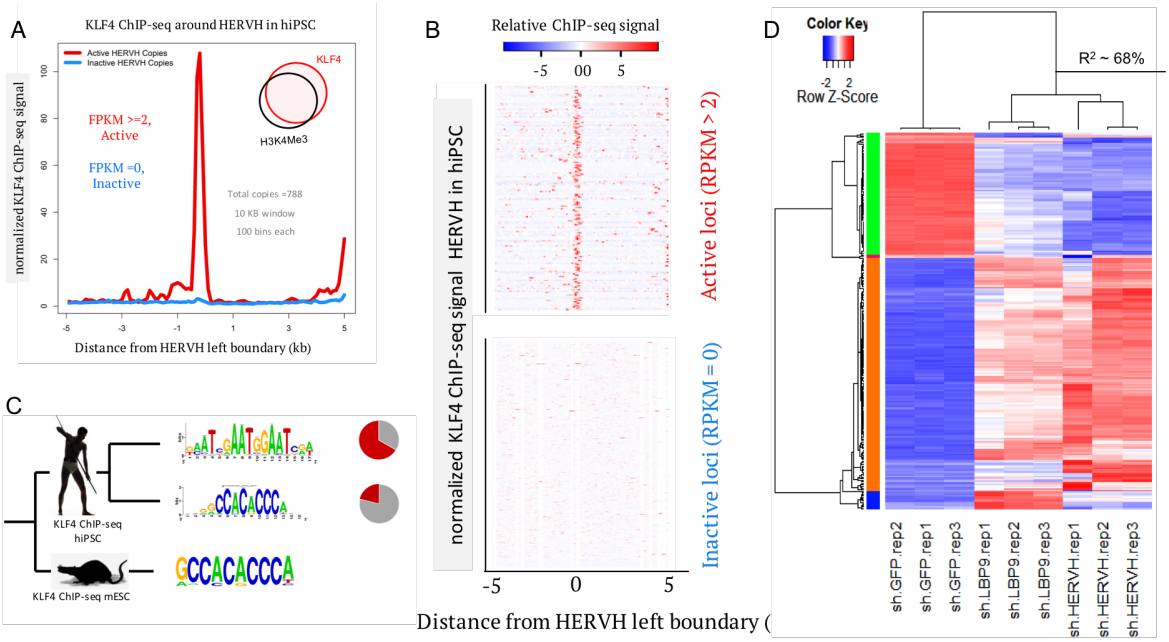


Figure 6.2:
A. Lineplot shows the ChIP-seq data analysis of KLF4 in H1ESC from [208]. Lines represent the average counts of KLF4 ChIP-seq signal around full-length HERV-H left boundary (5 KB on both sides from left boundary of HERV-H). ChIP-seq signal was plotted as mean of tags counted in 100 bins of 100 bps around each locus of Active (FPKM $>$ 2) and inactive (FPKM =0) HERV-H in the H1ESCs. Red line displays the signal over active loci whereas blue line is on inactive ones. Most of the KLF4 peaks overlap with H3K4Me3 peaks (Histone modification that marks promoter activity) Note: KLF4 specifically occupies the TSS of active HERV-H loci.
B. Similar to previous figure, only the signals are visualized as heatmap over individual HERV-H locus. Note: Every active locus harbours KLF4 ChIP-seq signal whereas inactive ones do not.
C. Motif discovery from KLF4 ChIP-seq peaks in mouse and human embryonic stem cells. Pie chart shows the fraction of ChIP-seq peaks (shaded in red) represent human sequence motif adjacent to it. Note: KLF4 binds one motif in mouse whereas, two in human and human-specific motif is around two-third of total motifs.
D. Heat map showing genome-wide gene expression in hESC-H9 after knockdown of GFP (shGFP), LBP9 (shLBP9) and HERV-H (shHERV-H). h9ESC's transcriptome is highly similar as the knockdown effect of LBP9 and HERV-H are highly similar in sharp contrast with conventional H9-ESCs ($\rho \sim 0.70$ from Spearman's correlation).

KLF4 peaks exclusively over transcriptional start sites of active HERV-H sequences. Strikingly, we did not observe any KLF4 binding event on inactive HERV-H loci (Figure 6.2B). Curiously, we have noticed that KLF4 binding sites ($\sim$2/3 of total peaks) have a human-specific binding motif in hiPSCs. This motif is different from the one ($\sim$1/3 of total peaks) annotated in the JASPAR database, and corresponds to the mouse motif, identified by the mouse ChIP-seq data analysis (Figure 6.2C).

6.3 KLF4 binding on HERV-H, an escape from repression

Curiously, the host can also activate HERV-H expression, suggesting that the regulation of HERV-H is more complex than simple repressing transposable element activity. Indeed, KLF4 was implicated in activating HERV-H in pluripotent stem cells by replacing KAP1 binding [208]. To test the substitution model, we analyse KLF4 ChIPseq and RNAseq data in human pluripotent stem cells hiPSCs [208]. We generate the subsets of HERV-H as active (log2 FPKM > 1) and inactive (FPKM =0) from the RNAseq datasets. We then count the average ChIPseq signal in a 5Kb window divided into 100 bins on the left boundary of each subset of HERV-H loci. We observe that KLF4 is exclusively enriched on active vs

inactive HERV-H, whereas, KAP1 was enriched on both active and inactive HERV-H loci (Figure 6.2A and 6.4A). The KLF4 peaks are detected around 500-600 bps upstream of KAP1 occupied site (Figure 6.2B and 6.4A), suggesting that an active HERV-H locus might be simultaneously occupied by both KLF4 and KAP1. We speculate that in absence of KLF4 occupancy, KAP1 represses HERV-H transcription in hPSCs. Since each active locus is simultaneously occupied by KLF4 and KAP1, so former one does not seem to be removing KAP1 but might be providing alternate transcription to HERV-H.

To test this hypothesis, we fetched CAGE peaks from FANTOM5 phase 2 consortium and mapped them on full-length HERV-H loci. We observe three discrete peaks over full-length HERV-H loci, indicating multiple promoter sites (Figure 6.4E-F). By mapping raw read counts to the left boundary (10 Kb window) of full-length HERV-H we confirm the existence of alternative HERV-H transcripts (Figure 6.4G). TSS-1 corresponds to the position of KLF4 binding on HERV-H, while the transcription factors driving messages from TSS-2/3 are yet to be identified. The alternative transcripts follow a specific developmental pattern (e.g. early, EGA and blastocyst) (Figure 6.4G). We find that from two HERV-H remodelled genes (e.g. ABHD12B and SCGB3A2), distinct chimeric transcripts are produced during early embryogenesis, suggesting that the stage-specific transcripts might modulate their function (Figure 6.4H). We conclude the interplay of host encoded transcription factors (including KLF4) over HERV-H loci regulates the expression of alternative transcripts, from different transcription start sites (TSSs). Did the gain of a KLF4 site on HERV-H occurred by chance? To address this question, we mapped the two types (e.g. murine and human) of KLF4 sites on HERV-H TSS-1. Our analysis reveals that HERV-H selectively recruited the human-specific KLF4 sites, suggesting that the evolution of HERV-H and KLF4 functions might be linked (Figure 6.2C).

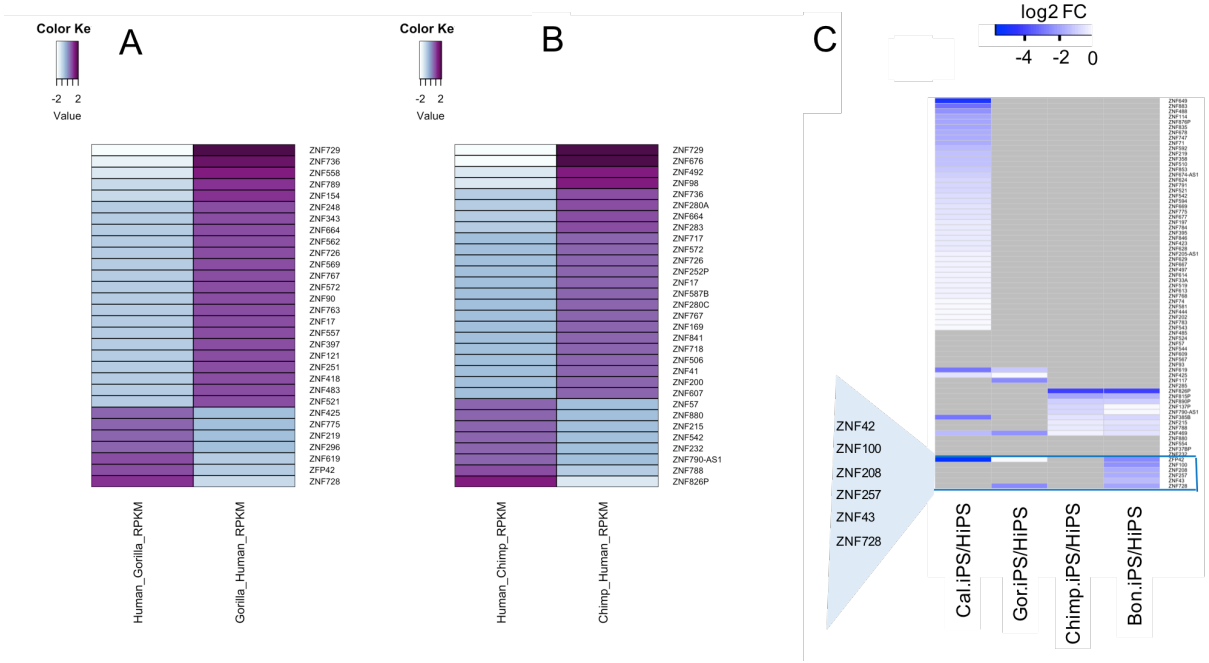


Figure 6.3:
Heatmap shows ZFPs exhibiting loss and gain of gene expression at each split in comparison with human iPSCs. **A**. human vs *Gorilla*. **B**. human vs *Chimpanzee* **C**. Heatmap showing the loss of ZFP gene expression in pairwise comparison between NHPs and hiPSCs. Blue color codes for downregulated ZFPs (False Discovery Rate < 0.05) and grey colors indicate the non-significant regulation (False Discovery Rate > 0.05). Star marked KZFPs (one in each heatmap) are shown in further sections to be binding on HERV-H loci and dynamically expressed during human embryogenesis and reprogramming.

6.4 KAP1 does not confer absolute repression of HERV-H

The KAP1/TRIM28 repressor complex, involved in regulating a plethora of transposable elements, including ERVs, is tethered to specific genomic loci by various KRAB-ZNF transcription factors (KZFPs) [183, 184, 102, 366]. Activation or inactivation of HERV-H is vaguely understood as TFCP2L1 [212] and KLF4 [208] are known host factors to drive its expression in pluripotent states, whereas KAP1 is involved in conditional repression of HERV-H [208, 366]. Nevertheless, KRAB-ZNF (KZFPs) were demonstrated to be targeting a vast variety of TrEs in the human genome via recruiting the transcriptional regulator KAP1 and hetero-chromatin forming subunits [103, 186, 446, 183, 447, 448, 102, 166, 449]. Consequently, the depletion in the level of KAP1 in human pluripotent cells leads to pervasive spectrum of TE activation [184, 183, 133]. Intriguingly, being the master regulator of pluripotency, the regulation of HERV-H was never studied in the previous context.

Our cross-species PSC transcriptome study supports the view that each primate has a specific set of expressed (K)ZFPs [103] (Figures 6.3 and 5.10C) that has been evolved to perform a role in controlling transposable elements. As HERV-H has been inactivated around 10 MYA as an ERV, we wondered if its expression is still regulated by KZFP/KAP1 repression. To enlist the KZFPs targeting HERV-H, we examined a vast array of ChIPexo datasets [103] and intersected their unique peaks with individual HERV-H loci. Out of 210 we found 6 KZFPs (e.g. ZNF534, ZNF90, ZNF75D, ZNF257, ZFP57 and ZFP69B) that recognized at least 50 HERV-H loci (Figure 6.4B). Upon comparing 1,200 full-length HERV-H loci versus 4,500 truncated ones, we observed that KZFPs were mostly recruited on the most probably active full-length loci (Figure 6.5B). The six KZFPs bind HERV-H simultaneously or the various combinations of HERV-H-specific KZFPs recognize different subsets of HERV-Hs. Except ZNF75D, the remaining five KFZPs have two footprints on HERV-H (Figure 6.5C). The ZNF534, ZNF90 and ZFP69B footprints are in close proximity of each other, and this trio targets either gag or pol region of HERV-H (Figure 6.4E). Our data mining on human preimplantation transcriptome data [297] indicates that the six KFZPs are expressed at different stages of development (Figure 6.4C). Specifically, the expression of ZNF90 rises with EGA, while and ZNF534 at morula stage along-with ZNF257 and ZFP57, albeit at the lower level (Figure 6.4C). Accordingly, KZFPs also have a specific expression pattern during the reprogramming process (Figure 6.4D).

The ICM/EPI-specific ZNF75D is expressed all the way through the reprogramming process (Figure 6.4C). ZNF90, ZNF534 and ZFP57 mark freshly stabilised cells, but the expression of the later two go down with passaging (Figure 6.4C). Finally, we find that the six KFZPs bind preferentially to full-length than truncated HERV-H loci (Figure 6.5B). Three of these six KZNFs *viz.* ZNF90, ZFP57 and ZNF257 showed species-specific higher expression (Figure 6.3). Additionally, ZNF534 and ZNF90 showed numerous common binding sites that also overlaps with the KAP1 suggesting its recruitment as co-factor (Figure 6.5). As KFZPs are implicated in targeting the KAP1 repression complex [102, 184, 183, 103], we determine the potential overlap between KFZP and KAP1 occupancy over HERV-H loci. Quite surprisingly, KAP1 binding was enriched on both active and inactive HERV-H loci (Figure 6.4A), suggesting that binding of KAP1 might not confer absolute repression of HERV-H. Indeed, we notice the sharp peak of KAP1 over HERV-H but still the expression of HERV-H is not significantly affected in KAP1 depleted hESC-h1 (Figure 6.4B). Furthermore, the substantial fraction of KZFPs peaks does not overlap with KAP1 peaks over HERV-H loci (Figure 6.5A), indicating that these KZFPs do not recruit the repression complex of KAP1 (Figure 6.5A), and might be involving the other processes

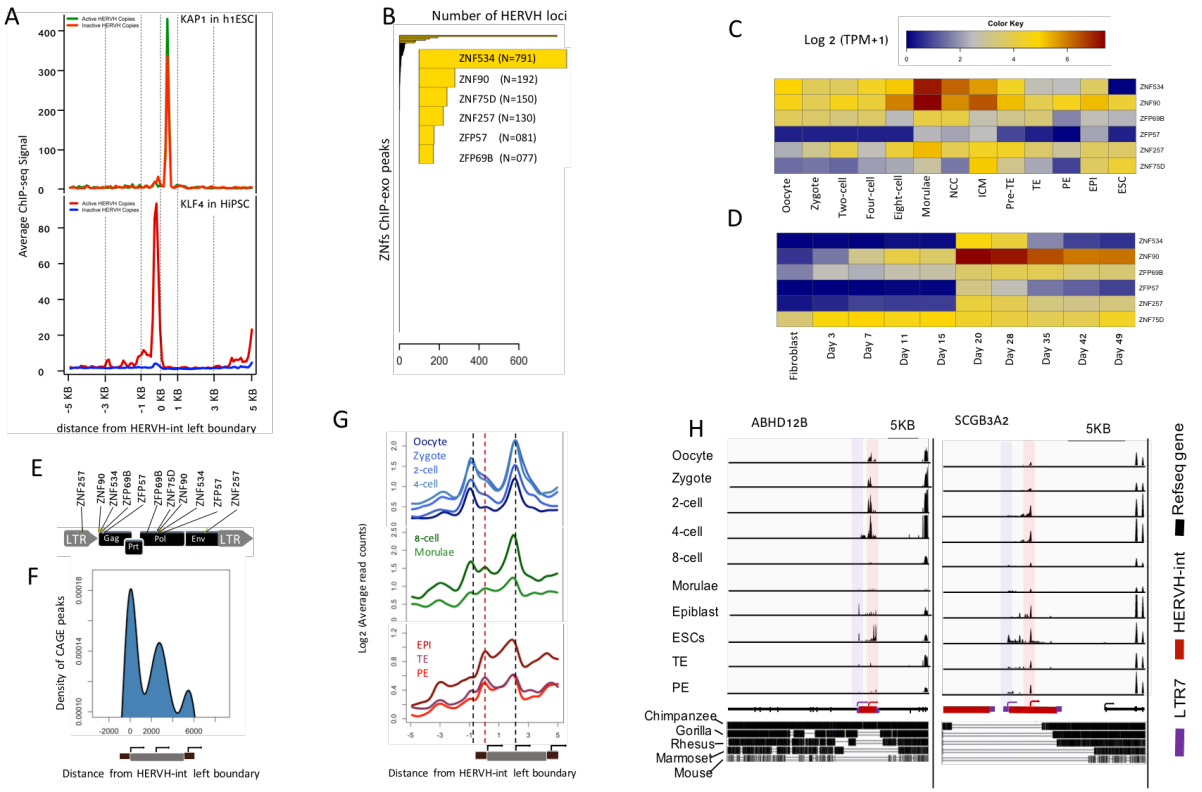


Figure 6.4:

A. Lineplot shows the ChIP-seq data analysis of KAP1 and KLF4 in H1ESC ref and HiPSCs ref respectively around full-length HERV-H left boundary (5 KB on both sides from left boundary of HERV-H). ChIP-seq signal was plotted as mean of tags counted in 100 bins of 100 bps around each locus of Active (FPKM $>$ 2) and inactive (FPKM =0) HERV-H in their respective cell lines.

B. Boxplot shows the distribution of HERV-H expression in H1ESCs in three different conditions, H1ESCs transfected with empty vector, non-specific construct and KAP1-KD construct. Note that HERV-H expression doesn't change significantly upon depleting KAP1 in H1ESCs. RNA-seq data used for this analysis was published [184]

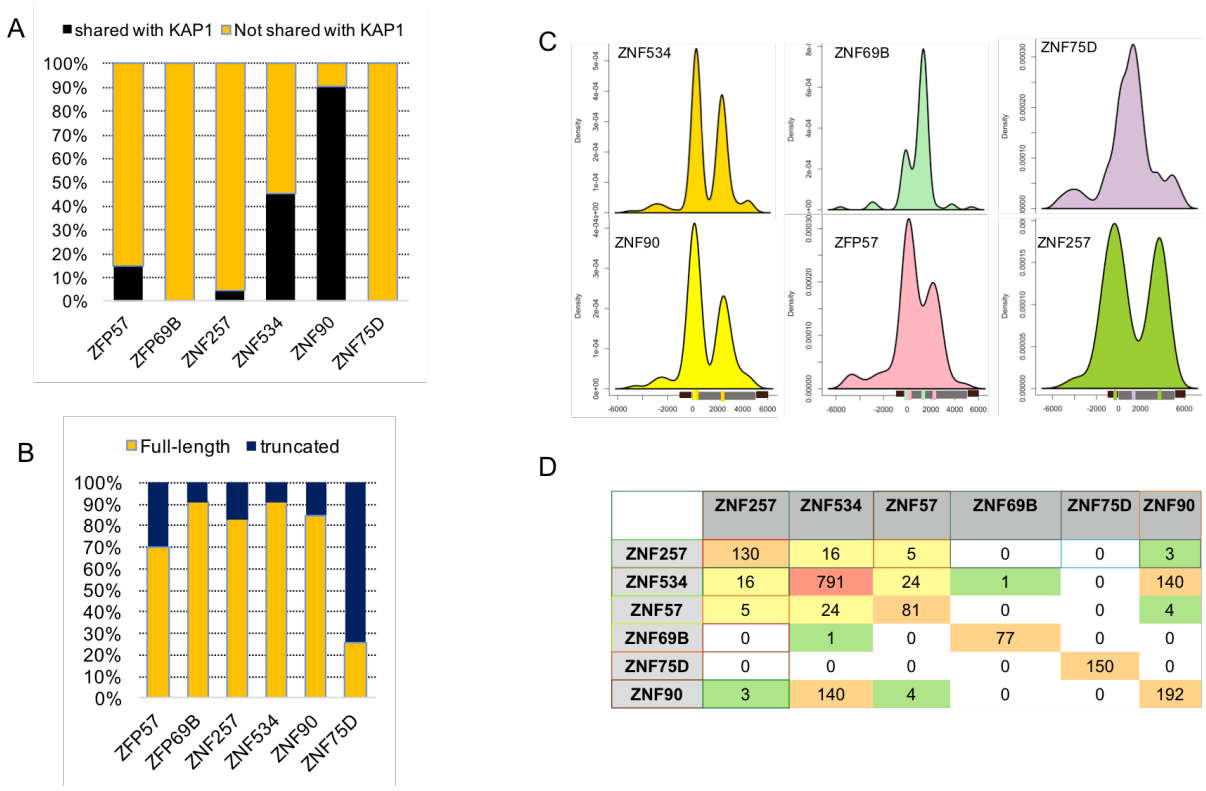
C. Barplot displaying number of significant ChIP-exo peaks (FDR $<$ 0.05) of 220 KRAB-ZNFs ref over all HERV-H loci annotated in hg19 version of human genome. Yellow bars represent ZNFs that at least occupies 1% of total HERV-H loci (e.g. 50 out of total 5000 loci). black bars have fewer peaks ($<$ 30 peaks over total HERV-H loci) so ignored for further analysis.

D-E. Heatmap showing the raw expression (log2 TPM) dynamics of Six ZNFs observed in Fig 3b-c during early embryogenesis and reprogramming. blue denotes no expression(0-2), gold as optimum (3-5) expression and darkred is denoted as higher expression($>$6).

F. Schematic illustrates the 11 binding sites of observed six ZNFs on full-length HERV-H locus. For detailed analysis of density of individual peaks over HERV-H loci (see. Supple. figure S8).

G. Density plot of Cap analysis gene expression (CAGE) peaks phase1 and 2 combined (FANTOM 5 project) over HERV-H loci displaying multiple TSS within full-length HERV-H. HERV-H schematic and TSS (black arrows) is shown beneath the density plot. **H**. Line plots displaying the comparison of absolute RNA-seq read coverage dynamics over HERV-H loci between distinct stages of early embryogenesis. Plot represent the average profiling of raw reads that were counted in 100 bins of 100 bps around each locus of full-length HERV-H. Plot was generated using smooth. spline function in R using spar=45. Dotted lines on plot shows the site over HERV-H that is enriched for RNA-seq reads at different stages of development.

I. IGV of uniquely mapped reads over the chosen genes (that has been shown to forming crimeric transcripts in our previous published work ref) and closest full-length HERV-H loci indicates the usage of multiple TSS arising from HERV-H at distinct stages. Both genes loose their annotated TSS and proximal exons to form HERV-H chimera. The chimeric ABHD12B transcript is expressed from zygote to EPI, but expression pauses in 8-cell/morula, exons of ABHD12B upstream of HERV-H/LTR7 are skipped. While ABHD12B HERV-H appears to be intact in *Chimpanzee*, it has several deletions compared to the human version (not shown). SCGB3A2, implicated in pluripotency, exhibit partially overlapping expression patterns, usage of distinct human-specific HERV-H TSS and loss of annotated TSS and proximate exons.



	ZNF257	ZNF534	ZNF57	ZNF69B	ZNF75D	ZNF90
ZNF257	130	16	5	0	0	3
ZNF534	16	791	24	1	0	140
ZNF57	5	24	81	0	0	4
ZNF69B	0	1	0	77	0	0
ZNF75D	0	0	0	0	150	0
ZNF90	3	140	4	0	0	192

Figure 6.5:
A. Stacked barplot demonstrates the percentage of full-length HERV-H locus bound by shown six KRAB-ZNFs and KAP1 in H1ESCs (black). Note: this figure is an extension to Fig 3B.
B. Stacked barplot display the fraction of ZNF peaks overlapping with KAP1 peaks over HERV-H loci. Note: KAP1 ChIP-exo was performed in H1ESCs.
C. Density plots illustrating the genome-wide relative densities of above mentioned six KRAB-ZNFs binding around the full-length HERV-H left boundary. Centre of the peaks are drawn on full-length HERV-H (LTR7 in black and HERV-H-int in grey beneath the plots) schematically to show the binding sites of mentioned KRAB-ZNFs.
D. Table contains the pairwise comparison/number of shared peaks of mentioned KRAB-ZNFs over HERV-H loci. White color of boxes show no overlapping peaks, green, yellow and orange denotes least, medium and maximum number of shared peaks respectively between any given pair of KRAB-ZNFs.

Chapter 7

HERV-H transcription defines naive-like stem cells

As opposed to primed or epiblast embryonic stem cells, naive embryonic stem cells (ESCs) are the true unprimed ESCs. While derivable in mice, they have proven difficult to isolate from human blastocysts and their existence is questioned [450]. To maintain the pluripotency, naive mESCs require leukaemia inhibitor factor (LIF), while the differentiation stimuli are suppressed by inhibiting the ERK and GSK3β signaling pathways [451, 335]. In contrast, human primed hESCs are dependent on bFGF and activin/TGFβ signalling [316]. The ability to derive and stably maintain ground-state human pluripotent stem cells (hPSCs) that resemble the cells seen *in-vivo* in the inner cell mass has the potential to be an invaluable tool for researchers developing stem cell–based therapies. To date, derivation of human naive-like pluripotent stem cell lines has been limited to a small number of lineages, and their long-term culturing remains problematic. We tag hPSCs by GFP, expressed by the long terminal repeat (LTR7) of HERV-H endogenous retrovirus. While the involvement of HERV-H explains many human-mouse differences in ESCs and naive cells, it is striking that ERVs have been recruited as pluripotency regulators in at least one other lineage. The mouse endogenous retrovirus mERVL expression of which is restricted to the zygote/2C stage, also regulates pluripotency [311]. The involvement of retroviruses in pluripotency supports the view that transposable elements (TEs) can govern numerous developmental processes [276, 211], often in a clade specific manner. The HERV-H-expressing cells have a similar, but nonidentical, expression pattern to other naive-like cells, suggesting that alternative pluripotent states might exist.

7.1 HERV-H genetically marks ground state naive human cells

ESC cultures are established from the inner cell mass (ICM) of preimplantation blastocyst embryos. Although, human and mouse ESCs share some features hESCs resemble mouse epiblast stem cells (mEpiSCs). In mouse, mESCs are rapidly proliferating cells that form dome-shaped colonies (3D), and maintain molecular/epigenetic features of ICM. Despite numerous attempts to isolate naive human pluripotent stem cells, a marker for ground state pluripotency in hESCs cultures1 remains elusive [380, 317, 357, 367, 359, 361, 366]. In fact, LBP9, NANOG and KLF4, clustered together on HERV-H, are highly expressed in the ICM during human embryonic development32, raising the possibility that HERV-H-driven expression marks naive–like stage in hPSC cultures.

To explore this possibility the reporter construct, pT2-HERV-H-GFP was transfected into either mouse or human ESCs (Figure 7.1A). Only about 4% of cells in each hESCs colony expressed GFP, indicating cellular heterogeneity (not shown).

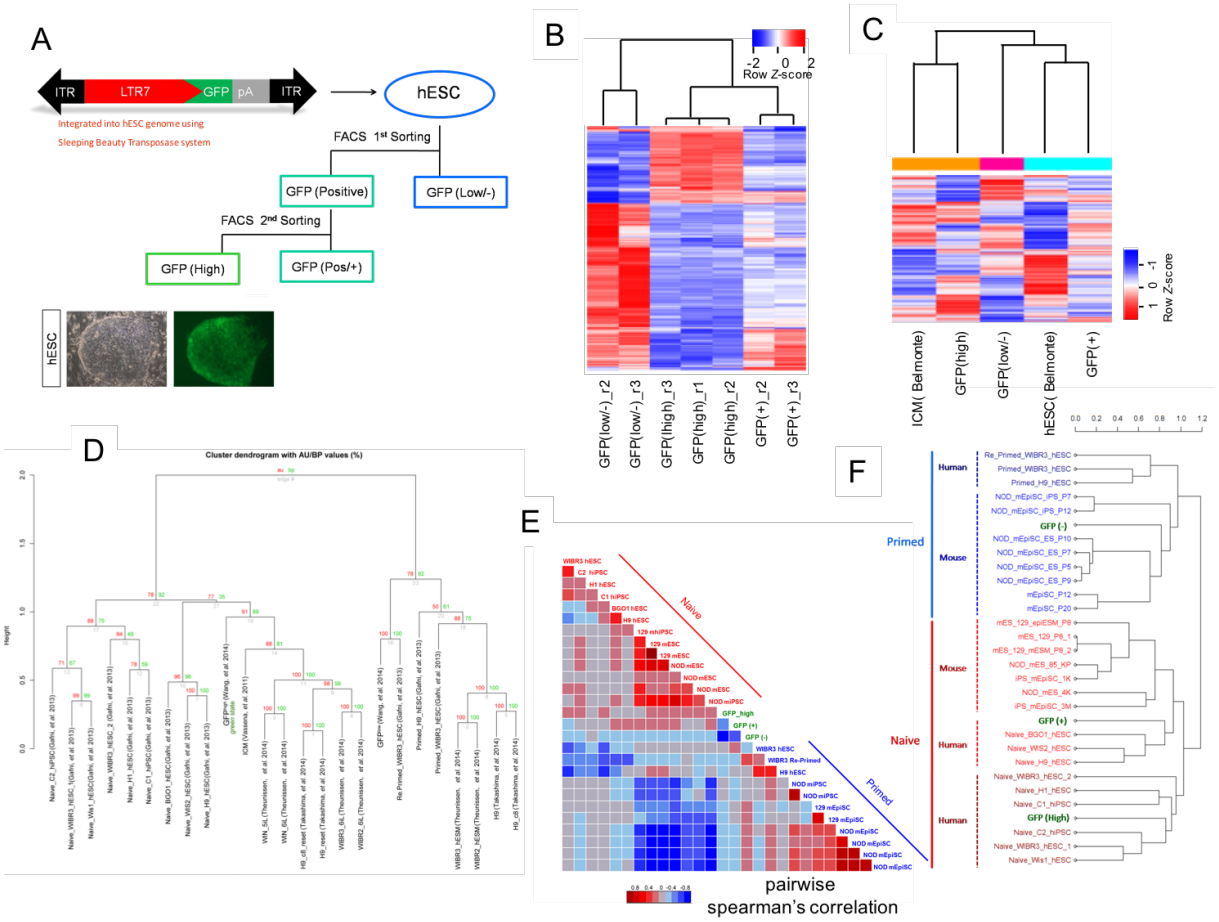


Figure 7.1:
A. Experimental scheme for isolating naive-like hPSCs. pT2-LTR7-GFP-rep-2-marked hESC-H9 were enriched by FACS sorting in multiple rounds and cultured in conventional hESC medium and in 2i/LIF medium, respectively; scale bar, 200 μm.
B. Heatmap showing the global expression comparison between GFPhigh, GFP+ and GFPlow cells. Hierarchical clustering of the mean expression values of global gene expression using Spearman's correlation (biological replicates are shown).
C. The expression of genes influenced by HERV-H expression in different human cell types, including GFPhigh, HERV-H-depleted hPSCs, published ICM and hESCs. The heat map shows the comparison of row-normalized relative expression levels at log2 scale. Genes shown are those differentially expressed within every pairwise comparison (differential expression defined by log2 modular change > 1, with FDR cutoff at 0.01). Isoform expression merged to single gene. Samples are represented in the order of euclidean distance and were clustered using Spearman's correlation and centroid linkage.
D. Global expression cluster dendrogram between GFPhigh, GFP+ and GFPlow hESC-H9, human inner cell mass (ICM) and previously established human naive and primed cell lines. Approximately unbiased (au) probability, bootstrap probability (bp) values and edge numbers at P value less than 0.01 are shown. ICM clusters closest with GFPhigh.
E. Correlation matrix displaying the unbiased and pairwise comparison of mouse–human orthologous gene expression between GFP-marked hESC-H9 (this study, green) and mouse and human naive as well as primed PSCs. Colour bar indicates Spearman's correlation strength.
F. Cluster analysis using the average distance method on the same data set as in E. GFPhigh, GFP+ and GFPlow cells in D–F were collected from hESC-H9 cells cultured in conventional human ESC medium by FACS sorting.

RNAseq data of hESCs from single cells [297] and hPSC lines confirm that pluripotent cultures exhibit variability in HERV-H expression (Figure 4.2C). To characterize GFP-positive (+) and negative (-) subpopulations, hESCs were FACS-sorted in two rounds, and separately re-plated on feeders (Figure 7.1A) [212]. GFP-positive cells were further divided into GFP(high) and GFP(+) categories. Strikingly, GFP(high) cells are capable of forming tight, uniformly expressing 3D colonies which is the characteristic of murine naive cells. In contrast, GFP(-) cells form flat colonies, resembling mouse mEpiSCs [212]. GFP(high) colonies cultured in hESC medium gradually loose both the reporter signal and their 3D morphology, except some GFP expressing groups of cells [212]. Their variable morphology is also reflected in the transcriptome pattern as GFP(high) cells display an unique pattern in sharp contrast with GFP(-) cells.

Next, we compared genome-wide expression profiles of GFP(high), GFP(+) vs GFP(-) cells. Unbiased hierarchical clustering of the expression profiles revealed that GFP(high) and GFP(+) cells have a similar, but not identical expression pattern, one that sharply contrasts with GFP(-) could be used to fetch distinct cell types within hESC cultures (Figure 7.1B). This analysis add an extra layer to the heterogeneity where an instrinsic retrovirus could be used to fetch distinct cell types within hESC cultures (Figure 7.1B). However, when cultured in LIF condition GFP(high) cells have improved single-cell cloning capacity, keep their pluripotency and differentiation potential, and could maintain their naive morphology for a longer period of time (Figure 7.1A) [212]. Thus, the GFP(high) subpopulation in human pluripotent stem cells appear thus to be enriched for cells resembling the ground state, i.e. naive hESCs. A hallmark of a naive cell population is also its resemblance to ICM of blastocyst. Thus, we compared the z-scores of quantile normalized intersected data of our genetically marked and cells derived either from cultured hESCs or directly from blastocysts (Figure 7.1B-C). GFP(high) and GFP(+) samples were distinguishable and clustered with the ICM and the hESCs samples respectively (Figure 7.1C). Despite the significant correlation, the human naive cultures [357], including our GFP(high), display some differences from ICM (Figure 7.1C).

To evaluate how the GFP sorted populations relate to previously characterised naive and primed cell populations from both human and mouse, we performed an unbiased cross-species hierarchical clustering of the Spearman's correlation of global gene expression (Figure 7.1D-F). Our cross-species and cross-platform pipeline included GFP(high), GFP(+) and GFP(-) samples as well as re-analysed gene expression data of 9,583 mouse–human orthologous genes for mouse and human naive as well as primed hPSCs. We found global transcriptional similarity between GFP(-) and all primed stage cells (Figure 7.1E). GFP(+) correlated better with mouse naive PSCs, than human naive PSCs (Figure 7.1F). Further dissection of the matrix reveals that GFP(high) cells show relatively high similarity to both mouse naive and human naive PSCs [357], supporting the significance of HERV-H-driven transcription defining naivety. Subjecting the same dataset to a cluster analysis using the average distance method revealed two major clusters of either primed or naive lineation (Figure 7.1F). The primed cluster included GFP(-) and transcriptomes of primed cells of either mouse or human origin (Figure 7.1F). The 'naive cluster' was further subdivided into two sub-clusters, represented by either GFP(+) or GFP(high).

