Human Reproduction is Regulated by Retrotransposons derived from Ancient Hominin-Specific Viral Infections

Yu Tao

UCLA

Xiinyu Xiang

Zhejiang University-University of Edinburgh Institute

Fei-Man Hsu

UCLA

Julien Pontis

École polytechnique fédérale de Lausanne (EPFL)

Didier Trono

EPFL

Wanlu Liu

Zhejiang University-University of Edinburgh Institute & School of Basic Medicine, Zhejiang University School of Medicine, Hangzhou

Amander Clark (✉ clarka@ucla.edu)

University of California, Los Angeles <https://orcid.org/0000-0003-2483-3982>

Article

Keywords: transposable elements, ancient viruses, retrotransposons

Posted Date: October 2nd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-80780/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Nature Communications on January 24th, 2022. See the published version at <https://doi.org/10.1038/s41467-022-28105-1>.

Human Reproduction is Regulated by Retrotransposons derived from Ancient Hominin-Specific Viral Infections

Yu Tao^{1,8}, Xinyu Xiang^{5,8}, Fei-Man Hsu¹, Julien Pontis⁷, Didier Trono⁷, Wanlu Liu^{5,6,9}, Amander T. Clark^{1,2,3,4,9}

¹Department of Molecular Cell and Developmental Biology, University of California, Los Angeles, Los Angeles, CA 90095, USA

²Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles, Los Angeles, CA 90095, USA

³Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90095, USA

⁴Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, CA 90095, USA

⁵Zhejiang University-University of Edinburgh Institute (ZJU-UoE Institute), Zhejiang University School of Medicine, International Campus, Zhejiang University, 718 East Haizhou Rd., Haining, 314400, China

⁶Department of Orthopedic of the Second Affiliated Hospital of Zhejiang University School of Medicine, Zhejiang University, Hangzhou, 310029, China

⁷School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland

⁸These authors contributed equally

⁹Lead Contact

Correspondence: wanlulu@intl.zju.edu.cn (W.L.), clarka@ucla.edu (A.C.)

Summary

Germ cells are essential to pass DNA from one generation to the next. In human reproduction, germ cell development begins with the specification of primordial germ cells (PGCs) and a failure to specify PGCs leads to human infertility. Recent studies have revealed that the transcription factor network required for PGC specification has diverged in mammals, and this has a significant impact on our understanding of human reproduction. Here, we evaluated the emerging epigenetic landscape during hPGC specification using a combination of *in vivo* and *in vitro* analysis of hPGCs/hPGC-like cells (hPGCLCs) and human embryonic stem cells (hESCs). Our data reveals that hominid restricted Transposable Elements (TEs) partly derived from ancient viruses are pre-bound by the transcription factors TFAP2C and NANOG in undifferentiated hESCs, become transcriptionally induced during PGC specification and undergo dynamic epigenetic reprogramming leading to increased chromatin accessibility, localized DNA demethylation and establishment of broad peaks of H3K27ac. Using KRAB mediated CRISPRi we show that blocking this remodeling has a significant impact on hPGC specification. In summary, our data reveals that human reproduction requires the establishment of an epigenetic landscape during hPGC specification driven by the acquisition of hominid-specific TEs that were derived from ancient viral infections that entered the hominid germline less than 5 million years ago.

Introduction

Successful reproduction enables the survival and population growth of modern humans. There are many requirements for successful human reproduction, however the germline is absolutely essential for the passage of DNA through the generations. The human germline begins each generation with the fertilization of a haploid oocyte and spermatozoan to create the one-cell diploid zygote. The zygote subsequently undergoes genome wide chromatin and epigenetic reprogramming to produce totipotent cells of the 2-cell and cleaving embryo, followed by formation of the blastocyst consisting of an inner cell mass and trophoblast. The inner cell mass cells transition into pre-implantation epiblast and primitive endoderm, and it is the pluripotent epiblast cells that ultimately differentiate into the amnion as well as all the cells of the human body. As the mammalian embryo implants into the uterine wall, the embryo's germ cells are specified, and these cells are called primordial germ cells (PGCs). Failure to specify PGCs leads to certain infertility as PGCs constitute the founder germline cells that ultimately give rise to oocytes and sperm in the adult.

In primates, PGCs originate from peri-implantation progenitors at day 11 (D11) post-fertilization during the time when amnion cells are specified from the pluripotent epiblast and before gastrulation (Chen et al., 2019; Sasaki et al., 2016). This is different from the mouse (m), where mPGCs are specified in the post-implantation posterior epiblast around the time of primitive streak formation (Saitou and Yamaji, 2012). Despite the different origins, both mouse and human (h) PGC specification depend on signaling by Bone Morphogenetic Proteins (BMPs), specifically BMP4 (Irie et al., 2015; Lawson et al., 1999; Ohinata et al., 2009; Saitou and Miyauchi, 2016; Sasaki et al., 2015). In contrast, to the conserved signaling network, the transcription factors required for PGC specification has diverged in placental mammals. For example, the transcription factor SOX17 is crucial for hPGC specification (Irie et al., 2015), whereas it is not expressed in mPGCs (Hara et al., 2009; Kanai-Azuma et al., 2002). Similarly, the gene regulatory elements have also diverged, with one example being the utilization of a naïve enhancer at the *POU5F1* (OCT4) locus in hPGCs (Chen et al., 2018), whereas this naïve enhancer sequence is not conserved in mouse (Pastor et al., 2018). Given that new transcription factors have evolved to specify hPGCs relative to other mammals, it could be hypothesized that additional gene regulatory sequences have also evolved to enable successful hPGC specification and therefore human reproduction.

Since the completion of the human genome project (Lander et al., 2001) and launch of the encyclopedia of DNA elements (ENCODE), significant knowledge has been gained regarding the biochemical activity of non-coding regions of the human genome, including areas which are primate specific. One important initial discovery was that RNAs are extensively generated from intergenic regions of the human genome, including transcription from TEs (Kazazian and Moran, 2017). Even though TEs are transcribed, most are transposition incompetent (in capable of movement), except a small number of young TE families including *Human Specific LINE1*'s (*L1_Hs*), *Arthrobacter luteus* restriction endonuclease sites (*Alu*), and composite *SINE-Variable*

Nucleotide Tandem Repeat (VNTR)-Alu (SVA) TEs each having recently moved in the human genome (Hancks and Kazazian, 2016; Kaer and Speek, 2013; Richardson et al., 2014).

Given the transcriptional activity and epigenetic marking of TEs in the genome, recent studies have more and more focus on the role of TEs in gene regulation. In the human pre-implantation embryo just prior to hPGC specification, pluripotent transcription factors including KLF4, KLF17 and OCT4 are hypothesized to bind to motif sequences in TEs, most notably of the human endogenous retrovirus (HERV) class (Britten and Davidson, 1971; Grow et al., 2015; Pontis et al., 2019). HERVs are ancient viruses that infected and integrated into the germline through primate evolution. Hominid-specific HERVs arose from viral germline infections that occurred less than 9 million years ago after the human and chimpanzee split (Barbulescu et al., 1999; Jha et al., 2011; Medstrand and Mager, 1998; Reus et al., 2001; Turner et al., 2001). An example of a hominid-specific HERV are the LTR5 (HERVK/HML2) group, which includes LTR5A, LTR5B and LTR5_Hs. Phylogeny analysis indicates that LTR5_Hs and LTR5A were derived from independent viral infections similar to the virus that created LTR5B (Subramanian et al., 2011). Of the three LTR5 family members, only LTR5_Hs is expressed by cells of the human pre-implantation embryo, where transcription and translation of the RNA is hypothesized to confer an immunoprotective effect as the embryo develops in the reproductive tract (Grow et al., 2015). Using human naïve hESCs and human embryonic carcinoma cell lines, LTR5_Hs and their SVA derivative's sequences are now known to serve as binding sites for the naive transcription factors KLF4 and KLF17 to regulate gene expression over long genomic ranges (Fuentes et al., 2018; Pontis et al., 2019). Given that hPGCs exhibit a naïve-like transcriptional identity, yet are not pluripotent, we were interested in the role of transposons in the specification of hPGCs. We discovered that during hPGC specification *in vitro* and *in vivo* the epigenome of LTR5_Hs sites are remodeled leading to increased chromatin accessibility and targeted DNA demethylation. Failure to open these LTR5_Hs sites compromises hPGC formation and also de-regulates germline gene expression. In summary, our results reveal that a subset of hominid-specific LTR5_Hs sequences are involved in hPGC specification, and therefore should be considered new loci in the regulation of human reproduction.

Results

In order to identify dynamically expressed TEs during hPGC development, we analyzed the RNA-seq data sets published from our lab (Chen et al., 2018), which included hPGCs isolated by Fluorescence Activated Cell Sorting (FACS) from the human embryo at day 59 (D59) to D103 post fertilization, as well as day 4 (D4) hPGC-like cells (hPGCLCs) differentiated from hESCs through incipient mesoderm like cells (iMeLC), an intermediate cell type between hESCs and hPGCLCs in the *in vitro* differentiation strategy (Fig 1A) (Sasaki et al., 2015). The D4 hPGCLCs are transcriptionally equivalent to early primate PGCs between D11-D21 post-fertilization (Chen et al., 2019). To overview the expression pattern of TEs during hPGCLC differentiation, the top 200 TE subfamilies with the highest cross-sample variation were visualized (Fig 1B). In general,

we observed a dynamic TE expression pattern, with the TEs expressed by hPGCLCs similar to hPGCs isolated during embryo development (Fig 1B). In order to determine the most up- and down-regulated TE subfamilies during hPGCLC differentiation, we calculated the percentages of differentially expressed TE (DETE) between hPGCLCs and hESCs for each TE subfamily. This analysis revealed that the most up-regulated TE subfamilies in hPGCLCs were the hominid-specific LTR5_Hs and HERVK-int. In contrast, the most down-regulated TE subfamilies were HERVH-int and a large number of the primate-specific LINE1 PA (L1PA) families which evolved as a single lineage during primate evolution (Fig 1C). Scatter plot of individual DETE copies for LTR5_Hs (n=175) and HERVK-int (n=65) subfamilies showed that the majority of differential expressed LTR5_Hs and HERVK-int were up-regulated in hPGCLCs compared to hESCs (Fig S1A).