7.2 HERV-H re-wired the primate-specific pluripotency network

While LBP9 (TFCP2L1) is key to the naive state in mice, HERV-H is primate specific. Experimentally, the depletion of either LBP9 or HERV-H results in loss of pluripotency as the transcriptional levels of

OCT4 and NANOG were reduced [212].

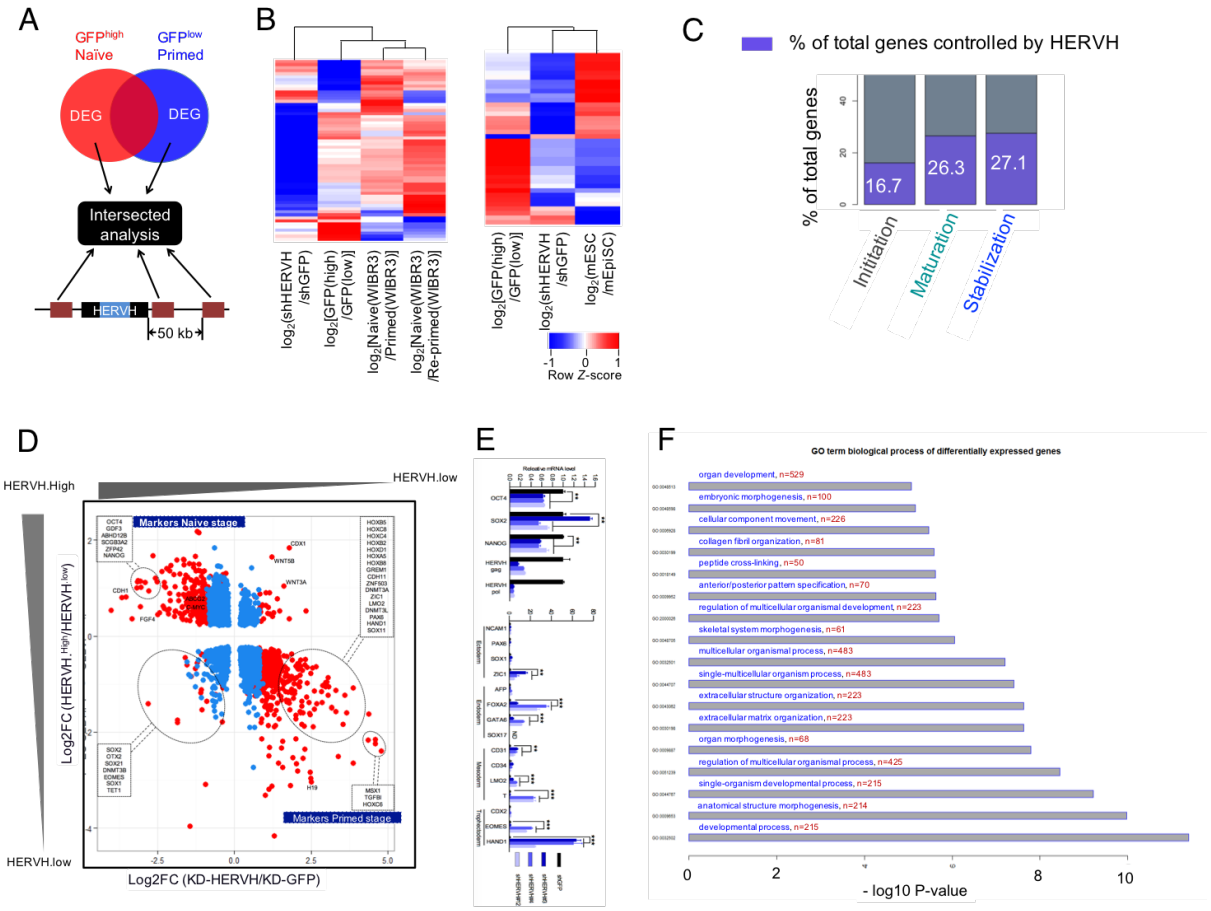


Figure 7.2:
A. The schematic shows the selection of genes for downstream analysis. Genes that commonly expressed in GFP(high) and naive cells; upregulated or downregulated genes compared with primed or GFP(-) cells. These genes were further mapped around HERV-H locus. Genes falling on downstream locus upto 50 KB was taken for further analysis as neighbour genes.
B. The expression of neighbouring genes to HERV-H in different human cell types, including GFPhigh, HERV-H-depleted hPSCs, published naive hPSCs (naive(WIBR3)) and primed hESCs (reprimed (WIBR3)) [357, 361]. The heat map shows the comparison of row-normalized differential expression levels at log2 scale of fold changes of GFPhigh versus GFPlow, shHERV-H versus shGFP, naive WIBR3 hESC versus primed and re-primed WIBR3 (GSE46872). Genes shown are those differentially expressed within every pairwise comparison (differential expression defined by log2 modular change > 1, with FDR cutoff at 0.01). In right panel the heat map showing the comparison with mESC versus mouse epiblast stem cells (mEpiSCs) of HERV-H neighbour genes. Log2 fold change values of orthologous genes were subjected to hierarchical clustering (Pearson correlation, centroid linkage, k = 3). Genes selected and clustered as in 7.2D.
C. Stacked barplot shows the number of genes involved in distinct stages of reprogramming (Figure 4.3) and blue shaded are those which are controlled by HERV-H expression in H9ESCs (Figure 7.2D). **D**. Scatter plot showing the differentially expressed genes between GFPhigh and GFPlow are negatively correlated with the ones between HERV-H-depleted hESCs and wild-type hESCs. The enlisted genes are enriched in GFPhigh versus GFPlow and are specific to naive state (upper right), while genes downregulated by HERV-H depletion are specific to primed hESCs or lineage commitment (lower). Red dots indicate differentially expressed genes. Representative clusters are shown.
E. qRT–PCR analyses to determine the relative expression level of pluripotency and differentiation markers after knockdown of LBP9 or HERV-H in hESC-H9. Data shown are representative of three independent experiments with biological triplicates per experiment. Error bars indicate s. d. , astericks represent the significant p-value (< 0.05) obtained by t-test.
F. Gene ontology (GO) categories for up-regulated and downregulated genes in GFPhigh compared to GFPlow.

Interestingly, the depletion of either of them pushes cells to differentiate into trophectoderm suggesting that LBP9/HERV-H netwrok might be serving as check-point at ICM-TE transition. Thus, the role of LBP9, comparable to HERV-H, is to support self-renewal, while transcriptional down-regulation of either LBP9 or HERV-H potentiates differentiation [212].

The above evidence is consistent with the possibility that HERV-H, which does not exist in rodents, has been integrated as a regulatory element into the transcriptional circuitry of pluripotency in humans, acting in turn as a marker of the naive state. To address how HERV-H-driven gene expression modulates pluripotency, we surveyed differentially regulated genes in GFP(high) vs GFP(+), intersected by HERV-H cis-regulation (Figure 7.2A). The differentially regulated genes located in the neighbourhood (+/-30 kb) of HERV-H display a similar expression pattern to those differentially expressed in GFP(high) vs GFP(-) and in human naive vs primed stages, derived under specific culture conditions [357] (Figure 7.2B, left panel). In contrast, a distinct pattern when compared to mouse naive vs epiblast (Figure 7.2B, right panel). Consistent with the observation that depletion of HERV-H expression compromises reprogramming process [212, 209], there are numerous dysregulated genes which contribute to distinct phases of reprogramming are controlled by HERV-H expression in hESCs (Figure 7.2 C and 4.2B).

As there is an antagonistic pattern of expression between genes defining naive stage [up in GFP(high/+) vs GFP(-)] and those that are down-regulated in HERV-H knockdowns and similar pattern was observed on the other way around (Figure 7.2D-E). The majority of differentially regulated genes upon depleting HERV-H expression in H1ESC was rescued by the activity of HERV-H regulatory region, provides the subtle regulatory network of HERV-H in hESCs (Figure 7.2D). The observed list of genes upregulated by HERV-H expression are putative markers of human pluripotency in naive state such as NANOG, POU5F1 etc. However the downregulated genes by HERV-H expression belonged to primed or artificial pluripotent states (Figure 7.2D-E). To examine this possibility, GFP(high) vs GFP(low) cells were subjected to further analyses. qRT-PCR confirmed significantly up-regulated transcripts in GFP(high) vs GFP(low), including TBX3, NANOG, OCT4, LBP9, FGF4, previously associated with naive pluripotency (Figure 7.2D-E). In contrast, up-regulated transcripts in GFP(low) relative to GFP(high) populations are associated with lineage-commitment, and include genes expressed in primed ESCs (e.g. ZIC1, SOX61, XIST, GATA6) (Figure 7.2E). The enlisted observations confirm an underlying the significance of HERV-H in regulating the native pluripotency in primates. Differentially expressed genes between GFP(high) vs GFP(-) populations were enriched for Gene Ontology (GO) terms of developmental processes, morphogenesis and organismal processes (Figure 7.2F).

7.3 Re-activation of XIST RNA is unusual in forced naive cells

An additional hallmark of naive pluripotent cells in female mice is two active X chromosomes (XaXa), while primed cells show X inactivation (XaXi), whereas, human naive cells orchestrate the X-chromosome inactivation [317, 361, 357]. X-chromosome inactivation (XCI) depends on XIST, a long noncoding transcript that guides PRC2 complex to catalyse the H3K27 methylation and silences the X-chromosome. X-chromosomal dynamics is unique in human preimplantation embryos in a way that both X-chromosomes are active in females and males express X and Y chromosomes. This unusual mode of dosage compensation eradicates the essence of XIST expression in the early embryos. The first strike on naive cells are over-activation of XIST RNA. XIST is under top-five most upregulated genes in forced naive cells following DNMT3L, KLF2 and KLF17 (not shown). We re-analyzed the transcriptomes of forced naive cells SSEA-

Neg, SSEA-Pos ("Primed" cells grown in 5iL media and sorted out on the basis SSEA, the pluripotency marker; SSEA-Neg is non-pluripotent naive cells whereas, SSEA-Pos is pluripotent naive cells), UCLA-20n is isolated blastocyst cells cultured in 5iL media whereas, UCLA-20n primed cells are the cells from same blastocyst but cultured in 2iL media [361, 356].

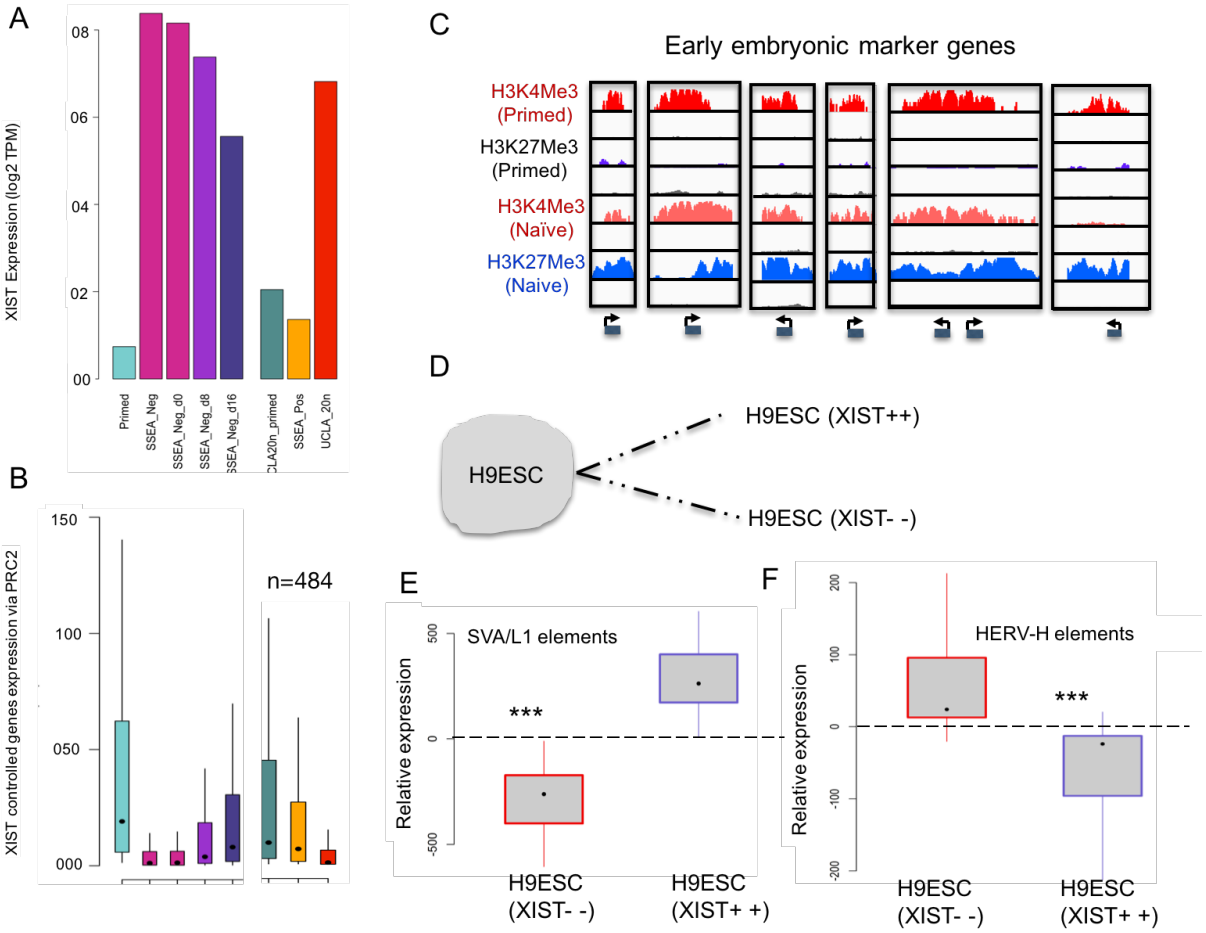


Figure 7.3:
A. Barplot shows the level of expression of XIST RNA at log scale (log2 TPM+1). Samples involved in this study are forced naive cells i. e SSEA-Neg, SSEA-Pos and UCLA-20n ("Primed" cells grown in 5iL media and sorted out on the basis SSEA, the pluripotency marker; SSEA-Neg is non-pluripotent naive cells whereas, SSEA-Pos is pluripotent naive cells). UCLA-20n is blastocyst cells cultured in 5iL media whereas, UCLA-20n primed cells are the cells from same blastocyst but grown in 2iL media [361, 356]. Reset cells are another naive cells resultant from forced expression of NANOG and KLF2 [359]. SSEA-Neg cells were also grown in differentiation media and RNA-seq was captured on day-0, day-8 and day-16 as labelled in the plot.
B. Boxplot showing the overall expression of genes regulated by XIST via recruitment of PRC2 complex.
C. IGV plot showing the visualization of normalized ChIP-seq reads over the promoter of genes involved in embryogenesis. Note: These genes are activated in primed cells (H3K4Me3 mark) but they are either in poised state (first four panel shows the H3K4Me3 and H3K27Me3 co-occupancy) or in repressed (only H3K27Me3 mark).
D. A Schematic diagram illustrating the design of study reanalyzed to obtain next figures [452]. The two populations were sorted out on the basis of XIST expression, named as XIST+ and XIST-
E. Boxplot showing relative expression of SVA/L1 elements. Significant enrichment (-log10(p-value) < 16) of mutagenic TrEs (SVA and L1-Hs) in XIST+ H9-ESCs compared with XIST- H9-ESCs.
F. Boxplot showing relative expression of HERV-H elements. Significant enrichment (-log10(p-value) < 16) of HERV-H in XIST- H9-ESCs compared with XIST+ H9-ESCs.

Additionally, included reset cells, are another naive cells resultant from forced expression of NANOG and KLF2 [359] in 2i Gö6983 media. SSEA-Neg cells were also grown in differentiation media and

RNA-seq was captured on day-0, day-8 and day-16 as labelled in the plot (Figure 7.3A). Upon surveying the XIST expression in mentioned naive and primed cells, we observed that the over expression of XIST in forced naive cells was not lost during their differentiation. This indicates the irreversible activation of XIST upon reverting the primed cells to naive state (Figure 7.3A). Furthermore, we checked the transcriptional dynamics of those genes that were repressed by XIST expression. For this, only those genes were chosen that gained H3K27Me3 mark on their promoter during the reversion of primed to naive state (Figure 7.3B). We found that XIST target genes (n=484) showing the antagonistic pattern of expression and moreover, these genes do not gain their expression upon converting back (naive to primed state) (Figure 7.3C).
Previous studies have shown that forced naive cells in 5iL media repress HERV-H expression and up-regulate the transcription of SVA elements as their molecular signature [366]. I checked the transposcriptome of XIST- and XIST+ cells sorted out of H9ESCs (Figure 7.3D). H9ESCs deficient in XIST expression (so does the early embryonic cells) showed the repressed state of SVA elements and activated HERV-H expression, whereas, antagonistic pattern was observed in XIST+ H9ES cells (Figure 7.3E-F).

7.4 In vitro culturing might compromise evolutionary fine-tuned, human specific features

We have shown in previous sections that HERV-H mediated re-wired transcriptional networks led the molecular evolution of biological processes to be human-specific. Generation of naive cells is confrontation of artificial pluripotent states against natural ones via inhibiting biological processes at molecular level. Our study suggests that the HERV-H driven transcriptional network has significantly modulated pluripotency during primate evolution. How current *in vitro* pluripotent stem cell cultures do reflect these features? We examine human naive cell cultures (e.g. 3D morphology) that are either converted from primed cells [359, 366] (e.g. 2D morphology) or freshly established from the human blastocyst [356], and compare them and to their primed counterparts (e.g. H9ESCs and UCLA primed cells). The naive and primed cells are cultured in 5iL or 2iL conditions, respectively. Clustering analysis of their trancriptomes reveals four distinct HERV-H expression clusters (Figure 7.4A), suggesting that their origin, culture conditions as well as their pluripotent state affects the genomic HERV-H expression. Cluster 1 identifies HERV-H loci activated in naive cells cultured in 5iL condition, but not expressed in primed cells. In Cluster 2, the majority of HERV-H loci is active in primed cells and in naive cells that is established from the blastocyst (UCLA-20n) [356]. In contrast, Cluster 3 identifies HERV-H loci that are activated in primed and pluripotent naive cells converted from primed (SSEA4-Pos) indicating the pluripotency-specific loci. Finally, Cluster 4 is specific to converted SSEA4-Pos populations grown in 5iL media could be used as a marker for pluripotent populations in 5iL media (Figure 7.4A).

Importantly, HERV-H is upregulated in both prime-converted (SSEA4-Pos) or blastocyst derived (UCLA20n) cells when compared to (SSEA-4-Neg) naive cells that are negative for the pluripotency marker, SSEA-4 (Figure 7.4B-C), suggesting that the lack of pluripotency is associated with the heavy down-regulation of HERV-H.

We hypothesized that along-with the loosing of HERV-H expression, several human-specific traits, evolved to fine-tune human pluripotency might be lost as well. Do the *in vitro* cultured naive cells express the human specific features properly in their transcriptome?

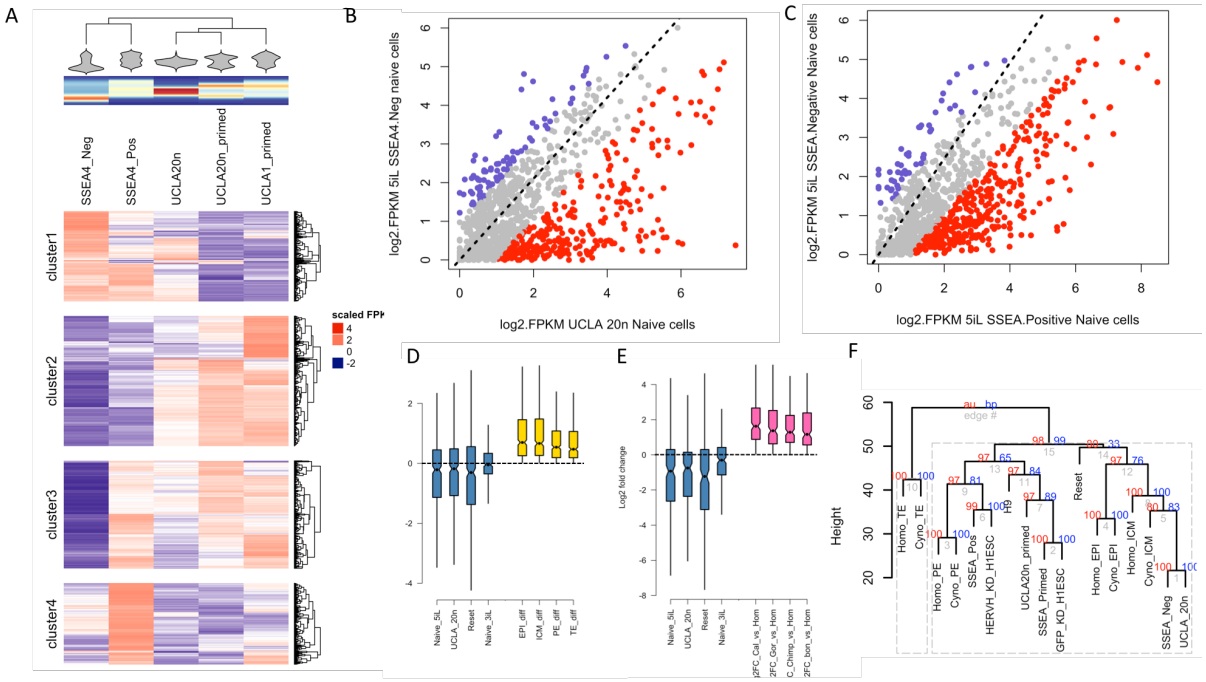


Figure 7.4:
A. Heat map shows the five distinct clusters of highly expressed (log2 FPKM $>$ 1) full-length HERV-H locus (n=250) in different naive and primed cells. Distribution of HERV-H expression shown as violin plots on top of heatmap. Dendrogram above cell type labels is clustering of scaled FPKM by euclidean distance and spearman's correlation. Naive cells in this analysis are SSEA-Negative 5-iL, SSEA-Positive 5-iL, UCLA20n and primed cells
B. Scatterplot shows the differential expression of individual HERV-H locus in 5iL Naive cells that are negative for SSEA4(the pluripotency marker) reverted from primed cells and blastocyst cells directly (UCLA20n). Each dot on the plot is single locus. Red ones are those that gained its expression in SSEA-Positive 5iL naive cells. Blue ones are those that gained its expression in SSEA-Negative 5iL naive cells.
C. Scatter-plot shows the differential expression of individual HERV-H locus in 5iL naive cells that are positive or negative for SSEA4, the pluripotency marker. Each dot on the plot is single locus. Red ones are those that gained its expression in SSEA-Positive 5iL naive cells. Blue ones are those that gained its expression in SSEA-Negative 5iL naive cells.
D. Notched boxplot shows the distribution of differential expression (Log2 Fold Change) forced Naive cells with respect to their respective primed cells (as described in Figure 6A) as shown in steel blue colored boxes, for those genes (n=246) that are commonly up-regulated (average difference $>$ 0) in all lineages of human blastocysts compared with their counter lineages in *Cynomolgus* blastocysts (ICM, EPI, PE and TE) as shown in gold colored boxes.
E. Notched boxplot shows the distribution of differential expression (Log2 Fold Change) forced Naive cells with respect to their respective primed cells (as descripted in Figure 6A) as shown in steel blue colored boxes, for those genes (n=197) that are commonly up-regulated (GFOLD $>$ 0) in human PSCs compared with all analyzed non-human PSCs (Callithrix, *Gorilla*, *Chimpanzee* and *Bonobo*) as shown in deep pink colored boxes.
F. Bootstrapped Dendrogram (1000 replicates) represent hierarchical clustering using euclidean distance and complete linkage method on averaged expression from the cell population's (mentioned in Figure 4A, 6A and HERV-H-KD cells) transcriptome pooled together. 1055 most variable genes (method) were chosen to construct the clustered Dendrogram. Height of Dendrogram represents the euclidian distance of dissimilarity matrix, numbers in red and blue indicate au and bp values from bootstrapping. Noticeable clusters in figure shows the compact cluster of cross-species EPI, PE and TE. Human and *Cynomolgus* TE form distinct cluster that is not shared with any of *in vitro* cells. Notably, HERV-H-KD cells and 5iL SSEA-Positive cells, Primed cells along with H1 and H9 cells share their cluster with PE albeit fall in to form distinct cluster. Interestingly, 5iL cells (SSEA-Negative and UCLA-20n) clusters closely with *Cynomolgus* ICM compared with human ones. Reset cells were found to be outlier in the same cluster.

As the human primed cells resemble more with the *Cynomolgus* post-implantation epiblast cells [385], and do not express the human pluripotent blastocyst features, we compared the fold-change of

naive vs primed cells with the fold-change of pairwise Homo and *Cynomolgus* blastocyst stages (ICM, EPI, PE and TE) (Figure 7.4D). Our analysis reveals that the human pluripotent blastocyst features are significantly downregulated at *in vitro* naive cells (Figure 7.4D). Similarly, expression of those genes that mark the human specific vs primate features of pluripotent stem cells (e.g. innate immune response, STAT3, IFITM1, etc.) are inhibited *in vitro* naive cultures 7.4E). In addition, *in vitro* naive cultures appear to express genes whose expression has been shifted during primate evolution to a different developmental lineage (Figure 8.3). Finally, 1K bootstrapping on 1200 MVG clustered SSEA4-neg naive cells close to monkey ICM, while SSEA4-pos cells plugged in Homo- *Cynomolgus* PE cluster (Figure 7.4F).

The up-regulated genes in each human stage showed significant down-regulation in naive cells suggests the loss of human blastocyst specificity 7.4D). Next, we dissected the up-regulated genes in human comparing cross-species pluripotent states and naive cells showed significant loss there too, pinning our hypothesis that human-specific pluripotency might have been lost while inhibiting the pathways that has been evolved to be human-specific 7.4E). Some of up-regulated genes in forced naive cells are those genes that mark pluripotent states in *Cynomolgus* and are expressed in human blastocyst but only to mark trophectoderm populations but not the pluripotent ones 8.3), indicates another dimension in the loss of human-specificity. Finally, 1K bootstrapping on 1200 MVG showed the clustering of SSEA4-negative naive cells close to monkey ICM compared with human ones, whereas SSEA4-positive cells plugged in human- *Cynomolgus* PE cluster 7.4F).

7.5 Resemblance between cultured naive-like cells and newly defined human pre-implantation cell types

Recently, a plethora of human cell lines have been described with properties that resemble naive murine pluripotent stem cells (e.g. rounded morphology of cultures) [358, 453, 357, 356, 359, 366, 361, 362, 212]. Most of these naive/naive-like lines, except the extracted cell type using a HERV-H-reporter (HERV-Hhigh) [212] are generated by overexpressing various transcription factors and/or cultured under conditions inhibiting certain developmental processes. Given our ability to identify genes particular to each stage in early human embryogenesis, we now ask about the resemblance between these various naive-like cell types, primed cells and stages in pre-implantation development. First, we construct a dendrogram based on relative gene expression profiles (microarrays) including all available naive-like cultures, their corresponding primed counterparts, and adding the transcriptome of blastocyst and morula 7.5A-C). As expected, naive-like cells and primed cells cluster separately. Within the naive cell grouping, a few naive-like lines cluster in proximity with both morula and blastocyst, while others form a different cluster distinct from morula/blastocyst. HERV-Hhigh is unique in clustering close to blastocyst but not morula and display significant level of correlation with the same 7.5A-B).

7.5.1 The majority of forced naive cultures are heterogeneous cell populations, expressing various lineage specific markers

Indeed, certain forced naive lines overexpress several human morula genes compared with their primed counterparts 7.5C). However, these cells are not clear morula either. To find more out about their identity, we examine genes that are significantly upregulated in at-least 80% of the naive-like lines when compared to their primed counterparts. We intersect the obtained genes with our lineage specific markers (AUC

cut-off > 0.85) 3.1A-C). This strategy reveals that although the studied naive/naive-like lines express EPI markers, they represent heterogeneous mixtures of cells. Curiously, some cells aberrantly present a variety of human preimplantation embryonic lineage markers, including 8-cell, morula, NCC and PE 7.7A). Thus, although a fraction of cells in naive-like cultures resemble real stages of development, the majority of the cells exhibit a disturbed expression profile.

We also integrate the scaled averaged expression profiles of morula, NCC, ICM, EPI and PE with certain naive-like lines [358, 356, 359], extract 5,870 variable genes (mean/dispersion < 1), and load them for Kernels k-means clustering. This strategy gives us four distinct clusters containing 1,210, 1,265, 1,181 and 2,214 genes, respectively 7.6A). The euclidian distances (dendrograms in Figure 7.6B-E) clearly separate developmental stages and naive-like cells with the four clusters. Based on ranked correlation coefficients, the clustering analysis further supports the heterogeneous existence of morula, NCC, ICM and EPI markers in any of the four clusters (Figure 7.6B-E).

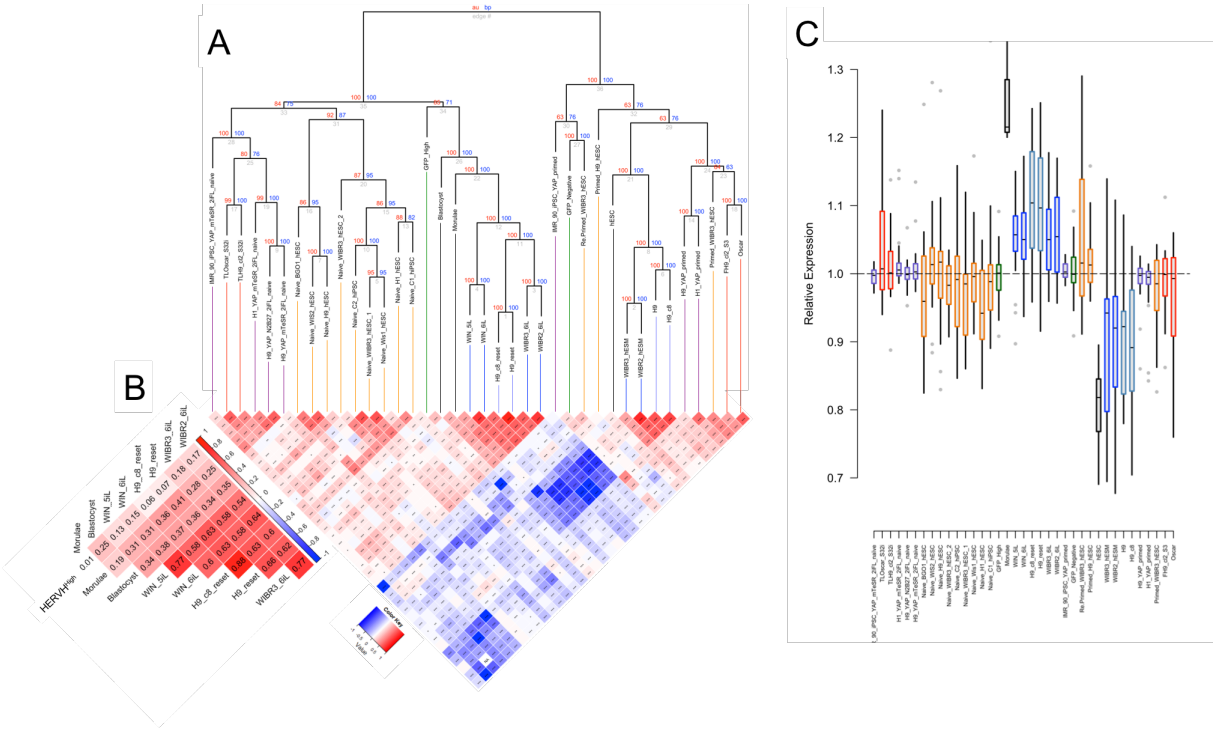


Figure 7.5:
A. Cross-platform clustering (ranked correlation) of 8346 genes. Distance measure shown as euclidian distance, genes, 1K Bootstrap and correlogram showing the cross-platform analysis of available microarray datasets of naive/naive-like and primed cells. Red color shows positive correlation and blue corresponding to negative correlation. Star marks the significance of correlation obtained as (p-value < 0.01).
B. Correlation of a subset (from A) of naive/naive-like cells to morula/early blastocyst. Pairwise ranked correlation values of relative expression are shown.
C. Relative expression of morula-specific genes in a subset of naive/naive-like cells with respect to their primed counterparts. Note color code and order of samples are as in A. Relative expression of morula and blastocysts were compared against hESCs.

7.5.2 Forced naive-like cells exhibit disturbed transcriptomes accompanied with upregulation of mutagenic transposable elements

The forced naive-like cells are also unusual in having more frequent generation of chimeric transcripts that are still expressed upon converting them to primed state (Figure 7.7B), and might help to explain why

these cells are not capable of proper conversion [356]. When compared to human blastocyst (E5-E7), forced naive cultures express around 352 genes that are not expressed in any cells of the blastocyst, and fail to express 67 'essential' genes that are expressed in every cell of the human blastocysts (E5-E7) (Figure 8.4A-B). These cells also acquire features of the 'maturation' stage of the reprogramming process (Figure 7.7D and 8.4D). This rather unstable stage is accompanied by erasure of DNA methylation over the TSSs of genes (imprinted/non-imprinted) and Young ERVs that are hypermethylated in oocytes (Figure 8.4E-F) and upregulation of transposable elements with mutagenic potential [366, 208]. Notably, forced naive conditions also suppress the expression of both the EPI-specific genes and HERV-H loci (Figure 7.7C).

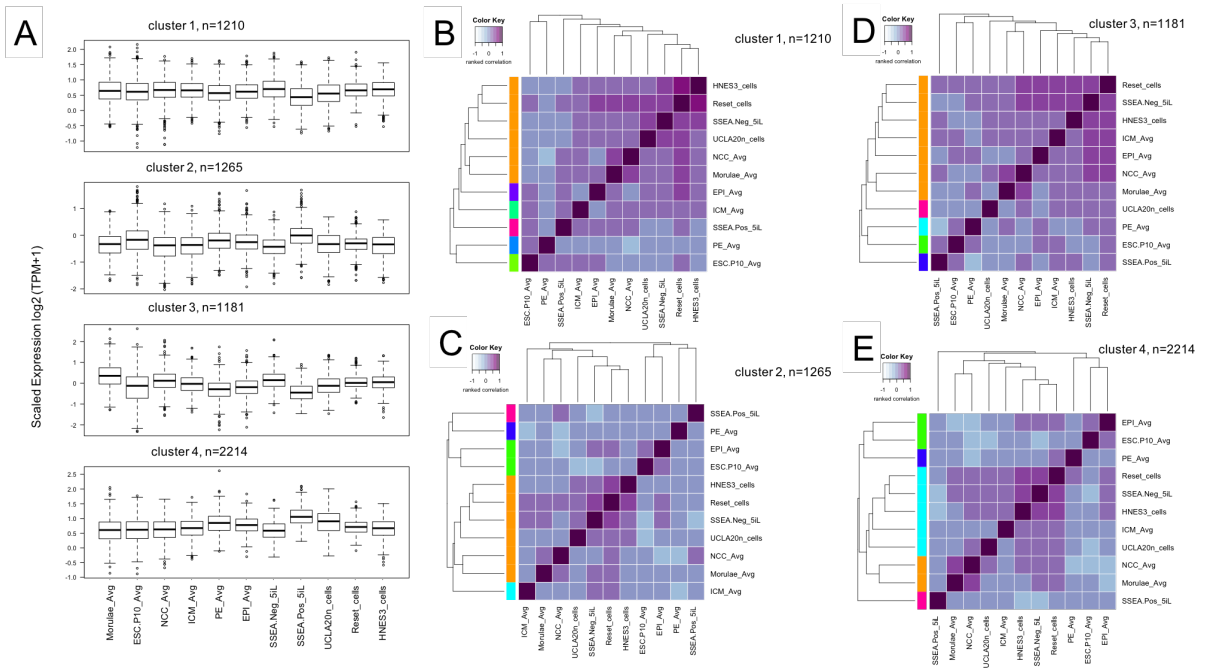


Figure 7.6:
A. Multiple boxplots showing the relative expression (scaled to mean) of genes clustered together in four clusters within the data frame comprises the transcriptome of samples labeled on x-axis. Clustering is based on kernel density clusters that divided genes into four groups each containes 1210, 1265, 1181 and 2214 genes, respectively.
B. Dendrogram shows the clustering of samples based on euclidian distance of dissimilarity matrix using 1210 genes from cluster 1. Heatmap of pairwise spearman's correlation showing purple as positive and light blue as negative correlation. Note: naive/naive-like cells cluster with NCC and morula.
C. As in B, but using 1265 genes from cluster 2. Note: naive/naive-like cells cluster with NCC and morula.
D. As in B, but using 1181 genes from cluster 3. Note: naive/naive-like cells clustering with EPI and ICM.
E. As in B, but using 2214 genes from cluster 4. Note: naive/naive-like cells clustering with ICM.

7.6 Human *in vitro* pluripotent cultures should express species-specific traits properly

Finally, our studies allow us to suggest a checklist that could be used to guide *in vitro* studies. The key properties of the pluripotent EPI, including self-renewal and the ability to maintain a relatively homogeneous transcriptome, would make this cell population as a best candidate to mimic *in vitro*. Thus, our checklist includes top genes that appear to have a unique expression status (+/-) in human EPI against the rest of the clustered cells (Figure 3.1B and 7.7C). These genes have a good overlap with those whose expression is continuously rising during the reprogramming process and plateau at the 'stabilisation stage'

(Figure 7.7E). We also provide a checklist to exclude genes that do not feature any real developmental stage in human, partially overlap with the 'maturation' stage of reprogramming (Figure 7.7D). The expression of these genes could induce instability and various aberrant processes that could compromise pluripotency. The checklists include ESRG, LEFTY1, HHLA1 and excludes H19, KLF2/17 expression (Figure 7.7C, lower panel). Notably, ESRG and HHLA1 genes carry full-length HERV-H sequences, suggesting that human pluripotent cultures should reflect the species-specific features properly, including the expression of HERV-H-remodelled genes that have been contributed to fine-tune pluripotency regulation in humans.

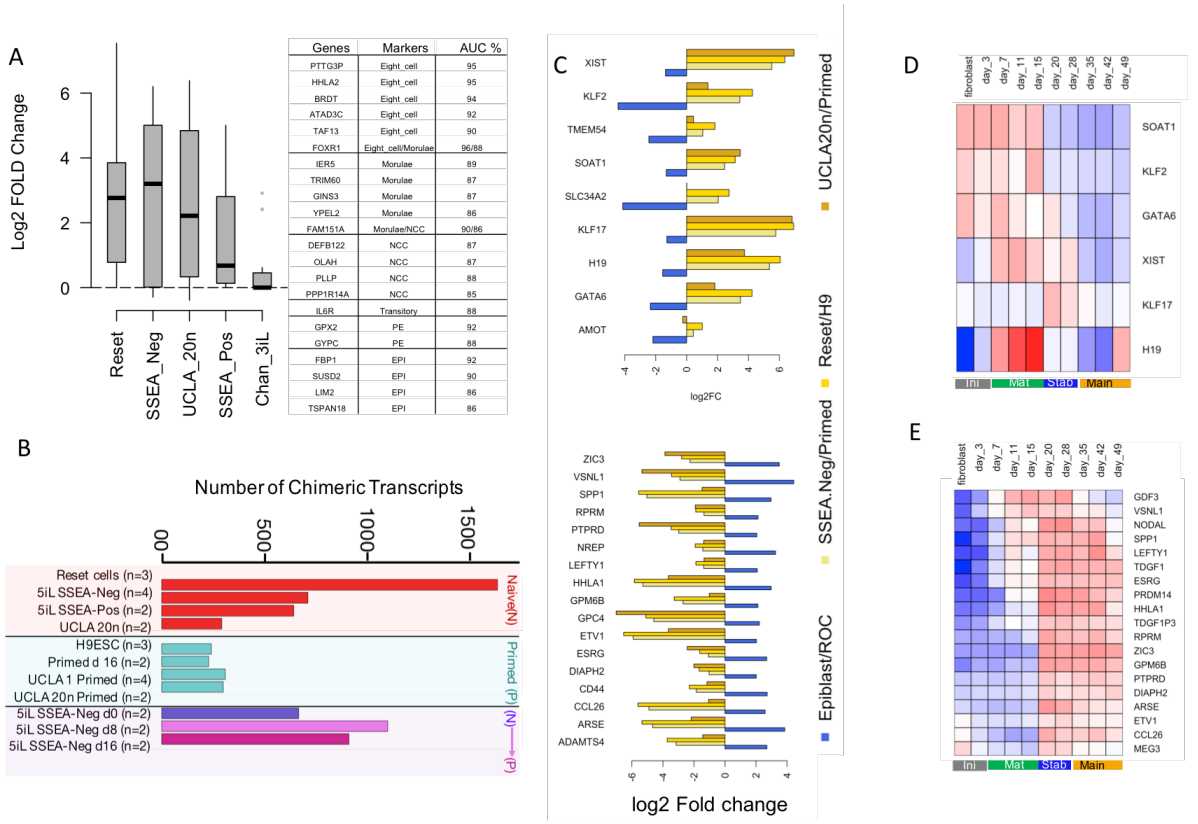


Figure 7.7:
A. Boxplot shows the upregulation of early embryonic lineage markers in naive cells with respect to their respective primed cells (GFOLD calculated on Reset cells/H9-ESCs [359], 5iL-SSEA-Neg/UCLA1-primed, UCLA-20n/UCLA-20n-primed, 5iL-SSEA-Pos/UCLA1-primed [356] and Chan-3iL/H1-ESCs [358]); The marker genes selected for the analysis flag distinct lineages with AUC cuttoff values (table next to the boxplot), and were chosen on the basis of the following criteria: (i). Should be putative markers of any distinct lineage in human preimplantation embryos (AUC cutoff > 0.85) after EGA; (ii) should be significantly upregulated in the majority of analyzed naive cells (3 out of 5).
B. Bar plot showing the number of chimeric transcripts detected from the cell populations representing naive, primed and differentiation of naive to primed stage.
C. Bar plot of EPI-like checklist. Log2-fold expression values of genes exhibiting expression status (+/-) left and right, respectively. EPI-like cells should express/upregulate 67 genes (See Figure). These genes are either specifically expressed in pre-EPI/EPI or upregulated in EPI vs rest of the cells (ROC) in the human embryos (selected examples, left panel). EPI-like cells should not express/upregulate 352 genes (see also Figure S). These genes are either specifically repressed in EPI and Pre-EPI or downregulated in EPI vs ROC (selected examples, right panel).
D-E. Heatmap showing expression levels of selected genes from the checklist at z-score during somatic reprogramming. Note the overlap of selected genes (as in Figure 7.7 with genes expressed in the stabilisation/maturation stages.

Chapter 8

Discussions

8.1 Fate of the human embryonic lineages at single cell resolution

Despite of recent speculations, our inferences are consistent with the assumption that the major features of pre-implantation embryogenesis, regarding cell types and trajectories, are basically conserved in mammals. Our strategy to analyse recent single cell transcriptome data employed a dimension reduction clustering methodology that allowed us to cluster highly similar cell populations, but also resolve small differences. Our approach assisted us to re-establish the lineages specific markers and an updated atlas of the early human development. We dissected the entire intra-uterine embryonic lineages, including the ICM, the holy grail of pluripotency. Beside similarities, we also identified multiple features that have been diverged during the evolution of early development.

We found that, following the morula stage, cells exhibit an unusual heterogeneity that could explain some of the difficulties to resolve the stages. The edges between developmental stages are not 'sharp', and a significant number of cells exist simulatneously in 'transitory' states, expressing more than one lineage marker. Furthermore, identification of a class of cells that fails to commit to any lineage could successfully unmask the ICM. A clear distinctive feature of human pre-implantation development is the presence of dynamically changing, hard-to-catch transitory zones between the stages. Although ICM is short-lived before it segregates to EPI and PE, it is clearly identifiable, supporting the hypothesis that blastocyst formation, similarly to other mammalian species, occurs in well-defined sequential steps. It was also necessary to unmask non-committed cells from a relatively large, previously unidentified cell population. These cells derive from morula, exhibit enormous transcriptome heterogeneity and fail to appropriately express lineage markers. These non-committed cells (NCCs) are subjected to programmed cell death and filtered out from the developmental process after E5 blastocyst stage.

Apoptosis is a process that is an essential part of normal development, and participates in eliminating certain types of cells, including aberrant (e.g. genetically damaged) cells from the developmental process. While the normally developing human embryos show no evidence of apoptosis before compaction, scattered cells with fragmented nuclei and DNA damage characteristic of cells undergoing apoptosis were found at the morula and blastocyst stages [454]. We observed relatively large numbers ($\sim$33%) of apoptotic cells at E5. Similarly, a large number of apoptotic cells were observed in cow morulas ($\sim$ 50% of embryos) [455]. In normally developing embryos, the activation of apoptosis is thought to be

associated with and detectable immediately after embryonic genome activation (EGA).

8.2 Formation of epiblast and self-renewal network by exaptation of HERV-H

Non-committed cells are likely to be filtered out from the developing blastocyst. These cells show elevated levels of potentially mutagenic transposing elements, suggesting a selective filter against damaged cells in the early embryo. On the other side, the cells that continue to participate in the developmental program are characteristically marked by the expression of ancient TrEs, particularly the human endogenous retrovirus H (HERV-H). While HERV-H appears to be an important player, there are many others unknown. Foremost is the question how HERV-H derived products are involved in the regulation of pluripotency. Second, are HERV-H-associated activities restricted to a certain cell type? Our new high-resolution atlas of early human development enables us to address these issues. Further, by close analysis of co-expressed gene partners we seek to determine the functional modules within which HERV-H is expressed. With the new cell classification we provide further insight into human early development. Out of all the cell types resolved, EPI is uniquely homogeneous in expression, consistently with its self-renewal ability, and is thus recommendable for *in vitro* culturing. Epiblast cells should be the reference frame while comparing in vitro pluripotent states. Even though ICM is taken out of the blastocyst to culture embryonic stem cells, it first differentiates to epiblast and then forms *in vitro* pluripotent states that mimic post-implantation blastocysts. Regarding epiblast, it is not clear whether it has self-renewal capacity, since self-renewal was implicated to be existing in cultured pluripotent cells. Defining the peculiarities of early human development is of relevance, not least because culture conditions for embryonic stem cells are classically derived from our understanding of early mouse development. Further, there is debate as to what are the real features of naive embryonic stem cells. In principle, naive cells should be *in vitro* model of pluripotent blastocyst. A checklist of gene expression changes during preimplantation development would be valuable to define such naive cells. Deciphering how pluripotency is modulated to reach its human shape would be helpful in defining *in vitro* culture conditions that could mimic blastocyst the closest and preserve genome stability and self-renewal network.

8.3 Human transposcriptome dynamics mark the distinct lineages of early embryo

EGA is accompanied by global epigenetic changes that modulate the transposcriptome, for instance, by de-repressing transposable elements. Notably, the timing of EGA (and thus the wave of TE activation) varies among different mammalian species. In the mouse, the major activation event occurs during the 2-cell stage [311], whereas in humans it occurs later, between the 4 and 8-cell stages [295]. Apparently, apoptosis enforces a selective filter against damaged cells emerging after EGA and fail to properly express lineage markers. We identified a difference between young and old TrEs, with young ones expressed post transcriptional reactivation. This distinction is consistent with continual cycles of invasion and TE activity, followed by suppression preventing expression and transposition. During the phase of embryonic transcriptional suppression, the transcripts from TrEs are detected, presumably as oocyte remnants alone. This continual evolutionary turnover is further reflected in the gene gain events of families associated with

TE control. Indeed, among the most dynamic gene families during primate evolution we identify ZFPs, implicated in regulating TE activity [166, 133].

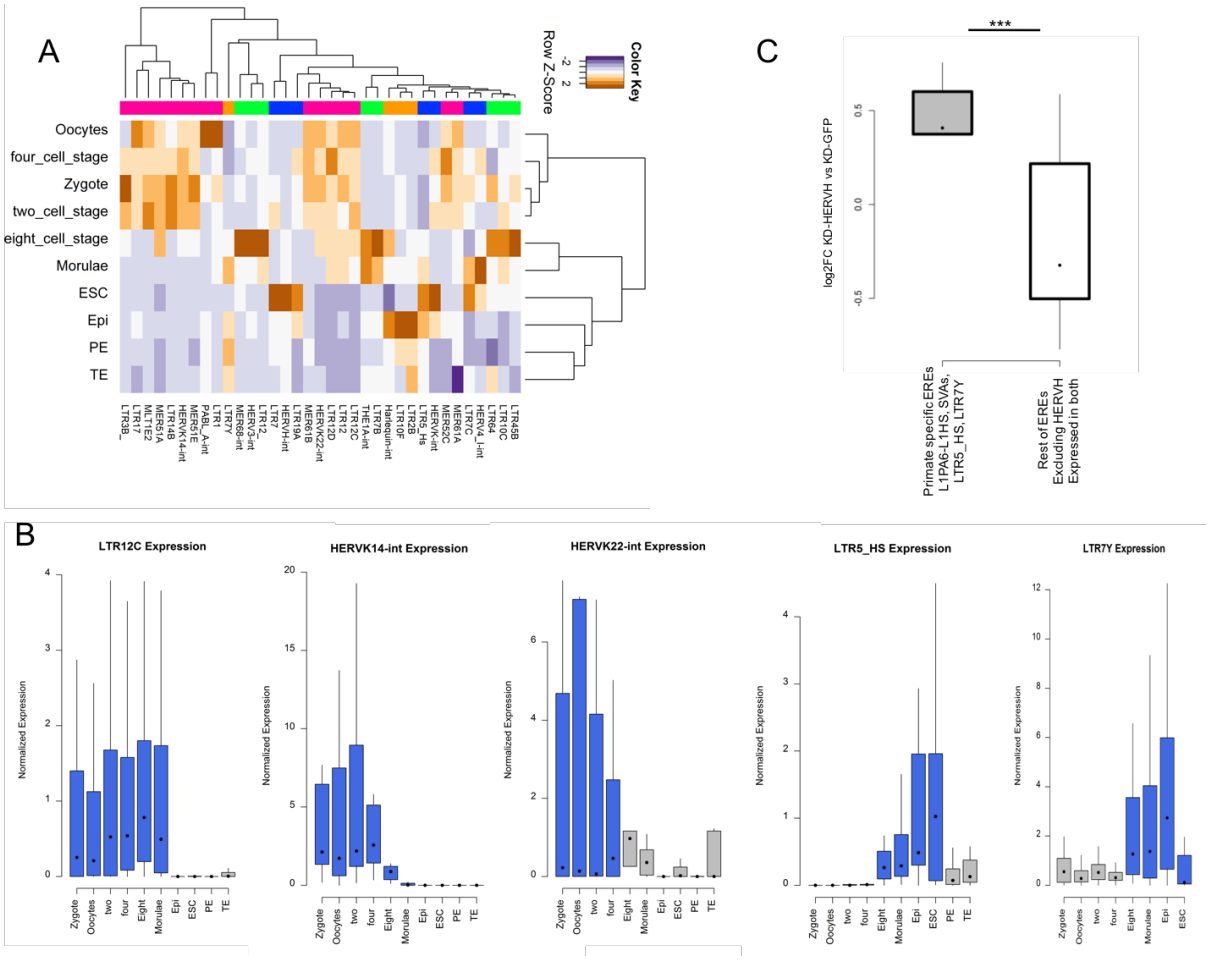


Figure 8.1:
A. Unbiased clustering of expression values of ERVs transcribed during early development (RPKM>1) using the Spearman's, rank correlation method. Data are re-analyzed from [297].
B. The box-plots show level of expression of the LTR7B/Y subfamilies during early development. Distribution of expression is shown as tag per million (tpm) values for each locus.
C. Box-plot displaying the relative expression of young and old retroelements upon depletion of HERV-H in human embryonic stem cells .

Given the apparent cycle of invasion then suppression of TrEs, it is perhaps to be expected that one class of TE (HERV-H) would be exceptional and become co-opted for control of early development (although this extension of pluripotency might also be directly beneficial to a TE [312]). Indeed, many of the most variable genes between the cells are altered by HERV-H insertions. More generally, the primate-specific HERV-H-enforced remodelling appears to underpin much of the primate/human specificity. Indeed, HERV-H is central to many examples of new genes, chimeric, *de-novo* and remodelled genes. It also recruits naive-associated TFs in a primate-specific manner. Apparently, following their inactivation as an endogenous retrovirus [456], some HERV-H-derived sequences have been gradually domesticated for cellular host functions. HERV-H-enforced gene structures are expressed in well-defined clusters, are incorporated into the developmental process and become an essential component throughout pre-implantation development (e.g. defining pluripotency). Human-specific protein-coding gene ESRG has emerged as the most remarkable of these HERV-H-enforced gene structures [312] that has been

integrated into the regulatory network of human pluripotency. In addition to ESRG, a few more HERV-H-enforced transcripts have been characterised and known to have functionality (e.g. lincROR, SCGB3A2) [286, 444, 312]. Still, as the domestication process of HERV-H is relatively young, it remains to be established what proportion of the rich diversity of HERV-H-associated transcripts seen in early embryos are functional.

We asked what is the rationale for HERV-H being one of the TrEs recruited to regulate the human pluripotency. We suggest that this might be due to HERV-H's abilities to suppress young TrEs and to generate a great degree of diversity via its heterogeneous transcripts. Moreover, HERV-H gains multiple TSS which makes it flexible to be expressed even in various stages of preimplantation development. In human, the expression of potentially mutagenic, phylogenetically young elements peak in morula. Upregulated young TrEs potentially contribute to the observed high heterogeneity of cells segregating from morula, giving rise to both committed (progenitor) and non-committed cells (NCCs). Progenitor cells express OCT4/NANOG, whereas NCCs show higher level of H2AX activity. Additionally, we detect Line1 transposase expression in NCCs. The Line1+ cells are also positive for H2AX staining, indicative of DNA damage (e.g. double strand breaks) likely to be generated by active transposition.
While NCCs are marked by young TrEs (e.g. Line1, SVA), progenitor cells, that passed quality control and continue to participate in the developmental program, characteristically express ancient but dominantly full-length ERVs, including HERV-H. Curiously, HERV-H also selectively marks EPI vs PE, suggesting that HERV-H might contribute to cell fate determination during blastocyst formation in humans. EPI expresses a number of HERV-H-enforced transcripts, several of them contributing to pluripotency regulation [286, 444, 312]. Here, we demonstrate how ESRG [312], a human specific HERV-H-driven gene, has been incorporated into the regulatory network of self-renewal in EPI. We find EPI as a uniquely homogeneous cell population, being consistent with its self-renewal ability, an optimal candidate for *in vitro* culturing.

8.4 Primate evolution of pluripotency

Domestication of transposable elements largely contributes to the diversification process. Indeed, a clearly distinctive feature of human blastocyst compared to non-human primates is the presence of the plethora of HERV-H-driven or remodelled genes, especially in EPI and TE, but not in PE, where HERV-H is not expressed. The activity of HERV-H-enforced transcripts in distinct blastocyst lineages are driven by different LTR7 versions. The domesticated versions of HERV-H in different blastocyst lineages might reflect the complex interaction between the host and HERV-H during the arms race of repeated invasion-escape cycles.
Our analysis using primate iPSCs as models of the pluripotent epiblast reveals a rapid lineage specification that occurred following HERV-H invasion in new world monkeys. I show that the major gene expression gain (19) and loss (29) events in regulating pluripotency between human and new world monkeys (Callithrix) are due to the HERV-H-governed gene expression. Nevertheless, the major remodelling of human pluripotency by HERV-H has occurred quite recently, following the split form the ancestors of *Gorilla*s. In essence, the primate evolution of the blastocyst is a fine-tuning of selected cellular mechanisms, including the innate immune response, metabolism and viral defence. We speculate that certain features rendered HERV-H particularly suitable for being domesticated, such as its ability to generate a great degree of diversity via its heterogeneous transcripts, its flexibility to be tightly controlled

even in various cell types of the blastocyst or its potential capability to suppress other transposable elements. Similarly to HERV-H, a TE derived gene L1TD1 (Line1-derived) has been domesticated and assumed to have a role in both pluripotency and host defence. Also other ERVs, including HERV-K, contributed to the primate evolution of the blastocyst, although in a less dominant way. Furthermore, the human-specific expression of HERV-V and HERV-FRD1 derived envelop genes in the TE suggests a very recent co-option event. Notably, various human diseases are reportedly associated with the out-of-context expression of the freshly established, human specific regulatory circuitries and genes (UCA1, preeclampsia, cancer, etc). The most important message of the evolutionary studies is that pluripotency is primate-specific, and might be even specific to humans. We propose that human pluripotent cultures should reflect the species-specific features properly. By artificially inhibiting cellular mechanisms that were evolved to fine-tune the development of blastocyst, the human-specificity of regulatory networks results to be compromised.

8.5 Human-specific features of blastocyst

Our current study aims at deciphering the evolution of the blastocyst, with a special focus on pluripotency regulation in primates. In agreement with the hourglass model of development, we find that multiple features of early development are massively divergent between humans and non-human primates (NHPs). Nevertheless, despite of the general belief, we propose that the trajectories and the cell types in the blastocyst are basically conserved in primates. Our comparative single cell high-resolution analysis of human vs *Cynomolgus* blastocyst can clearly identify the human ICM (marked by NANOG, POU5F1 expression), breaking down the dogma that it does not exist in human. Following the identification of the blastocyst lineages, we established a precise atlas of lineage specific gene expression in human versus *Cynomolgus*. The least diverged lineage of the blastocyst is the primitive endoderm (PE). While the top three markers in both comparators are common, the pluripotent EPI, and TE appear to be the fastest evolving cell types in primates. By contrast, the underlying genetic architecture of the blastocyst is highly divergent between human and other primates, including apes. We found that such divergence is owed to rewired regulation of genes, remodelled genes and even novel human-specific genes. In the first scenario, the gene structure is conserved, but gene expression is shifted, due to a change in regulation (e.g. ATP12A). The shift could occur from one blastocyst lineage to another, but the affected gene can change expression even in a larger dimension (pre vs postimplantation) (e.g. GAGE13, KLK4). Second, the gene structure is slightly changed, resulting in gene expression in a different cell type (e.g. AluSx exonization in MAP4K1). Third, the gene model is robustly different, and it is not even obvious to predict the altered function of the affected gene (e.g. SCGB3A2, ABHD12B). Finally, we also found the examples of *de-novo* genes that are derived from transposable element (e.g. ESRG, UCA1).

8.6 Reactivation of retroelements in human early embryos

To preserve genomic stability, the repression of actively transposing retrotransposons by the host is selectively favored during embryogenesis [165]. Indeed, the host has evolved several layers of regulatory mechanisms to control TE activities. Transcription from a TE locus is epigenetically regulated by DNA methylation and histone modification. The KRAB-ZNF proteins can specifically recognize TE families and recruit the TRIM28/KAP1 repression complex and induce heterochromatin formation [103]. At the

portranscriptional level, TrEs can be also controlled by small RNA-mediated silencing or by APOBEC-mediated gene editing [93]. Although these regulatory mechanisms were established during the arms race between TrEs and the host [102], several of them are still functional (e.g. control the transcription). Initial stages of embryonic development are governed by maternal effects [299], with embryonic genome activation (EGA) in the human embryo occurring between the 4 and 8 cell stages of development, later than in mice (2 cell). Notably, a massive activation of TrEs in both species is observed at the switch from maternal to embryonic genome activation [263, 312, 427]. In human, DNA transposon-derived transcripts are relatively abundant in zygotes and 2-cells stage, but their levels, together with other phylogenetically Old (> 7 MY) TrEs, gradually decline as development proceeds. This might reflect decay of remnant RNAs expressed in oocytes. In human, the transcriptional activation of the young (< 7 MY) elements, including L1 (L1_Hs) and SVA (SVA_D, E and F), capable of retrotransposition is substantial from 8-cell stage, peaking at morula with a contrasting dynamic to DNMT3A and 3B, but declining in the blastocyst. Old LTR7-HERV-H peaking at the blastocyst is the only exception to the old/young difference, and curiously contrasts the expression pattern of young TrEs, including the mutagenic SVA or the human specific HERV-K_Hs elements.

8.7 HERV-H-enforced transcripts modulate throughout pre-implantation development

Regarding HERV-H, a key question is the extent to which HERV-H contributed to modulate human early development. HERV-H invaded the primate genome in the lineages post-dating the human-macaque common ancestor and so is quite specific to a crown group of primates. Its high expression levels in certain stages of early human embryos, specifically in pluripotent stem cells, is functionally relevant. Indeed, cells expressing from an HERV-H reporter gene form a mass of naive-like pluripotent stem cells. Notably, HERV-H elements are driven by several variants of its LTR, the large LTR7 group. While LTR7B,HERV-H elements peak at the eight,cell stage, LTR7,HERV-Hs are expressed in the blastocyst, and feature early passage, *in vitro* embryonic stem cell cultures [263]. The youngest subset, LTR7Y,HERV-Hs are expressed from the eight,cell to blastocyst-stage embryos. SVA and Line1 (L1) elements, still capable of retrotransposition in the human genome are also reactivated during early development. SVA is highly expressed, but restricted to a developmental window at morula stage [426]. The transcription level of primate,specific L1 reflects their evolutionary age: the younger L1s are more highly expressed than the older ones at the blastocyst stage [426].

Contrary to other TrEs, transcriptional activation of HERV-H (with various intensity) occurs throughout early human development. Based on the LTR type, the expression of LTR7, and LTR7Y,driven HERV-H peaks at the blastocyst, LTR7B-driven HERV-H is transcribed at the 8-cell stage, while HERV-H associated LTR7 sequence with a certain 38 bp deletion is activated even before the 8-cell stage of human preimplantation embryogenesis. Thus HERV-H, driven by distinct LTR variants, is expressed during the entire human preimplantation embryogenesis. Germline integration of a retrovirus is a prerequisite of a successful endogenization process. As pluripotent stem cells can contribute to the germline, an extension of activity to pluripotent stem cells might be also be directly beneficial to a retrovirus [312]. Apparently, following their inactivation as an endogenous retrovirus (ERV) [456], certain HERVH-derived transcripts have been gradually domesticated for modulating pluripotency [427]. Although domestication is an

exceptionally rare event, our current study suggests that HERVH might have been repeatedly co-opted. HERV-H-enforced gene structures are expressed in well-defined clusters, have incorporated into the developmental process throughout pre-implantation development. Indeed, HERV-H is central to many examples of new genes – chimeric, *de-novo* and remodelled genes (Figure 8.2). I show several examples, when the novel gene model is specific to us, humans, thus the HERV-H-enforced remodelling appears to underpin certain human specific aspects of early development. As the domestication process of HERV-H is relatively young, it remains to be established what proportion of the rich diversity of HERV-H-associated transcripts seen in early embryos are functional. Is there any rationale as to why HERV-H might be the TE that was recruited? We suggest that this might be owing to HERV-H's flexibility to be tightly controlled even in various stages of preimplantation development or its ability to generate a great degree of diversity via its heterogeneous transcripts, and perhaps to suppress Young TrEs. While, NCCs are marked by young TrEs (e.g. Line1, SVA), progenitor cells that passed quality control and continue to participate in the developmental program are characteristically express ancient, but dominantly full-length ERVs, including HERV-H. Curiously, HERV-H also selectively marks EPI vs PE, suggesting that HERV-H might contribute to cell fate determination during blastocyst formation in humans. EPI expresses a number of HERV-H-enforced transcripts, several of them contribute to regulate pluripotency [286, 444, 312]. Most remarkable of these is the human-specific protein-coding gene ESRG [312] that has been integrated into the regulatory network of human pluripotency. Here, we demonstrate how ESRG [212, 209], a human specific, HERV-H-driven gene has been incorporated into the regulatory network of self-renewal in EPI. We find EPI as a uniquely homogeneous cell population, being consistent with its self-renewal ability, an optimal candidate for *in vitro* culturing. In addition to ESRG, a few more HERV-H-enforced transcripts have been characterised and known to have functionality (e.g. lincROR, SCGB3A2) [286, 444]. Still, as the domestication process of HERV-H is relatively young, it remains to be established what proportion of the rich diversity of HERV-H-associated transcripts seen in early embryos are functional.

8.8 HERV-H expression in human pluripotent stem cells

In vitro models of human pluripotency maintenance and differentiation are known as pluripotent stem cells, including human embryonic stem cells (hESCs) [295] and induced pluripotent stem cells (hiPSCs) [328]. While hESC cultures are established from the pluripotent blastocysts, iPSCs are reprogrammed somatic cells that have regained a pluripotent state. Both, hESCs and hiPSCs are enriched for LTR7-driven HERV-H and associated transcripts (Figure 8.2) [312, 212, 209], indicating that HERV-H derived transcripts are hallmarks of human pluripotent stem cells (hPSCs). ERVs in human can be classified into many different families or groups, and independent members have the expected structure of an integrated retrovirus. Namely, independent ERVs have the canonical retroviral genes, gag, pol, and env, flanked by long terminal repeats (LTRs), which contain all the transcriptional regulatory signals necessary for retroviral expression. While a low number of HERV-H like elements occur in the New World monkeys, the major expansion of this family occurred in the Old World branch (Figure 1.4), and most HERV-H insertions in human have orthologous loci in Old World monkeys such as Rhesus macaque (Figure 1.6). Thus, over 80% of HERV-H integrations into the human genome occurred within the last 30MY (Figure 1.6). Notably, a common partially deleted form, which amplified to several hundred copies in the Old World monkeys, and which is associated primarily with LTRs annotated as LTR7 or LTR7B (originally termed Type I and Type II LTRs, [219, 220], was estimated by genomic Southern analysis to be present in

no more than 50 copies in two species of the New World monkeys [218]. A later expansion in hominoids of approximately 100 elements with a variant LTR (termed Type Ia and annotated LTR7Y in Repbase [7], has also been documented. In comparison with LTR7 or LTR7B, LTR7Y is represented by fewer copies in the human genome, and has higher promoter activity in transfection assays [219, 220]. HERV-H transcripts are highly abundant and account for 2% of total RNA in hPSCs [212] and are concentrated in the nucleus [209]. The transcription of HERV-H is supported by open chromatin, characterized by markers for transcriptionally active promoters like H3K4me1/2/3, H3K9ac, H3K27ac, H3K36me3 and H3K79me2 [212, 209]. By contrast, repressive chromatin marks of H3K27me3 and H3K9me3 at the transcriptionally active HERV-H loci are rare [212]. Thus, LTR7 has an active function as promoter and enhancer in hESCs. We have shown that HERV-H is regulated by key pluripotency factors that also regulate human embryonic development, and the extremely high levels of HERV-H RNA in hESC cells suggest that HERV-H contributes to pluripotency in human cells. The retention of the active chromatin structure recruited to the site of the integrated HERV-H provirus, may have acted as hot-spot within the germ-line of the primates. Now, after 30 years of its discovery [241], it is widely accepted that HERV-H RNA provides a specific marker for pluripotency in human cells.

8.9 HERV-H provides a platform to key pluripotency transcription factors

How is it that HERV-H is transcriptionally active in pluripotent stem cells? In the human genome, one fourth of the all the binding sites for transcription factors modulating pluripotency are provided by TrEs. Active HERV-H copies have binding sites for four key transcription factors driving pluripotency, such as OCT4, SOX2, LBP9 (TFCP2L1) and NANOG (Figure 8.2) [212, 209]. Specifically, NANOG has been shown to bind the 5'LTR of HERV-H, while OCT4 and SOX2 have binding sites in the HERV-H internal sequence towards its 5' end [212]. Functioning as a platform for alternative transcription factor binding sites and as long-range enhancers, LTR7/HERV-H is expected to affect the transcriptional regulatory network of pluripotency. In fact, HERV-H is indispensable for pluripotency maintenance, as knocking down of HERV-H in hPSCs results in loss of pluripotency [212, 209]. Curiously, the global knockdown effect of HERV-H and LBP9 are correlated, suggesting that the functions of LBP9 binding to HERV-H and modulating pluripotency might have evolved together [212]. Besides regulating pluripotency, the observed correlation argues for an additional function of LBP9 in retroviral control present only in primates (hence the name LTR-binding protein 9).

8.10 HERV-H domestication by host proteins containing krueppel domains

Indeed, the domestication process of transposable elements largely contributes to the diversification process. A clearly distinctive feature of human blastocyst compared to non-human primates is the presence of the plethora of HERVH-driven or remodelled genes, especially in EPI and TE, but not in PE, where HERVH is not expressed. The HERVH-enforced transcripts in distinct blastocyst lineages are driven by variants of LTR7. The various domesticated versions of HERVH might reflect the complex interaction between the host and HERVH during the arms race of repeated invasion-escape cycles. Apparently,

following their inactivation as an endogenous retrovirus, some HERVH-derived sequences have been gradually domesticated for cellular host functions. The domestication process of HERVH involved several host-encoded factors. Although HERVH is recognized by six different ZFPs, likely inherited from the arms race period, not all ZFPs are involved in recruiting the TRIM28/KAP1 repressor complex. Furthermore, KAP1 binds both active and inactive HERVH loci, suggesting that KAP1 does not confer absolute repression of HERVH. Besides ZFPs, HERVH provides a binding surface to various transcription factors, whose functions might even co-evolve with host defence and pluripotency (e.g. KLF4, LBP9). HERVH primarily contributes to the generation of lincRNA, but has a potency to alter host gene regulation or restructure a gene model. HERVH could also provide alternative transcription signals to generate various isoforms of a gene expressed in distinct developmental stages.