To investigate the expressions of the hominid specific TEs during hPGCLC specification more precisely, we reanalyzed the single cell RNA-seq data of hESC differentiate to hPGCLCs published by our lab (Chen et al., 2019). Using uniform manifold approximation and projection (UMAP) to visualize all the cells during the whole process, we clearly see the progression of different cells in each stage (Fig 1D). We use NANOS3 to indicate the hPGCLC population as previously validated (Fig 1E and (Chen et al., 2019)), and show that LTR5_Hs is detectable in hESC, becoming upregulated during lineage differentiation into NANOS3+ hPGCLCs (Fig 1F).

Next we asked whether HERVK RNAs are required for hPGC development. This hypothesis was based on the observation that HERVs can function as long noncoding RNAs to regulate cell biology (Lu et al., 2014). To achieve this, we used short hairpin RNA (shRNA) and scrambled control sequences previously validated to knockdown HERVK/HML2 (Zhou et al., 2016). Following lentiviral infection and antibiotic selection, hPGCLC differentiation was performed, and the hPGCLCs were isolated at D4 using FACS for EPCAM/ITGA6 (Fig S1B). Semi-quantitative real time PCR analysis of the sorted hPGCLCs revealed a significant reduction of HERVK mRNA in the shHERVK group compared to the control (Fig S1C). However, the knockdown of HERVK showed no effect on the percentage of hPGCLCs specified from hESCs (Fig S1D). These data indicated that although HERVK RNAs are up-regulated during hPGCLC specification, reducing its RNA level has no effect on the percentage of hPGCLCs specified by D4.

Given the potential enhancer role of TEs in regulating gene expression (Fuentes et al., 2018; Pontis et al., 2019), we next evaluated changes in chromatin accessibilities during hPGCLC specification with our previously published Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) data (Chen et al., 2018). Comparing to ATAC-seq data in hESCs and iMeLCs, we identified 31,276 ATAC-seq peaks that become specifically more accessible in D4 hPGCLCs (Fig 2A). For the remainder of this work, we will refer to these regions as “hPGCLC-specific” open chromatin.

To uncover specific TE families enriched in the hPGCLC-specific open chromatin regions, we annotated those regions and investigated their TE enrichment. Our analysis revealed that LTR and SVAs families were specifically enriched over the hPGCLC-

specific open chromatin regions (Fig 2B). Further analysis on LTR subfamilies showed that ERVK were the top enriched LTR subfamilies in hPGCLC specific open chromatin regions (Fig 2C). We also counted the copy number of TE elements in each LTR subfamily, and found that although ERVK only represents about 1.5% of all LTRs, 5.4% of LTR copies overlapped with hPGCLC-specific open chromatin regions (Fig 2D). Analysis of the ERVK subfamilies showed that the LTR5_Hs TEs are particularly highly represented in hPGCLC-specific open chromatin (Fig S2A), and of the 697 LTR5_Hs family copies in the human genome, majority of them become more accessible at chromatin level when hPGCLCs are differentiated from hESCs (Fig 2E and 2F). Consistent with LTR5_Hs, we observed a similar pattern for the SVA_D family which contains LTR5 sequences (Fig S2B). The SVA_D copy number corresponds to 26.9% all SVAs in the human genome, while SVA_D contributed to 53.7% of SVAs overlapping with hPGCLC-specific open chromatin (Fig S2C). Heatmap of the chromatin states of the 1546 SVA_D copies showed that they become significantly more accessible in hPGCLCs with differentiation from hESCs (Fig S2D). However when we mapped the expression of SVA_D in the hPGCLC differentiation time course by the single cell RNA-seq data, we noticed SVA_D is ubiquitously expressed in majority of cells, including undifferentiated hESCs, which explains why it was not one of the top DETE in the bulk RNA-seq analysis (Fig S2E and Fig 1D).

Given the finding that hominid enriched LTR5_Hs sequences become more accessible in hPGCLC specification, DNA methylation levels were also evaluated. Using the hPGCLC D4 whole Genome Bisulfite Sequencing (WGBS) data set published by us (Gell et al., 2020), we calculated the average CpG methylation in hESCs and hPGCLCs, and showed that the majority of CpGs in hESCs and hPGCLCs were highly methylated with an average of >70% CpG methylation. Focusing on the CpG sequences of all LTRs (n=457,205) revealed similarly high levels of CpG DNA methylation in both hESCs and hPGCLCs (Fig 3A and (Gell et al., 2020)). In contrast, analysis of the LTR5_Hs sequences revealed a significant reduction in average CpG DNA methylation in hPGCLCs compared to hESCs (Fig 3B). Analysis at SVA_Ds also revealed lower average CpG DNA methylation at the 3' end where the LTR5_Hs like sequence is located in this element (Fig S3A). The targeted decrease in DNA methylation coupled with the increase in accessibility at these sites could imply the repurposing of LTR5_Hs and SVA_D sequences may serve as gene regulatory elements in hPGC specification.

To probe this further, we performed a transcription factor motif search for hPGCLC-specific open chromatin regions that overlapped with LTR5_Hs, and discovered AP2 as one of the top candidates (Fig 3C). One of the highly expressed AP2 family proteins in hPGCLCs is TFAP2C (Fig S3C and (Chen et al., 2018)). Previous studies have shown that TFAP2C is required for hPGCLC specification, and that TFAP2C is involved in opening naïve enhancers including the naïve enhancer at the OCT4 locus in hPGCLCs (Chen et al., 2018; Irie et al., 2015). To determine whether TFAP2C could bind to LTR5_Hs in hPGCLCs, we evaluated binding of TFAP2C to the 697 LTR5_Hs sites using the ChIP-Seq data of (Chen et al., 2019). Our results showed that the majority of LTR5_Hs copies were occupied by TFAP2C in both hESCs and hPGCLCs (Fig 3D).

Similarly, H3K27ac, as one of the active enhancer markers, displayed clear enrichment over LTR5_Hs in hPGCLCs, while interestingly, about half of them are also enriched with sharp peaks of H3K27ac in hESCs (Fig 3D). We also analyzed the TFAP2C and H3K27ac ChIP-seq data on SVA_D, which showed clear binding of TFAP2C and enrichment of H3K27ac over SVA_Ds in hPGCLC but not in hESC (Fig S3B). Because the transcriptome of human seminoma-like cell Tcam-2 is similar to human PGCs, and knockdown of SOX17 in Tcam-2 cells has been reported to affect the expression of other important germ cell transcriptional factors, like TFAP2C and NANOG (Irie et al., 2015), we investigated the binding of SOX17 on LTR5_Hs in germline lineage using SOX17 ChIP-seq data in Tcam-2 cell line (Jostes et al., 2020). We observed strong binding of SOX17 on LTR5_Hs as we expected from the binding of TFAP2C (Fig 3D). We also performed NANOG ChIP-seq in hPGCLCs, and showed that NANOG binds to most of the LTR5_Hs in both hESCs and hPGCLCs (Fig 3D). Taken together, during hPGCLC specification, almost all of the LTR5_Hs sites become more accessible in hPGCLC chromatin, these sites are already associated with TFAP2C and NANOG binding and many are also marked by H3K27ac enrichment in the undifferentiated state. However, upon hPGCLC specification where TFAP2C and NANOG become specific to the germline together with SOX17 (Fig S3C-E), these LTR5_Hs sites become more accessible, undergo DNA demethylation while developing more broad peaks of H3K27ac.

Given that H3K27ac marks promoters and enhancers (Heintzman et al., 2009; Rada-Iglesias et al., 2011; Whyte et al., 2013), and LTR5_Hs serves as a TE Enhancer (TEENhancer) in pre-implantation embryo development (Pontis et al., 2019), we evaluated the association between differentially expressed genes in hPGCLCs compared to undifferentiated hESCs, and distance from a nearby hPGCLC-specific LTR5_Hs. This data showed genes within 50kb to LTR5_Hs in hPGCLC are statistically more likely to be up-regulated in hPGCLCs, whereas down-regulated genes have no such correlation (Figure 3E). For instance, one example is the transcription factor KLF7, which is up-regulated in hPGCLCs compared to hESCs and is within 50kb of a dynamically reprogrammed LTR5_Hs (Fig S3F). Taken together, we propose that the KLF4-KLF17 TEENhancers that operate to regulate gene expression in the pre-implantation embryo are also utilized in hPGCLCs by TFAP2C, NANOG and SOX17 to regulate germline specification from pluripotent cells.

In order to evaluate whether chromatin accessibility of LTR5_Hs / SVA_D is essential for hPGC specification, we used dCAS9-KRAB CRISPRi to inhibit accessibility over TEENhancers corresponding to LTR5_Hs and SVA_D sequences. To achieve this we transduced hESCs with lentivirus encoding dCas9-KRAB fusion protein together with validated gRNAs targeting LTR5_Hs and SVA_D (Pontis et al., 2019). This was followed by hPGCLC differentiation of the CRISPRi and control lines (Fig 4A). Using FACS to quantify the percentage of hPGCLCs in targeted and control CRISPRi experiments, our data shows that CRISPRi with targeted gRNAs results in a significant reduction in the percentage of hPGCLCs compared to controls (Fig 4B and 4C). To determine the specificity of the CRISPRi, we collected the hPGCLCs by FACS at D4 using ITGA6 and EPCAM to isolate the hPGCLCs, and performed RNA-seq of the

FACS isolated hPGCLCs from the control and CRISPRi groups. This analysis revealed that LTR5_Hs, HERVK and SVA are significantly reduced in hPGCLCs compared to other TEs (Fig 4D). To evaluate how CRISPRi affected neighboring gene expression to the LTR5_Hs, we calculated the numbers and percentages of up- and down-regulated genes upon CRISPRi, and discovered almost all genes within 100kb LTR5_Hs are down-regulated upon reducing accessibility at LTR5_Hs (Fig 4E and 4F). Interestingly, with reduced accessibility of LTR5_Hs, we saw a corresponding increase in the expression of HERVH, a retrotransposon more closely associated with primed pluripotency (Fig 4F). This data suggested that the accessibility of LTR5_Hs and SVA are not only transcriptionally, but also phenotypically involved in hPGC differentiation.