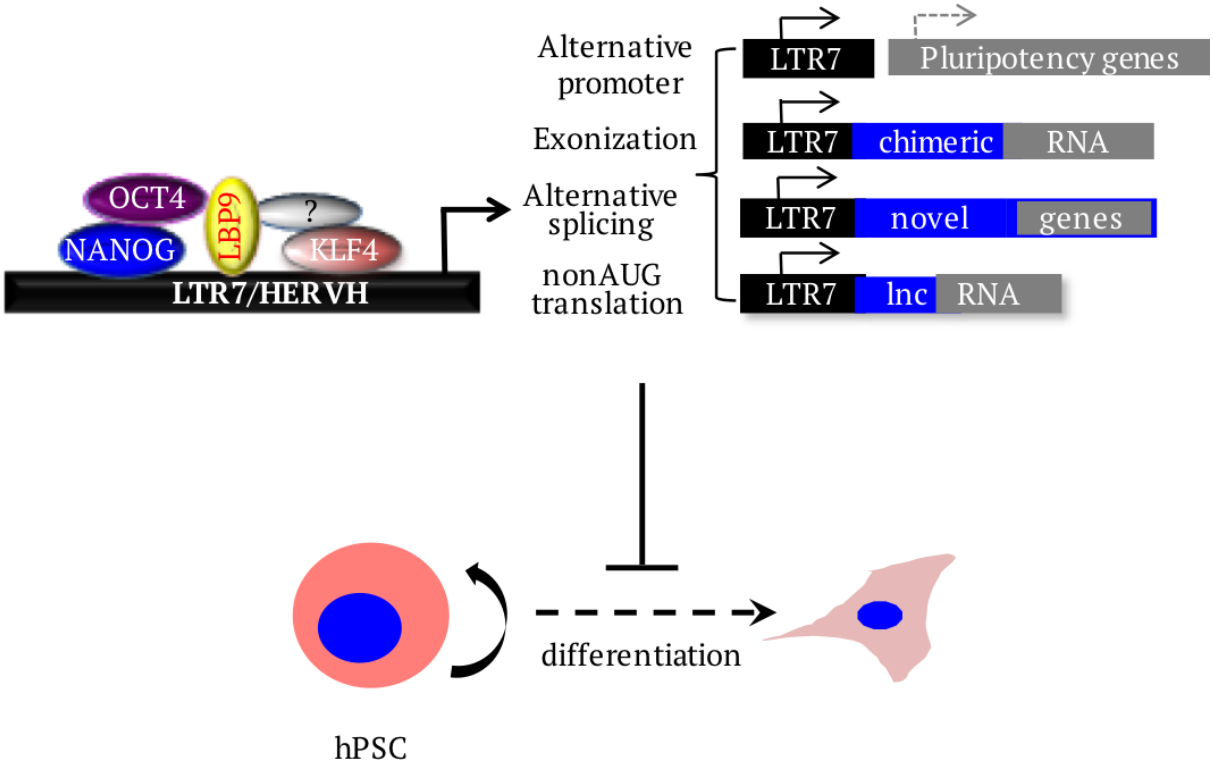


Figure 8.2:
LTR7/HERV-H clusters naive TF binding sites, NANOG, OCT4, KLF4, and LBP9. Transcription from LTR7/HERV-H forces diversification of transcripts in hPSCs. Activated HERV-Hs generate numerous novel, stem cell specific alternative gene products. HERV-H incorporates a set of regulatory lncRNAs into the network and defines novel pluripotent genes through alternative splicing or alternative non-AUG usage. HERV-H inhibits differentiation, while HERV-H-derived products contribute to maintain pluripotency. HERV-H regulation might have evolved in conjunction with host defense.

These host factors in conjunction with alternative transcription signals generate alternative transcripts from HERV-H. The utilization of the alternative transcriptional signals follows a clear developmental pattern. HERV-H primarily contributes to the generation of lincRNA, but has a potency to alter host gene regulation or restructure a gene. HERV-H could also provide alternative transcription signals to generate various isoforms of a gene expressed in distinct developmental stages. The list of characterised HERV-H-enforced transcripts with functionality is gradually growing (e.g. ESRG, lincROR, SCGB3A2, UCA1). We speculate that certain features rendered HERVH particularly suitable for being domesticated,

such as its ability to generate a great degree of diversity via its heterogeneous transcripts, its flexibility to be tightly controlled even in various cell types of the blastocyst or its potential capability to suppress other transposable elements. Similarly to HERVH, the Line1-derived domesticated L1TD1 gene has been also reported to have a role in both pluripotency and host defence [12]. Less dominantly, but HERV-K also contributed to the primate evolution of the blastocyst [263]. Furthermore, the human-specific expression of HERV-V and HERV-FRD1 derived envelop genes in TE suggest a recent co-option event. Basically, the primate evolution of the blastocyst is the fine-tuning of selected cellular mechanisms, including the innate immune response, metabolism and viral defence. Notably, various human diseases are reportedly associated with the out-of-context expression of the freshly established, human specific regulatory circuitries and genes (UCA1 in preeclampsia, cancer, etc)

8.11 HERV-H produces pluripotency-specific ncRNAs

How might HERV-H impact on pluripotency? Noncoding RNAs (ncRNAs) have been frequently found to be involved in regulating developmental processes including pluripotency and differentiation [444]. While TrEs contribute very little to protein coding sequences [282], they give rise to numerous ncRNAs, including long non-coding RNAs (lncRNAs), long intergenic non,coding RNAs (lincRNAs), microRNAs, etc [211]. Intriguingly, TrEs are integral to 83% of lincRNAs [282] and thus likely shaped their evolution. Once upregulated, HERV-H affects the expression of genes within a 40 kb window [134]. In hPSCs, HERV-H serves as a major source of alternative transcripts, regulating pluripotency. These HERV-H-derived transcripts were recruited in a primate or even human-specific manner. Transcriptionally active HERV-Hs are driving the production of numerous lncRNAs and lincRNAs. LTR7/HERV-H sequences alone account for more than 43% of hESC,specific lncRNA promoters and give rise to over a hundred (127) lincRNAs which are robustly and specifically transcribed in hPSCs. Transcription of HERV-H derived lincRNAs requires prior binding of SP1, OCT4 or NANOG to the 5'LTR of HERV-H. Curiously, HERV-H lncRNAs share a conserved core domain, which is capable of recruiting RNA-binding proteins, pluripotency factors and histone modifiers. As a scaffold, HERV-H,derived nuclear lncRNA interacts with pluripotency factors (e.g. OCT4) and transcriptional co,activators (e.g. p300, mediator subunits MED6 and MED12) [209]. This scaffold acts as a feedback regulator, modulates LTR7 enhancer function and the expression of neighbouring genes that are essential for hPSCs identity. An interesting example of HERV-H lincRNA is the linc-Regulator of Reprogramming (linc,RoR) that supports pluripotency by functioning as a miRNA sponge. Linc-RoR shares miRNA response elements with core pluripotency transcription factors OCT4, SOX2 and NANOG, thus protecting them from miRNA,mediated decay.

8.12 HERV-H produces chimeric and novel transcripts

Another means by which HERV-H regulates pluripotency maintenance in hPSCs is via the generation of various chimeric transcripts thereby combining cellular and viral sequences. HERV-H contains a conserved splice donor site that can connect the retroviral element with splice acceptor sites of cellular protein-coding genes. With a transcriptional start site (TSS) between the 5'LTR and HERV-H sequence, these chimeric transcripts often lack their 5' exon(s) of the canonical version, while part of HERV-H can be exonized. Examples of HERV-H enforced chimeras include SCGB3A2, RPL39L, NCR1 and KLKB1 [212]. We assume that the function of novel HERV-H-enforced chimeras may be related to the

original gene, but modified. However, in certain cases the LTR7/HERV-H enforced chimeric gene model is so robustly altered that it is not even possible to predict its new function. An interesting example of a HERV-H enforced novel transcript, part HERV-H, part host DNA, is ESRG that has a putative open reading frame (or frames) only in humans. ESRG is expressed in hPSCs and has been shown to promote the reprogramming process. A knockdown of ESRG hampers the self-renewal potential and pluripotency of hPSCs. The case of the ESRG gene is extreme as it consists almost entirely of repetitive sequences. The non-HERV-H sequence recruited is intronic DNA from the host gene within which ESRG resides. The ability of HERV-H to generate a great degree of diversity via its heterogeneous transcripts can in part explain why selection may have favoured the preservation of some HERV-H-associated transcripts. With such a diversity it is more likely that a few will evolve new functions in the cells in which they are easily expressed.

8.13 HERV-H promotes somatic cell reprogramming

Human somatic cells can be reprogrammed to become pluripotent via different routes. The classical way is the forced expression of pluripotency factors OCT4, SOX2, KLF4 and c-MYC (referred to as the OSKM factors) [328]. By providing specific binding sites for pluripotency factors LTR7/HERV-H is good source material to evolve functions that influence the process of somatic cell reprogramming. Indeed, exogenous NANOG can activate HERV-H transcription in fibroblasts and promote the acquisition of pluripotency in somatic cells. Furthermore, the ectopic expression of OCT4 would only increase reprogramming efficiency when certain HERV-H transcripts are present. Thus, besides serving as a binding platform for pluripotency specific transcription factors, HERV-H driven transcripts facilitate the hiPSC generation. HERV-H is expressed in distinct pattern through the various steps of reprogramming. HERV-H loci which is known to be promoting reprogramming starts expressing from third on-wards following the over-expression of OSKM. Intriguingly, around 300 loci of HERV-H not only express during maturation and stabilisation phase but drive the expression of their neighbour genes. This re-wires the existing transcriptional networks to promote and stabilise the human-specific reprogramming process. In consistency with the reverse mode of embryogenesis in reprogramming, many of these loci are also activated during preimplantation embryogenesis in reverse mode. Indeed, certain LTR7/HERV-H products (e. g ESRG, linc-RoR), implicated in promoting reprogramming are expressed during the process. Only when the reprogramming is completed and cells have acquired the pluripotent state do the levels of HERV-H and its flanking LTR7 drop to those observed in hESCs. Collectively, in addition to the role in supporting self-renewal and inhibiting differentiation, LTR7/HERV-H plays a role in pluripotency acquisition, and perhaps in stabilisation of the pluripotent state, underlining the various cellular functions of HERV-H that might affect pluripotency.

8.14 HERV-H as a marker of naive-like pluripotency

Ground-state or naive pluripotency is an ability of the pluripotent blastocyst for unbiased differentiation. During mammalian development, the epiblast cells in the inner cell mass (ICM) of blastocysts enter the developmental “ground state”, and these pluripotent cells can give rise to all somatic lineages and the germline [335, 336, 337, 338]. Though the ground state pluripotency in the embryo is a transient condition, in principle, its nature can be captured indefinitely *in vitro*, through derivation of embryonic stem

cells (ESCs) from the ICM. Naturally, one would expect that a mechanism as important as pluripotency would be conserved in different species of mammals. However, it does not seem to be the case. Thus far, long-term, ground state naive cell cultures could be successfully established only in mice and rats [307, 295, 309, 308, 346, 296, 314]. Mouse naive cultures express naive transcription factors fairly homogenously, and maintain their resemblance to the inner cell mass (ICM) long term. Like their mouse counterparts, human ESCs (hESCs) can be also differentiated into three germ layers both *in vitro* and *in vivo*. However, hESCs are more similar to primed mouse epiblast stem cells than to mESCs [335, 357, 362, 366]. Most problematically, the expression profiles of hESC lines are heterogeneous, and are further from those of cells in the ICM. Over the past decade considerable efforts have been devoted to generate or identify the "holy grail" of the field, the human ground state, naive stem cells [336, 337, 377, 378, 379]. In the last few years, several naive-like lineages have been generated, mostly by overexpressing key TFs, and/or improving culture conditions using certain chemical compounds [380, 317, 357, 367, 359, 361, 366]. Perhaps significantly, a recent study implicates HERV-H as having a potential role in defining naive-like stem cells. Notably, an LTR7/HERV-H-based GFP reporter system can be used to track and enrich human pluripotent cultures that express HERV-H [212]. These HERV-H-marked cells resemble the very early, pluripotent state, forming dome shaped colonies, and they uniformly express naive pluripotent markers, similarly to mouse naive cells. Importantly, naive cells are naturally present as a small fraction in almost all embryonic stem cells. A specific HERV-H expression pattern characteristic of human pluripotent stem cells recapitulates features of the ICM of blastocyst. High-levels of HERV-H expression not only mark cells in a naive-like state, but apparently play a role in maintenance of this state, while inhibiting differentiation *in vitro*. However, in hESCs, both the number and the intensity of HERV-H-expression are higher in cultured conditions when compared to ICM. In fact, too high level of HERV-H expression in the ICM may inhibit differentiation. Indeed, high level of expression from LTR7/HERV-H is associated with a reversible differentiation failure phenotype in human induced pluripotent stem cells (hiPSCs) [208]. However, by tuning down HERV-H expression the cells differentiate normally. The lack of diapause behavior in human embryos suggests that hyperactivation of HERV-H/LTR7 after the ICM stage would not be adaptive. Reporters, such as the LTR7-GFP system, are powerful tools not only in derivation, but also in optimization of naive-like hPSC culture conditions. The next, so far unresolved, task is to maintain these cells in long-term cultures that allow for their proliferation. In vitro hPSC cultures consist of a heterogeneous cell population with only around 4% naive-like cells exhibiting naive-like pluripotency. Importantly, these naive-like cells can be identified based on elevated LTR7/HERV-H expression compared to primed hESCs. Higher expression levels of HERV-H in naive-like hPSCs have been associated with a subset of HERV-H possessing a binding site for the pluripotency transcription factor LBP9. This initiates transcription of specific lncRNAs and chimeric transcripts as a part of the primate-specific pluripotency network. Others have reported LTR7Y/HERV-H, HERV-K or SVA as being a more precise marker for naive-like hPSCs. This discrepancy could be explained with the existence of various naive-like cell lines. Which naive-like cell line mimics real developmental conditions most precisely, remains currently a matter of debate. Importantly, however, SVA is potentially mutagenic, and in contrast to HERV-H, to date there is no evidence of HERV-K function in pluripotency.

8.15 Degree of human-specificity in naive pluripotent states

We pushed this frontier little further to investigate the mode of sequence-conserved genes changing their expression profile during the evolution of pluripotency. We hypothesized that numerous key genes might be remodelled or new genes or chimeric genes are involved in driving the evolution of pluripotency. Furthermore, given the involvement of HERV-H, we ask how important TrEs are in making the difference between species and, more particularly, whether HERV-H is in any manner exceptional as it is in human pluripotency [212, 312, 208, 209]. Moreover, we also attempted to dissect the fast-evolving biological processes and molecular functions that are coupled with HERV-H.

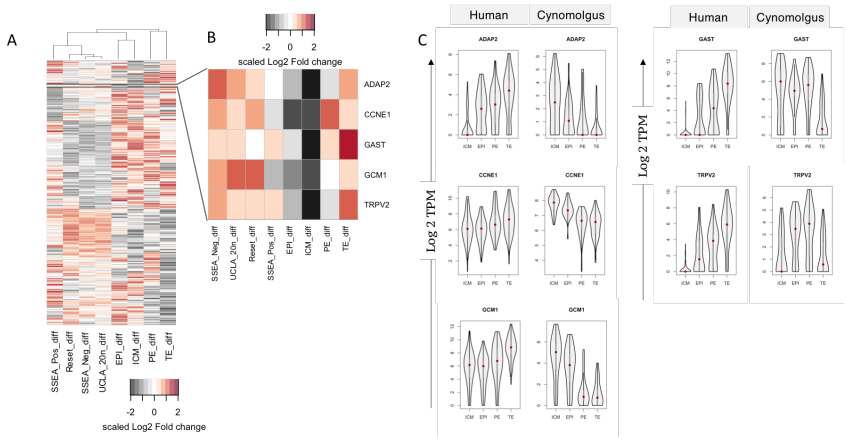


Figure 8.3:
A. Heatmap showing the row-wise Z-score of log2 Fold change of most variable genes (n=948) that defined human and cynomolgus lineage segregation (Fig 1A) in forced naive cells versus their respective primed cells along-with human blastocyst lineages against their counterparts in cynomolgus. Red color shows upregulation of gene and black corresponds to down-regulation. Clustered dendrogram represents spearman cor-relation and euclidian distance. Note the contrasting pattern of genes between Naive pluripotent states (SSEA-neg, Reset and UCLA-20n cells) and human Native pluripotent states (ICM and EPI) compared with their counter samples. and Star marked KZFPs (one in each heatmap) are shown in further section
B. Heatmap showing the row-wise Z-score of log2 Fold change of selected five genes from previous heatmap whose higher expres-sion marks pluripotent states in cynomolgus but TE populations in human and they are upregulated in forced naive cells compared with their primed counterparts.
C. Multiple violin plots visualize the density and distribution of gene expression mentioned in previ-ous figure that is highly expressed in TE lineage of human blastocysts but in the case of cynomolgus blastocyst, it is enriched in EPI/ICM at higher level.

These questions are in principle resolvable by the strategical analysis of the plethora of single cell and bulk transcriptomic data. In addition, we employed transcriptome analysis of primate pluripotent stem cells (PSCs) to dissect how the transcriptional networks of pluripotency has been re-wired to exist in humans. The most important message of our evolutionary studies is that certain components of human pluripotency is primate specific, and few amongst them might be even specific to humans. Together with trophectoderm, the pluripotent epiblast is the fastest evolving cell type of the blastocyst. While HERV-H-enforced remodelling appears to underpin much of the human specificity, the major HERV-H-driven remodelling of pluripotency has occurred quite recently, following the split of the Gorilla-human common ancestor.

We propose that human *in vitro* pluripotent stem cell cultures should reflect the human-specific features properly. These include the expression of the remodelled genes that have been contributed to fine-tune pluripotency regulation in humans. However, we find that many *in vitro* naive cultures are heterogeneous

and are both evolutionary and developmentally "confused". Indeed, under *in vitro* naive conditions the transcriptomes display similarity to various cell types of *Cynomolgus* pre-implantation pluripotent cells instead of human ones. Our list of the human pluripotent EPI could help researchers to establish culture conditions, where the human specific features of the pluripotent cells, capable of self-renewal, are properly expressed. Moreover, we have also shown the similarity and dissimilarity of each lineages of blastocysts of human and *Cynomolgus* based on orthologous and human-specific genes independently. Finally, we characterized the behaviour of conserved and remodelled biological process *in-silico* upon the reversion of primed cells to claimed naive cells, from an evolutionary point of view in order to set-up another reliable layer in the regenerative medicine grounds.

8.16 Lessons for *in vitro* cell lines

Our analysis holds lessons for establishing self-renewing, pluripotent stem cell cultures *in vitro* (e.g. naive-like). The human pluripotent EPI forms a relatively homogenous cell cluster, has self-renewal capacity (Figure 3.1B and 5.6) and attenuated young TE activity, potentially making a good choice for naive-like culturing. By contrast, both morula and NCC appear to be transitory labile stages. In morula-like naive cells, hypomethylation of imprinted genes is not reversible, and young, potentially transposing, mobile elements have peak transcriptional activity [366, 356]. The boost of TE activity might determine the fate of the embryo, whether it proceeds to the blastocyst stage or is selected out in a process reminiscent of attrition in fetal oocytes, involving programmed cell death. Altogether, a long-term maintenance of cells in a rather unsettled and potentially mutagenic phase would make morula/ICM a poor target to mimic.

Surprisingly, many of the forced naive-like stem cells resemble morula stage [366]. Our analysis on existing human naive-like cultures revealed that these lines are highly heterogeneous, consisting of a mixture of cells of various identity (both naturally existing or artificial). While being artificial is not problematic in itself (e.g. these cells might be easily maintained, etc.), nevertheless, they should mimic early human development. In order to transform the alternative cell types for therapeutic application, it is crucial to discern whether potentially damaging transposition is happening or not. It should also be of concern that these cells maintain expression of certain genes that are not expressed during pre-implantation embryogenesis at all, since they might possibly interfere with proper implantation, regulation of imprinting, maintenance of genome stability.

By contrast, the strategy of enriching human cultures for naive-like human pluripotent stem cells (rather than engineering) based on LTR7/HERV-H seems to be feasible, although HERV-H levels in these cells exceeds those observed in EPI. Conditions enabling long-term culturing remain to be solved. HERV-H marked cells are also not uniform, therefore it might be necessary to distinguish between different HERV-H grouped cells. We see a dynamics of different subsets during development or reprogramming. If HERV-H remodelled network is ignored, we ignore the evolutionary selection of networks that was fine tuning the pre-implantation development in humans.

Why do some naive-like cell types resemble morula/ICM? We can speculate that the cells have been forced, either through the genetic manipulation or culture conditions, into a rather artificial, cell type non-existent during normal development. Indeed, one might consider that as it has proven hard to define the key transitory phases, it is also quite challenging to accurately set the transcriptome of the cells. Our checklist also highlights the necessity to consider the human-specific features of human development in

naive stem cells culturing. Forced expression of a transcription factor that does not match the correct developmental stage (e.g. KLF2) could generate artificial cell types, possibly due to inhibition of pathways that define human-specificity. The human-specific checklist should help us to improve human naive stem cell derivation and culturing, as well as our understanding of early human development and its potential aberrations.

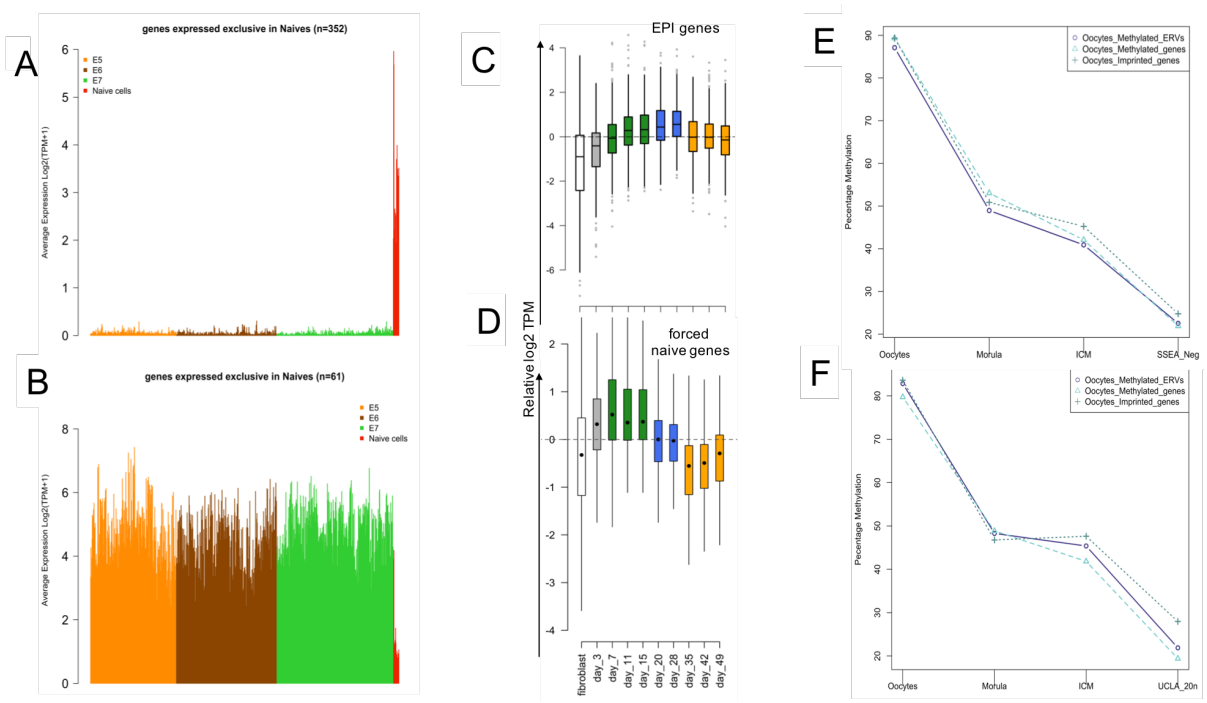


Figure 8.4:
A. Barplot showing the mean expression of 352 genes expressed in all analyzed naive/naive-like cells, but are not expressed in any cells during human preimplantation embryonic development.
B. Barplot showing the mean expression of 67 genes expressed in each single cell during human preimplantation embryonic development (essential genes), but are not expressed in any of the analysed naive/naive-like cells **C**. Boxplot shows the dynamic expression of 311 EPI-marker genes expressed during the reprogramming process. Note that the expression of EPI genes plateaus in the 'stabilisation' stage with p-value < 0.0005. **D**. Boxplot shows the dynamic expression of 242 upregulated genes over-expressed during the reverting primed to naive process (commonly upregulated in all naive cells compared with their respective primed cells). Note that the expression of naive genes rises and plateau in the 'maturation stage' stage with p-value < 0.0025. **E-F**. (Upper panel) The average level of DNA methylation in morula, ICM and UCLA20n naive cells over the TSS of Young ERVs and imprinted genes hypermethylated in oocytes. (Lower panel) Same as in the upper panel, but for the naive cell type SSEA-Neg 6iL.

Summary

Human genome has been expanded to approximately 8% by multiple waves of retroviral invasions, followed by lateral expansion, mainly in germ cells. Recent advancements provide substantial evidence that few families of Human Endogenous Retro-Viruses (HERVs) are co-opted to regulate cell-type specific physiological networks. Following up, I show that diverse modes of HERV-H activity have played a pivotal role in the evolution and development of human-specific embryogenesis including pluripotency. HERV-H provides binding sites for pluripotency transcription factors and serves as a major source for ncRNAs and chimeric transcripts in hPSCs. The interplay of host kruepel box proteins with HERV-H transcription determines the stability of human pluripotency. We updated the transcriptomic and transposcriptomic atlas of Human Preimplantation Embryogenesis (HPE) that re-define the progression of HPE at single-cell resolution. I characterize an unattended cell population in HPE that did not commit to any of the known lineages to form a stable preimplantation blastocyst. ICM being considered as reference frame for the purity of hESCs in naive/primed state conflict, we redefine ICM which adds another layer to the understanding of basic and regenerative biology. Committed and non-committed cells are distinguished by the reactivation of full-length HERVs and mutagenic retrotransposons respectively. I show that HERV-H transcripts control younger mutagenic elements with the mechanisms not known to us yet. Additionally, we also show that HERV-H expression is also prominent during fertilization and early stage of embryogenesis. We find the contrasting pattern of distinct loci of HERV-H during HPE, few of them driving the stage-specific markers that might determine the commitment of host cells to pluripotent lineages. Upon comparing the embryonic lineages and pluripotent states between primates, we decipher the probable mechanisms shaping the human-specific nature of embryogenesis. Much divergence within primate early development is owed to *de-novo* genes and chimeric transcripts that were remodeled by HERV-H. Moreover, HERV-H orchestrates the human-specific role in stabilizing self-renewal/pluripotency during HPE as inferred from transcriptome-wide cross-species comparison of primate pluripotent cells. Nevertheless, the HERV-H-derived regulatory network has been incorporated, and appears to be an essential feature of human pluripotency in native state. I show that HERV-H derived genes distinguishes the different pluripotent states in human embryos which plays pivotal role in the self-renewal during embryonic development and somatic reprogramming too. I show that the human pluripotency is required to maintain human-specificity *in vitro* to concur the native biological states with high fidelity. Finally, I show that HERV-H displays human-specific cross-talk with host-factors in distinct developmental lineages, including *in-vivo* and *in vitro* human pluripotent states. Finally our check-list of markers would be advantageous in order to sort *in vitro* pluripotent cells resembling *in-vivo* cells.

Zusammenfassung

Das menschliche Genom hat durch mehrere Wellen von retroviraler Invasion eine Ausdehnung von 8% erfahren, gefolgt von einer lateralen Ausdehnung hauptsächlich in Keimzellen. Neuere Forschung hat ergeben, dass nur wenige Familien der Humanen Endogenen Retro-Viren (HERV) kooptiert werden, um zelltypenspezifische physiologische Netzwerke zu regulieren. Davon ausgehend zeigen wir, dass verschiedene Modi der HERV-H-Aktivität eine herausragende Rolle in der Evolution und Entwicklung der humanspezifischen Embryogenese, einschließlich der Pluripotenz, innehaben. HERV-H liefert Andockstellen für Pluripotenz-Transkriptions-Faktoren und dient als eine Quelle für ncRNAs und chimäre Transkripte in hPSCs. Das Zusammenspiel der Krüppel-Box-Proteine des Wirts mit der HERV-H-Transkription determiniert die humane Pluripotenz. Wir haben den transkriptomischen Atlas der humanen Präimplantations-Embryogenese (HPE) aktualisiert, was den Fortschritt der HPE auf Ebene der Einzelzellen-Resolution neu definiert. Wir beschreiben eine bisher vernachlässigte Zellpopulation bei der HPE, die sich bei der Formung einer stabilen Präimplantations-Bastocyste für keine der bekannten Zelltypen determinieren ließ. Während ICM als Referenzrahmen für die Reinheit von hESCs einen Konflikt von naivem und präpariertem Zustand konstatiert, definieren wir ICM neu, was für die regenerative Biologie eine weitere Ebene eröffnet. Determinierte und nicht-determinierte Zellen unterscheiden sich durch die Reaktivierung von sowohl langen HERVs als auch mutagenen Retrotransposons. Wir zeigen, dass HERV-Transkripte jüngere mutagene Elemente kontrollieren. Darüber hinaus zeigen wir, dass der HERV-H-Ausdruck auch während der Befruchtung und der frühen Embryogenese prominent sind. Wir haben ein kontrastierendes Muster bestimmter loci von HERV-H während der HPE gefunden, wovon wenige die phasenspezifischen Marker antreiben, welche die Determinierung der Wirtszellen hin zu pluripotenten Zelltypen bestimmt. Beim Vergleich von embryonischen Zelltypen und pluripotenten Phasen bei Primaten haben wir den möglichen Mechanismus entschlüsselt, der die humanspezifische Natur der Embryogenese hervorbringt. Einige Divergenz innerhalb der frühen Entwicklung der Primaten verdankt sich de-novo-Genen und chimären Transkripten, die von HERV-H umgestaltet wurden. Vielmehr gestaltet HERV-H die humanspezifische Rolle bei der Selbsterneuerung/Pluripotenz während der HPE, wie Rückschlüsse vom transkriptomweiten, speziesübergreifenden Vergleich von pluripotenten Zellen bei Primaten zulassen. Dennoch wurde das HERV-H-abhängige regulierende Netzwerk inkorporiert und scheint für den ursprünglichen Zustand der humanen Pluripotenz essentiell zu sein. Ich zeige, dass HERV-H eine humanspezifische Rolle innehat bezüglich der Selbsterneuerung/Pluripotenz sowohl während der embryonischen Entwicklung als auch der somatischen Reprogrammierung. Des Weiteren zeige ich, dass humane Pluripotenz notwendig ist für die Humanspezifizität in vitro, um in hoher Genauigkeit

kongruent mit den ursprünglichen biologischen Phasen zu sein. Schließlich zeige ich, dass HERV-H einen humanspezifischen Dialog mit den Wirtfaktoren in bestimmten entwicklungsgemäßen Zelltypen aufweist, wie u.a. in in-vivo und in vitro pluripotenten Zuständen, und somit eine verlässliche Grundlage für klinische Anwendungen liefert.

Bibliography

[1] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[2] Prescott L Deininger, John V Moran, Mark A Batzer, and Haig H Kazazian. Mobile elements and mammalian genome evolution. *Current opinion in genetics & development*, 13(6):651–658, 2003.

[3] Christian Biémont and Cristina Vieira. Genetics: junk dna as an evolutionary force. *Nature*, 443(7111):521–524, 2006.

[4] Bartley G Thornburg, Valer Gotea, and Wojciech Makałowski. Transposable elements as a significant source of transcription regulating signals. *Gene*, 365:104–110, 2006.

[5] Jean-Nicolas Volff. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays*, 28(9):913–922, 2006.

[6] Astrid Böhne, Frédéric Brunet, Delphine Galiana-Arnoux, Christina Schultheis, and Jean-Nicolas Volff. Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Research*, 16(1):203–215, 2008.

[7] Weidong Bao, Kenji K Kojima, and Oleksiy Kohany. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6(1):11, 2015.

[8] B McClintock. The significance of responses of the genome to challenge. *Science*, 226(4676):792–801, 1984.

[9] Barbara McClintock. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, 36(6):344–355, 1950.

[10] Roy J. Britten and Eric H. Davidson. Gene regulation for higher cells: A theory. *Science*, 165(3891):349–357, 1969.

[11] Phillip SanMiguel, Alexander Tikhonov, Young-Kwan Jin, Natasha Motchoulskaia, Dmitrii Zakharov, Admasu Melake-Berhan, Patricia S Springer, Keith J Edwards, Michael Lee, Zoya Avramova, et al. Nested retrotransposons in the intergenic regions of the maize genome. *Science*, 274(5288):765–768, 1996.

[12] Edward B Chuong, Nels C Elde, and Cédric Feschotte. Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics*, 2016.

[13] Barbara McClintock. Controlling elements and the gene. In *Cold Spring Harbor symposia on quantitative biology*, volume 21, pages 197–216. Cold Spring Harbor Laboratory Press, 1956.

[14] Roy J Britten and Eric H Davidson. Gene regulation for higher cells: a theory. *Science*, 165(3891):349–357, 1969.

[15] W Ford Doolittle and Carmen Sapienza. Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284(5757):601–603, 1980.

[16] Leslie E Orgel, FHC Crick, and C Sapienza. Selfish dna. *Nature*, 288(5792):645–646, 1980.

[17] Rita Rebollo, Mark T Romanish, and Dixie L Mager. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annual review of genetics*, 46:21–42, 2012.

[18] Grindley NDF. Craig nl, craigie r, geller m, lambowitz am. mobile dna ii, 2002.

[19] John K Pace and Cédric Feschotte. The evolutionary history of human dna transposons: evidence for intense activity in the primate lineage. *Genome research*, 17(4):422–432, 2007.

[20] Bethaney J Vincent, Jeremy S Myers, Huei Jin Ho, Gail E Kilroy, Jerilyn A Walker, W Scott Watkins, Lynn B Jorde, and Mark A Batzer. Following the lines: an analysis of primate genomic variation at human-specific line-1 insertion sites. *Molecular biology and evolution*, 20(8):1338–1348, 2003.

[21] Cédric Feschotte and Ellen J Pritham. Dna transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.*, 41:331–368, 2007.

[22] Cédric Feschotte, Lakshmi Swamy, and Susan R Wessler. Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (mites). *Genetics*, 163(2):747–758, 2003.

[23] Guojun Yang, Dawn Holligan Nagel, Cédric Feschotte, C Nathan Hancock, and Susan R Wessler. Tuned for transposition: molecular determinants underlying the hyperactivity of a stowaway mite. *science*, 325(5946):1391–1394, 2009.

[24] Vladimir V Kapitonov and Jerzy Jurka. Rolling-circle transposons in eukaryotes. *Proceedings of the National Academy of Sciences*, 98(15):8714–8719, 2001.

[25] Ivana Grabundzija, Simon A Messing, Jainy Thomas, Rachel L Cosby, Ilija Bilic, Csaba Miskey, Andreas Gogol-Döring, Vladimir Kapitonov, Tanja Diem, Anna Dalda, et al. A helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. *Nature communications*, 7, 2016.