Discussion

TEs are considered parasite genomic elements that are capable of changing genomic sequence through transposition. Although many are permanently silenced by defensive mechanisms, some TE sequences are co-opted by the host during evolution, and are therefore kept under selective pressure (Chuong et al., 2017). Hominid specific TEs do not exist in other mammalian model organism, like mouse, new world-monkeys or even old-world monkey's such as the rhesus or cynomolgous macaque. These relatively new hominid TEs have evolved species specific functions which are unique to apes, and in some cases are unique to humans. In the current study, we show that a specific subfamily of HERV and SVA sequences have been kept under selective pressure to regulate hPGC specification, and therefore could be considered essential to successful human reproduction.

Human PGC development has been mostly studied using extrapolation from other mammalian scientific models including the mouse. In recent years combining the differentiation of human pluripotent stem cells into hPGCLCs together with the analysis of human fetal tissue containing hPGCs, it is now appreciated that transcription factors involved in hPGC development are different from the mouse (Chen et al., 2018; Irie et al., 2015; Kojima et al., 2017). Here, we show that specification of hPGCs is associated with increased RNA expression at TEs that are not even found in the mouse genome, and this is associated with changes to the epigenome most notably over the hominid-specific LTR5_Hs subfamilies. LTR5_Hs was previously reported to function as TE enhancers during pre-implantation embryo development and regulation of the pluripotent naïve ground state (Grow et al., 2015; Pontis et al., 2019). In the naïve ground state, LTR5_Hs sequences are bound by KLF family members, most notably KLF4 and KLF17, as well as OCT4 (Grow et al., 2015; Pontis et al., 2019). OCT4 is expressed in both primed and naïve pluripotency, whereas KLF17 and KLF4 are naïve-specific transcription factors. Over expression of KLF4-KLF17 in the primed state of pluripotency is sufficient to open the LTR5_Hs sequences and regulate neighboring gene expression (Pontis et al., 2019). Human PGCs do not express KLF17, and KLF4 has no role in hPGCLC transcriptional identity or hPGC development (Hancock et al., 2020). This would suggest that a different transcription factor serves to regulate LTR5_Hs in human PGC specification.

Despite this lack of KLF4 activity and KLF17 expression in hPGC development, hPGCs *in vitro* and *in vivo* also exhibit a naïve-like pluripotent molecular program that has similarities to naïve state in pre-implantation embryo. This includes two active X chromosomes in females, genome-wide DNA demethylation and expression of naïve pluripotent transcription factors including KLF4 and TFAP2C (Chen et al., 2018; Gkoutela et al., 2015; Guo et al., 2015; Tang et al., 2015). Similar to KLF4-KLF17, TFAP2C also regulates transcription and the identity of naïve pluripotent stem cells by opening naïve-specific enhancers to regulate neighboring gene expression (Pastor et al., 2018; Pontis et al., 2019). Our results imply that opening LTR5_Hs sequences during hPGC development is a task conducted by TFAP2C, whereas in the pre-implantation embryo either KLF4-KLF17 or TFAP2C can open naïve enhancers. Taken together, given the finding that TFAP2C and NANOG are bound to LTR5_Hs in undifferentiated hESCs, our data established a model whereby TFAP2C and NANOG may serve as pioneer factors in hPGC specification, and upon the signal to undergo differentiation, with the recruitment of SOX17 which is barely expressed in hESC but highly expressed in hPGCLCs/hPGCs, TFAP2C and NANOG cooperate to recruit chromatin remodeling complexes to open chromatin in order to promote DNA demethylation at LTR5_Hs and hPGCLC specification (Fig 5). Our results also demonstrate that site-specific DNA demethylation at LTR5_Hs precedes the global DNA demethylation which is a hall mark of hPGC development in the embryo (Gkoutela et al., 2015; Guo et al., 2015; Tang et al., 2015).

In addition to gene regulation at enhancers and promoters. HERVs are also known to regulate 3-D chromatin architecture in pluripotent stem cells. Specifically, HERVH is highly expressed in primed pluripotent stem cells, and is involved in maintaining 3-D chromatin structure (Zhang et al., 2019). Given that HERVH is down-regulated and LTR5_Hs sequences are up-regulated during hPGCLC specification, it's possible that LTR5_Hs could also be required for the 3D architecture of hPGCs and therefore the upregulation of HERVH in the CRISPRi.

Modern and archaic human began to diverge about 500,000 years ago with modern humans becoming the dominant surviving human species around 50,000 years ago (Sankararaman et al., 2016; Wolf and Akey, 2018). Extinction occurs when reproduction fails. Considering the contributions of TEs to the renewal of the genetic pool during evolution, one hypothesis could be that human specific TEs, like LTR5_Hs became beneficial to hPGC specification and consequently improved human reproductive fitness. This question will be better answered with more information about the genomic sequences of ancient humans.

Figure 1

Figure 1. Transposons are dynamically expressed during hPGC/hPGCLC development.

A. Schematic illustration of hPGCLCs specification *in vitro* differentiation.

B. Heatmap of top 200 variable TEs in hESCs, iMeLCs, hPGCLCs and hPGCs.

C. Top differentially expressed TE (DETE) subfamilies in hPGCLCs compared to hESCs. X axis shows DETE percentage over total TE copy number. DETEs with at least 4-fold changes and FDR less than 0.05 in hPGCLC compared to hESC were used for this analysis.

D-F. UMAP of scRNA-seq data display cell subpopulations during UCLA2 hESC differentiation to hPGCLC (**D**), expression of NANOS3 (hPGCLC marker) (**E**) and LTR5_Hs (**F**). Biological replicates (r1, r2) as well as day 1 (D1) to day 4 (D4) hPGCLC were displayed. Single cell RNA-seq data were from (Chen et al., 2019).

Figure 2

Figure 2. HERVK/LTR5_Hs become uniquely accessible with hPGCLC differentiation.

A. Metaplot and heatmap of chromatin accessibility over the 31,276 hPGCLC-specific open chromatin regions. ATAC-seq data is from (Chen et al., 2018).

B. Enrichment of hPGCLC-specific open regions over different TE families. “Observed” is occurrences in hPGCLC-specific open chromatin regions, “expected” is occurrences in random shuffled regions.

C. LTR subfamily enrichment in hPGCLC-specific open chromatin regions.

D. Pie charts showing the distribution of LTR subfamilies in the human genome (left), and the distribution of LTR subfamilies identified in hPGCLC-specific open chromatin regions (right).

E. Scatter plot of ATAC-seq comparing LTR5_Hs accessibility in hESCs and hPGCLCs. Red dots indicate n=697 LTR5_Hs, grey dots are 100,000 random shuffled transposable elements.

F. Metaplot and heatmap of ATAC-seq accessibility over LTR5_Hs sites in hESC, iMeLC, hPGCLC and hPGC.

Figure 3

Figure 3. Epigenetic reprogramming of LTR5_Hs during hPGCLC development.

A. CG methylation level over genome-wide and LTRs in hESCs and hPGCLCs. The x axis is methylation level binned in increments of 10%, y axis indicates the percentage of total mCG sites.

B. Metaplot of average mCG levels over LTR5_Hs in hESCs and hPGCLCs.

C. Motif enrichment analysis of hPGCLC-specific open chromatin regions overlapping with LTR5_Hs. Observed (occurrences in hPGCLC-specific open chromatin regions) over expected (occurrences in random shuffled regions) were calculated. Top 20 TF motifs with highest observed over expected ratios were plotted.

D. NANOG/TFAP2C/H3K27ac ChIP-seq over LTR5_Hs (n=697) in hESCs and hPGCLCs and SOX17 ChIP-seq over LTR5_Hs in Seminoma. NANOG ChIP-seq in hPGCLC was from this paper. NANOG ChIP-seq in hESC was from GSE105028 (Lyu et al., 2018), TFAP2C ChIP-seq in hESC and hPGCLC, H3K27ac ChIP-seq in hPGCLC were from GSE140021 (Chen et al., 2019), H3K27ac ChIP-seq in hESC was from GSE69646 (Hikabe et al., 2016) and SOX17 ChIP-seq in Seminoma-like cells was from GSE130519 (Jostes, S. V. Et al., 2020).

E. Barplot for the enrichment of LTR5_Hs associated DEGs. X-axis is the distance of LTR5_Hs to the TSS of neighboring genes. DEGs with at least 4-fold changes and FDR less than 0.05 in hPGCLC compared to hESC were used to for this analysis. * indicates *p-value* <0.05 from hypergeometric test.

Figure 4

Figure 4. LTR5_Hs chromatin accessibility is involved in hPGCLC differentiation.

A. Schematic illustration depicting transduction of dCas9-KRAB-gRNAs targeting LTR5_Hs (Pontis et al., 2019) in hESCs followed by induced hESC differentiation to hPGCLCs.

B. Representative flow cytometry plots of hPGCLC differentiation in control (Ctrl (22)) and CRISPRi (55+57) groups. The hPGCLC population corresponds to the highlighted ITGA6/EPCAM double positive cells.

C. Statistical analysis comparing percentage of hPGCLCs in control relative to CRISPRi groups (biological replicates $n=3$; ** means $p=0.0035$ from unpaired two-tailed t test; error bar is standard deviation/SD).

D. Expression level of LTR5_Hs, HERVK, SVA and Other TEs in hPGCLC upon the CRISPRi treatment. The red lines represent the median values of each TE groups. Minimum and maximum values are shown with both ends of the whisker. The box ends represent the first (25%) and third (75%) quartiles. For LTR5_Hs: *** indicates $p=7.4 \times 10^{-83}$ from one sample t-test. For HERVK: *** indicates $p=4.9 \times 10^{-21}$ from one sample t-test. For SVA: *** indicates $p=2.9 \times 10^{-6}$ from one sample t-test.

E. Count for neighboring DEGs of LTR5_Hs in CRISPRi hPGCLC compared to control cells. Cut-off for DEG is FDR less than 0.05 and fold change > 10% between paired replicates.

F. Expression level of HERVK, LTR5_Hs, HERVH and representative neighboring genes (*ST6GAL1*, *PRODH*) in control and CRISPRi hPGCLCs from RNA-seq. (For HERVK, LTR5_Hs and HERVH, *** indicates $p < 0.001$ from paired two-tailed t-test; for *ST6GAL1*, *PRODH*, *** indicates $p < 0.001$ from unpaired two-tailed t-test; error bar is standard deviation/SD)

Figure 5

Figure 5. Proposed model for the roles of LTR5_Hs in determining germ cell fate.

Based on our data, we proposed that in pluripotent cells, LTR5_Hs DNA is hypermethylated, slightly opened and occupied by TFAP2C and NANOG, which primes germ cell specification. Upon germline cell differentiation, the chromatin of LTR5_Hs becomes significantly more accessible, SOX17 binds with NANOG and TFAP2C resulting in the DNA becoming hypomethylated and the TEENhancers are utilized to promote expression of neighboring protein coding genes (PCGs) to regulate the germ cell fate.

Methods

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Amander Clark (clarka@ucla.edu). All unique/stable reagents generated in this study are available from the Lead Contact with a completed Materials Transfer Agreement.

hESC culture

UCLA2 hESCs are cultured as previously described (Chen et al, 2018), briefly the hESCs are cultured in hESC media, which was composed of 20% knockout serum replacement (KSR) (Life Technologies, A3181502), 1x MEM Non-Essential Amino Acids (NEAA) (Fisher Scientific, 25-025-CI), 1x Penicillin/Streptomycin/Glutamine (Thermo Fisher, 10378016), 55μM 2- Mercaptoethanol (Life Technologies, 21985-023), 10ng/mL recombinant human FGF basic (Proteintech, HZ1285), and 50ng/mL primocin (InvivoGen, ant-pm-2) in DMEM/F12 media (GIBCO, 11330-032). The primed hESCs were split by 1mg/ml Collagenase type IV (GIBCO, 17104-019) and maintained routinely on mitomycin C (MMC)-inactivated mouse embryonic fibroblasts (MEFs). The hESCs were split every 7 days using Collagenase type IV (GIBCO, 17104-019). All hESC lines used in this study are registered with the National Institute of Health Human Embryonic Stem Cell Registry and are available for research use with NIH funds. Mycoplasma test (Lonza, LT07-418) was performed every month. All experiments were approved by the UCLA Embryonic Stem Cell Research Oversight Committee.

hPGCLC differentiation

Using the UCLA2 hESC line, the differentiation of hPGCLCs in vitro was performed as described in (Chen et al., 2017; Sasaki et al., 2015). Specifically, 0.05% trypsin-EDTA (Thermo Fisher Scientific, 25300120) was used to digest confluent hESCs cultured on mitomycin C inactivated mouse embryonic fibroblasts (MEFs) into single cells, followed by plating onto a 12-well-plate which had previously been coated with human plasma fibronectin (Life Technologies, 33016-015) for at least 1 hour. Cells were plated at cell density of 200,000 cells/well in 2 ml/well of incipient mesoderm-like cells (iMeLCs) medium, which is composed of 15% knockout serum replacement (KSR, Life Technologies, A3181502), 1x MEM Non-Essential Amino Acids (NEAA) (Fisher Scientific, 25-025-CI), 55μM 2- Mercaptoethanol (Life Technologies, 21985-023), 1xPenicillin/Streptomycin/Glutamine (Thermo Fisher, 10378016), 1 mM sodium pyruvate (Life Technologies, 11360070), 50 ng/mL Activin A (PeproTech, AF-120-14E), 3 mM CHIR99021 (Reprocell, 04-0004-10), 10 mM ROCKi (Y27632, Stemgent, 04-0012-10), and 50 ng/mL primocin (InvivoGen, ant-pm-2) in Glasgow's minimal essential medium (GMEM) (Life Technologies, 11710035). After 24 hours, iMeLCs were dissociated into single cells by 0.05% trypsin-EDTA (Thermo Fisher Scientific, 25300120), then plated into ultra-low cell attachment U-bottom 96-well plates (Corning, 7007) at a density of 3,000 cells/well in 200 μl/well of hPGCLC medium, which is composed of 15% KSR (Life Technologies, A3181502), 1x MEM Non-Essential Amino Acids (NEAA) (Fisher Scientific, 25-025-CI), 55μM 2- Mercaptoethanol (Life

Technologies, 21985-023), 1 x Penicillin/Streptomycin/Glutamine (Thermo Fisher, 10378016), 1 mM sodium pyruvate (Life Technologies, 11360070), 10 ng/mL recombinant human leukemia inhibitory factor (Sigma-Aldrich, LIF1010), 200 ng/mL human BMP4 (R&D systems, 314-BP), 50 ng/mL human epidermal growth factor (Fisher Scientific, 236EG200), 10 mM of ROCKi (Y27632, Stemgent, 04-0012-10), and 50 ng/mL primocin in GMEM (Life Technologies, 11710035). Day-4 hPGCLC aggregates were collected for further analysis.

Flow cytometry and fluorescence activated cell sorting

hPGCLC aggregates were dissociated with 0.05% Trypsin-EDTA 0.05% trypsin-EDTA (Thermo Fisher Scientific, 25300120) for 10 min at 37 C. The dissociated cells were then stained with conjugated antibodies, washed with FACS buffer (1% BSA in PBS) and resuspended in FACS buffer with 7-AAD (BD PharMingen, 559925) as viability dye. The single cell suspension was sorted for further experiments. The antibodies used in this study are: BV421 conjugated anti-human/mouse CD49f (ITGA6) (BioLegend; Cat#313624; RRID: AB_2562244; Lot#B274314); APC-conjugated anti-human CD326 (EPCAM) (BioLegend; Cat#324208; RRID: AB_756082; Lot#B284158).

shRNA knockdown

shRNA sequence was from (Zhou et al., 2016). The shRNA sequences were subcloned into a lentiviral shRNA vector (Ivanova et al., 2006). The shRNA vector was transfected into HEK-293T cells to package virus. After collecting the supernatant ultra-centrifugation, the virus was used to infect at least 200k primed hESCs in suspension. The transduced hESCs were re-plated on DR4 inactivated MEFs are selected using 200 µg/ml hygromycin (Thermo Fisher Scientific, 10687010) for successfully integration of shRNA constructs. hPGCLC differentiation was performed under puromycin selection as described above.

Quantitative realtime PCR

Cell pellets were lysed using 350 µL RLT buffer and RNA was extracted using RNeasy micro kit (QIAGEN, 220006-800). cDNA was synthesized using SuperScript II Reverse Transcriptase (Invitrogen, 18064-014). Quantitative real-time PCR was performed using PowerUp SYBR Green Master Mix (ThermoFisher, A25741). Expression levels of genes of interest were normalized to the housekeeping gene GAPDH. Three biological replicates were performed for each experiment and two technical replicates were conducted for each biological replicate. The real-time PCR primers for HERVK are from (Zhou et al., 2016).

ChIP-seq protocol

The ChIP-seq was performed as previously described (Pastor et al., 2018). Whole aggregates of day 4 hPGCLC were collected for NANOG ChIPseq. Cells were dissociated with 0.05% Tris-EDTA for 10 minutes at 37°C. Cells gathered from at least 3 96-well plates were then fixed with 1% paraformaldehyde (Sigma) and incubated with rotation for 10 min at room temperature. The paraformaldehyde was quenched by adding glycine to a final concentration of 0.14 M and rotated another 10 min at room

temperature. The cells were then centrifuged at 500g for 5 min and then flash-frozen with liquid nitrogen and stored at -80°C .

Cells were thawed and resuspended with 1 ml lysis buffer (10 mM TrisHCl pH 8.0, 0.25% Triton X-100, 10 mM EDTA, 0.5 mM EGTA, 1 $\times$ protease inhibitors (Roche) and 1 mM PMSF), then rotated for 15 min at room temperature. Nuclei were pelleted by centrifugation at 1,500g for 5 min at 4°C . Nuclei were then resuspended with 1 ml 10 mM TrisHCl pH 8.0, 200 mM NaCl, 10 mM EDTA, 0.5 mM EGTA, 1 $\times$ protease inhibitor, 1 mM PMSF and rotated for 10 min at 4°C . Nuclei were then pelleted and resuspended in 650 μl 10 mM Tris-HCl pH 8.0, 10 mM EDTA, 0.5 mM EGTA, 1 $\times$ protease inhibitor, 1 mM PMSF and sonicated in a 12 mm $\times$ 12 mm sonication tube (Covaris) in Covaris S2 (intensity = 5; cycles per burst = 200; duty cycle = 5%; 8 $\times$ (30s on /30s off) for 4 min effective sonication). The sonicated lysate was then centrifuged for 10 min at maximal speed, and the supernatant were collected. 10% of the supernatant was saved as 'Input' and the rest was used for ChIP.

Protein A Dynabeads (30 μl / sample, ThermoFisher) were washed three times with ChIP buffer (16.7 mM TrisHCl pH 8.0, 0.01% SDS, 1.1 Triton X-100, 1.2 mM EDTA, 167 mM NaCl). The 30 μl of beads were then resuspended in 650 μl of ChIP buffer and combined with the ChIP sample to pre-clear the sample. Beads and chromatin were rotated for 2 h at 4°C , and collect the supernatant. Anti-NANOG antibody (5 μg , R&D, AF1997) was added to each ChIP sample. The samples were then rotated overnight at 4°C .