[26] Ellen J Pritham, Tasneem Putliwala, and Cédric Feschotte. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to dna viruses. *Gene*, 390(1):3–17, 2007.

[27] Hugh M Robertson. Evolution of dna transposons in eukaryotes. In *Mobile DNA ii*, pages 1093–1110. American Society of Microbiology, 2002.

[28] Yongming Wang, Diana Pryputniewicz-Dobrinska, Enikö Éva Nagy, Christopher D Kaufman, Manvendra Singh, Steve Yant, Jichang Wang, Anna Dalda, Mark A Kay, Zoltán Ivics, et al. Regulated complex assembly safeguards the fidelity of sleeping beauty transposition. *Nucleic acids research*, 45(1):311–326, 2017.

[29] Jason T Huff, Daniel Zilberman, and Scott W Roy. Mechanism for dna transposons to generate introns on genomic scales. *Nature*, 538(7626):533–536, 2016.

[30] Todd A Schlenke and David J Begun. Strong selective sweep associated with a transposon insertion in drosophila simulans. *Proceedings of the National Academy of Sciences of the United States of America*, 101(6):1626–1631, 2004.

[31] Paul N Nelson, PR Carnegie, J Martin, H Davari Ejtehadi, Paul Hooley, D Roden, S Rowland-Jones, Phil Warren, J Astley, and Paul G Murray. Demystified... human endogenous retroviruses. *Molecular Pathology*, 56(1):11, 2003.

[32] Patric Jern and John M Coffin. Effects of retroviruses on host genome function. *Annual review of genetics*, 42:709–732, 2008.

[33] Jonathan P Stoye. Endogenous retroviruses: still active after all these years? *Current biology*, 11(22):R914–R916, 2001.

[34] Arian FA Smit. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current opinion in genetics & development*, 9(6):657–663, 1999.

[35] Emma Whitelaw and David IK Martin. Retrotransposons as epigenetic mediators of phenotypic variation in mammals. *Nature genetics*, 27(4):361–365, 2001.

[36] Jennifer F Hughes and John M Coffin. Human endogenous retrovirus k solo-ltr formation and insertional polymorphisms: implications for human and viral evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 101(6):1668–1672, 2004.

[37] Robert Belshaw, Jason Watson, Aris Katzourakis, Alexis Howe, John Woolven-Allen, Austin Burt, and Michael Tristem. Rate of recombinational deletion among human endogenous retroviruses. *Journal of virology*, 81(17):9437–9442, 2007.

[38] Catherine A Dunn, Mark T Romanish, Leanne E Gutierrez, Louie N van de Lagemaat, and Dixie L Mager. Transcription of two human genes from a bidirectional endogenous retrovirus promoter. *Gene*, 366(2):335–342, 2006.

[39] Amit Kapoor, Peter Simmonds, and W Ian Lipkin. Discovery and characterization of mammalian endogenous parvoviruses. *Journal of virology*, 84(24):12628–12635, 2010.

[40] C Leib-Mosch, W Seifarth, and U Schon. Influence of human endogenous retroviruses on cellular gene expression. *Retroviruses and Primate Genome Evolution: Landes Bioscience*, 2005.

[41] Vladimir V Kapitonov and Jerzy Jurka. The long terminal repeat of an endogenous retrovirus induces alternative splicing and encodes an additional carboxy-terminal sequence in the human leptin receptor. *Journal of molecular evolution*, 48(2):248–251, 1999.

[42] Nancy A Jenkins, Neal G Copeland, Benjamin A Taylor, and Barbara K Lee. Dilute (d) coat colour mutation of dba/2j mice is associated with the site of integration of an ecotropic mulv genome. *Nature*, 293(5831):370–374, 1981.

[43] Zoltán Ivics. Genomic parasites and genome evolution. *Genome biology*, 10(4):306, 2009.

[44] Norbert Bannert and Reinhard Kurth. Retroelements and the human genome: new perspectives on an old relation. *Proceedings of the National Academy of Sciences*, 101(suppl 2):14572–14579, 2004.

[45] Timothy J Crow. Left brain, retrotransposons, and schizophrenia. *British medical journal (Clinical research ed.)*, 293(6538):3, 1986.

[46] Timothy J Crow. Schizophrenia as the price that homo sapiens pays for language: a resolution of the central paradox in the origin of the species. *Brain research reviews*, 31(2):118–129, 2000.

[47] Brian A Mozer and Seymour Benzer. Ingrowth by photoreceptor axons induces transcription of a retrotransposon in the developing drosophila brain. *Development*, 120(5):1049–1058, 1994.

[48] Timothy J Crow. A re-evaluation of the viral hypothesis: is psychosis the result of retroviral integration at a site close to the cerebral dominance gene? *The British Journal of Psychiatry*, 145(3):243–253, 1984.

[49] S Haahr, M Sommerlund, Tove Christensen, AW Jensen, HJ Hansen, and A MØLLER-LARSEN. A putative new retrovirus associated with multiple sclerosis and the possible involvement of epstein-barr virus in this disease. *Annals of the New York Academy of Sciences*, 724(1):148–156, 1994.

[50] Girolama La Mantia, Domenico Maglione, Gina Pengue, Antonio Di Cristofano, Antonio Simeone, Luisa Lanfrancone, and Luigi Lania. Identification and characterization of novel human endogenous retroviral sequences prefentially expressed in undifferentiated embryonal carcinoma cells. *Nucleic acids research*, 19(7):1513–1520, 1991.

[51] Sha Mi, Xinhua Lee, Xiang-ping Li, Geertruida M Veldman, Heather Finnerty, Lisa Racie, Edward Lavallie, Xiang-Yang Tang, Philippe Edouard, Steve Howes, et al. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*, 403(6771):785–789, 2000.

[52] Paromita Deb-Rinker, Timothy A Klempan, Richard L O'Reilly, E Fuller Torrey, and Shiva M Singh. Molecular characterization of a msrv-like sequence identified by rda from monozygotic twin pairs discordant for schizophrenia. *Genomics*, 61(2):133–144, 1999.

[53] Dimitri Lavillette, Mariana Marin, Alessia Ruggieri, François Mallet, François-Loïc Cosset, and David Kabat. The envelope glycoprotein of human endogenous retrovirus type w uses a divergent family of amino acid transporters/cell surface receptors. *Journal of virology*, 76(13):6442–6452, 2002.

[54] Kai W Wucherpfennig and Jack L Strominger. Molecular mimicry in t cell-mediated autoimmunity: viral peptides activate human t cell clones specific for myelin basic protein. *Cell*, 80(5):695–705, 1995.

[55] Michael C Levin, Sang Min Lee, Franck Kalume, Yvette Morcos, F Curtis Dohan, Karen A Hasty, Joseph C Callaway, Joseph Zunt, Dominic M Desiderio, and John M Stuart. Autoimmunity due to molecular mimicry as a cause of neurological disease. *Nature medicine*, 8(5):509–513, 2002.

[56] James B Johnston, Claudia Silva, Janet Holden, Kenneth G Warren, Arthur W Clark, and Christopher Power. Monocyte activation and differentiation augment human endogenous retrovirus expression: implications for inflammatory brain diseases. *Annals of neurology*, 50(4):434–442, 2001.

[57] Akio Takahashi, Noorbibi K Day, Voravich Luangwedchakarn, Robert A Good, and Soichi Haraguchi. A retroviral-derived immunosuppressive peptide activates mitogen-activated protein kinases. *The Journal of Immunology*, 166(11):6771–6775, 2001.

[58] Bernard Conrad, Richard Nicolas Weissmahr, Jürg Böni, Rosanna Arcari, Jörg Schüpbach, and Bernard Mach. A human endogenous retroviral superantigen as candidate autoimmune gene in type i diabetes. *Cell*, 90(2):303–313, 1997.

[59] Marlies Sauter, Klaus Roemer, Barbara Best, Matthias Afting, Stefanie Schommer, Gerhard Seitz, Michael Hartmann, and Nikolaus Mueller-Lantzsch. Specificity of antibodies directed against env protein of human endogenous retroviruses in patients with germ cell tumors. *Cancer research*, 56(19):4362–4365, 1996.

[60] Carl W Schmid. Alu: structure, origin, evolution, significance, and function of one-tenth of human dna. *Progress in nucleic acid research and molecular biology*, 53:283–319, 1996.

[61] Wojciech Makalowski. Genomic scrap yard: how genomes utilize all that junk. *Gene*, 259(1):61–67, 2000.

[62] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

[63] Prescott L Deininger and Mark A Batzer. Alu repeats and human disease. *Molecular genetics and metabolism*, 67(3):183–193, 1999.

[64] Xiang H-F Zhang and Lawrence A Chasin. Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *Proceedings of the National Academy of Sciences*, 103(36):13427–13432, 2006.

[65] Noa Sela, Britta Mersch, Nurit Gal-Mark, Galit Lev-Maor, Agnes Hotz-Wagenblatt, and Gil Ast. Comparative analysis of transposed element insertion within human and mouse genomes reveals alu's unique role in shaping the human transcriptome. *Genome biology*, 8(6):R127, 2007.

[66] Rotem Sorek. The birth of new exons: mechanisms and evolutionary consequences. *Rna*, 13(10):1603–1608, 2007.

[67] Galit Lev-Maor, Oren Ram, Eddo Kim, Noa Sela, Amir Goren, Erez Y Levanon, and Gil Ast. Intronic alus influence alternative splicing. *PLoS genetics*, 4(9):e1000204, 2008.

[68] Kathi Zarnack, Julian König, Mojca Tajnik, Iñigo Martincorena, Sebastian Eustermann, Isabelle Stévant, Alejandro Reyes, Simon Anders, Nicholas M Luscombe, and Jernej Ule. Direct competition between hnrnp c and u2af65 protects the transcriptome from the exonization of alu elements. *Cell*, 152(3):453–466, 2013.

[69] Andranik Ivanov, Sebastian Memczak, Emanuel Wyler, Francesca Torti, Hagit T Porath, Marta R Orejuela, Michael Piechotta, Erez Y Levanon, Markus Landthaler, Christoph Dieterich, et al. Analysis of intron sequences reveals hallmarks of circular rna biogenesis in animals. *Cell reports*, 10(2):170–177, 2015.

[70] Haig H Kazazian and John V Moran. The impact of l1 retrotransposons on the human genome. *Nature genetics*, 19(1):19–24, 1998.

[71] R Keith Slotkin and Robert Martienssen. Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, 8(4):272–285, 2007.

[72] Dongmei D Luan, Malka H Korman, John L Jakubczak, and Thomas H Eickbush. Reverse transcription of r2bm rna is primed by a nick at the chromosomal target site: a mechanism for non-ltr retrotransposition. *Cell*, 72(4):595–605, 1993.

[73] Marie Dewannieux, Cécile Esnault, and Thierry Heidmann. Line-mediated retrotransposition of marked alu sequences. *Nature genetics*, 35(1):41–48, 2003.

[74] Dustin C Hancks, John L Goodier, Prabhat K Mandal, Ling E Cheung, and Haig H Kazazian Jr. Retrotransposition of marked sva elements by human l1s in cultured cells. *Human molecular genetics*, 20(17):3386–3400, 2011.

[75] Brook Brouha, Joshua Schustak, Richard M Badge, Sheila Lutz-Prigge, Alexander H Farley, John V Moran, and Haig H Kazazian. Hot l1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences*, 100(9):5280–5285, 2003.

[76] Beth A Dombroski, Stephen L Mathias, Elizabeth Nanthakumar, Alan F Scott, and Haig H Kazazian Jr. Isolation of an active human transposable element. *Science*, 254(5039):1805–1808, 1991.

[77] Dustin C Hancks and Haig H Kazazian. Active human retrotransposons: variation and disease. *Current opinion in genetics & development*, 22(3):191–203, 2012.

[78] Alan F Scott, Barbara J Schmeckpeper, Mona Abdelrazik, Catherine Theisen Comey, Bruce O'Hara, Judith Pratt Rossiter, Tim Cooley, Peter Heath, Kirby D Smith, and Louise Margolet. Origin of the human l1 elements: proposed progenitor genes deduced from a consensus dna sequence. *Genomics*, 1(2):113–125, 1987.

[79] Ahmet M Denli, Iñigo Narvaiza, Bilal E Kerman, Monique Pena, Christopher Benner, Maria CN Marchetto, Jolene K Diedrich, Aaron Aslanian, Jiao Ma, James J Moresco, et al. Primate-specific orf0 contributes to retrotransposon-mediated diversity. *Cell*, 163(3):583–593, 2015.

[80] GARY D Swergold. Identification, characterization, and cell specificity of a human line-1 promoter. *Molecular and cellular biology*, 10(12):6718–6729, 1990.

[81] Mart Speek. Antisense promoter of human l1 retrotransposon drives transcription of adjacent cellular genes. *Molecular and cellular biology*, 21(6):1973–1985, 2001.

[82] Vladimir O Kolosha and Sandra L Martin. In vitro properties of the first orf protein from mouse line-1 support its role in ribonucleoprotein particle formation during retrotransposition. *Proceedings of the National Academy of Sciences*, 94(19):10155–10160, 1997.

[83] John V Moran, Susan E Holmes, Thierry P Naas, Ralph J DeBerardinis, Jef D Boeke, and Haig H Kazazian. High frequency retrotransposition in cultured mammalian cells. *Cell*, 87(5):917–927, 1996.

[84] John L Goodier, Lili Zhang, Melissa R Vetter, and Haig H Kazazian. Line-1 orf1 protein localizes in stress granules with other rna-binding proteins, including components of rna interference rna-induced silencing complex. *Molecular and cellular biology*, 27(18):6469–6483, 2007.

[85] Stephen L Mathias, Alan F Scott, et al. Reverse transcriptase encoded by a human transposable element. *Science*, 254(5039):1808, 1991.

[86] Sarah J Wheelan, Yasunori Aizawa, Jeffrey S Han, and Jef D Boeke. Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. *Genome research*, 15(8):1073–1078, 2005.

[87] Hazel A Cruickshanks and Cristina Tufarelli. Isolation of cancer-specific chimeric transcripts induced by hypomethylation of the line-1 antisense promoter. *Genomics*, 94(6):397–406, 2009.

[88] Ruchi Shukla, Kyle R Upton, Martin Muñoz-Lopez, Daniel J Gerhardt, Malcolm E Fisher, Thu Nguyen, Paul M Brennan, J Kenneth Baillie, Agnese Collino, Serena Ghisletti, et al. Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell*, 153(1):101–111, 2013.

[89] Szilvia Solyom, Adam D Ewing, Eric P Rahrmann, Tara Doucet, Heather H Nelson, Michael B Burns, Reuben S Harris, David F Sigmon, Alex Casella, Bracha Erlanger, et al. Extensive somatic l1 retrotransposition in colorectal tumors. *Genome research*, 22(12):2328–2338, 2012.

[90] Jennifer A Erwin, Maria C Marchetto, and Fred H Gage. Mobile dna elements in the generation of diversity and complexity in the brain. *Nature Reviews Neuroscience*, 15(8):497–506, 2014.

[91] Matthew T Reilly, Geoffrey J Faulkner, Joshua Dubnau, Igor Ponomarev, and Fred H Gage. The role of transposable elements in health and diseases of the central nervous system. *Journal of Neuroscience*, 33(45):17577–17586, 2013.

[92] Tammy A Morrish, José Luis Garcia-Perez, Thomas D Stamato, Guillermo E Taccioli, JoAnn Sekiguchi, and John V Moran. Endonuclease-independent line-1 retrotransposition at mammalian telomeres. *Nature*, 446(7132):208–212, 2007.

[93] Maria CN Marchetto, Iñigo Narvaiza, Ahmet M Denli, Christopher Benner, Thomas A Lazzarini, Jason L Nathanson, Apuã CM Paquola, Keval N Desai, Roberto H Herai, Matthew D Weitzman, et al. Differential l1 regulation in pluripotent stem cells of humans and apes. *Nature*, 503(7477):525–529, 2013.

[94] Kathleen H Burns. Transposable elements in cancer. *Nature Reviews Cancer*, 2017.

[95] Jinchuan Xing, Hui Wang, Victoria P Belancio, Richard Cordaux, Prescott L Deininger, and Mark A Batzer. Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proceedings of the National Academy of Sciences*, 103(47):17608–17613, 2006.

[96] Masao Ono, Masaya Kawakami, and Toshiyuki Takezawa. A novel human nonviral retroposon derived from an endogenous retrovirus. *Nucleic acids research*, 15(21):8725–8737, 1987.

[97] Liming Shen, Lai Chu Wu, Salih Sanlioglu, Ruju Chen, Anna R Mendoza, Andrew W Dangel, Michael C Carroll, William B Zipf, and Chack-Yung Yu. Structure and genetics of the partially duplicated gene rp located immediately upstream of the complement c4a and the c4b genes in the hla class iii region. molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. *Journal of Biological Chemistry*, 269(11):8466–8476, 1994.

[98] Liora Z Strichman-Almashanu, Richard S Lee, Patrick O Onyango, Elizabeth Perlman, Folke Flam, Matthew B Frieman, and Andrew P Feinberg. A genome-wide screen for normally methylated human cpg islands that can identify novel imprinted genes. *Genome research*, 12(4):543–554, 2002.

[99] Eric M Ostertag, John L Goodier, Yue Zhang, and Haig H Kazazian. Sva elements are nonautonomous retrotransposons that cause disease in humans. *The American Journal of Human Genetics*, 73(6):1444–1451, 2003.

[100] Dustin C Hancks, Adam D Ewing, Jesse E Chen, Katsushi Tokunaga, and Haig H Kazazian. Exon-trapping mediated by the human retrotransposon sva. *Genome research*, 19(11):1983–1991, 2009.

[101] Hui Wang, Jinchuan Xing, Deepak Grover, Dale J Hedges, Kyudong Han, Jerilyn A Walker, and Mark A Batzer. Sva elements: a hominid-specific retroposon family. *Journal of molecular biology*, 354(4):994–1007, 2005.

[102] Frank MJ Jacobs, David Greenberg, Ngan Nguyen, Maximilian Haeussler, Adam D Ewing, Sol Katzman, Benedict Paten, Sofie R Salama, and David Haussler. An evolutionary arms race between krab zinc-finger genes znf91/93 and sva/l1 retrotransposons. *Nature*, 516(7530):242–245, 2014.

[103] Michaël Imbeault, Pierre-Yves Helleboid, and Didier Trono. Krab zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*, 543(7646):550–554, 2017.

[104] Annette Damert, Julija Raiz, Axel V Horn, Johannes Löwer, Hui Wang, Jinchuan Xing, Mark A Batzer, Roswitha Löwer, and Gerald G Schumann. 5'-transducing sva retrotransposon groups spread efficiently throughout the human genome. *Genome research*, 19(11):1992–2008, 2009.

[105] Olga Vasieva, Sultan Cetiner, Abigail Savage, Gerald G Schumann, Vivien J Bubb, and John P Quinn. Potential impact of primate-specific sva retrotransposons during the evolution of human cognitive function. *Trends in Evolutionary Biology*, 6(1), 2017.

[106] Stefan A Rensing, Daniel Lang, Andreas D Zimmer, Astrid Terry, Asaf Salamov, Harris Shapiro, Tomoaki Nishiyama, Pierre-François Perroud, Erika A Lindquist, Yasuko Kamisugi, et al. The physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, 319(5859):64–69, 2008.

[107] Jo Ann Banks, Tomoaki Nishiyama, Mitsuyasu Hasebe, John L Bowman, Michael Gribskov, Victor A Albert, Naoki Aono, Tsuyoshi Aoyama, Barbara A Ambrose, Neil W Ashton, et al. The selaginella genome identifies genetic changes associated with the evolution of vascular plants. *science*, 332(6032):960–963, 2011.

[108] Sanjida H Rangwala and Eric J Richards. The structure, organization and radiation of sadhu non-long terminal repeat retroelements in arabidopsis species. *Mobile DNA*, 1(1):10, 2010.

[109] Guojun Yang, Feng Zhang, C Nathan Hancock, and Susan R Wessler. Transposition of the rice miniature inverted repeat transposable element mping in arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, 104(26):10962–10967, 2007.

[110] Smriti Gupta, Andrea Gallavotti, Gabrielle A Stryker, Robert J Schmidt, and Shailesh K Lal. A novel class of helitron-related transposable elements in maize contain portions of multiple pseudogenes. *Plant molecular biology*, 57(1):115–127, 2005.

[111] Barbara McClintock. Chromosome organization and genic expression. In *Cold Spring Harbor symposia on quantitative biology*, volume 16, pages 13–47. Cold Spring Harbor Laboratory Press, 1951.

[112] Nathalie Picault, Christian Chaparro, Benoit Piegu, Willfried Stenger, Damien Formey, Cristel Llauro, Julie Descombin, Francois Sabot, Eric Lasserre, Donaldo Meynard, et al. Identification of an active ltr retrotransposon in rice. *The Plant Journal*, 58(5):754–765, 2009.

[113] Zijun Xu and Wusirika Ramakrishna. Retrotransposon insertion polymorphisms in six rice genes and their evolutionary history. *Gene*, 412(1):50–58, 2008.

[114] Chapman Jarrod A, Kirkness Ewen F, Simakov Oleg, Hampson Steven E, Mitros Therese, Weinmaier Thomas, Rattei Thomas, Balasubramanian Prakash G, Borman Jon, Busam Dana, et al. The dynamic genome of hydra. *Nature*, 464(7288), 2010.

[115] Eric W Ganko, Vikram Bhattacharjee, Paul Schliekelman, and John F McDonald. Evidence for the contribution of ltr retrotransposons to c. elegans gene evolution. *Molecular biology and evolution*, 20(11):1925–1931, 2003.

[116] Michael F Palopoli, Matthew V Rockman, Aye TinMaung, Camden Ramsay, Stephen Curwen, Andrea Aduna, Jason Laurita, and Leonid Kruglyak. Molecular basis of the copulatory plug polymorphism in c. elegans. *Nature*, 454(7207):1019, 2008.

[117] Cheng Ran Lisa Huang, Kathleen H Burns, and Jef D Boeke. Active transposition in genomes. *Annual review of genetics*, 46:651–675, 2012.

[118] Samuel Aparicio, Jarrod Chapman, Elia Stupka, Nik Putnam, Jer-ming Chia, Paramvir Dehal, Alan Christoffels, Sam Rash, Shawn Hoon, Arian Smit, et al. Whole-genome shotgun assembly and analysis of the genome of fugu rubripes. *Science*, 297(5585):1301–1310, 2002.

[119] Peter A Novick, Holly Basta, Mark Floumanhaft, Marcella A McClure, and Stéphane Boissinot. The evolutionary dynamics of autonomous non-ltr retrotransposons in the lizard anolis carolinensis shows more similarity to fish than mammals. *Molecular biology and evolution*, 26(8):1811–1822, 2009.

[120] Oliver Piskurek, Hidenori Nishihara, and Norihiro Okada. The evolution of two partner line/sine families and a full-length chromodomain-containing ty3/gypsy ltr element in the first reptilian genome of anolis carolinensis. *Gene*, 441(1):111–118, 2009.

[121] Wesley C Warren, LaDeana W Hillier, Jennifer A Marshall Graves, Ewan Birney, Chris P Ponting, Frank Grützner, Katherine Belov, Webb Miller, Laura Clarke, Asif T Chinwalla, et al. Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, 453(7192):175, 2008.

[122] John K Pace, Clément Gilbert, Marlena S Clark, and Cédric Feschotte. Repeated horizontal transfer of a dna transposon in mammals and other tetrapods. *Proceedings of the National Academy of Sciences*, 105(44):17023–17028, 2008.

[123] Ellen J Pritham and Cédric Feschotte. Massive amplification of rolling-circle transposons in the lineage of the bat myotis lucifugus. *Proceedings of the National Academy of Sciences*, 104(6):1895–1900, 2007.

[124] David A Ray, Cedric Feschotte, Heidi JT Pagan, Jeremy D Smith, Ellen J Pritham, Peter Arensburger, Peter W Atkinson, and Nancy L Craig. Multiple waves of recent dna transposon activity in the bat, myotis lucifugus. *Genome Research*, 2008.

[125] RA Gibbs, GM Weinstock, ML Metzker, DM Muzny, EJ Sodergren, S Scherer, G Scott, D Steffen, KC Worley, PE Burch, et al. Celera, ra holt, md adams, pg amanatides, h. *Baden-Tillson, M. Barnstead, S. Chin, CA Evans, S. Ferriera, and C. Fosler. Genome sequence of the brown norway rat yields insights into mammalian evolution. Nature*, 428:493–521, 2004.

[126] Alexander Kirilyuk, Genrich V Tolstonog, Annette Damert, Ulrike Held, Silvia Hahn, Roswitha Löwer, Christian Buschmann, Axel V Horn, Peter Traub, and Gerald G Schumann. Functional endogenous line-1 retrotransposons are expressed and mobilized in rat chloroleukemia cells. *Nucleic acids research*, 36(2):648–665, 2007.

[127] John M Sedivy, Jill A Kreiling, Nicola Neretti, Marco De Cecco, Steven W Criscione, Jeffrey W Hofmann, Xiaoai Zhao, Takahiro Ito, and Abigail L Peterson. Death by transposition–the enemy within? *Bioessays*, 35(12):1035–1043, 2013.

[128] Martín Muñoz-López and José L García-Pérez. Dna transposons: nature and applications in genomics. *Current genomics*, 11(2):115–128, 2010.

[129] Nemanja Rodić and Kathleen H Burns. Long interspersed element–1 (line-1): passenger or driver in human neoplasms? *PLoS genetics*, 9(3):e1003402, 2013.

[130] Salima Hacein-Bey-Abina, C Von Kalle, M Schmidt, MP McCormack, N Wulffraat, Pet al Leboulch, A Lim, CS Osborne, R Pawliuk, E Morillon, et al. Lmo2-associated clonal t cell proliferation in two patients after gene therapy for scid-x1. *science*, 302(5644):415–419, 2003.

[131] Gregory S Payne, J Michael Bishop, and Harold E Varmus. Multiple arrangements of viral dna and an activated host oncogene in bursal lymphomas. *Nature*, 295(5846):209–214, 1982.

[132] Florence Cammas, Manuel Mark, Pascal Dollé, Andrée Dierich, Pierre Chambon, and Régine Losson. Mice lacking the transcriptional corepressor tif1beta are defective in early postimplantation development. *Development*, 127(13):2955–2963, 2000.

[133] Helen M Rowe, Johan Jakobsson, Daniel Mesnard, Jacques Rougemont, Séverine Reynard, Tugce Aktas, Pierre V Maillard, Hillary Layard-Liesching, Sonia Verp, Julien Marquis, et al. Kap1 controls endogenous retroviruses in embryonic stem cells. *Nature*, 463(7278):237, 2010.

[134] Marc Friedli and Didier Trono. The developmental control of transposable elements and the evolution of higher species. *Annual review of cell and developmental biology*, 31:429–451, 2015.

[135] Axel Roers, Björn Hiller, and Veit Hornung. Recognition of endogenous nucleic acids by the innate immune system. *Immunity*, 44(4):739–754, 2016.

[136] Xi Shi, Andrei Seluanov, and Vera Gorbunova. Cell divisions are required for l1 retrotransposition. *Molecular and cellular biology*, 27(4):1264–1270, 2007.

[137] Shuji Kubo, Maria del Carmen Seleme, Harris S Soifer, José Luis Garcia Perez, John V Moran, Haig H Kazazian, and Noriyuki Kasahara. L1 retrotransposition in nondividing and primary human somatic cells. *Proceedings of the National Academy of Sciences*, 103(21):8036–8041, 2006.

[138] Jeffrey A Yoder, Colum P Walsh, and Timothy H Bestor. Cytosine methylation and the ecology of intragenomic parasites. *Trends in genetics*, 13(8):335–340, 1997.

[139] Peter A Jones and Shirley M Taylor. Cellular differentiation, cytidine analogs and dna methylation. *Cell*, 20(1), 1980.

[140] James P Jackson, Anders M Lindroth, Xiaofeng Cao, and Steven E Jacobsen. Control of cpnpg dna methylation by the kryptonite histone h3 methyltransferase. *Nature*, 416(6880):556, 2002.

[141] Robert A Rollins, Fatemeh Haghighi, John R Edwards, Rajdeep Das, Michael Q Zhang, Jingyue Ju, and Timothy H Bestor. Large-scale structure of genomic methylation patterns. *Genome research*, 16(2):157–163, 2006.

[142] Albert Jeltsch, Wolfgang Nellen, and Frank Lyko. Two substrates are better than one: dual specificities for dnmt2 methyltransferases. *Trends in biochemical sciences*, 31(6):306–308, 2006.

[143] Déborah Bourc'his and Timothy H Bestor. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking dnmt3l. *Nature*, 431(7004):96, 2004.

[144] Kenichiro Hata, Masaki Okano, Hong Lei, and En Li. Dnmt3l cooperates with the dnmt3 family of de novo dna methyltransferases to establish maternal imprints in mice. *Development*, 129(8):1983–1993, 2002.

[145] Keith D Robertson, Eva Uzvolgyi, Gangning Liang, Cathy Talmadge, Janos Sumegi, Felicidad A Gonzales, and Peter A Jones. The human dna methyltransferases (dnmts) 1, 3a and 3b: coordinate mrna expression in normal tissues and overexpression in tumors. *Nucleic acids research*, 27(11):2291–2298, 1999.

[146] Masaki Okano, Daphne W Bell, Daniel A Haber, and En Li. Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257, 1999.

[147] Giedrius Vilkaitis, Isao Suetake, Saulius Klimašauskas, and Shoji Tajima. Processive methylation of hemimethylated cpg sites by mouse dnmt1 dna methyltransferase. *Journal of Biological Chemistry*, 280(1):64–72, 2005.

[148] Jafar Sharif, Masahiro Muto, Shin-ichiro Takebayashi, Isao Suetake, Akihiro Iwamatsu, Takaho A Endo, Jun Shinga, Yoko Mizutani-Koseki, Tetsuro Toyoda, Kunihiro Okamura, et al. The sra protein np95 mediates epigenetic inheritance by recruiting dnmt1 to methylated dna. *Nature*, 450(7171):908, 2007.

[149] Paula M Vertino, RW Yen, Jin Gao, and Stephen B Baylin. De novo methylation of cpg island sequences in human fibroblasts overexpressing dna (cytosine-5-)-methyltransferase. *Molecular and cellular biology*, 16(8):4555–4565, 1996.

[150] Ina Rhee, Kam-Wing Jair, Ray-Whay Chiu Yen, Christoph Lengauer, et al. Cpg methylation is maintained in human cancer cells lacking dnmt1. *Nature*, 404(6781):1003, 2000.

[151] Chun-Chang Chen, Keh-Yang Wang, and Che-Kun James Shen. Dna 5-methylcytosine demethylation activities of the mammalian dna methyltransferases. *Journal of Biological Chemistry*, 288(13):9084–9091, 2013.

[152] Kyeoung-Hwa Kim and Kyung-Ah Lee. Maternal effect genes: Findings and effects on mouse embryo development. *Clinical and experimental reproductive medicine*, 41(2):47–61, 2014.

[153] Yuzuru Kato, Masahiro Kaneda, Kenichiro Hata, Kenji Kumaki, Mizue Hisano, Yuji Kohara, Masaki Okano, En Li, Masami Nozaki, and Hiroyuki Sasaki. Role of the dnmt3 family in de novo methylation of imprinted and repetitive sequences during male germ cell development in the mouse. *Human molecular genetics*, 16(19):2272–2280, 2007.

[154] T Yokomine, K Hata, M Tsudzuki, and H Sasaki. Evolution of the vertebrate dnmt3 gene family: a possible link between existence of dnmt3l and genomic imprinting. *Cytogenetic and genome research*, 113(1-4):75–80, 2006.

[155] Heming Zhu, Theresa M Geiman, Sichuan Xi, Qiong Jiang, Anja Schmidtmann, Taiping Chen, En Li, and Kathrin Muegge. Lsh is involved in de novo methylation of dna. *The EMBO journal*, 25(2):335–345, 2006.

[156] Jiaqiang Huang, Tao Fan, Qingsheng Yan, Heming Zhu, Stephen Fox, Haleem J Issaq, Lionel Best, Lisa Gangi, David Munroe, and Kathrin Muegge. Lsh, an epigenetic guardian of repetitive elements. *Nucleic acids research*, 32(17):5019–5028, 2004.

[157] Rabindranath De La Fuente, Claudia Baumann, Tao Fan, Anja Schmidtmann, Ina Dobrinski, and Kathrin Muegge. Lsh is required for meiotic chromosome synapsis and retrotransposon silencing in female germ cells. *Nature cell biology*, 8(12):1448, 2006.

[158] Annika Wylie, Amanda E Jones, Alejandro D'Brot, Wan-Jin Lu, Paula Kurtz, John V Moran, Dinesh Rakheja, Kenneth S Chen, Robert E Hammer, Sarah A Comerford, et al. p53 genes function to restrain mobile elements. *Genes & development*, 30(1):64–77, 2016.

[159] Swaminathan Venkatesh and Jerry L Workman. Histone exchange, chromatin structure and the regulation of transcription. *Nature reviews. Molecular cell biology*, 16(3):178, 2015.

[160] Yutaka Kondo and Jean-Pierre J Issa. Enrichment for histone h3 lysine 9 methylation at alu repeats in human cells. *Journal of Biological Chemistry*, 278(30):27658–27662, 2003.

[161] Marius Walter, Aurélie Teissandier, Raquel Pérez-Palacios, and Déborah Bourc'his. An epigenetic switch ensures transposon repression upon dynamic loss of dna methylation in embryonic stem cells. *Elife*, 5:e11418, 2016.

[162] Joost HA Martens, Roderick J O'Sullivan, Ulrich Braunschweig, Susanne Opravil, Martin Radolf, Peter Steinlein, and Thomas Jenuwein. The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *The EMBO journal*, 24(4):800–812, 2005.

[163] Tarjei S Mikkelsen, Manching Ku, David B Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, William Brockman, Tae-Kyung Kim, Richard P Koche, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553, 2007.

[164] Colum P Walsh, J Richard Chaillet, and Timothy H Bestor. Transcription of iap endogenous retroviruses is constrained by cytosine methylation. *Nature genetics*, 20(2):116–117, 1998.

[165] Helen M Rowe and Didier Trono. Dynamic control of endogenous retroviruses during development. *Virology*, 411(2):273–287, 2011.

[166] Toshiyuki Matsui, Danny Leung, Hiroki Miyashita, Irina A Maksakova, Hitoshi Miyachi, Hiroshi Kimura, Makoto Tachibana, Matthew C Lorincz, and Yoichi Shinkai. Proviral silencing in embryonic stem cells requires the histone methyltransferase eset. *Nature*, 464(7290):927, 2010.

[167] Martin Leeb, Diego Pasini, Maria Novatchkova, Markus Jaritz, Kristian Helin, and Anton Wutz. Polycomb complexes act redundantly to repress genomic repeats and genes. *Genes & development*, 24(3):265–276, 2010.

[168] Michelle A Carmell, Angélique Girard, Henk JG van de Kant, Deborah Bourc'his, Timothy H Bestor, Dirk G de Rooij, and Gregory J Hannon. Miwi2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Developmental cell*, 12(4):503–514, 2007.