Protein A Dynabeads (60 μl / sample) were added to the ChIP samples and rotated for 2h at 4°C . The beads were then washed 2 $\times$ 4min with 500 μl wash buffer A (50 mM HEPES pH 7.9, 1% TritonX-100, 0.1% deoxycholate, 1 mM EDTA, 140 mM NaCl), 500 μl wash buffer B (50 mM HEPES pH 7.9, 0.1% SDS, 1% Triton X100, 0.1% deoxycholate, 1 mM EDTA, 500 mM NaCl) and 500 μl TE buffer (10 mM TrisHCl pH 8.0, 1 mM EDTA). Each wash consisted of resuspension in 500 μl buffer and rotation for 4 min, followed by collection of beads and removal of supernatant. DNA was eluted in 100 μl elution buffer (50 mM TrisHCl pH 8.0, 1 mM EDTA, 1% SDS) at 65°C for 10 min in a ThermoMixer (Eppendorf) shaking at 1,400 r.p.m. The eluant was collected and the beads were subjected to a second round of elution with 150 μl elution buffer. The ChIP eluants were pooled, and the input sample was diluted to 250 μl with elution buffer. The samples were incubated at 65°C overnight to promote decrosslinking. The samples were then allowed to cool to room temperature, 15 μg of RNase A (Purelink, ThermoFisher) was added, and the samples were incubated for 30 min at 37°C to degrade RNA.

Proteinase K (100 μg) was then added and the samples were incubated at 56°C for 2 h. DNA was purified using a MinElute PCR purification kit (Qiagen) and libraries were generated using the Ovation Ultralow Library System V2 (TECAN).

CRISPR/dCas9-kRAB assay

gRNAs that target LTR5_Hs were from (Pontis et al., 2019). The gRNA-CRISPR/dCas9-KRAB vectors were transfected into HEK-293T cells to produce lentivirus. Supernatants that contain lentivirus were then collected and ultracentrifuged. Confluent hESC were trypsinized with 0.05% trypsin at 37°C for 5 minutes, then 200k cells were counted and collected to mix with the concentrated lentivirus. After mixture on nutator for two hours at room temperature, cells were transferred onto mitomycin C treated MEFs in hESC

media with 10 mM ROCKi. After transduction into hESC, 1ug/ml puromycin was used to screen for positive cells for at least 5 days. Survived cells were then used to perform downstream assays.

RNA-sequencing

RNA-seq was performed as previously described (Tao et al., 2018). Briefly the hPGCLCs were directly sorted into 350ul RLT lysis buffer (QIAGEN RNeasy micro kit, 220006-800). Total RNA was then extracted by RNeasy micro kit (Qiagen RNeasy micro kit, 220006-800). Total RNA was reverse transcribed and cDNA was amplified using Ovation RNA-Seq System V2 (Tecan, 7102-32) according to the manufacturer's instructions. Amplified cDNA was then sheared to ~200 bp length by Covaris S220 Focused ultrasonicator. RNA-seq libraries were constructed by using Ovation Rapid Library Systems (Tecan, 0319-32) and quantified by a KAPA library quantification kit (Kapa Biosystems, kk4824). Libraries were then subjected to pair-end sequencing on Nova-seq sequencer.

RNA-seq analysis

RNA-seq data from previous publications GSE93126 (Chen et al., 2018) were downloaded and performed quality control by fastqc v0.11.8 (Andrews, 2010). Then the sequences were aligned to human reference genome GRCh38 with STAR v2.7.0e (Dobin et al., 2013) allowing maximal 1000 multiple mapped reads and the best hit was kept (--outFilterMultimapNmax 1000 --outSAMmultNmax 1). Annotation for genes were obtained from Ensembl (Howe et al., 2020) GRCh38.97 while annotation for transposable elements used were based on repeatmasker (<http://repeatmasker.org/>) GRCh38. Read count over transposable elements were calculated using FeatureCounts v2.0.0 (Liao et al., 2014a) while read count over transposable elements were counted using bedtools v2.17.0 (Quinlan and Hall, 2010) multicov function. Heatmap for transposable elements expression were generated in R using R package ComplexHeatmap v2.5.1 (Gu et al., 2016a). Differential expressed genes (DEG) or TE (DETE) were processed using R package DESeq2 v1.26.0 (Love et al., 2014b) with at least 4 fold change and FDR less than 0.05.

ATAC-seq analysis

ATAC-seq data from previous publications GSE120648 (Chen et al., 2018) were downloaded and performed quality control by fastqc v0.11.8 (Andrews, 2010). Then the sequences were aligned to human reference genome GRCh38 with STAR v2.7.0e (Dobin et al., 2013) allowing maximal 1000 multiple mapped reads and only one read were kept (--outFilterMultimapNmax 1000 --outSAMmultNmax 1). Splice junction is neglected by building STAR index without general feature format file and not allowing intron length (--alignIntronMax 1). PCR duplicates were removed using Samtools v1.2 (Li et al., 2009b) rmdup function. Since ATAC-seq is also able to capture long reads derived from nucleosome signals, only the open chromatin reads pair with length less than 100bp were kept for further analysis. ATAC-seq peaks were defined using the MACS2 v2.2.7.1 (Zhang et al., 2008b) callpeaks function. Here we only kept peaks with a fold change enrichment greater than 4. In order to identify peaks specific to one cell type (for instance, hPGCLC specific

peaks), we used bedtools v2.29.2 (Quinlan and Hall, 2010) multiinter function with Ryan Layers's clustering algorithm. Bigwig tracks were generated using deepTools v3.2.1 (Ramírez et al., 2014b) by normalizing to RPKM using binsize of 10bp. ATAC-seq signal over cell type specific regions or retrotransposon regions were plotted using deepTools v3.2.1 (Ramírez et al., 2014b). ATAC-seq reads counts over LTR5_Hs, SVA_D and 100,000 random TE regions were calculated by bedtools v2.29.2 (Quinlan and Hall, 2010) multicov function. TE annotation over hPGCLC specific peaks is used Homer v4.7 (Heinz et al., 2010a) annotatePeaks.pl function by using TE general feature format file based on repeatmasker. TF motifs distribution on GRCh38 were calculated using Homer v4.7 (Heinz et al., 2010a) scanMotifGenomeWide.pl function. Motifs specifically enriched over hPGCLC specific open region were calculated by bedtools v2.29.2 (Quinlan and Hall, 2010) intersect function (without -wa).

WGBS analysis

WGBS datasets were downloaded from GEO (Gell et al., 2020). Duplicated reads were first removed, and adaptors were trimmed with Cutadapt. Pre-processed reads were aligned to hg38 genome using BS Seeker2 (Guo et al., 2015) with default setting. When calculating methylation level over each cytosine, only reads uniquely mapped were kept, and only CG sites with coverage ≥ 4 are preserved. Methylation level at CG sites were then calculated by $\#C/(\#C+\#T)$.

Transcription factor (TF) motif enrichment analysis

Motif file for over 400 transcriptional factors were collected from Homer and the position of each motif in GRCh38 genome were calculated using Homer scanMotifGenome.pl function. As control, random shuffled regions with same number and length as the hPGCLC specific ATAC-seq peaks were generated using bedtools shuffle function. To analyze the enriched TF motifs over hPGCLC specific open chromatin which overlapped with LTR5_Hs, the frequency of occurrences for TF motifs in hPGCLC specific ATAC-seq peaks and shuffled control were processed using bedtools intersect function. Observed (occurrences in hPGCLC specific ATAC-seq peaks) over expected (occurrences in random shuffled regions) were calculated. Top 20 TF motifs with highest observed over expected ratios were plotted.

ChIP-seq analysis

NANOG ChIP-seq in hPGCLC was from this paper. NANOG ChIP-seq in hESC from GSE105028 (Lyu et al., 2018), TFAP2C ChIP-seq in hESC and hPGCLC, H3K27ac ChIP-seq in hPGCLC were from previous publications GSE140021 (Chen et al., 2019), H3K27ac ChIP-seq was in hESC hPGCLC from previous publications GSE69646 (Hikabe et al., 2016) and SOX17 ChIP-seq in Seminoma-like cells was from GSE130519 (Jostes et al., 2020). For all ChIP-seq data, quality control was performed by fastqc v0.11.8 (Andrews, 2010). Then the sequences were aligned to human reference genome GRCh38 with STAR v2.7.0e (Dobin et al., 2013) allowing maximal 1000 multiple mapped reads and only one read were kept (`--outFilterMultimapNmax 1000 --outSAMmultNmax 1`). Splice junction is neglected by building STAR index without general feature format file and not allowing intron length (`--alignIntronMax 1`). PCR duplicates were removed using Samtools v1.2 (Li et al., 2009b) rmdup function.

Bigwig tracks were generated using deeptools v3.2.1 (Ramírez et al., 2014b) by normalizing to RPKM using binsize of 10bp. ChIP-seq signal over TFAP2C and H3K27ac were plotted using deeptools v3.2.1(Ramírez et al., 2014b).

PADEG (Peak Associated Differential Genes) analysis

LTR5Hs associated DEG (PADEG) analysis were processed as described before(Pastor et al., 2018). Gene list within different distance to LTR5_Hs were calculated using bedtools v 2.29.2 (Quinlan and Hall, 2010) intersect function. Then number of differential expressed genes in each distance category were counted and the observed/expected ratio over genomic average (all genes) were plotted. P value were calculated by hypergenometric test package phyper (v 3.6.0) in R (Team, 2019).

DATA AND SOFTWARE AVAILABILITY

The accession number for the RNA-seq, ChIP-seq reported in this paper is GEO: xxxxx. Custom codes for all the high throughout sequencing analysis are available upon request.

Acknowledgement

The authors would also like to thank Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research University of California, Los Angeles Flow Cytometry Core Resource, especially Felicia Codrea, Jessica Scholes, and Jeffery Calimlim for FACS, Jinghua Tang for banking and culturing of the UCLA hESC lines, Suhua Feng in assistance with high throughput sequencing, Jonathan DiRusso in reviewing and providing feedback for the manuscript. This study was supported by grant from the NIH/NICHD R01HD079546, and the Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at UCLA Research Award Program.

Author Contributions

A.T.C. and Y.T. conceived the study, designed experiments and wrote the manuscript. Y.T. performed the experiments. X.X., W.L. and F.H. performed the bioinformatics analyses. J.P. and D.T. designed the CRISPRi gRNAs. All authors contributed to the review and corrections of the manuscript.

References

- Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166-169.
- Andrews, S. (2010). FASTQC. A quality control tool for high throughput sequence data.
- Barbulescu, M., Turner, G., Seaman, M.I., Deinard, A.S., Kidd, K.K., and Lenz, J. (1999). Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. *Current Biology* 9, 861-S861.

644 Britten, R.J., and Davidson, E.H. (1971). Repetitive and Non-Repetitive DNA Sequences and a
 645 Speculation on the Origins of Evolutionary Novelty. *The Quarterly Review of Biology* 46, 111-
 646 138.

647 Chen, D., Liu, W., Lukianchikov, A., Hancock, G.V., Zimmerman, J., Lowe, M.G., Kim, R., Galic, Z.,
 648 Irie, N., Surani, M.A., *et al.* (2017). Germline competency of human embryonic stem cells
 649 depends on eomesodermin⁺. *Biology of Reproduction* 97, 850-861.

650 Chen, D., Liu, W., Zimmerman, J., Pastor, W.A., Kim, R., Hosohama, L., Ho, J., Aslanyan, M., Gell,
 651 J.J., Jacobsen, S.E., *et al.* (2018). The TFAP2C-Regulated OCT4 Naive Enhancer Is Involved in
 652 Human Germline Formation. *Cell Reports* 25, 3591-3602.e3595.

653 Chen, D., Sun, N., Hou, L., Kim, R., Faith, J., Aslanyan, M., Tao, Y., Zheng, Y., Fu, J., Liu, W., *et al.*
 654 (2019). Human Primordial Germ Cells Are Specified from Lineage-Primed Progenitors. *Cell*
 655 *Reports* 29, 4568-4582.e4565.

656 Chuong, E.B., Elde, N.C., and Feschotte, C. (2017). Regulatory activities of transposable
 657 elements: from conflicts to benefits. *Nature Reviews Genetics* 18, 71-86.

658 Diaz Perez, S.V., Kim, R., Li, Z., Marquez, V.E., Patel, S., Plath, K., and Clark, A.T. (2011).
 659 Derivation of new human embryonic stem cell lines reveals rapid epigenetic progression in vitro
 660 that can be prevented by chemical modification of chromatin. *Human Molecular Genetics* 21,
 661 751-764.

662 Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M.,
 663 and Gingeras, T.R. (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15-21.

664 Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M.,
 665 and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15-21.

666 Fuentes, D.R., Swigut, T., and Wysocka, J. (2018). Systematic perturbation of retroviral LTRs
 667 reveals widespread long-range effects on human gene regulation. *eLife* 7, e35989.

668 Gell, J.J., Liu, W., Sosa, E., Chialastri, A., Hancock, G., Tao, Y., Wamaitha, S.E., Bower, G., Dey,
 669 S.S., and Clark, A.T. (2020). An Extended Culture System that Supports Human Primordial Germ
 670 Cell-like Cell Survival and Initiation of DNA Methylation Erasure. *Stem Cell Reports*.

671 Gkoutela, S., Zhang, Kelvin X., Shafiq, Tiasha A., Liao, W.-W., Hargan-Calvopiña, J., Chen, P.-Y.,
 672 and Clark, Amander T. (2015). DNA Demethylation Dynamics in the Human Prenatal Germline.
 673 *Cell* 161, 1425-1436.

674 Grow, E.J., Flynn, R.A., Chavez, S.L., Bayless, N.L., Wossidlo, M., Wesche, D.J., Martin, L., Ware,
 675 C.B., Blish, C.A., Chang, H.Y., *et al.* (2015). Intrinsic retroviral reactivation in human
 676 preimplantation embryos and pluripotent cells. *Nature* 522, 221-225.

677 Gu, Z., Eils, R., and Schlesner, M. (2016a). Complex heatmaps reveal patterns and correlations in
 678 multidimensional genomic data. *Bioinformatics* 32, 2847-2849.

679 Gu, Z., Eils, R., and Schlesner, M. (2016b). Complex heatmaps reveal patterns and correlations
 680 in multidimensional genomic data. *Bioinformatics* 32, 2847-2849.

681 Guo, F., Yan, L., Guo, H., Li, L., Hu, B., Zhao, Y., Yong, J., Hu, Y., Wang, X., Wei, Y., *et al.* (2015).
 682 The Transcriptome and DNA Methylome Landscapes of Human Primordial Germ Cells. *Cell* 161,
 683 1437-1452.

684 Guo, W., Fizev, P., Yan, W., Cokus, S., Sun, X., Zhang, M.Q., Chen, P.-Y., and Pellegrini, M.
 685 (2013). BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics*
 686 14, 774.

687 Hancks, D.C., and Kazazian, H.H. (2016). Roles for retrotransposon insertions in human disease.
688 Mobile DNA 7, 9.

689 Hancock, G., Liu, W., Peretz, L., Chen, D., Gell, J., Collier, A., Zamudio, J., Plath, K., and Clark, A.
690 (2020). Divergent roles for KLF4 and TFCP2L1 in Naive Ground State Pluripotency and Human
691 Primordial Germ Cell Development (Research Square).

692 Hara, K., Kanai-Azuma, M., Uemura, M., Shitara, H., Taya, C., Yonekawa, H., Kawakami, H.,
693 Tsunekawa, N., Kurohmaru, M., and Kanai, Y. (2009). Evidence for crucial role of hindgut
694 expansion in directing proper migration of primordial germ cells in mouse early embryogenesis.
695 Developmental Biology 330, 427-439.

696 Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K.,
697 Stuart, R.K., Ching, C.W., *et al.* (2009). Histone modifications at human enhancers reflect global
698 cell-type-specific gene expression. Nature 459, 108-112.

699 Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H.,
700 and Glass, C.K. (2010a). Simple Combinations of Lineage-Determining Transcription Factors
701 Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell
702 38, 576-589.

703 Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H.,
704 and Glass, C.K. (2010b). Simple combinations of lineage-determining transcription factors prime
705 cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38, 576-589.

706 Hikabe, O., Hamazaki, N., Nagamatsu, G., Obata, Y., Hirao, Y., Hamada, N., Shimamoto, S.,
707 Imamura, T., Nakashima, K., Saitou, M., *et al.* (2016). Reconstitution in vitro of the entire cycle
708 of the mouse female germ line. Nature *advance online publication*.

709 Howe, K.L., Contreras-Moreira, B., De Silva, N., Maslen, G., Akanni, W., Allen, J., Alvarez-Jarreta,
710 J., Barba, M., Bolser, D.M., Cambell, L., *et al.* (2020). Ensembl Genomes 2020-enabling non-
711 vertebrate genomic research. Nucleic Acids Research 48, D689-D695.

712 Irie, N., Weinberger, L., Tang, Walfred W.C., Kobayashi, T., Viukov, S., Manor, Y.S., Dietmann, S.,
713 Hanna, Jacob H., and Surani, M.A. (2015). SOX17 Is a Critical Specifier of Human Primordial
714 Germ Cell Fate. Cell 160, 253-268.

715 Ivanova, N., Dobrin, R., Lu, R., Kotenko, I., Levorse, J., DeCoste, C., Schafer, X., Lun, Y., and
716 Lemischka, I.R. (2006). Dissecting self-renewal in stem cells with RNA interference. Nature 442,
717 533-538.

718 Jha, A.R., Nixon, D.F., Rosenberg, M.G., Martin, J.N., Deeks, S.G., Hudson, R.R., Garrison, K.E.,
719 and Pillai, S.K. (2011). Human Endogenous Retrovirus K106 (HERV-K106) Was Infectious after
720 the Emergence of Anatomically Modern Humans. PLOS ONE 6, e20234.

721 Ji, X., Dadon, Daniel B., Powell, Benjamin E., Fan, Zi P., Borges-Rivera, D., Shachar, S.,
722 Weintraub, Abraham S., Hnisz, D., Pegoraro, G., Lee, Tong I., *et al.* (2016). 3D Chromosome
723 Regulatory Landscape of Human Pluripotent Cells. Cell Stem Cell 18, 262-275.

724 Jostes, S.V., Fellermeier, M., Arévalo, L., Merges, G.E., Kristiansen, G., Nettersheim, D., and
725 Schorle, H. (2020). Unique and redundant roles of SOX2 and SOX17 in regulating the germ cell
726 tumor fate. International Journal of Cancer 146, 1592-1605.

727 Kaer, K., and Speek, M. (2013). Retroelements in human disease. Gene 518, 231-241.

728 Kanai-Azuma, M., Kanai, Y., Gad, J.M., Tajima, Y., Taya, C., Kurohmaru, M., Sanai, Y., Yonekawa,
729 H., Yazaki, K., Tam, P.P.L., *et al.* (2002). Depletion of definitive gut endoderm in
730 *Sox17*-null mutant mice. Development 129, 2367-2379.