[169] Alexei A Aravin, Ravi Sachidanandam, Angelique Girard, Katalin Fejes-Toth, and Gregory J Hannon. Developmentally regulated pirna clusters implicate mili in transposon control. *Science*, 316(5825):744–747, 2007.

[170] Michael Reuter, Shinichiro Chuma, Takashi Tanaka, Thomas Franz, Alexander Stark, and Ramesh S Pillai. Loss of the mili-interacting tudor domain–containing protein-1 activates transposons and alters the mili-associated small rna profile. *Nature structural & molecular biology*, 16(6):639–646, 2009.

[171] Masanobu Shoji, Takashi Tanaka, Mihoko Hosokawa, Michael Reuter, Alexander Stark, Yuzuru Kato, Gen Kondoh, Katsuya Okawa, Takeshi Chujo, Tsutomu Suzuki, et al. The tdrd9-miwi2 complex is essential for pirna-mediated retrotransposon silencing in the mouse male germline. *Developmental cell*, 17(6):775–787, 2009.

[172] Lang Ma, Gregory M Buchold, Michael P Greenbaum, Angshumoy Roy, Kathleen H Burns, Huifeng Zhu, Derek Y Han, R Alan Harris, Cristian Coarfa, Preethi H Gunaratne, et al. Gasz is essential for male meiosis and suppression of retrotransposon expression in the male germline. *PLoS genetics*, 5(9):e1000635, 2009.

[173] Kazuko Nishikura. Functions and regulation of rna editing by adar deaminases. *Annual review of biochemistry*, 79:321–349, 2010.

[174] Donna A MacDuff, Zachary L Demorest, and Reuben S Harris. Aid can restrict l1 retrotransposition suggesting a dual role in innate and adaptive immunity. *Nucleic acids research*, 37(6):1854–1867, 2009.

[175] Stephen L Gasior, Astrid M Roy-Engel, and Prescott L Deininger. Ercc1/xpf limits l1 retrotransposition. *DNA repair*, 7(6):983–989, 2008.

[176] Sandeep Satapathy. Provirus silencing in stem cells: The forbidden regulators of cell fate. *Signal Transduction Insights*, 5:15, 2016.

[177] Diego E Montoya-Durango, Yongqing Liu, Ivo Teneng, Ted Kalbfleisch, Mary E Lacy, Marlene C Steffen, and Kenneth S Ramos. Epigenetic control of mammalian line-1 retrotransposon by retinoblastoma proteins. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 665(1):20–28, 2009.

[178] N Zamudio and D Bourc'his. Transposable elements in the mammalian germline: a comfortable niche or a deadly trap? *Heredity*, 105(1):92, 2010.

[179] Nehmé Saksouk, Elisabeth Simboeck, and Jérôme Déjardin. Constitutive heterochromatin formation and transcription in mammals. *Epigenetics & chromatin*, 8(1):3, 2015.

[180] Stuart Huntley, Daniel M Baggott, Aaron T Hamilton, Mary Tran-Gyamfi, Shan Yang, Joomyeong Kim, Laurie Gordon, Elbert Branscomb, and Lisa Stubbs. A comprehensive catalog of human krab-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. *Genome research*, 16(5):669–677, 2006.

[181] Hui Liu, Li-Hsin Chang, Younguk Sun, Xiaochen Lu, and Lisa Stubbs. Deep vertebrate roots for mammalian zinc finger transcription factor subfamilies. *Genome biology and evolution*, 6(3):510–525, 2014.

[182] Margaret G Kidwell and Damon R Lisch. Perspective: transposable elements, parasitic dna, and genome evolution. *Evolution*, 55(1):1–24, 2001.

[183] Nathaly Castro-Diaz, Gabriela Ecco, Andrea Coluccio, Adamandia Kapopoulou, Benyamin Yazdanpanah, Marc Friedli, Julien Duc, Suk Min Jang, Priscilla Turelli, and Didier Trono. Evolutionally dynamic l1 regulation in embryonic stem cells. *Genes & development*, 28(13):1397–1409, 2014.

[184] Priscilla Turelli, Nathaly Castro-Diaz, Flavia Marzetta, Adamandia Kapopoulou, Charlene Raclot, Julien Duc, Vannary Tieng, Simon Quenneville, and Didier Trono. Interplay of trim28 and dna methylation in controlling human endogenous retroelements. *Genome research*, 24(8):1260–1270, 2014.

[185] Helen M Rowe, Marc Friedli, Sandra Offner, Sonia Verp, Daniel Mesnard, Julien Marquis, Tugce Aktas, and Didier Trono. De novo dna methylation of endogenous retroviruses is shaped by krab-zfps/kap1 and eset. *Development*, 140(3):519–529, 2013.

[186] Gernot Wolf, Peng Yang, Annette C Füchtbauer, Ernst-Martin Füchtbauer, Andreia M Silva, Chungoo Park, Warren Wu, Anders L Nielsen, Finn S Pedersen, and Todd S Macfarlan. The krab zinc finger protein zfp809 is required to initiate epigenetic silencing of endogenous retroviruses. *Genes & development*, 29(5):538–554, 2015.

[187] Masamichi Muramatsu, Kazuo Kinoshita, Sidonia Fagarasan, Shuichi Yamada, Yoichi Shinkai, and Tasuku Honjo. Class switch recombination and hypermutation require activation-induced cytidine deaminase (aid), a potential rna editing enzyme. *Cell*, 102(5):553–563, 2000.

[188] Kate N Bishop, Rebecca K Holmes, Ann M Sheehy, and Michael H Malim. Apobec-mediated editing of viral rna. *Science*, 305(5684):645–645, 2004.

[189] Dennis DY Kim, Thomas TY Kim, Thomas Walsh, Yoshifumi Kobayashi, Tara C Matise, Steven Buyske, and Abram Gabriel. Widespread rna editing of embedded alu elements in the human transcriptome. *Genome research*, 14(9):1719–1725, 2004.

[190] Eli Eisenberg, Sergey Nemzer, Yaron Kinar, Rotem Sorek, Gideon Rechavi, and Erez Y Levanon. Is abundant a-to-i rna editing primate-specific? *Trends in Genetics*, 21(2):77–81, 2005.

[191] Weidong Yang, Thimmaiah P Chendrimada, Qingde Wang, Miyoko Higuchi, Peter H Seeburg, Ramin Shiekhattar, and Kazuko Nishikura. Modulation of microrna processing and expression through rna editing by adar deaminases. *Nature structural & molecular biology*, 13(1):13, 2006.

[192] Hal P Bogerd, Heather L Wiegand, Brian P Doehle, Kira K Lueders, and Bryan R Cullen. Apobec3a and apobec3b are potent inhibitors of ltr-retrotransposon function in human cells. *Nucleic acids research*, 34(1):89–95, 2006.

[193] Cecile Esnault, Jean Millet, Olivier Schwartz, and Thierry Heidmann. Dual inhibitory effects of apobec family proteins on retrotransposition of mammalian endogenous retroviruses. *Nucleic acids research*, 34(5):1522–1531, 2006.

[194] Eric W Refsland and Reuben S Harris. The apobec3 family of retroelement restriction factors. In *Intrinsic Immunity*, pages 1–27. Springer, 2013.

[195] Darren J Obbard, Karl HJ Gordon, Amy H Buck, and Francis M Jiggins. The evolution of rnai as a defence against viruses and transposable elements. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1513):99–115, 2009.

[196] Anne E Peaston, Alexei V Evsikov, Joel H Graber, Wilhelmine N De Vries, Andrea E Holbrook, Davor Solter, and Barbara B Knowles. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Developmental cell*, 7(4):597–606, 2004.

[197] Petr Svoboda, Paula Stein, Martin Anger, Emily Bernstein, Gregory J Hannon, and Richard M Schultz. Rnai and expression of retrotransposons muerv-l and iap in preimplantation mouse embryos. *Developmental biology*, 269(1):276–285, 2004.

[198] Toshiaki Watanabe, Yasushi Totoki, Atsushi Toyoda, Masahiro Kaneda, Satomi Kuramochi-Miyagawa, Yayoi Obata, Hatsune Chiba, Yuji Kohara, Tomohiro Kono, Toru Nakano, et al. Endogenous sirnas from naturally formed dsrnas regulate transcripts in mouse oocytes. *Nature*, 453(7194):539, 2008.

[199] Andrew Grimson, Mansi Srivastava, Bryony Fahey, Ben J Woodcroft, H Rosaria Chiang, Nicole King, Bernard M Degnan, Daniel S Rokhsar, and David P Bartel. The early origins of microrna and piwi-interacting rnas in animals. *Nature*, 455(7217), 2008.

[200] Alexei A Aravin, Ravi Sachidanandam, Deborah Bourc'his, Christopher Schaefer, Dubravka Pezic, Katalin Fejes Toth, Timothy Bestor, and Gregory J Hannon. A pirna pathway primed by individual transposons is linked to de novo dna methylation in mice. *Molecular cell*, 31(6):785–799, 2008.

[201] Satomi Kuramochi-Miyagawa, Toshiaki Watanabe, Kengo Gotoh, Yasushi Totoki, Atsushi Toyoda, Masahito Ikawa, Noriko Asada, Kanako Kojima, Yuka Yamaguchi, Takashi W Ijiri, et al. Dna methylation of retrotransposon genes is regulated by piwi family members mili and miwi2 in murine fetal testes. *Genes & development*, 22(7):908–917, 2008.

[202] John L Goodier. Restricting retrotransposons: a review. *Mobile DNA*, 7(1):16, 2016.

[203] S Mehdi Belgnaoui, Roger G Gosden, O John Semmes, and Abdelali Haoudi. Human line-1 retrotransposon induces dna damage and apoptosis in cancer cells. *Cancer cell international*, 6(1):13, 2006.

[204] Cédric Feschotte and Clément Gilbert. Endogenous viruses: insights into viral evolution and impact on host biology. *Nature reviews. Genetics*, 13(4):283, 2012.

[205] Sandra Blaise, Nathalie de Parseval, Laurence Bénit, and Thierry Heidmann. Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene

conserved on primate evolution. *Proceedings of the National Academy of Sciences*, 100(22):13013–13018, 2003.

[206] Aurelie Kapusta, Zev Kronenberg, Vincent J Lynch, Xiaoyu Zhuo, LeeAnn Ramsay, Guillaume Bourque, Mark Yandell, and Cedric Feschotte. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding rnas. *PLoS genetics*, 9(4):e1003470, 2013.

[207] Pierre-Etienne Jacques, Justin Jeyakani, and Guillaume Bourque. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS genetics*, 9(5):e1003504, 2013.

[208] Mari Ohnuki, Koji Tanabe, Kenta Sutou, Ito Teramoto, Yuka Sawamura, Megumi Narita, Michiko Nakamura, Yumie Tokunaga, Masahiro Nakamura, Akira Watanabe, et al. Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proceedings of the National Academy of Sciences*, 111(34):12426–12431, 2014.

[209] Xinyi Lu, Friedrich Sachs, LeeAnn Ramsay, Pierre-Étienne Jacques, Jonathan Göke, Guillaume Bourque, and Huck-Hui Ng. The retrovirus hervh is a long noncoding rna required for human embryonic stem cell identity. *Nature structural & molecular biology*, 21(4):423–425, 2014.

[210] Galih Kunarso, Na-Yu Chia, Justin Jeyakani, Catalina Hwang, Xinyi Lu, Yun-Shen Chan, Huck-Hui Ng, and Guillaume Bourque. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature genetics*, 42(7):631–634, 2010.

[211] Jens Durruthy-Durruthy, Vittorio Sebastiano, Mark Wossidlo, Diana Cepeda, Jun Cui, Edward J Grow, Jonathan Davila, Moritz Mall, Wing H Wong, Joanna Wysocka, et al. The primate-specific noncoding rna hpat5 regulates pluripotency during human preimplantation development and nuclear reprogramming. *Nature genetics*, 48(1):44–52, 2016.

[212] Jichang Wang, Gangcai Xie, Manvendra Singh, Avazeh T Ghanbarian, Tamás Raskó, Attila Szvetnik, Huiqiang Cai, Daniel Besser, Alessandro Prigione, Nina V Fuchs, et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature*, 516(7531):405, 2014.

[213] Louie N van de Lagemaat, Josette-Renée Landry, Dixie L Mager, and Patrik Medstrand. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends in Genetics*, 19(10):530–536, 2003.

[214] Vasavi Sundaram, Yong Cheng, Zhihai Ma, Daofeng Li, Xiaoyun Xing, Peter Edge, Michael P Snyder, and Ting Wang. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome research*, 24(12):1963–1976, 2014.

[215] Patrick Gemmell, Jotun Hein, and Aris Katzourakis. Phylogenetic analysis reveals that ervs" die young" but herv-h is unusually conserved. *PLoS computational biology*, 12(6):e1004964, 2016.

[216] Gkikas Magiorkinis, Robert Belshaw, and Aris Katzourakis. 'there and back again': revisiting the pathophysiological roles of human endogenous retroviruses in the post-genomic era. *Phil. Trans. R. Soc. B*, 368(1626):20120504, 2013.

[217] Dixie L Mager and J DOUGLAS Freeman. Human endogenous retroviruslike genome with type c pol sequences and gag sequences related to human t-cell lymphotropic viruses. *Journal of virology*, 61(12):4060–4066, 1987.

[218] DIXIE L MAGER and J DOUGLAS FREEMAN. Herv-h endogenous retroviruses: presence in the new world branch but amplification in the old world primate lineage. *Virology*, 213(2):395–404, 1995.

[219] Nancy L Goodchild, David A Wilkinson, and Dixie L Mager. Recent evolutionary expansion of a subfamily of rtvl-h human endogenous retrovirus-like elements. *Virology*, 196(2):778–788, 1993.

[220] Sølvi Anderssen, Eva Sjøttem, Gunbjørg Svineng, and Terje Johansen. Comparative analyses of ltrs of the erv-h family of primate-specific retrovirus-like elements isolated from marmoset, african green monkey, and man. *Virology*, 234(1):14–30, 1997.

[221] Nancy L Goodchild, J Douglas Freeman, and Dixie L Mager. Spliced herv-h endogenous retroviral sequences in human genomic dna: evidence for amplification via retrotransposition. *Virology*, 206(1):164–173, 1995.

[222] Luisa Robbez-Masson and Helen M Rowe. Retrotransposons shape species-specific embryonic stem cell gene expression. *Retrovirology*, 12(1):45, 2015.

[223] DAVID T NELSON, NANCY L GOODCHILD, and DIXIE L MAGER. Gain of sp1 sites and loss of repressor sequences associated with a young, transcriptionally active subset of herv-h endogenous long terminal repeats. *Virology*, 220(1):213–218, 1996.

[224] Patrik Medstrand and Dixie L Mager. Human-specific integrations of the herv-k endogenous retrovirus family. *Journal of virology*, 72(12):9782–9787, 1998.

[225] John M Coffin, Stephen H Hughes, and Harold E Varmus. *The interactions of retroviruses and their hosts*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 1997.

[226] Nathalie de Parseval, Jean-François Casella, Laetitia Gressin, and Thierry Heidmann. Characterization of the three herv-h proviruses with an open envelope reading frame encompassing the immunosuppressive domain and evolutionary history in primates. *Virology*, 279(2):558–569, 2001.

[227] Patric Jern, Göran O Sperber, and Jonas Blomberg. Definition and variation of human endogenous retrovirus h. *Virology*, 327(1):93–110, 2004.

[228] Mats Lindeskog, Dixie L Mager, and Jonas Blomberg. Isolation of a human endogenous retroviral herv-h element with an open env reading frame. *Virology*, 258(2):441–450, 1999.

[229] Carla J Cohen, Wynne M Lock, and Dixie L Mager. Endogenous retroviral ltrs as promoters for human genes: a critical assessment. *Gene*, 448(2):105–114, 2009.

[230] C Stocking and CA Kozak. Endogenous retroviruses. *Cellular and molecular life sciences*, 65(21):3383–3398, 2008.

[231] Michael Tristem. Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *Journal of virology*, 74(8):3715–3730, 2000.

[232] Norbert Bannert and Reinhard Kurth. The evolutionary dynamics of human endogenous retroviral families. *Annu. Rev. Genomics Hum. Genet.*, 7:149–173, 2006.

[233] Aris Katzourakis and Michael Tristem. Phylogeny of human endogenous and exogenous retroviruses. *), Retroviruses and Primate Genome Evolution. Landes Bioscience, Georgetown*, pages 186–203, 2005.

[234] Robert Belshaw, Aris Katzourakis, Jan Paces, Austin Burt, and Michael Tristem. High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection. *Molecular biology and evolution*, 22(4):814–817, 2005.

[235] Robert Belshaw, Vini Pereira, Aris Katzourakis, Gillian Talbot, Jan Pačes, Austin Burt, and Michael Tristem. Long-term reinfection of the human genome by endogenous retroviruses. *Proceedings of the National Academy of Sciences of the United States of America*, 101(14):4894–4899, 2004.

[236] Aris Katzourakis, Andrew Rambaut, and Oliver G Pybus. The evolutionary dynamics of endogenous retroviruses. *Trends in microbiology*, 13(10):463–468, 2005.

[237] Laurence Bénit, Jean-Baptiste Lallemand, Jean-François Casella, Hervé Philippe, and Thierry Heidmann. Erv-l elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. *Journal of virology*, 73(4):3301–3308, 1999.

[238] Steven Xijin Ge. Exploratory bioinformatics investigation reveals importance of "junk" dna in early embryo development. *BMC genomics*, 18(1):200, 2017.

[239] Yong Jin Choi, Chao-Po Lin, Davide Risso, Sean Chen, Thomas Aquinas Kim, Meng How Tan, Jin Billy Li, Yalei Wu, Caifu Chen, Zhenyu Xuan, et al. Deficiency of microrna mir-34a expands cell fate potential in pluripotent stem cells. *Science*, 355(6325):eaag1927, 2017.

[240] Jon Schoorlemmer, Raquel Pérez-Palacios, María Climent, Diana Guallar, and Pedro Muniesa. Regulation of mouse retroelement muerv-l/mervl expression by rex1 and epigenetic control of stem cell potency. *Frontiers in oncology*, 4, 2014.

[241] Dixie L Mager and Paula S Henthorn. Identification of a retrovirus-like repetitive element in human dna. *Proceedings of the National Academy of Sciences*, 81(23):7510–7514, 1984.

[242] Marianne Mangeney, Nathalie de Parseval, Gilles Thomas, and Thierry Heidmann. The full-length envelope of an herv-h human endogenous retrovirus has immunosuppressive properties. *Journal of General Virology*, 82(10):2515–2518, 2001.

[243] Robert Belshaw, Anna LA Dawson, John Woolven-Allen, Joanna Redding, Austin Burt, and Michael Tristem. Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family herv-k (hml2): implications for present-day activity. *Journal of virology*, 79(19):12507–12514, 2005.

[244] Heui-Soo Kim, Osamu Takenaka, and Timothy J Crow. Isolation and phylogeny of endogenous retrovirus sequences belonging to the herv-w family in primates. *Journal of general virology*, 80(10):2613–2619, 1999.

[245] Javier Costas. Characterization of the intragenomic spread of the human endogenous retrovirus family herv-w. *Molecular biology and evolution*, 19(4):526–533, 2002.

[246] Adam Pavlíček, Jan Pačes, Daniel Elleder, and Jiří Hejnar. Processed pseudogenes of human endogenous retroviruses generated by lines: their integration, stability, and distribution. *Genome research*, 12(3):391–399, 2002.

[247] Antoinette C van der Kuyl. Hiv infection and herv expression: a review. *Retrovirology*, 9(1):6, 2012.

[248] Ravi P Subramanian, Julia H Wildschutte, Crystal Russo, and John M Coffin. Identification, characterization, and comparative genomic distribution of the herv-k (hml-2) group of human endogenous retroviruses. *Retrovirology*, 8(1):90, 2011.

[249] Madalina Barbulescu, Geoffrey Turner, Michael I Seaman, Amos S Deinard, Kenneth K Kidd, and Jack Lenz. Many human endogenous retrovirus k (herv-k) proviruses are unique to humans. *Current Biology*, 9(16):861–S1, 1999.

[250] Geoffrey Turner, Madalina Barbulescu, Mei Su, Michael I Jensen-Seaman, Kenneth K Kidd, and Jack Lenz. Insertional polymorphisms of full-length endogenous retroviruses in humans. *Current Biology*, 11(19):1531–1535, 2001.

[251] Jens Mayer, Marlies Sauter, Alexander Rácz, Daniela Scherer, Nikolaus Mueller-Lantzsch, and Eckart Meese. An almost-intact human endogenous retrovirus k on human chromosome 7. *Nature genetics*, 21(3):257–258, 1999.

[252] Masao Ono, T Yasunaga, T Miyata, and H Ushikubo. Nucleotide sequence of human endogenous retrovirus genome related to the mouse mammary tumor virus genome. *Journal of virology*, 60(2):589–598, 1986.

[253] MASAO Ono. Molecular cloning and long terminal repeat sequences of human endogenous retrovirus genes related to types a and b retrovirus genes. *Journal of virology*, 58(3):937–944, 1986.

[254] Marie Dewannieux, Francis Harper, Aurélien Richaud, Claire Letzelter, David Ribet, Gérard Pierron, and Thierry Heidmann. Identification of an infectious progenitor for the multiple-copy herv-k human endogenous retroelements. *Genome research*, 16(12):1548–1556, 2006.

[255] Young Nam Lee and Paul D Bieniasz. Reconstitution of an infectious human endogenous retrovirus. *PLoS pathogens*, 3(1):e10, 2007.

[256] Jennifer F Hughes and John M Coffin. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. *Nature genetics*, 29(4):487, 2001.

[257] Kathryn E Plant, Samantha JE Routledge, and Nick J Proudfoot. Intergenic transcription in the human β-globin gene cluster. *Molecular and cellular biology*, 21(19):6507–6514, 2001.

[258] SJE Routledge and NJ Proudfoot. Definition of transcriptional promoters in the human β globin locus control region. *Journal of molecular biology*, 323(4):601–611, 2002.

[259] Eugene D Sverdlov. Retroviruses and primate evolution. *Bioessays*, 22(2):161–171, 2000.

[260] Laurence Lavie, Patrik Medstrand, Werner Schempp, Eckart Meese, and Jens Mayer. Human endogenous retrovirus family herv-k (hml-5): status, evolution, and reconstruction of an ancient betaretrovirus in the human genome. *Journal of virology*, 78(16):8788–8798, 2004.

[261] Yuri B Lebedev, Oksana S Belonovitch, Natalia V Zybrova, Paul P Khil, Sergey G Kurdyukov, Tatyana V Vinogradova, Gerhard Hunsmann, and Eugene D Sverdlov. Differences in herv-k ltr insertions in orthologous loci of humans and great apes. *Gene*, 247(1):265–277, 2000.

[262] Patricia Gerdes, Sandra R Richardson, Dixie L Mager, and Geoffrey J Faulkner. Transposable elements in the mammalian embryo: pioneers surviving through stealth and service. *Genome biology*, 17(1):100, 2016.

[263] Jonathan Göke, Xinyi Lu, Yun-Shen Chan, Huck-Hui Ng, Lam-Ha Ly, Friedrich Sachs, and Iwona Szczerbinska. Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell stem cell*, 16(2):135–141, 2015.

[264] Peter E Warburton, Joti Giordano, Fanny Cheung, Yefgeniy Gelfand, and Gary Benson. Inverted repeat structure of the human genome: the x-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome research*, 14(10a):1861–1869, 2004.

[265] Arthur D Riggs. Marsupials and mechanisms of x-chromosome inactivation. *Australian Journal of Zoology*, 37(3):419–441, 1989.

[266] Jennifer C Chow, Ziny Yen, Sonia M Ziesche, and Carolyn J Brown. Silencing of the mammalian x chromosome. *Annu. Rev. Genomics Hum. Genet.*, 6:69–92, 2005.

[267] DL Mager and JP Stoye. Mammalian endogenous retroviruses. microbiol spectr 2015, 3 (1). doi: 10.1128/microbiolspec. Technical report, MDNA3-0009-2014, 2015.

[268] Jürgen Brosius and Stephen Jay Gould. On" genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other" junk dna". *Proceedings of the National Academy of Sciences*, 89(22):10706–10710, 1992.

[269] Wolfgang J Miller, John F McDonald, and Wilhelm Pinsker. Molecular domestication of mobile elements. In *Evolution and Impact of Transposable Elements*, pages 261–270. Springer, 1997.

[270] Edward B Chuong, Nels C Elde, and Cédric Feschotte. Regulatory activities of transposable elements: from conflicts to benefits. *Nature reviews. Genetics*, 18(2):71, 2017.

[271] Ken Naito, Feng Zhang, Takuji Tsukiyama, Hiroki Saito, C Nathan Hancock, Aaron O Richardson, Yutaka Okumoto, Takatoshi Tanisaka, and Susan R Wessler. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature*, 461(7267):1130, 2009.

[272] Craig B Lowe and David Haussler. 29 mammalian genomes reveal novel exaptations of mobile elements for likely regulatory functions in the human genome. *PLoS One*, 7(8):e43128, 2012.

[273] Guillaume Bourque, Bernard Leong, Vinsensius B Vega, Xi Chen, Yen Ling Lee, Kandhadayar G Srinivasan, Joon-Lin Chew, Yijun Ruan, Chia-Lin Wei, Huck Hui Ng, et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome research*, 18(11):1752–1762, 2008.

[274] Björn Lamprecht, Korden Walter, Stephan Kreher, Raman Kumar, Michael Hummel, Dido Lenze, Karl Köchert, Mohamed Amine Bouhlel, Julia Richter, Eric Soler, et al. Derepression of an endogenous long terminal repeat activates the csf1r proto-oncogene in human lymphoma. *Nature medicine*, 16(5):571–579, 2010.

[275] Vedran Franke, Sravya Ganesh, Rosa Karlic, Radek Malik, Josef Pasulka, Filip Horvat, Maja Kuzman, Helena Fulka, Marketa Cernohorska, Jana Urbanova, et al. Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes. *Genome Research*, pages gr–216150, 2017.

[276] Edward J Grow, Ryan A Flynn, Shawn L Chavez, Nicholas L Bayless, Mark Wossidlo, Daniel Wesche, Lance Martin, Carol Ware, Catherine A Blish, Howard Y Chang, et al. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature*, 522(7555):221, 2015.

[277] Ricardo CH del Rosario, Nirmala Arul Rayan, and Shyam Prabhakar. Noncoding origins of anthropoid traits and a new null model of transposon functionalization. *Genome research*, 24(9):1469–1484, 2014.

[278] Dorothy Tuan and Wenhu Pi. In human beta-globin gene locus, erv-9 ltr retrotransposon interacts with and activates beta-but not gamma-globin gene, 2014.

[279] Deena Emera, Claudio Casola, Vincent J Lynch, Derek E Wildman, Dalen Agnew, and Günter P Wagner. Convergent evolution of endometrial prolactin expression in primates, mice, and elephants through the independent recruitment of transposable elements. *Molecular biology and evolution*, 29(1):239–247, 2011.

[280] Leonardo MR Ferreira, Torsten B Meissner, Tarjei S Mikkelsen, William Mallard, Charles W O'Donnell, Tamara Tilburgs, Hannah AB Gomes, Raymond Camahort, Richard I Sherwood, David K Gifford, et al. A distant trophoblast-specific enhancer controls hla-g expression at the maternal–fetal interface. *Proceedings of the National Academy of Sciences*, 113(19):5364–5369, 2016.

[281] Vincent J Lynch, Mauris C Nnamani, Aurélie Kapusta, Kathryn Brayer, Silvia L Plaza, Erik C Mazur, Deena Emera, Shehzad Z Sheikh, Frank Grützner, Stefan Bauersachs, et al. Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell reports*, 10(4):551–561, 2015.

[282] David Kelley and John Rinn. Transposable elements reveal a stem cell-specific class of long noncoding rnas. *Genome biology*, 13(11):R107, 2012.

[283] Georges St Laurent, Claes Wahlestedt, and Philipp Kapranov. The landscape of long noncoding rna classification. *Trends in Genetics*, 31(5):239–251, 2015.

[284] Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki, et al. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455, 2014.

[285] Jonathan Göke and Huck Hui Ng. Ctrl+ insert: retrotransposons and their contribution to regulation and innovation of the transcriptome. *EMBO reports*, page e201642743, 2016.

[286] Sabine Loewer, Moran N Cabili, Mitchell Guttman, Yuin-Han Loh, Kelly Thomas, In Hyun Park, Manuel Garber, Matthew Curran, Tamer Onder, Suneet Agarwal, et al. Large intergenic non-coding rna-ror modulates reprogramming of human induced pluripotent stem cells. *Nature genetics*, 42(12):1113–1117, 2010.

[287] Ali Zhang, Nanjiang Zhou, Jianguo Huang, Qian Liu, Koji Fukuda, Ding Ma, Zhaohui Lu, Cunxue Bai, Kounosuke Watabe, and Yin-Yuan Mo. The human long non-coding rna-ror is a p53 repressor in response to dna damage. *Cell research*, 23(3):340, 2013.

[288] Ross J Flockhart, Dan E Webster, Kun Qu, Nicholas Mascarenhas, Joanna Kovalski, Markus Kretz, and Paul A Khavari. Brafv600e remodels the melanocyte transcriptome and induces bancr to regulate melanoma cell migration. *Genome research*, 22(6):1006–1014, 2012.

[289] Eleonora Leucci, Elizabeth A Coe, Jean-Christophe Marine, and Keith W Vance. The emerging role of long non-coding rnas in cutaneous melanoma. *Pigment cell & melanoma research*, 2016.

[290] Fan Wang, Xu Li, XiaoJuan Xie, Le Zhao, and Wei Chen. Uca1, a non-protein-coding rna up-regulated in bladder carcinoma and embryo, influencing cell growth and promoting invasion. *FEBS letters*, 582(13):1919–1927, 2008.

[291] Dixie L Mager and Jonathan P Stoye. Mammalian endogenous retroviruses. In *Mobile DNA III*, pages 1079–1100. American Society of Microbiology, 2015.

[292] DA Wilkinson, NL Goodchild, TM Saxton, S Wood, and DL Mager. Evidence for a functional subclass of the rtvl-h family of human endogenous retrovirus-like sequences. *Journal of virology*, 67(6):2981–2989, 1993.

[293] THIERRY Tchenio and THIERRY Heidmann. Defective retroviruses can disperse in the human genome by intracellular transposition. *Journal of virology*, 65(4):2113–2118, 1991.

[294] Federico A Santoni, Jessica Guerra, and Jeremy Luban. Herv-h rna is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology*, 9(1):111, 2012.

[295] Kathy K Niakan, Jinnuo Han, Roger A Pedersen, Carlos Simon, and Renee A Reijo Pera. Human pre-implantation embryo development. *Development*, 139(5):829–841, 2012.

[296] Sophie Petropoulos, Daniel Edsgärd, Björn Reinius, Qiaolin Deng, Sarita Pauliina Panula, Simone Codeluppi, Alvaro Plaza Reyes, Sten Linnarsson, Rickard Sandberg, and Fredrik Lanner. Single-cell rna-seq reveals lineage and x chromosome dynamics in human preimplantation embryos. *Cell*, 165(4):1012–1026, 2016.

[297] Liying Yan, Mingyu Yang, Hongshan Guo, Lu Yang, Jun Wu, Rong Li, Ping Liu, Ying Lian, Xiaoying Zheng, Jie Yan, et al. Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*, 20(9):1131–1139, 2013.

[298] Paul Blakeley, Norah ME Fogarty, Ignacio Del Valle, Sissy E Wamaitha, Tim Xiaoming Hu, Kay Elder, Philip Snell, Leila Christie, Paul Robson, and Kathy K Niakan. Defining the three cell lineages of the human blastocyst by single-cell rna-seq. *Development*, 142(18):3151–3165, 2015.

[299] Peter Braude, Virginia Bolton, and Stephen Moore. Human gene expression first occurs between the four-and eight-cell stages of preimplantation development. *Nature*, 332(6163):459–461, 1988.

[300] Jan Tesařík, Václav Kopečnỳ, Michelle Plachot, and Jacqueline Mandelbaum. Ultrastructural and autoradiographic observations on multinucleated blastomeres of human cleaving embryos obtained by in-vitro fertilization. *Human reproduction*, 2(2):127–136, 1987.

[301] Deborah M Taylor, Pierre F Ray, Asangla Ao, Robert ML Winston, and Alan H Handyside. Paternal transcripts for glucose-6-phosphate dehydrogenase and adenosine deaminase are first detectable in the human preimplantation embryo at the three-to four-cell stage. *Molecular reproduction and development*, 48(4):442–448, 1997.

[302] Anthony T Dobson, Rajiv Raja, Michael J Abeyta, Theresa Taylor, Shehua Shen, Christopher Haqq, and Renee A Reijo Pera. The unique transcriptome through day 3 of human preimplantation development. *Human molecular genetics*, 13(14):1461–1470, 2004.

[303] Gin Flach, MH Johnson, PR Braude, RA Taylor, and VN Bolton. The transition from maternal to embryonic control in the 2-cell mouse embryo. *The EMBO journal*, 1(6):681, 1982.

[304] Q Tian Wang, Karolina Piotrowska, Maria Anna Ciemerych, Ljiljana Milenkovic, Matthew P Scott, Ronald W Davis, and Magdalena Zernicka-Goetz. A genome-wide study of gene activity reveals developmental signaling pathways in the preimplantation mouse embryo. *Developmental cell*, 6(1):133–144, 2004.

[305] Polani B Seshagiri, Debbie I McKenzie, Barry D Bavister, Judy L Williamson, and Judd M Aiken. Golden hamster embryonic genome activation occurs at the two-cell stage: Correlation with major developmental changes. *Molecular reproduction and development*, 32(3):229–235, 1992.

[306] Rita Vassena, Stéphanie Boué, Eva González-Roca, Begoña Aran, Herbert Auer, Anna Veiga, and Juan Carlos Izpisua Belmonte. Waves of early transcriptional activation and pluripotency program initiation during human preimplantation development. *Development*, 138(17):3699–3709, 2011.

[307] Ewart W Kuijk, Leni TA van Tol, Hilde Van de Velde, Richard Wubbolts, Maaike Welling, Niels Geijsen, and Bernard AJ Roelen. The roles of fgf and map kinase signaling in the segregation of the epiblast and hypoblast cell lineages in bovine and human embryos. *Development*, 139(5):871–882, 2012.

[308] Mila Roode, Kathryn Blair, Philip Snell, Kay Elder, Sally Marchant, Austin Smith, and Jennifer Nichols. Human hypoblast formation is not dependent on fgf signalling. *Developmental biology*, 361(2):358–363, 2012.

[309] Kathy K Niakan and Kevin Eggan. Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. *Developmental biology*, 375(1):54–64, 2013.

[310] Gloria S Kwon, Manuel Viotti, and Anna-Katerina Hadjantonakis. The endoderm of the mouse embryo arises by dynamic widespread intercalation of embryonic and extraembryonic lineages. *Developmental cell*, 15(4):509–520, 2008.

[311] Todd S Macfarlan, Wesley D Gifford, Shawn Driscoll, Karen Lettieri, Helen M Rowe, Dario Bonanomi, Amy Firth, Oded Singer, Didier Trono, and Samuel L Pfaff. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*, 487(7405):57–63, 2012.