731 Kazazian, H.H., and Moran, J.V. (2017). Mobile DNA in Health and Disease. *New England Journal*
 732 *of Medicine* 377, 361-370.
 733 Kojima, Y., Sasaki, K., Yokobayashi, S., Sakai, Y., Nakamura, T., Yabuta, Y., Nakaki, F., Nagaoka,
 734 S., Woltjen, K., Hotta, A., *et al.* (2017). Evolutionarily Distinctive Transcriptional and Signaling
 735 Programs Drive Human Germ Cell Lineage Specification from Pluripotent Stem Cells. *Cell Stem*
 736 *Cell* 21, 517-532.e515.
 737 Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K.,
 738 Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome.
 739 *Nature* 409, 860-921.
 740 Lawson, K.A., Dunn, N.R., Roelen, B.A.J., Zeinstra, L.M., Davis, A.M., Wright, C.V.E., Korving,
 741 J.P.W.F.M., and Hogan, B.L.M. (1999). Bmp4 is required for the generation of primordial germ
 742 cells in the mouse embryo. *Genes & Development* 13, 424-436.
 743 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and
 744 Durbin, R. (2009a). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25,
 745 2078-2079.
 746 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and
 747 Durbin, R. (2009b). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25,
 748 2078-2079.
 749 Liao, Y., Smyth, G.K., and Shi, W. (2014a). FeatureCounts: An efficient general purpose program
 750 for assigning sequence reads to genomic features. *Bioinformatics* 30, 923-930.
 751 Liao, Y., Smyth, G.K., and Shi, W. (2014b). featureCounts: an efficient general purpose program
 752 for assigning sequence reads to genomic features. *Bioinformatics* 30, 923-930.
 753 Love, M.I., Huber, W., and Anders, S. (2014a). Moderated estimation of fold change and
 754 dispersion for RNA-seq data with DESeq2. *Genome Biology* 15, 550.
 755 Love, M.I., Huber, W., and Anders, S. (2014b). Moderated estimation of fold change and
 756 dispersion for RNA-seq data with DESeq2. *Genome Biology* 15.
 757 Lu, X., Sachs, F., Ramsay, L., Jacques, P.-É., Göke, J., Bourque, G., and Ng, H.-H. (2014). The
 758 retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity.
 759 *Nat Struct Mol Biol* 21, 423-425.
 760 Lyu, X., Rowley, M.J., and Corces, V.G. (2018). Architectural Proteins and Pluripotency Factors
 761 Cooperate to Orchestrate the Transcriptional Response of hESCs to Temperature Stress. *Mol*
 762 *Cell* 71, 940-955.e947.
 763 Medstrand, P., and Mager, D.L. (1998). Human-Specific Integrations of the HERV-K Endogenous
 764 Retrovirus Family. *Journal of Virology* 72, 9782-9787.
 765 Ohinata, Y., Ohta, H., Shigeta, M., Yamanaka, K., Wakayama, T., and Saitou, M. (2009). A
 766 Signaling Principle for the Specification of the Germ Cell Lineage in Mice. *Cell* 137, 571-584.
 767 Pastor, W.A., Liu, W., Chen, D., Ho, J., Kim, R., Hunt, T.J., Lukianchikov, A., Liu, X., Polo, J.M.,
 768 Jacobsen, S.E., *et al.* (2018). TFAP2C regulates transcription in human naive pluripotency by
 769 opening enhancers. *Nature Cell Biology* 20, 553-564.
 770 Pontis, J., Planet, E., Offner, S., Turelli, P., Duc, J., Coudray, A., Theunissen, T.W., Jaenisch, R.,
 771 and Trono, D. (2019). Hominoid-Specific Transposable Elements and KZFPs Facilitate Human
 772 Embryonic Genome Activation and Control Transcription in Naive Human ESCs. *Cell Stem Cell*
 773 24, 724-735.e725.

774 Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic
 775 features. *Bioinformatics* 26, 841-842.
 776 Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A
 777 unique chromatin signature uncovers early developmental enhancers in humans. *Nature* 470,
 778 279-283.
 779 Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014a). deepTools: a flexible
 780 platform for exploring deep-sequencing data. *Nucleic Acids Res* 42, W187-191.
 781 Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014b). DeepTools: A flexible
 782 platform for exploring deep-sequencing data. *Nucleic Acids Research* 42, W187-W191.
 783 Reus, K., Mayer, J., Sauter, M., Zischler, H., Müller-Lantzsch, N., and Meese, E. (2001). HERV-
 784 K(OLD): Ancestor Sequences of the Human Endogenous Retrovirus Family HERV-K(HML-2).
 785 *Journal of Virology* 75, 8917-8926.
 786 Richardson, S.R., Morell, S., and Faulkner, G.J. (2014). L1 Retrotransposons and Somatic
 787 Mosaicism in the Brain. *Annual Review of Genetics* 48, 1-27.
 788 Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for
 789 differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139-140.
 790 Saitou, M., and Miyauchi, H. (2016). Gametogenesis from Pluripotent Stem Cells. *Cell Stem Cell*
 791 18, 721-735.
 792 Saitou, M., and Yamaji, M. (2012). Primordial Germ Cells in Mice. *Cold Spring Harbor*
 793 *Perspectives in Biology* 4.
 794 Sankararaman, S., Mallick, S., Patterson, N., and Reich, D. (2016). The Combined Landscape of
 795 Denisovan and Neanderthal Ancestry in Present-Day Humans. *Current Biology* 26, 1241-1247.
 796 Sasaki, K., Nakamura, T., Okamoto, I., Yabuta, Y., Iwatani, C., Tsuchiya, H., Seita, Y., Nakamura,
 797 S., Shiraki, N., Takakuwa, T., *et al.* (2016). The Germ Cell Fate of Cynomolgus Monkeys Is
 798 Specified in the Nascent Amnion. *Developmental Cell* 39, 169-185.
 799 Sasaki, K., Yokobayashi, S., Nakamura, T., Okamoto, I., Yabuta, Y., Kurimoto, K., Ohta, H.,
 800 Moritoki, Y., Iwatani, C., Tsuchiya, H., *et al.* (2015). Robust In Vitro Induction of Human Germ
 801 Cell Fate from Pluripotent Stem Cells. *Cell Stem Cell* 17, 178-194.
 802 Subramanian, R.P., Wildschutte, J.H., Russo, C., and Coffin, J.M. (2011). Identification,
 803 characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human
 804 endogenous retroviruses. *Retrovirology* 8, 90.
 805 Tang, Walfred W.C., Dietmann, S., Irie, N., Leitch, Harry G., Floros, Vasileios I., Bradshaw,
 806 Charles R., Hackett, Jamie A., Chinnery, Patrick F., and Surani, M.A. (2015). A Unique Gene
 807 Regulatory Network Resets the Human Germline Epigenome for Development. *Cell* 161, 1453-
 808 1467.
 809 Tao, Y., Yen, M.-R., Chitiashvili, T., Nakano, H., Kim, R., Hosohama, L., Tan, Y.C., Nakano, A.,
 810 Chen, P.-Y., and Clark, A.T. (2018). TRIM28-Regulated Transposon Repression Is Required for
 811 Human Germline Competency and Not Primed or Naive Human Pluripotency. *Stem Cell Reports*
 812 10, 243-256.
 813 Team, R.C. (2019). R: A Language and Environment for Statistical Computing.
 814 Turner, G., Barbulescu, M., Su, M., Jensen-Seaman, M.I., Kidd, K.K., and Lenz, J. (2001).
 815 Insertional polymorphisms of full-length endogenous retroviruses in humans. *Current Biology*
 816 11, 1531-1535.

Whyte, Warren A., Orlando, David A., Hnisz, D., Abraham, Brian J., Lin, Charles Y., Kagey, Michael H., Rahl, Peter B., Lee, Tong I., and Young, Richard A. (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* 153, 307-319.

Wolf, A.B., and Akey, J.M. (2018). Outstanding questions in the study of archaic hominin admixture. *PLOS Genetics* 14, e1007349.

Zhang, Y., Li, T., Preissl, S., Amaral, M.L., Grinstein, J.D., Farah, E.N., Destici, E., Qiu, Y., Hu, R., Lee, A.Y., *et al.* (2019). Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nature Genetics* 51, 1380-1388.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., *et al.* (2008a). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* 9, R137.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., *et al.* (2008b). Model-based analysis of ChIP-Seq (MACS). *Genome Biology* 9.

Zhou, F., Li, M., Wei, Y., Lin, K., Lu, Y., Shen, J., Johanning, G.L., and Wang-Johanning, F. (2016). Activation of HERV-K Env protein is essential for tumorigenesis and metastasis of breast cancer cells. *Oncotarget* 7, 84093-84117.

Figure S1

Figure S1. HERVK/LTR5_Hs RNAs are not required for hPGCLC specification, related to figure 1.

A. Scatter plot of HERVK-int (blue dots) and solo LTR5_Hs (red dots) in hESCs compared to hPGCLCs.

B. Representative flow cytometry plots of hPGCLCs following transduction with short hairpin Scramble control (shScramble) or HERVK (shHERVK) lentiviral vectors. Circle indicates hPGCLC population.

C. RT-qPCR quantitative analysis of HERVK expression in shScramble related to shHERVK groups. (biological replicates n=3; ** means $p=0.0029$ from unpaired two-tailed t test; error bar is standard deviation/SD).

D. Statistical analysis of percent hPGCLCs, comparing shScramble to shHERVK groups (biological replicates n=3; ns indicates not significant, $p=0.6157$ from unpaired two-tailed t test; error bar is standard deviation/SD).

Figure S2

Figure S2. SVA_D are specifically more accessible in hPGCLCs than hESCs, related to figure 2.

A,B. Barplot of $\log_2(\text{Observed/Expected})$ enrichment for hPGCLC-specific open ERVK (A) and SVA (B) subfamilies .

C. Pie chart for the distribution of SVA subfamily members in the human genome, and distribution of enriched SVA subfamilies in hPGCLC-specific open regions.

D. Metaplot and heatmap of ATAC-seq data over SVA_D in hESCs, iMeLC, hPGCLCs and hPGCs.

E. Expression pattern of SVA_D in the hPGCLC differentiation scRNA-seq dataset.

Figure S3

Figure S3. related to figure 3. Epigenetic reprogramming of SVA_D in hPGCLCs.

A. Metaplot of mCG levels over SVA_D in hESC and hPGCLC.

B. NANOG/TFAP2C/H3K27ac ChIP-seq over SVA_D (n=1546) in hESCs and hPGCLCs and SOX17 ChIP-seq over SVA_D in Seminoma.

C-E. Expression pattern of *TFAP2C* (C), *SOX17* (D) and *NANOG* (E) in the hPGCLC differentiation scRNA-seq dataset.