[312] Zsuzsanna Izsvák, Jichang Wang, Manvendra Singh, Dixie L Mager, and Laurence D Hurst. Pluripotency and the endogenous retrovirus hervh: Conflict or serendipity? *BioEssays*, 38(1):109–117, 2016.

[313] Claire Chazaud and Yojiro Yamanaka. Lineage specification in the mouse preimplantation embryo. *Development*, 143(7):1063–1074, 2016.

[314] Anna Sahakyan and Kathrin Plath. Transcriptome encyclopedia of early human development. *Cell*, 165(4):777–779, 2016.

[315] Martin J Evans and Matthew H Kaufman. Establishment in culture of pluripotential cells from mouse embryos. *Nature*, 292(5819):154–156, 1981.

[316] James A Thomson, Joseph Itskovitz-Eldor, Sander S Shapiro, Michelle A Waknitz, Jennifer J Swiergiel, Vivienne S Marshall, and Jeffrey M Jones. Embryonic stem cell lines derived from human blastocysts. *science*, 282(5391):1145–1147, 1998.

[317] Jacob H Hanna, Krishanu Saha, and Rudolf Jaenisch. Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. *Cell*, 143(4):508–525, 2010.

[318] Gail R Martin. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proceedings of the National Academy of Sciences*, 78(12):7634–7638, 1981.

[319] Paul J Tesar, Josh G Chenoweth, Frances A Brook, Timothy J Davies, Edward P Evans, David L Mack, Richard L Gardner, and Ronald DG McKay. New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature*, 448(7150):196, 2007.

[320] I Gabrielle M Brons, Lucy E Smithers, Matthew WB Trotter, Peter Rugg-Gunn, Sun Bowen, Susana M Chuva de Sousa Lopes, Sarah K Howlett, Amanda Clarkson, Lars Ahrlund-Richter, Roger A Pedersen, et al. Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature*, 448(7150):191, 2007.

[321] Harry G Leitch, Kate Blair, William Mansfield, Harold Ayetey, Peter Humphreys, Jennifer Nichols, M Azim Surani, and Austin Smith. Embryonic germ cells from mice and rats exhibit properties consistent with a generic pluripotent ground state. *Development*, 137(14):2279–2287, 2010.

[322] Yasuhisa Matsui, Krisztina Zsebo, and Brigid LM Hogan. Derivation of pluripotential embryonic stem cells from murine primordial germ cells in culture. *Cell*, 70(5):841–847, 1992.

[323] Kinarm Ko, Marcos J Araúzo-Bravo, Natalia Tapia, Julee Kim, Qiong Lin, Christof Bernemann, Dong Wook Han, Luca Gentile, Peter Reinhardt, Boris Greber, et al. Human adult germline stem cells in question. *Nature*, 465(7301):E1–E1, 2010.

[324] Michael J Shamblott, Joyce Axelman, Shunping Wang, Elizabeth M Bugg, John W Littlefield, Peter J Donovan, Paul D Blumenthal, George R Huggins, and John D Gearhart. Derivation of pluripotent stem cells from cultured human primordial germ cells. *Proceedings of the National Academy of Sciences*, 95(23):13726–13731, 1998.

[325] James L Resnick, Lynn S Bixler, Linzhao Cheng, and Peter J Donovan. Long-term proliferation of mouse primordial germ cells in culture. *Nature*, 359(6395):550–551, 1992.

[326] John B Gurdon. The developmental capacity of nuclei taken from intestinal epithelium cells of feeding tadpoles. *Development*, 10(4):622–640, 1962.

[327] Ian Wilmut, N Beaujean, PA De Sousa, A Dinnyes, et al. Somatic cell nuclear transfer. *Nature*, 419(6907):583, 2002.

[328] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *cell*, 126(4):663–676, 2006.

[329] Thomas Vierbuchen, Austin Ostermeier, Zhiping P Pang, Yuko Kokubu, Thomas C Südhof, and Marius Wernig. Direct conversion of fibroblasts to functional neurons by defined factors. *Nature*, 463(7284):1035, 2010.

[330] Masaki Ieda, Ji-Dong Fu, Paul Delgado-Olguin, Vasanth Vedantham, Yohei Hayashi, Benoit G Bruneau, and Deepak Srivastava. Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell*, 142(3):375–386, 2010.

[331] Megan Scudellari. How ips cells changed the world. *Nature*, 534(7607):310, 2016.

[332] Carmen Lorenz, Pierre Lesimple, Raul Bukowiecki, Annika Zink, Gizem Inak, Barbara Mlody, Manvendra Singh, Marcus Semtner, Nancy Mah, Karine Auré, et al. Human ipsc-derived neural progenitors are an effective drug discovery model for neurological mtdna disorders. *Cell Stem Cell*, 20(5):659–674, 2017.

[333] Erin A Kimbrel and Robert Lanza. Current status of pluripotent stem cells: moving the first therapies to the clinic. *Nature reviews. Drug discovery*, 14(10):681, 2015.

[334] Evgenios Neofytou, Connor Galen O'Brien, Larry A Couture, and Joseph C Wu. Hurdles to clinical translation of human induced pluripotent stem cells. *The Journal of clinical investigation*, 125(7):2551, 2015.

[335] Jennifer Nichols and Austin Smith. Naive and primed pluripotent states. *Cell stem cell*, 4(6):487–492, 2009.

[336] Jamie A Hackett and M Azim Surani. Regulatory principles of pluripotency: from the ground state up. *Cell stem cell*, 15(4):416–430, 2014.

[337] Jun Wu, Daiji Okamura, Mo Li, Keiichiro Suzuki, Chongyuan Luo, Li Ma, Yupeng He, Zhongwei Li, Chris Benner, Isao Tamura, et al. An alternative pluripotent state confers interspecies chimaeric competency. *Nature*, 521(7552):316, 2015.

[338] Yoji Kojima, Keren Kaufman-Francis, Joshua B Studdert, Kirsten A Steiner, Melinda D Power, David AF Loebel, Vanessa Jones, Angelyn Hor, Gustavo de Alencastro, Grant J Logan, et al. The transcriptional and functional properties of mouse epiblast stem cells resemble the anterior primitive streak. *Cell Stem Cell*, 14(1):107–120, 2014.

[339] Christoph Bock, Evangelos Kiskinis, Griet Verstappen, Hongcang Gu, Gabriella Boulting, Zachary D Smith, Michael Ziller, Gist F Croft, Mackenzie W Amoroso, Derek H Oakley, et al. Reference maps of human es and ips cell variation enable high-throughput characterization of pluripotent cell lines. *Cell*, 144(3):439–452, 2011.

[340] Matthew G Guenther, Garrett M Frampton, Frank Soldner, Dirk Hockemeyer, Maya Mitalipova, Rudolf Jaenisch, and Richard A Young. Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. *Cell stem cell*, 7(2):249–257, 2010.

[341] Aaron M Newman and James B Cooper. Lab-specific gene expression signatures in pluripotent stem cells. *Cell stem cell*, 7(2):258–262, 2010.

[342] Mark H Chin, Mike J Mason, Wei Xie, Stefano Volinia, Mike Singer, Cory Peterson, Gayane Ambartsumyan, Otaren Aimiuwu, Laura Richter, Jin Zhang, et al. Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. *Cell stem cell*, 5(1):111–123, 2009.

[343] Zhumur Ghosh, Kitchener D Wilson, Yi Wu, Shijun Hu, Thomas Quertermous, and Joseph C Wu. Persistent donor cell gene expression among human induced pluripotent stem cells contributes to differences with human embryonic stem cells. *PloS one*, 5(2):e8975, 2010.

[344] Jie Deng, Robert Shoemaker, Bin Xie, Athurva Gore, Emily M LeProust, Jessica Antosiewicz-Bourget, Dieter Egli, Nimet Maherali, In-Hyun Park, Junying Yu, et al. Targeted bisulfite sequencing reveals changes in dna methylation associated with nuclear reprogramming. *Nature biotechnology*, 27(4):353–360, 2009.

[345] Jiho Choi, Soohyun Lee, Kendell Clement, William Mallard, Guidantonio Malagoli Tagliazucchi, Hotae Lim, In Young Choi, Francesco Ferrari, Alex Tsankov, Ramona Pop, et al. A comparison of genetically matched cell lines reveals the equivalence of human ipscs and escs. *Nature biotechnology*, 33(11):1173, 2015.

[346] Thomas O'leary, Björn Heindryckx, Sylvie Lierman, David Van Bruggen, Jelle J Goeman, Mado Vandewoestyne, Dieter Deforce, Susana M Chuva De Sousa Lopes, and Petra De Sutter. Tracking the progression of the human inner cell mass during embryonic stem cell derivation. *Nature biotechnology*, 30(3):278, 2012.

[347] Yali Huang, Rodrigo Osorno, Anestis Tsakiridis, and Valerie Wilson. In vivo differentiation potential of epiblast stem cells revealed by chimeric embryo formation. *Cell Reports*, 2(6):1571–1578, 2012.

[348] Rutger-Jan Swijnenburg, Sonja Schrepfer, Johannes A Govaert, Feng Cao, Katie Ransohoff, Ahmad Y Sheikh, Munif Haddad, Andrew J Connolly, Mark M Davis, Robert C Robbins, et al. Immunosuppressive therapy mitigates immunological rejection of human embryonic stem cell xenografts. *Proceedings of the National Academy of Sciences*, 105(35):12991–12996, 2008.

[349] Qi-Long Ying, Jason Wray, Jennifer Nichols, Laura Batlle-Morera, Bradley Doble, James Woodgett, Philip Cohen, and Austin Smith. The ground state of embryonic stem cell self-renewal. *Nature*, 453(7194):519, 2008.

[350] Aliaksandra Radzisheuskaya, Gloryn Le Bin Chia, Rodrigo L Dos Santos, Thorold W Theunissen, L Filipe C Castro, Jennifer Nichols, and José CR Silva. A defined oct4 level governs cell state transitions of pluripotency entry and differentiation into all embryonic lineages. *Nature cell biology*, 15(6):579, 2013.

[351] Athurva Gore, Zhe Li, Ho-Lim Fung, Jessica Young, Suneet Agarwal, Jessica Antosiewicz-Bourget, Isabel Canto, Alessandra Giorgetti, Mason Israel, Evangelos Kiskinis, et al. Somatic coding mutations in human induced pluripotent stem cells. *Nature*, 471(7336):63, 2011.

[352] Junfeng Ji, Siemon H Ng, Vivek Sharma, Dante Neculai, Samer Hussein, Michelle Sam, Quang Trinh, George M Church, John D Mcpherson, Andras Nagy, et al. Elevated coding mutation rate during the reprogramming of human somatic cells into induced pluripotent stem cells. *Stem cells*, 30(3):435–440, 2012.

[353] Prajna Guha, John W Morgan, Gustavo Mostoslavsky, Neil P Rodrigues, and Ashleigh S Boyd. Lack of immune response to differentiated cells derived from syngeneic induced pluripotent stem cells. *Cell stem cell*, 12(4):407–412, 2013.

[354] K Kim, A Doi, B Wen, K Ng, R Zhao, P Cahan, J Kim, MJ Aryee, H Ji, L Ehrlich, et al. Epigenetic memory in induced pluripotent stem cells. *Nature*, 467(7313):285, 2010.

[355] Gulsah Altun, Jeanne F Loring, and Louise C Laurent. Dna methylation in embryonic stem cells. *Journal of cellular biochemistry*, 109(1):1–6, 2010.

[356] William A Pastor, Di Chen, Wanlu Liu, Rachel Kim, Anna Sahakyan, Anastasia Lukianchikov, Kathrin Plath, Steven E Jacobsen, and Amander T Clark. Naive human pluripotent cells feature a methylation landscape devoid of blastocyst or germline memory. *Cell Stem Cell*, 18(3):323–329, 2016.

[357] Ohad Gafni, Leehee Weinberger, Abed AlFatah Mansour, Yair S Manor, Elad Chomsky, Dalit Ben-Yosef, Yael Kalma, Sergey Viukov, Itay Maza, Asaf Zviran, et al. Derivation of novel human ground state naive pluripotent stem cells. *Nature*, 504(7479):282, 2013.

[358] Yun-Shen Chan, Jonathan Göke, Jia-Hui Ng, Xinyi Lu, Kevin Andrew Uy Gonzales, Cheng-Peow Tan, Wei-Quan Tng, Zhong-Zhi Hong, Yee-Siang Lim, and Huck-Hui Ng. Induction of a human

pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast. *Cell stem cell*, 13(6):663–675, 2013.

[359] Yasuhiro Takashima, Ge Guo, Remco Loos, Jennifer Nichols, Gabriella Ficz, Felix Krueger, David Oxley, Fatima Santos, James Clarke, William Mansfield, et al. Resetting transcription factor control circuitry toward ground-state pluripotency in human. *Cell*, 158(6):1254–1269, 2014.

[360] Bahram Valamehr, Megan Robinson, Ramzey Abujarour, Betsy Rezner, Florin Vranceanu, Thuy Le, Amanda Medcalf, Tom Tong Lee, Michael Fitch, David Robbins, et al. Platform for induction and maintenance of transgene-free hipscs resembling ground state pluripotent stem cells. *Stem cell reports*, 2(3):366–381, 2014.

[361] Thorold W Theunissen, Benjamin E Powell, Haoyi Wang, Maya Mitalipova, Dina A Faddah, Jessica Reddy, Zi Peng Fan, Dorothea Maetzel, Kibibi Ganz, Linyu Shi, et al. Systematic identification of culture conditions for induction and maintenance of naive human pluripotency. *Cell stem cell*, 15(4):471–487, 2014.

[362] Carol B Ware, Angelique M Nelson, Brigham Mecham, Jennifer Hesson, Wenyu Zhou, Erica C Jonlin, Antonio J Jimenez-Caliani, Xinxian Deng, Christopher Cavanaugh, Savannah Cook, et al. Derivation of naive human embryonic stem cells. *Proceedings of the National Academy of Sciences*, 111(12):4484–4489, 2014.

[363] Galbha Duggal, Sharat Warrier, Sabitri Ghimire, Dorien Broekaert, Margot Van der Jeught, Sylvie Lierman, Tom Deroo, Luc Peelman, Ann Van Soom, Ria Cornelissen, et al. Alternative routes to induce naive pluripotency in human embryonic stem cells. *Stem Cells*, 33(9):2686–2698, 2015.

[364] Yuanyuan Yang, Xiaobai Zhang, Li Yi, Zhenzhen Hou, Jiayu Chen, Xiaochen Kou, Yanhong Zhao, Hong Wang, Xiao-Fang Sun, Cizhong Jiang, et al. Naïve induced pluripotent stem cells generated from β-thalassemia fibroblasts allow efficient gene correction with crispr/cas9. *Stem cells translational medicine*, 5(1):8–19, 2016.

[365] MG Carter, BJ Smagghe, AK Stewart, JA Rapley, E Lynch, KJ Bernier, KW Keating, VM Hatzi-ioannou, EJ Hartman, and Cynthia C Bamdad. A primitive growth factor, nme7ab, is sufficient to induce stable naive state human pluripotency; reprogramming in this novel growth factor confers superior differentiation. *Stem Cells*, 34(4):847–859, 2016.

[366] Thorold W Theunissen, Marc Friedli, Yupeng He, Evarist Planet, Ryan C O'Neil, Styliani Markoulaki, Julien Pontis, Haoyi Wang, Alexandra Iouranova, Michaël Imbeault, et al. Molecular criteria for defining the naive human pluripotent state. *Cell Stem Cell*, 19(4):502–515, 2016.

[367] Ge Guo, Ferdinand von Meyenn, Fatima Santos, Yaoyao Chen, Wolf Reik, Paul Bertone, Austin Smith, and Jennifer Nichols. Naive pluripotent stem cells derived directly from isolated cells of the human inner cell mass. *Stem cell reports*, 6(4):437–446, 2016.

[368] Richard L Gardner. Contributions of blastocyst micromanipulation to the study of mammalian development. *Bioessays*, 20(2):168–180, 1998.

[369] Andrzej K Tarkowski and Joanna Wróblewska. Development of blastomeres of mouse eggs isolated at the 4-and 8-cell stage. *Development*, 18(1):155–180, 1967.

[370] CA Ziomek and MH Johnson. Cell surface interaction induces polarization of mouse 8-cell blastomeres at compaction. *Cell*, 21(3):935–942, 1980.

[371] N Hillman, MI Sherman, and Chris Graham. The effect of spatial arrangement on cell determination during mouse development. *Development*, 28(2):263–278, 1972.

[372] RL Gardner and RSP Beddington. Multi-lineage 'stem'cells in the mammalian embryo. *J Cell Sci*, 1988(Supplement 10):11–27, 1988.

[373] Keisuke Kaji, Jennifer Nichols, and Brian Hendrich. Mbd3, a component of the nurd co-repressor complex, is required for development of pluripotent cells. *Development*, 134(6):1123–1132, 2007.

[374] Kazuki Kurimoto, Yukihiro Yabuta, Yasuhide Ohinata, Yukiko Ono, Kenichiro D Uno, Rikuhiro G Yamada, Hiroki R Ueda, and Mitinori Saitou. An improved single-cell cdna amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic acids research*, 34(5):e42–e42, 2006.

[375] FA Brook and RL Gardner. The origin and efficient derivation of embryonic stem cells in the mouse. *Proceedings of the National Academy of Sciences*, 94(11):5709–5712, 1997.

[376] Jennifer Nichols and Austin Smith. The origin and identity of embryonic stem cells. *Development*, 138(1):3–8, 2011.

[377] Jun Wu and Juan Carlos Izpisua Belmonte. Dynamic pluripotent stem cell states and their applications. *Cell Stem Cell*, 17(5):509–525, 2015.

[378] Leehee Weinberger, Muneef Ayyash, Noa Novershtern, and Jacob H Hanna. Dynamic stem cell states: naive to primed pluripotency in rodents and humans. *Nature reviews. Molecular cell biology*, 17(3):155, 2016.

[379] Victoria L Mascetti and Roger A Pedersen. Naivete of the human pluripotent stem cell. *Nature biotechnology*, 32(1):68–70, 2014.

[380] Christa Buecker, Hsu-Hsin Chen, Jose Maria Polo, Laurence Daheron, Lei Bu, Tahsin Stefan Barakat, Patricia Okwieka, Andrew Porter, Joost Gribnau, Konrad Hochedlinger, et al. A murine esc-like state facilitates transgenesis and homologous recombination in human pluripotent stem cells. *Cell stem cell*, 6(6):535–546, 2010.

[381] Hideki Masaki, Megumi Kato-Itoh, Ayumi Umino, Hideyuki Sato, Sanae Hamanaka, Toshihiro Kobayashi, Tomoyuki Yamaguchi, Ken Nishimura, Manami Ohtaka, Mahito Nakanishi, et al. Interspecific in vitro assay for the chimera-forming ability of human pluripotent stem cells. *Development*, 142(18):3222–3230, 2015.

[382] RL Gardner and J Rossant. Investigation of the fate of 4· 5 day post-coitum mouse inner cell mass cells by blastocyst injection. *Development*, 52(1):141–152, 1979.

[383] KIRSTIE A Lawson, Juanito J Meneses, and RA Pedersen. Clonal analysis of epiblast fate during germ layer formation in the mouse embryo. *Development*, 113(3):891–911, 1991.

[384] Janet Rossant. Mouse and human blastocyst-derived stem cells: vive les differences. *Development*, 142(1):9–12, 2015.

[385] Tomonori Nakamura, Ikuhiro Okamoto, Kotaro Sasaki, Yukihiro Yabuta, Chizuru Iwatani, Hideaki Tsuchiya, Yasunari Seita, Shinichiro Nakamura, Takuya Yamamoto, and Mitinori Saitou. A developmental coordinate of pluripotency among mice, monkeys and humans. *Nature*, 537(7618):57–62, 2016.

[386] Guangmei Yan, Guojie Zhang, Xiaodong Fang, Yanfeng Zhang, Cai Li, Fei Ling, David N Cooper, Qiye Li, Yan Li, Alain J Van Gool, et al. Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and chinese rhesus macaques. *Nature biotechnology*, 29(11):1019–1023, 2011.

[387] Maynard V Olson and Ajit Varki. Sequencing the chimpanzee genome: insights into human evolution and disease. *Nature reviews. Genetics*, 4(1):20, 2003.

[388] Laura Hewitson, Crista Martinovich, Calvin Simerly, Diana Takahashi, and Gerald Schatten. Rhesus offspring produced by intracytoplasmic injection of testicular sperm and elongated spermatids. *Fertility and sterility*, 77(4):794–801, 2002.

[389] DE Roberts, RJ Killiany, and DL Rosene. Neuron numbers in the hypothalamus of the normal aging rhesus monkey: stability across the adult lifespan and between the sexes. *Journal of Comparative Neurology*, 520(6):1181–1197, 2012.

[390] John Duncan. An adaptive coding model of neural function in prefrontal cortex. *Nature reviews. Neuroscience*, 2(11):820, 2001.

[391] Seung Hwan Han, Ichiro Sakuma, Eak Kyun Shin, and Kwang Kon Koh. Antiatherosclerotic and anti-insulin resistance effects of adiponectin: basic and clinical studies. *Progress in cardiovascular diseases*, 52(2):126–140, 2009.

[392] Alex T Kalinka, Karolina M Varga, Dave T Gerrard, Stephan Preibisch, David L Corcoran, Julia Jarrells, Uwe Ohler, Casey M Bergman, and Pavel Tomancak. Gene expression divergence recapitulates the developmental hourglass model. *Nature*, 468(7325):811, 2010.

[393] Henrik Kaessmann. Origins, evolution, and phenotypic impact of new genes. *Genome research*, 20(10):1313–1326, 2010.

[394] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470, 2008.

[395] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495–502, 2015.

[396] Michael L Metzker. Sequencing technologies–the next generation. *Nature reviews. Genetics*, 11(1):31, 2010.

[397] LeeAnn Ramsay, Maria C Marchetto, Maxime Caron, Shu-Huang Chen, Stephan Busche, Tony Kwan, Tomi Pastinen, Fred H Gage, and Guillaume Bourque. Conserved expression of transposon-derived non-coding transcripts in primate stem cells. *BMC genomics*, 18(1):214, 2017.

[398] Michael Lynch. The evolution of genetic networks by non-adaptive processes. *Nature reviews. Genetics*, 8(10):803, 2007.

[399] Gavin C Conant and Kenneth H Wolfe. Turning a hobby into a job: how duplicated genes find new functions. *Nature reviews. Genetics*, 9(12):938, 2008.

[400] Susumu Ohno. The enormous diversity in genome sizes of fish as a reflection of nature's extensive experiments with gene duplication. *Transactions of the American Fisheries Society*, 99(1):120–130, 1970.

[401] Flávio SJ de Souza, Lucía F Franchini, and Marcelo Rubinstein. Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? *Molecular biology and evolution*, 30(6):1239–1251, 2013.

[402] Nuno A Fonseca, Johan Rung, Alvis Brazma, and John C Marioni. Tools for mapping high-throughput sequencing data. *Bioinformatics*, 28(24):3169–3177, 2012.

[403] Ruiqiang Li, Yingrui Li, Karsten Kristiansen, and Jun Wang. Soap: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714, 2008.

[404] Heng Li, Jue Ruan, and Richard Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851–1858, 2008.

[405] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10(3):R25, 2009.

[406] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[407] Stephen M Rumble, Phil Lacroute, Adrian V Dalca, Marc Fiume, Arend Sidow, and Michael Brudno. Shrimp: accurate mapping of short color-space reads. *PLoS computational biology*, 5(5):e1000386, 2009.

[408] David Weese, Anne-Katrin Emde, Tobias Rausch, Andreas Döring, and Knut Reinert. Razers—fast read mapping with sensitivity control. *Genome research*, 19(9):1646–1654, 2009.

[409] Can Alkan, Jeffrey M Kidd, Tomas Marques-Bonet, Gozde Aksay, Francesca Antonacci, Fereydoun Hormozdiari, Jacob O Kitzman, Carl Baker, Maika Malig, Onur Mutlu, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics*, 41(10):1061–1067, 2009.

[410] Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[411] Kin Fai Au, Hui Jiang, Lan Lin, Yi Xing, and Wing Hung Wong. Detection of splice junctions from paired-end rna-seq data by splicemap. *Nucleic acids research*, 38(14):4570–4578, 2010.

[412] Kai Wang, Darshan Singh, Zheng Zeng, Stephen J Coleman, Yan Huang, Gleb L Savich, Xiaping He, Piotr Mieczkowski, Sara A Grimm, Charles M Perou, et al. Mapsplice: accurate mapping of rna-seq reads for splice junction discovery. *Nucleic acids research*, 38(18):e178–e178, 2010.

[413] Songbo Huang, Jinbo Zhang, Ruiqiang Li, Wenqian Zhang, Zengquan He, Tak-Wah Lam, Zhiyu Peng, and Siu-Ming Yiu. Soapsplice: genome-wide ab initio detection of splice junctions from rna-seq data. *Frontiers in genetics*, 2, 2011.

[414] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[415] Vikas Bansal. *Computational Analysis of High-Throughput Sequencing Data in Cardiac Disease and Skeletal Muscle Development*. PhD thesis, Freie Universität Berlin, 2016.

[416] Anton Valouev, David S Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M Myers, and Arend Sidow. Genome-wide analysis of transcription factor binding sites based on chip-seq data. *Nature methods*, 5(9):829–834, 2008.

[417] Jianxing Feng, Tao Liu, and Yong Zhang. Using macs to identify peaks from chip-seq data. *Current protocols in bioinformatics*, pages 2–14, 2011.

[418] Thomas Müller, Gesine Fleischmann, Katja Eildermann, Kerstin Mätz-Rensing, Peter A Horn, Erika Sasaki, and Rüdiger Behr. A novel embryonic stem cell line derived from the common marmoset monkey (callithrix jacchus) exhibiting germ cell-like characteristics. *Human Reproduction*, 24(6):1359–1372, 2009.

[419] Stephanie Wunderlich, Martin Kircher, Beate Vieth, Alexandra Haase, Sylvia Merkert, Jennifer Beier, Gudrun Göhring, Silke Glage, Axel Schambach, Eliza C Curnow, et al. Primate ips cells as tools for evolutionary analyses. *Stem cell research*, 12(3):622–629, 2014.

[420] Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2013.

[421] Aaron R Quinlan and Ira M Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[422] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):13, 2016.

[423] Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 2012.

[424] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740–742, 2014.

[425] Jianxing Feng, Clifford A Meyer, Qian Wang, Jun S Liu, X Shirley Liu, and Yong Zhang. Gfold: a generalized fold change for ranking differentially expressed genes from rna-seq data. *Bioinformatics*, 28(21):2782–2788, 2012.

[426] Hongshan Guo, Ping Zhu, Liying Yan, Rong Li, Boqiang Hu, Ying Lian, Jie Yan, Xiulian Ren, Shengli Lin, Junsheng Li, et al. The dna methylation landscape of human early embryos. *Nature*, 511(7511):606, 2014.

[427] Christine Römer, Manvendra Singh, Laurence D Hurst, and Zsuzsanna Izsvák. How to tame an endogenous retrovirus: Hervh and the evolution of human pluripotency. *Current Opinion in Virology*, 25:49–58, 2017.

[428] MZ Ratajczak, B Machalinski, W Wojakowski, J Ratajczak, and M Kucia. A hypothesis for an embryonic origin of pluripotent oct-4+ stem cells in adult bone marrow and other tissues. *Leukemia*, 21(5):860, 2007.

[429] Kazutoshi Takahashi and Shinya Yamanaka. A developmental framework for induced pluripotency. *Development*, 142(19):3274–3285, 2015.

[430] Thorsten Boroviak and Jennifer Nichols. Primate embryogenesis predicts the hallmarks of human naïve pluripotency. *Development*, 144(2):175–186, 2017.

[431] Margot Van der Jeught, Björn Heindryckx, Thomas O'leary, Galbha Duggal, Sabitri Ghimire, Sylvie Lierman, Nadine Van Roy, Susana M Chuva de Sousa Lopes, Tom Deroo, Dieter Deforce, et al. Treatment of human embryos with the tgfβ inhibitor sb431542 increases epiblast proliferation and permits successful human embryonic stem cell derivation. *Human Reproduction*, 29(1):41–48, 2013.

[432] Thorsten Boroviak, Remco Loos, Patrick Lombard, Junko Okahara, Rüdiger Behr, Erika Sasaki, Jennifer Nichols, Austin Smith, and Paul Bertone. Lineage-specific profiling delineates the emergence and progression of naive pluripotency in mammalian embryogenesis. *Developmental cell*, 35(3):366–382, 2015.

[433] Zhigang Xue, Kevin Huang, Chaochao Cai, Lingbo Cai, Chun-yan Jiang, Yun Feng, Zhenshan Liu, Qiao Zeng, Liming Cheng, Yi E Sun, et al. Genetic programs in human and mouse early embryos revealed by single-cell rna sequencing. *Nature*, 500(7464):593, 2013.

[434] Karolina Tykwinska, Roland Lauster, Petra Knaus, and Mark Rosowski. Growth and differentiation factor 3 induces expression of genes related to differentiation in a model of cancer stem cells and protects them from retinoic acid-induced apoptosis. *PloS one*, 8(8):e70612, 2013.

[435] Lalage M Wakefield and Caroline S Hill. Beyond tgf [beta]: roles of other tgf [beta] superfamily members in cancer. *Nature Reviews. Cancer*, 13(5):328, 2013.

[436] Yongjing Tian, Xiuying Zhang, Yinghua Hao, Zhengyu Fang, and Yanling He. Potential roles of abnormally expressed long noncoding rna uca1 and malat-1 in metastasis of melanoma. *Melanoma research*, 24(4):335–341, 2014.

[437] J Huang, N Zhou, K Watabe, Z Lu, F Wu, M Xu, and YY Mo. Long non-coding rna uca1 promotes breast tumor growth by suppression of p27 (kip1). *Cell death & disease*, 5(1):e1008, 2014.

[438] Matthew Gormley, Katherine Ona, Mirhan Kapidzic, Tamara Garrido-Gomez, Tamara Zdravkovic, and Susan J Fisher. Preeclampsia: novel insights from global rna profiling of trophoblast subpopulations. *American Journal of Obstetrics and Gynecology*, 2017.

[439] Julianna Zadora, Manvendra Singh, Florian Herse, Lukasz Przybyl, Nadine Haase, Michaela Golic, Hong Wa Yung, Berthold Huppertz, Judith E Cartwright, Guy S Whitley, et al. Disturbed placental imprinting in preeclampsia leads to altered expression of dlx5, a human-specific early trophoblast marker. *Circulation*, pages CIRCULATIONAHA–117, 2017.

[440] Shijun Hu, Kitchener D Wilson, Zhumur Ghosh, Leng Han, Yongming Wang, Feng Lan, Katherine J Ransohoff, Paul Burridge, and Joseph C Wu. Microrna-302 increases reprogramming efficiency via repression of nr2f2. *Stem Cells*, 31(2):259–268, 2013.

[441] Maria Suntsova, Elena V Gogvadze, Sergey Salozhin, Nurshat Gaifullin, Fedor Eroshkin, Sergey E Dmitriev, Natalia Martynova, Kirill Kulikov, Galina Malakhova, Gulnur Tukhbatova, et al. Human-specific endogenous retroviral insert serves as an enhancer for the schizophrenia-linked gene prodh. *Proceedings of the National Academy of Sciences*, 110(48):19472–19477, 2013.

[442] Julia Halo Wildschutte, Zachary H Williams, Meagan Montesion, Ravi P Subramanian, Jeffrey M Kidd, and John M Coffin. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. *Proceedings of the National Academy of Sciences*, 113(16):E2326–E2334, 2016.

[443] Ivana Grabundzija, Jichang Wang, Attila Sebe, Zsuzsanna Erdei, Robert Kajdi, Anantharam Devaraj, Doris Steinemann, Karoly Szuhai, Ulrike Stein, Tobias Cantz, et al. Sleeping beauty transposon-based system for cellular reprogramming and targeted gene insertion in induced pluripotent stem cells. *Nucleic acids research*, 41(3):1829–1847, 2012.

[444] Shi-Yan Ng, Rory Johnson, and Lawrence W Stanton. Human long non-coding rnas promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *The EMBO journal*, 31(3):522–533, 2012.

[445] Jacob Hanna, Albert W Cheng, Krishanu Saha, Jongpil Kim, Christopher J Lengner, Frank Soldner, John P Cassady, Julien Muffat, Bryce W Carey, and Rudolf Jaenisch. Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse escs. *Proceedings of the National Academy of Sciences*, 107(20):9222–9227, 2010.

[446] Daniel Wolf and Stephen P Goff. Embryonic stem cells use zfp809 to silence retroviral dnas. *Nature*, 458(7242):1201, 2009.

[447] David C Schultz, Kasirajan Ayyanathan, Dmitri Negorev, Gerd G Maul, and Frank J Rauscher. Setdb1: a novel kap-1-associated histone h3, lysine 9-specific methyltransferase that contributes to hp1-mediated silencing of euchromatic genes by krab zinc-finger proteins. *Genes & development*, 16(8):919–932, 2002.

[448] David C Schultz, Josh R Friedman, and Frank J Rauscher. 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 & development*, 15(4):428–443, 2001.

[449] Simon Quenneville, Priscilla Turelli, Karolina Bojkowska, Charlène Raclot, Sandra Offner, Adamandia Kapopoulou, and Didier Trono. The krab-zfp/kap1 system contributes to the early embryonic establishment of site-specific dna methylation patterns maintained during development. *Cell reports*, 2(4):766–773, 2012.

[450] Maaike Welling and Niels Geijsen. Uncovering the true identity of naive pluripotent stem cells. *Trends in cell biology*, 23(9):442–448, 2013.

[451] Xiaoxia Qi, Teng-Guo Li, Jing Hao, Jie Hu, Jing Wang, Holly Simmons, Shigeto Miura, Yuji Mishina, and Guang-Quan Zhao. Bmp4 supports self-renewal of embryonic stem cells by inhibiting mitogen-activated protein kinase pathways. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6027–6032, 2004.

[452] Céline Vallot, Christophe Huret, Yann Lesecque, Alissa Resch, Noufïssa Oudrhiri, Annelise Bennaceur, Laurent Duret, and Claire Rougeulle. Xact, a long non-coding transcript coating the active x chromosome in human pluripotent cells. *Epigenetics & Chromatin*, 6(1):O33, 2013.

[453] Hongwei Chen, Irène Aksoy, Fabrice Gonnot, Pierre Osteil, Maxime Aubry, Claire Hamela, Cloé Rognard, Arnaud Hochard, Sophie Voisin, Emeline Fontaine, et al. Reinforcement of stat3 activity reprogrammes human embryonic stem cells to naive-like pluripotency. *Nature communications*, 6, 2015.

[454] Kate Hardy. Apoptosis in the human embryo. *Reviews of Reproduction*, 4(3):125–134, 1999.

[455] Dušan Fabian, Juraj Koppel, and Poul Maddox-Hyttel. Apoptotic processes during mammalian preimplantation development. *Theriogenology*, 64(2):221–231, 2005.