F. Screenshot for one representative human specific LTR5_Hs loci adjacent to a hPGCLC up-regulated gene (*KLF7*).

839 Supplement table 1. Resource table
840

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
BV421 conjugated anti-human/mouse CD49f (ITGA6)	BioLegend	Cat#313624; RRID: AB_2562244; Lot#B274314
APC-conjugated anti-human CD326 (EPCAM)	BioLegend	Cat#324208; RRID: AB_756082; Lot#B284158
NANOG	R&D	Cat#AF1997; Lot#KKJ0617071
Bacterial and Virus Strains		
CRISPRi - pLV hU6-sgRNA hUbC-dCas9-KRAB-T2a-Puro	Addgene	#71236; RRID: Addgene_71236
hH1P-shRNA-pUbqC-HygroEGFP	(Ivanova et al., 2006)	N/A
Chemicals, Peptides, and Recombinant Proteins		
Y27632	Stemgent	Cat# 04-0012-10
Puromycin	Invivogen	Cat# ant-pr-1
Hygromycin	Thermo Fisher Scientific	Cat# 10687010
Recombinant Human FGF basic Protein	R&D system	Cat# 233-FB
CHIR99021	Stemgent	Cat# 04-0004
Recombinant Activin A	Peptotech	Cat# AF-120-14E
Recombinant Human BMP4	R&D system	Cat# 314-BP
Recombinant Human EGF	R&D system	Cat# 236-EG
Recombinant Human LIF	Millipore	Cat# LIF1005
Critical Commercial Assays		
Ovation Ultralow System V2	Tecan	Cat# 7102-32
Lipofectamine 3000 Reagent	Thermo Fisher Scientific	Cat# L3000015
PowerUp SYBR Green Master Mix	Thermo Fisher Scientific	Cat# A25742
One Shot Stbl3 Chemically Competent <i>E. coli</i>	Thermo Fisher Scientific	Cat# C737303
Deposited Data		
RNA-seq data	(Chen et al., 2018; Chen et al., 2019) & This paper	GSE93126 GSE140021
ATAC-seq data	(Chen et al., 2018)	GSE120648
WGBS data	(Gell et al., 2020)	GSE139115
ChIP-seq data	(Chen et al., 2019; Ji et al., 2016; Jostes et al., 2020; Lyu et al., 2018) & This paper	GSE140021 GSE105028 GSE69646 GSE105028
Experimental Models: Cell Lines		
UCLA2	(Diaz Perez et al., 2011)	N/A

HEK293T	Female - Embryonic Kidney	N/A
Oligonucleotides		
shScramble sequence: GAATTCTTAACGACTACCA	(Zhou et al., 2016)	N/A
shHERVK sequence: CCTGAACATCCAGAATTAT	(Zhou et al., 2016)	N/A
GAPDH-F primer for RT-PCR: AGCCACATCGCTCAGACAC	(Zhou et al., 2016)	N/A
GAPDH-R primer for RT-PCR: GCCCAATACGACCAAATCC	(Zhou et al., 2016)	N/A
HERVK-F primer for RT-PCR: GGCCATCAGAGTCTAAACCACG	(Zhou et al., 2016)	N/A
HERVK-R primer for RT-PCR: CTGACTTTCTGGGGGTGGCCG	(Zhou et al., 2016)	N/A
gRNA/55 sequence target LTR5_Hs/SVA: CTCCCTAATCTCAAGTACCC	(Pontis et al., 2019)	N/A
gRNA/57 sequence target LTR5_Hs/SVA: TGTTTCAGAGAGCACGGGGT	(Pontis et al., 2019)	N/A
Software and Algorithms		
FlowJo	FlowJo, LLC	N/A
GraphPad Prism	GraphPad	N/A
GRCh38 - Genome Assembly		N/A
fastqc	Andrews, 2010	v0.11.8
STAR – Alignment	(Dobin et al., 2012)	v2.7.0e
Samtools - Processing post-alignment	(Li et al., 2009a)	V1.2
FeatureCounts – Counting reads over genes	(Liao et al., 2014b)	v2.0.0
Bedtools multicov - Counting reads over transposable elements	(Quinlan and Hall, 2010)	v2.17.0
Bedtools multiinter & intersect – identifying specific region	(Quinlan and Hall, 2010)	v2.17.0
Bedtools shuffle – region shuffling	(Quinlan and Hall, 2010)	v2.17.0
ComplexHeatmap – Heatmap plotting	(Gu et al., 2016b)	v2.5.1
DESeq2 – DEG calling	(Love et al., 2014a)	v1.26.0
MACS2 – Peak calling	(Zhang et al., 2008a)	v2.2.7.1
Homer – Peak annotation	(Heinz et al., 2010b)	v4.7
DeepTools - Bigwig tracks generating; metaplot and heatmap plotting	(Ramírez et al., 2014a)	v3.2.1
Phyper - hypergenomic test	R Core Team, 2019	v3.6.0
HTSeq	(Anders et al., 2015)	
edgeR	(Robinson et al., 2010)	
BS-Seeker 2	(Guo et al., 2013)	

841

842

Supplemental information. FACS gating strategy

Figures

Figure 1

Transposons are dynamically expressed during hPGC/hPGCLC development. A. Schematic illustration of hPGCLCs specification in vitro differentiation. B. Heatmap of top 200 variable TEs in hESCs, iMeLCs, hPGCLCs and hPGCs. C. Top differentially expressed TE (DETE) subfamilies in hPGCLCs compared to hESCs. X axis shows DETE percentage over total TE copy number. DETEs with at least 4-fold changes

and FDR less than 0.05 in hPGCLC compared to hESC were used to for this analysis. D-F. UMAP of scRNA-seq data display cell subpopulations during UCLA2 hESC differentiation to hPGCLC (D), expression of NANOS3 (hPGCLC marker) (E) and LTR5_Hs (F). Biological replicates (r1, r2) as well as day 1 (D1) to day 4 (D4) hPGCLC were displayed. Single cell RNA-seq data were from (Chen et al., 2019).

Figure 2

HERVK/LTR5_Hs become uniquely accessible with hPGCLC differentiation. A. Metaplot and heatmap of chromatin accessibility over the 31,276 hPGCLC-specific open chromatin regions. ATAC-seq data is from (Chen et al., 2018). B. Enrichment of hPGCLC-specific open regions over different TE families. “Observed”

Epigenetic reprogramming of LTR5_Hs during hPGCLC development. A. CG methylation level over genome-wide and LTRs in hESCs and hPGCLCs. The x axis is methylation level binned in increments of 10%, y axis indicates the percentage of total mCG sites. B. Metaplot of average mCG levels over LTR5_Hs in hESCs and hPGCLCs. C. Motif enrichment analysis of hPGCLC-specific open chromatin regions overlapping with LTR5_Hs. Observed (occurrences in hPGCLC-specific open chromatin regions) over expected (occurrences in random shuffled regions) were calculated. Top 20 TF motifs with highest observed over expected ratios were plotted. D. NANOG/TFAP2C/H3K27ac ChIP-seq over LTR5_Hs (n=697) in hESCs and hPGCLCs and SOX17 ChIP-seq over LTR5_Hs in Seminoma. NANOG ChIP-seq in hPGCLC was from this paper. NANOG ChIP-seq in hESC was from GSE105028 (Lyu et al., 2018), TFAP2C ChIP-seq in hESC and hPGCLC, H3K27ac ChIP-seq in hPGCLC were from GSE140021 (Chen et al., 2019), H3K27ac ChIP-seq in hESC was from GSE69646 (Hikabe et al., 2016) and SOX17 ChIP-seq in Seminoma-like cells was from GSE130519 (Jostes, S. V. Et al., 2020). E. Barplot for the enrichment of LTR5_Hs associated DEGs. X-axis is the distance of LTR5_Hs to the TSS of neighboring genes. DEGs with at least 4-fold changes and FDR less than 0.05 in hPGCLC compared to hESC were used for this analysis. * indicates p-value <0.05 from hypergeometric test.

Figure 4

LTR5_Hs chromatin accessibility is involved in hPGCLC differentiation. A. Schematic illustration depicting transduction of dCas9-KRAB-gRNAs targeting LTR5_Hs (Pontis et al., 2019) in hESCs followed by induced hESC differentiation to hPGCLCs. B. Representative flow cytometry plots of hPGCLC differentiation in control (Ctrl (22)) and CRISPRi (55+57) groups. The hPGCLC population corresponds to the highlighted ITGA6/EPCAM double positive cells. C. Statistical analysis comparing percentage of hPGCLCs in control relative to CRISPRi groups (biological replicates n=3; ** means p=0.0035 from unpaired two-tailed t test; error bar is standard deviation/SD). D. Expression level of LTR5_Hs, HERVK, SVA and Other TEs in

hPGCLC upon the CRISPRi treatment. The red lines represent the median values of each TE groups. Minimum and maximum values are shown with both ends of the whisker. The box ends represent the first (25%) and third (75%) quartiles. For LTR5_Hs: *** indicates $p=7.4 \times 10^{-83}$ from one sample t-test. For HERVK: *** indicates $p=4.9 \times 10^{-21}$ from one sample t-test. For SVA: *** indicates $p=2.9 \times 10^{-6}$ from one sample t-test. E. Count for neighboring DEGs of LTR5_Hs in CRISPRi hPGCLC compared to control cells. Cut-off for DEG is FDR less than 0.05 and fold change > 10% between paired replicates. F. Expression level of HERVK, LTR5_Hs, HERVH and representative neighboring genes (ST6GAL1, PRODH) in control and CRISPRi hPGCLCs from RNA-seq. (For HERVK, LTR5_Hs and HERVH, *** indicates $p < 0.001$ from paired two-tailed t-test; for ST6GAL1, PRODH, *** indicates $p < 0.001$ from unpaired two-tailed t-test; error bar is standard deviation/SD)

Figure 5

Proposed model for the roles of LTR5_Hs in determining germ cell fate. Based on our data, we proposed that in pluripotent cells, LTR5_Hs DNA is hypermethylated, slightly opened and occupied by TFAP2C and NANOG, which primes germ cell specification. Upon germline cell differentiation, the chromatin of LTR5_Hs becomes significantly more accessible, SOX17 binds with NANOG and TFAP2C resulting in the DNA becoming hypomethylated and the TEEnhancers are utilized to promote expression of neighboring protein coding genes (PCGs) to regulate the germ cell fate.