[456] Gkikas Magiorkinis, Daniel Blanco-Melo, and Robert Belshaw. The decline of human endogenous retroviruses: extinction and survival. *Retrovirology*, 12(1):8, 2015.

[457] Parsa Hosseini, Arianne Tremblay, Benjamin F Matthews, and Nadim W Alkharouf. An efficient annotation and gene-expression derivation tool for illumina solexa datasets. *BMC research notes*, 3(1):183, 2010.

[458] Simon Andrews et al. Fastqc: a quality control tool for high throughput sequence data, 2010.

[459] David S DeLuca, Joshua Z Levin, Andrey Sivachenko, Timothy Fennell, Marc-Danie Nazaire, Chris Williams, Michael Reich, Wendy Winckler, and Gad Getz. Rna-seqc: Rna-seq metrics for quality control and process optimization. *Bioinformatics*, 28(11):1530–1532, 2012.

[460] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):pp–10, 2011.

[461] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[462] Davide Risso, Sandrine Dudoit, Maintainer Davide Risso, Depends Biobase, Suggests BiocStyle, and Preprocessing biocViews DifferentialExpression. Package 'ruvseq'. *Bioconductor*, 2014.

[463] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. Htseq—a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 2015.

[464] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, 7(3):562, 2012.

[465] Bo Li and Colin N Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323, 2011.

[466] Simon Anders and Wolfgang Huber. Differential expression of rna-seq data at the gene level–the deseq package. *Heidelberg, Germany: European Molecular Biology Laboratory (EMBL)*, 2012.

[467] Yarden Katz, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nature methods*, 7(12):1009–1015, 2010.

[468] Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

[469] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. Model-based analysis of chip-seq (macs). *Genome biology*, 9(9):R137, 2008.

[470] Liguo Wang, Junsheng Chen, Chen Wang, Liis Uusküla-Reimand, Kaifu Chen, Alejandra Medina-Rivera, Edwin J Young, Michael T Zimmermann, Huihuang Yan, Zhifu Sun, et al. Mace: model based analysis of chip-exo. *Nucleic acids research*, 42(20):e156–e156, 2014.

[471] Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, et al. De novo transcript sequence reconstruction from rna-seq: reference generation and analysis with trinity. *Nature protocols*, 8(8), 2013.

[472] Marcel H Schulz, Daniel R Zerbino, Martin Vingron, and Ewan Birney. Oases: robust de novo rna-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–1092, 2012.

[473] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57, 2012.

[474] Fuchou Tang, Catalin Barbacioru, Siqin Bao, Caroline Lee, Ellen Nordman, Xiaohui Wang, Kaiqin Lao, and M Azim Surani. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell rna-seq analysis. *Cell stem cell*, 6(5):468–478, 2010.

[475] Qiaolin Deng, Daniel Ramsköld, Björn Reinius, and Rickard Sandberg. Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–196, 2014.

[476] Han Qin, Miroslav Hejna, Yanxia Liu, Michelle Percharde, Mark Wossidlo, Laure Blouin, Jens Durruthy-Durruthy, Priscilla Wong, Zhongxia Qi, Jingwei Yu, et al. Yap induces human naive pluripotency. *Cell reports*, 14(10):2301–2312, 2016.

[477] Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, et al. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75, 2012.

Appendix

Table .4: Gene ontology of Most Variable Genes (MVGs) in human preimplantation embryos

GO biological process complete Human embryonic MVGs	**Total genes in ontology**	**MVGs in ontology**	**expected count**	**fold Enrichment**	***P-value***
reproductive system development (GO:0061458)	422	83	31.55	2.63	*7.57E-11*
reproductive structure development (GO:0048608)	418	82	31.25	2.62	*1.26E-10*
embryo development (GO:0009790)	925	137	69.15	1.98	*7.53E-10*
in utero embryonic development (GO:0001701)	324	67	24.22	2.77	*3.36E-09*
developmental process in reproduction (GO:0003006)	631	103	47.17	2.18	*4.49E-09*
circulatory development (GO:0072359)	807	122	60.33	2.02	*5.00E-09*
regulation of cell proliferation (GO:0008284)	859	127	64.21	1.98	*7.25E-09*
vasculature development (GO:0001944)	485	84	36.26	2.32	*3.72E-08*
blood vessel development (GO:0001568)	462	81	34.54	2.35	*5.03E-08*
anatomical structure in morphogenesis (GO:0048646)	833	121	62.27	1.94	*7.84E-08*
system development (GO:0072358)	495	84	37	2.27	*1.02E-07*
blood vessel morphogenesis (GO:0048514)	377	68	28.18	2.41	*7.96E-07*
gland development (GO:0048732)	404	71	30.2	2.35	*9.30E-07*
placenta development (GO:0001890)	147	37	10.99	3.37	*3.91E-06*
embryonic morphogenesis (GO:0048598)	558	86	41.71	2.06	*5.98E-06*
tube development (GO:0035295)	560	86	41.86	2.05	*7.06E-06*
regulation of cell migration (GO:0030334)	702	101	52.48	1.92	*7.18E-06*
embryonic placenta development (GO:0001892)	87	26	6.5	4	*5.17E-05*
regulation of cell differentiation (GO:0045596)	637	91	47.62	1.91	*6.94E-05*
embryo development ending in birth (GO:0009792)	572	84	42.76	1.96	*8.05E-05*
angiogenesis (GO:0001525)	294	53	21.98	2.41	*9.40E-05*
chordate embryonic development (GO:0043009)	567	83	42.39	1.96	*1.13E-04*
response to drug (GO:0042493)	412	66	30.8	2.14	*1.43E-04*
response to growth factor (GO:0070848)	497	75	37.15	2.02	*1.73E-04*
response to growth factor stimulus (GO:0071363)	470	72	35.13	2.05	*1.86E-04*
urogenital system development (GO:0001655)	304	53	22.73	2.33	*2.72E-04*
embryonic organ development (GO:0048568)	421	66	31.47	2.1	*3.13E-04*
digestive system development (GO:0055123)	139	32	10.39	3.08	*4.29E-04*
morphogenesis involved in differentiation (GO:0000904)	500	74	37.38	1.98	*4.56E-04*
regulation of neuron differentiation (GO:0045664)	568	81	42.46	1.91	*4.97E-04*

Table .1: Top 5 genes expressed to mark the specific lineage of human early embryos

Lineage	*Gene*	Fold Enrichment (Log scale)
8-cell	*LEUTX*	8.614026583
8-cell	*GDF9*	5.785531319
8-cell	*MBD3L3*	6.866271663
8-cell	*NLRP13*	6.300821806
8-cell	*NLRP4*	6.457231726
Morulae	*MAGEB10*	4.863758933
Morulae	*IL13RA2*	4.496449198
Morulae	*BTN1A1*	4.437536663
Morulae	*DPRX*	4.290961165
Morulae	*FUT3*	4.217968529
Early Blastocyst	*GDPD2*	2.929824136
Early Blastocyst	*DHRS3*	2.704300443
Early Blastocyst	*MX1*	2.412899315
Early Blastocyst	*ATP6V0D2*	2.775312352
Early Blastocyst	*SLC17A5*	2.265813466
Mid Blastocyst	*RGS13*	2.669801474
Late Blastocyst	*CGA*	6.941708535
Late Blastocyst	*H19*	4.744489046
Late Blastocyst	*CCKBR*	2.74498718
Late Blastocyst	*PGF*	3.634187473
Late Blastocyst	*LGALS3*	3.420044108
Inner cell mass	*IL6R*	5.9217384
Inner cell mass	*SPIC*	4.528223891
Inner cell mass	*NANOGNB*	4.197172229
Inner cell mass	*CLDN19*	2.904596878
Inner cell mass	*PRSS3*	2.207050448
Trophectoderm	*ABCG2*	2.71907
Trophectoderm	*DLX3*	3.228201
Trophectoderm	*GATA2*	2.101643
Trophectoderm	*S100A6*	2.504125
Trophectoderm	*STS*	2.932098
epiblast	*NODAL*	4.344342
epiblast	*GDF3*	2.735747
epiblast	*MEG3*	2.880511
epiblast	*TDGF1*	3.327821
epiblast	*ATP12A*	3.0891
Primitive Endoderm	*APOA1*	5.628510
Primitive Endoderm	*RSPO3*	4.834071
Primitive Endoderm	*COL4A1*	4.956160
Primitive Endoderm	*GATA4*	4.312327
Primitive Endoderm	*LINC00261*	4.413228

Table .2: List of tools used in this study

Bioinformatic tools	***Employed for***	**Reference**
Extraction		
CASAVA	*Base calling from Illumina raw intensity to fastq files*	[457]
sratools	*Online binary files to fastq files*	NCBI-suit
Quality Control		
FASTQC	*Quality of fatsq files*	[458]
SAMtools	*Statistics of mapped sequencing reads*	[406]
RSeQC	*Visualization of RNA-seq mapping quality*	[459]
Trimming		
cutadapt	*Trimming unwanted nucleotides from sequences*	[460]
FASTX	*post-processing of nucleotides from sequences*	[461]
Normalization		
RUV	*Removal of Unwanted variable genes/samples*	[462]
sva	*Normalization of batch effect*	[422, 423]
short-read aligner		
Bowtie	*Alignment of ChIP-seq/RNA-seq data*	[405]
BWA	*Alignment of DNA-seq data*	[406]
de-novo splice aligner		
MapSplice	*Chimera detection*	[412]
RNA-Seq aligner		
TopHat	*splice junction mapper for RNA-Seq reads*	[410]
STAR	*Alignment of RNA-seq reads to a reference genome*	[414]
raw count calculator		
featureCount	*Uniquely mapped reads over gene body*	[420]
HTSeq-count	*Uniquely mapped reads over exons and transcripts*	[463]
Quantification		
Cufflinks	*de-novo Transcritpt quantification*	[464]
RSEM	*calculating counts, FPKM, TPM*	[465]
in-house	*TPM calculation, cross-platform and cross-species normalization*	
Differential expression		
DESeq	*Differential expression at counts level (replicated data)*	[466]
GFOLD	*Differential expression at RPKM level (unreplicated data)*	[425]
in-house	*Relative expression on unbalanced data*	
Splicing		
MISO	*differental splicing on unreplicated data*	[467]
DEXSeq	*differental splicing on replicated data*	[466]
in-house	*Alternate/aberrant splicing*	
Microarray		
lumi	*Converting Intensity to expression set*	[468]
limma	*differential expression, normalization*	[468]
in-house	*cross-platform and cross-species analysis*	
Methylation		
in-house	*model based methylation analysis*	
ChIP-seq/ChiP-exo		
macs2	*peak calling ChiP-seq/ChIP-exo*	[469]
MACE	*ChiP-exo processing*	[470]
in-house	*Annotation of peaks and calculation of signal*	
Single-cell RNA-seq		
Seurat	*Clusters and Markers discovery and visualization*	[395]
SCDE	*Differenttial expression*	[424]
in-house	*Processing and Plotting*	
de-novo assembly		
Trinity	*Transcriptome assembly*	[471]
Oases	*Genome or Transcriptome assembly*	[472]

Table .3: List of dataset used in this study with their accession ID

Bulk RNA-seq (Human)	**Accession-ID**	***Number of samples***	***Reference***
5iL (SSEA-Neg, SSEA-Pos), UCLA20n	GSE76970	*16*	*[356]*
Reset Cells RNA-seq	E-MTAB-2857	*6*	*[359]*
HNES cells	E-MTAB-4461	*9*	*[426]*
Chimpanzee, *Bonobo* and Human PSC	GSE47626	*12*	*[93]*
Gorilla and *Callithrix* PSC	This study	*4*	*This study*
KAP1-KD	GSE58323	*10*	*[183]*
HERV-H-KD	GSE38993	*4*	*[209]*
Encode	GSE33480	*24*	*[473]*
CAGE dataset	GSE34448	*78*	*[284]*
Single-cell RNA-seq (Human)			
Embryonic stages (Stages)	GSE36552	*124*	*[297]*
Embryonic days (E3,E4,E5,E6 and E7)	E-MTAB-3929	*1285*	*[296]*
Blastocyst lineages (EPI, PE and TE)	GSE66507	*30*	*[298]*
Single-cell RNA-seq (Mouse)			
Mouse embryogenesis (Stages)	GSE57249	*16*	*[474]*
Mouse embryogenesis (Stages)	GSE45719	*229*	*[475]*
Single-cell RNA-seq (*Cynomolgus*)			
Cynomolgus embryogenesis	GSE74767	*492*	*[385]*
RRBS (Human)			
human early embryos	GSE49828	*42*	*[426]*
5iL SSEA-Neg and UCLA 20n	GSE76970	*6*	*[356]*
Microarray (Human)			
Reset cells	E-MTAB-2856	*6*	*[359]*
5iL, 6iL WIBR cells	GSE59435	*12*	*[361]*
Blastocyst, Morulae, ES cells	GSE29397	*9*	*[306]*
STAT3, OSCAR Naive cells	GSE55708	*12*	*[453]*
YAP activated Naive cells	GSE69200	*6*	*[476]*
WIBR3,C1,C2,BGO1, H9 Naive cells	GSE46872	*24*	*[357]*
GFPHigh and GFPLow	GSE54726	*6*	*This study [212]*
ChIP-sequencing (Human)			
KLF4 in HiPEC	GSE56567	*8*	*[208]*
KAP1 in HiPSC	GSE84382	*8*	*[183]*
Histone marks hESCs	GSE54471	*53*	*[477]*
KRAB-ZNF (ChIP-exo)	GSE78099	*230*	*[103]*
Encode TFs	GSE31477	*426*	*[473]*
Encode Chromatin	GSE35583	*171*	*[473]*

Table .5: Gene ontology of Most Variable Genes (MVGs) in *Cynomolgus* preimplantation embryos

GO biological process complete Cynomolgus embryonic MVGs	**Total genes in ontology**	**MVGs in ontology**	**expected count**	**fold Enrichment**	***P-value***
anatomical structure morphogenesis (GO:0009653)	2003	151	56.17	2.69	*1.26E-25*
cell differentiation (GO:0030154)	3514	205	98.55	2.08	*1.86E-22*
tissue development (GO:0009888)	1651	124	46.3	2.68	*7.22E-20*
regulation of developmental process (GO:0050793)	2340	153	65.63	2.33	*1.09E-19*
animal organ development (GO:0048513)	2981	177	83.6	2.12	*3.96E-19*
embryonic morphogenesis (GO:0048598)	558	66	15.65	4.22	*2.39E-18*
embryo development (GO:0009790)	925	86	25.94	3.32	*4.09E-18*
nervous system development (GO:0007399)	2229	138	62.51	2.21	*3.85E-15*
animal organ morphogenesis (GO:0009887)	880	77	24.68	3.12	*2.26E-14*
regulation of developmental process (GO:0051094)	1213	91	34.02	2.68	*2.04E-13*
anatomical formation in morphogenesis (GO:0048646)	833	72	23.36	3.08	*5.97E-13*
epithelium development (GO:0060429)	1068	83	29.95	2.77	*8.82E-13*
circulatory system development (GO:0072359)	807	69	22.63	3.05	*4.78E-12*
neurogenesis (GO:0022008)	1513	99	42.43	2.33	*4.04E-11*
regulation of cell differentiation (GO:0045595)	1648	104	46.22	2.25	*7.26E-11*
tube development (GO:0035295)	560	54	15.71	3.44	*8.37E-11*
generation of neurons (GO:0048699)	1415	94	39.68	2.37	*9.28E-11*
regulation of cell proliferation (GO:0042127)	1570	100	44.03	2.27	*1.51E-10*
gastrulation (GO:0007369)	158	28	4.43	6.32	*3.15E-10*
endoderm development (GO:0007492)	74	20	2.08	9.64	*8.31E-10*
tube morphogenesis (GO:0035239)	335	39	9.4	4.15	*1.89E-09*
regulation of cell proliferation (GO:0008284)	859	66	24.09	2.74	*2.79E-09*
cell development (GO:0048468)	1489	93	41.76	2.23	*4.45E-09*
regionalization (GO:0003002)	312	37	8.75	4.23	*4.64E-09*
central nervous system development (GO:0007417)	891	67	24.99	2.68	*4.75E-09*
anterior/posterior pattern specification (GO:0009952)	197	29	5.52	5.25	*1.01E-08*
positive regulation of gene expression (GO:0010628)	1789	104	50.17	2.07	*1.14E-08*
skeletal system development (GO:0001501)	465	45	13.04	3.45	*1.38E-08*
epithelial tube morphogenesis (GO:0060562)	298	35	8.36	4.19	*2.60E-08*
embryonic organ development (GO:0048568)	421	42	11.81	3.56	*3.06E-08*
pattern specification process (GO:0007389)	404	41	11.33	3.62	*3.30E-08*
mesenchyme development (GO:0060485)	194	28	5.44	5.15	*3.89E-08*

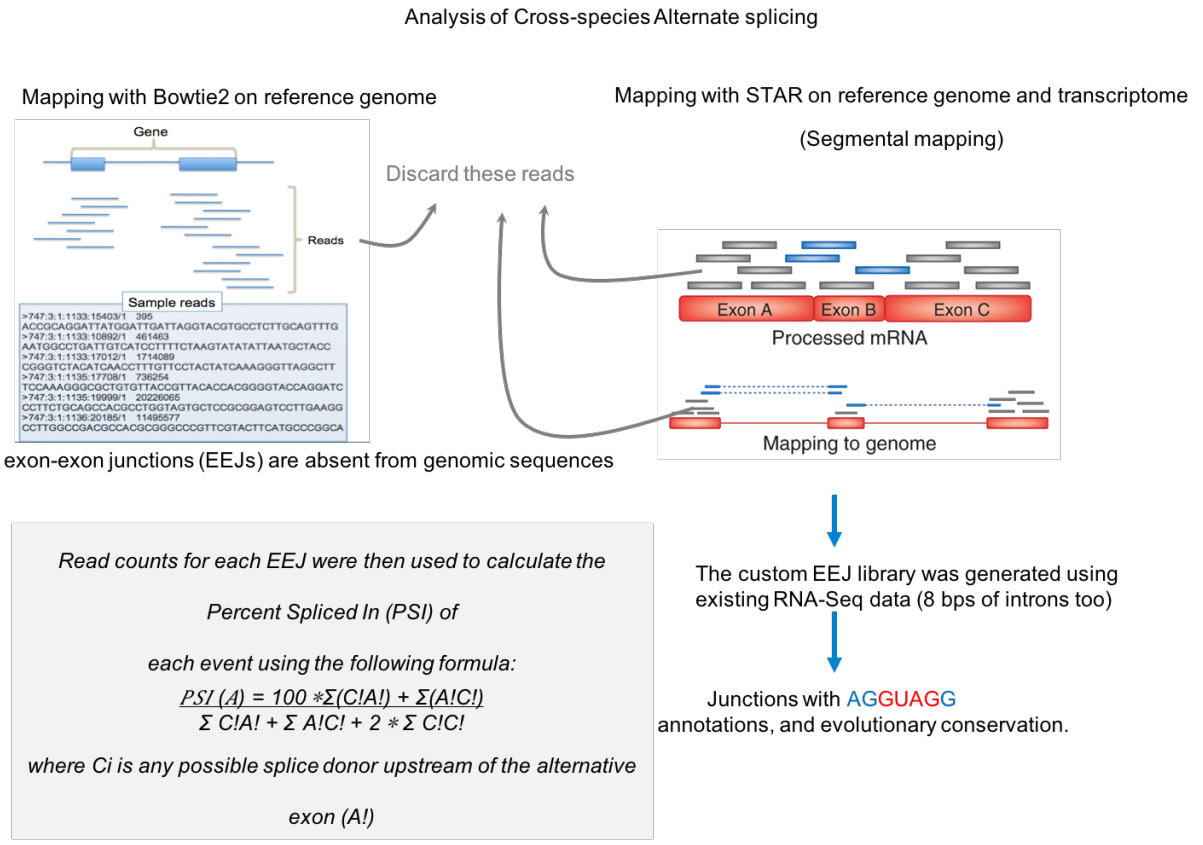


Figure .5:
This schematic represents the pipeline junction reads splicing events

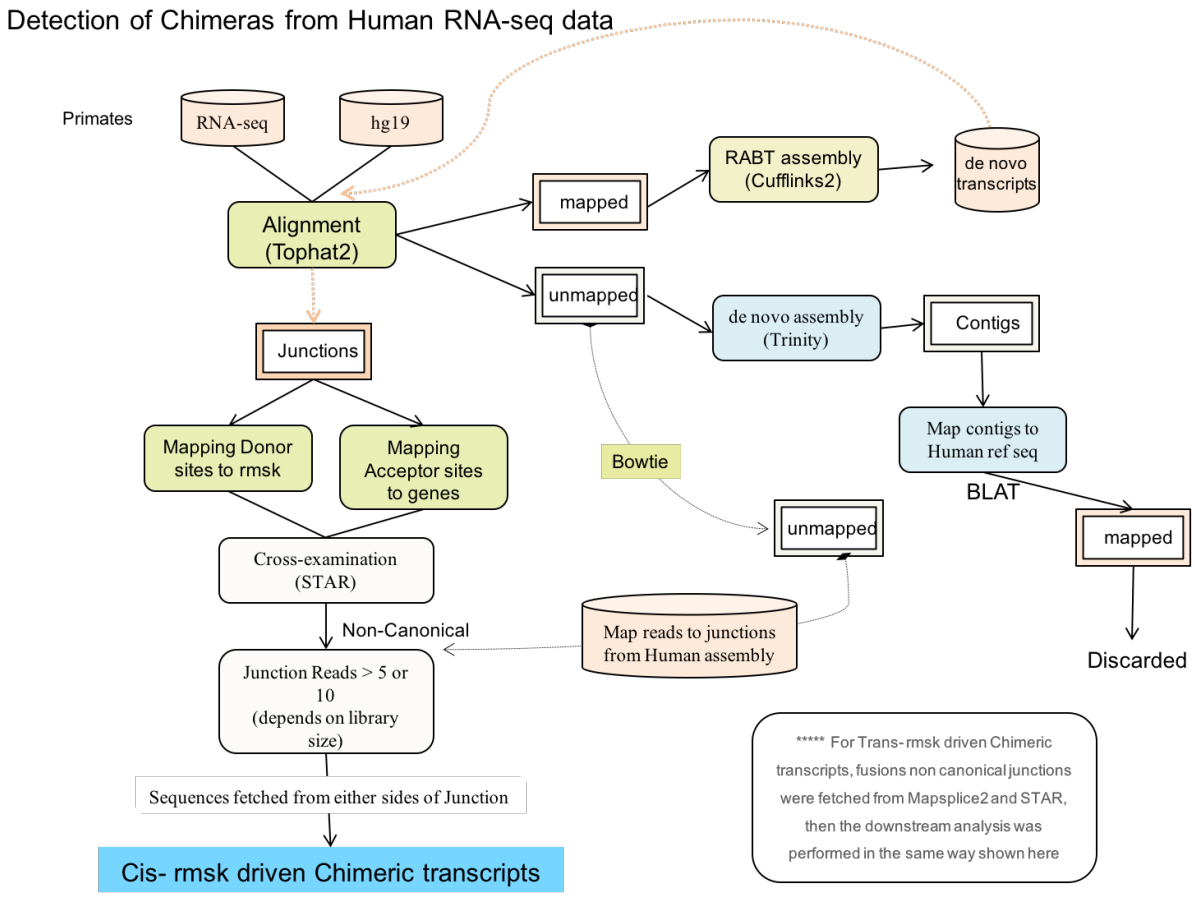


Figure .6:
This schematic represents the pipeline junction reads splicing events chimeric transcripts with repeat elements

Glossary

Alu An Alu element is a short stretch of DNA originally characterized by the action of the Arthrobacter luteus (Alu) restriction endonuclease. Alu elements are the most abundant transposable elements, containing over one million copies dispersed throughout the human genome 17

CAGE-seq Cap Analysis of Gene Expression followed by sequencing; A method used to precisely map the transcription start sites of capped RNAs genome-wide 33

Euchromatin Euchromatin a loosely packed form of chromatin that is gene rich and usually (but not always) associated with active transcription Source element the parental transposon insertion that gave rise to the copy in the new location 24

Heterochromatin a tightly packed form of chromatin that results in gene silencing. This is in contrast to euchromatin 24

In vitro fertilization The fertilization of the oocyte by sperm in a Petri dish 35

L1 LINE1 (also L1 and LINE-1) are transposable elements in the DNA of some organisms and belong to the group of Long interspersed nuclear elements (LINEs). L1 comprise approximately 17% of the human genome. The majority of L1 in the human genome are inactive; however, some retained the ability to retrotranspose 17

Latent hypothesis A few researchers advance the hypothesis that TEs are both mutualists and extreme parasites. TEs have the same pros and cons as any regulatory mechanism, the negative effects being comparable to losses of cells during meiosis or lethal errors in DNA replication, perhaps harmful for an individual but mainly beneficial for a population 16

Lateral Gene Transfer the transmission of DNA (deoxyribonucleic acid) between different genomes 19

Neutral selection changes in allele pools that are the result of random events instead of any advantageous or deleterious effect on the species 19

Orthologous homologous (similar) sequences in two different species; shared in an ancestor and separated by the speciation event 27

Rolling circle a replication process used by many eukaryotic viruses to multiply their circular genome; the replication intermediates are circular with a tail of newly made DNA 18

SVA SINE-VNTR-Alu (SVA) elements are present in hominoid primates and are divided into 6 subfamilies (SVA-A to SVA-F) and active in the human population 17

Transdifferentiation conversion of a cell type present in one tissue or organ into a cell type from another tissue or organ without going through a pluripotent cell state. Transdifferentiation between some cell types can occur naturally in response to injury and can be induced experimentally 39

Transposase Transposase is an enzyme that binds to the end of a transposon and catalyzes the movement of the transposon to another part of the genome by a cut and paste mechanism or a replicative transposition mechanism 18

aneuploidy Chromosomal abnormality characterized by an abnormal chromosome number 35

assisted reproductive technology ART encompasses clinical procedures including stimulation of ovulation via hormonal induction, intrauterine insemination (IUI), IVF and intracytoplasmic sperm injection (ICSI), a variation of IVF in which the sperm is injected directly into the oocyte cytoplasm 37

blastomeres one of the cells that are produced during cleavage of a zygote and that form the morula 41

cis-regulatory sequences Segments of DNA that regulate the transcription of adjacent genes 31

cleavage divisions A series of cell divisions after fertilization in which the net size of the embryo remains the same, but following DNA synthesis mitosis results in cells of approximately equal, decreased size. In humans, there are three cleavage divisions from 1 cell to 2 cells, 2 cells to 4 cells and 4 cells to 8 cells. 36

compaction A process during early embryo development, when blastomeres adhere to each other to form a cluster of cells (the morula) 36

embryonic genome activation The process during which the embryonic genome is activated, i.e. when transcription is evident (day 3 of human embryo development, at the 4- to 8-cell stage 35–37

embryonic transition The process of transferring embryos from in vitro culture to the uterus. This is often done at day 3 (at the 4- to 8-cell stage), but is now increasingly performed at day 5 (blastocyst stage) 36

endonucleases an enzyme that breaks down a nucleotide chain into two or more shorter chains by cleaving the internal covalent bonds linking nucleotides 16

enhancer enhancer sequences are regulatory DNA sequences that, when bound by specific proteins called transcription factors, enhance the transcription of an associated gene 26

epiblast The part of the embryo containing pluripotent cells that are able to give rise to all the tissues of the fetus 36, 37

genetic and epigenetic reprogramming The reversal of cell fate from a differentiated state to an embryonic state. In vivo, this occurs during embryogenesis with the innate reprogramming of the germ cell pronuclei to an embryonic fate. Differentiated somatic cells can also be reprogrammed by

somatic cell nuclear transfer (SCNT) to an oocyte, or in vitro by transgenically expressing a set of pluripotency-associated transcription factors (induced pluripotency) 36

genetic drift A process by which mutations become fixed in the population only by chance 32

genome shock hypothesis TEs when triggered can bring about large-scale chromosomal rearrangements which might collectively shape the stressed host genome and facilitate adaptive evolution, as demonstrated in cereal genomes 15

implantation In embryology, implantation refers specifically to the attachment of the fertilized egg to the uterine lining, which occurs approximately 6 or 7 days after conception (fertilization) 36

indels insertion and deletion. This is a class of sequence variation that results in a net gain or loss of nucleotides at the position 31

inner cell mass Comprises pluripotent cells that are able to give rise to all cells of the fetus 36, 37

insulator An insulator is a genetic boundary element that blocks the interaction between enhancers and promoters 27

molecular domestication inserted TEs can become fixed in the genome of a species and serve as a source for novel genetic loci. In some cases, accumulated mutations have caused neofunctionalization of inserted TEs 15

petri dish A Petri dish (alternatively known as a Petri plate or cell-culture dish), named after the German bacteriologist Julius Richard Petri, is a shallow cylindrical glass or plastic lidded dish that biologists use to culture cells 38

pluripotent One of the cells that are self-replicating, are derived from human embryos or human fetal tissue, and are known to develop into cells and tissues of the three primary germ layers. Although human pluripotent stem cells may be derived from embryos or fetal tissue, such stem cells are not themselves embryos 38

polymerase enzymes that synthesize DNA or RNA molecules from deoxyribonucleotides, the building blocks of DNA or RNA respectively 16

primitive endoderm Extra-embryonic cells that do not contribute to the fetus; instead, they give rise to extra-embryonic endoderm cells that will form the yolk sac 36, 37

promoter promoter is a region of DNA that initiates transcription of a particular gene. Promoters are located near the transcription start sites of genes, on the same strand and upstream on the DNA 26

provirus a form of a virus that is integrated into the genetic material of a host cell and by replicating with it can be transmitted from one cell generation to the next without causing lysis 19

purifying selection Advantageous selection against mutations that are deleterious to the fitness of the individual 31

reporter assay A putative cis-regulatory DNA sequence is cloned upstream of a reporter gene (such as luciferase) either in an episomal vector or as a chromosomally integrated construct and tested for its ability to enhance transcription of the reporter gene 33

reverse transcriptase a polymerase especially of retroviruses that catalyzes the formation of DNA using RNA as a template 16

reverse transcription the process of synthesizing DNA using RNA as a template and reverse transcriptase as a catalyst 16

self-renewal Self-renewal is the process of giving rise to indefinitely more cells of the same cell type. All stem cells have the capacity to self-renew by dividing 38

structural and enzymatic proteins The group antigens form the viral core structure, RNA genome binding proteins, and are the major proteins comprising the nucleoprotein core particle. Reverse transcriptase is the essential enzyme that carries out the reverse transcription process that take the RNA genome to a double-stranded DNA preintegrate form. The molecular gymnastics of the latter process are outlined below. The reverse transcriptase gene also encodes an Integrase activity and an RNase H activity that functions during genome reverse transscription 17

structural variation Genomic variation resulting from large-scale DNA mutations such as deletions, insertions, or rearrangements 29

t-SNE tSNE is a non-linear dimensionality reduction method. Note that in tSNE, the perplexity parameter is an estimate of the number of effective neighbors 51

totipotent capable of developing into a complete organism or differentiating into any of its cells or tissues 40

trans-regulatory sequences Segments of DNA that regulate the transcription of regions distant from adjacent genes; mostly the genes on different chromosome 32

transposase an enzyme that binds to the end of a transposon and catalyzes the movement of the transposon to another part of the genome by a cut and paste mechanism or a replicative transposition mechanism 16

transposed elements Numerous transposable elements have lost their ability to transpose so I think that they should rather be called "Transposed elements" 37

trophectoderm Extra-embryonic cells that surround the ICM and, upon implantation, give rise to the placental cytotrophoblast, syncytiotrophoblast and extravillous trophoblast 36, 37

xeno-pluripotency capability of PSCs from one species to enter into the early embryonic developmental program of another species and contribute to chimera formation 40

Acronyms

3C Chromatin conformation capture 33

ADAR Adenosine Deaminase, RNA Specific 25

AMD age-related macular degeneration 40

APOBEC apolipoprotein B mRNA editing enzyme 25

ATM Ataxia Telangiectasia Mutated (Serine/Threonine Kinase) 26

BMP bone morphogenetic protein 56, 57, 59, 66

C-GATE Catalogue of genes affected by TEs 32

ChIP-seq Chromatin immunoprecipitation followed by sequencing 33

CPM Counts per million 60, 62, 64

DNMT DNA methyltransferase (DNA MTase) family of enzymes catalyze the transfer of a methyl group to DNA. DNA methylation serves a wide variety of biological functions 24

EGC embryonic germ cells 39

Env Env in the envelope protein 17

Epi Epiblast 37

ERK extracellular signal regulated kinase 93

ERV Endogenous Retro-Viral Sequences 19

ESET ERG-associated protein with SET domain 24

FGF fibroblast growth factor 97

FPKM Fragment per million per kilobase 62–65, 83, 100

Gag Gag is a polyprotein and is an acronym for Group Antigens (ag) 17

GSK glycogen synthase kinase 93

HERVs Human Endogenous Retro-Viruses 26

hESCs human embryonic stem cells 38

HPAT human pluripotency-associated transcript 33

IAP intracisternal A-type particle 26

ICM inner cell mass 37

ICR Internal Coding Region 17

iPSCs induced Pluripotent Stem Cells 39

KAP1 Krüppel-associated box 25, 88–92, 109, 113, 164

KB Kilobases 17

KRAB Krueppel-associated box 25

LIF leukemia inhibitory factor 95

LINE Long Interspersed Nuclear Element 17

lncRNA long intergenic noncoding RNA 33

LTRs Long Terminal Repeats 17

MAPK mitogen-activated protein kinase 59

miRNA microRNA 114

MITEs Miniature inverted repeat transposable elements 18

MYA Million Years Ago 17

NWM New World Monkeys 80

ORF Open Reading Frame 20

OSKM octamer-binding protein 3/4 (OCT3/4), SOX2, Krüppel-like factor 4 (KLF4) and MYC)-induced reprogramming 86

OWM Old World Monkeys 80

PBS primer binding site 29

PGC primordial germ cells 39

PIWI Piwi Like RNA-Mediated Gene Silencing 26

Pol Pol is the reverse transcriptase 17

PrE Primitive endoderm 37

RACE Rapid amplification of cDNA ends 33

RISC RNA-induced silencing complex 26

RNA-seq High-throughput sequencing of complementary DNAs 33

RPE retinal pigment epithelial 40

SCNT somatic cell nuclear transfer 39

SINE Short Interspersed Nuclear Element 17

single cell RNA-seq High-throughput sequencing of complementary DNAs of one cell at a time 35

SVA SINE–VNTR–Alu 17

TE Trophectoderm 37

TES Transcription End Site 33

TFs Transcription Factors 19

TIR Terminal Inverted Repeats 18

TPM Transcript per million 51, 55–58, 62, 64, 67, 73, 74, 77

transposcriptome genome-wide transcriptional repertoire of transposable elements 106

TrEs Transposable Elements 15

TSS Transcription Start Site 33

UHRF1 Ubiquitin like with PHD nnd Ring Finger Domains 24

UTR Untranslated Region 17, 20, 33

XRV Exoogenous Retro-Viral Sequences 19

ZGA zygotic genome activation 36, 37