




# Human reproduction is regulated by retrotransposons derived from ancient *Hominidae*-specific viral infections

Xinyu Xiang<sup>1,10</sup>, Yu Tao<sup>2,10</sup>, Jonathan DiRusso<sup>2,3</sup>, Fei-Man Hsu<sup>2</sup>, Jinchun Zhang<sup>1</sup>, Ziwei Xue<sup>1</sup>, Julien Pontis<sup>4</sup>, Didier Trono<sup>1,4</sup> , Wanlu Liu<sup>1,5,6,7</sup>  & Amander T. Clark<sup>1,2,3,8,9</sup> 

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 reveal that the *Hominidae*-specific Transposable Elements (TEs) LTR5Hs, may serve as TEEnhancers (TE Embedded eEnhancers) to facilitate PGC specification. LTR5Hs TEEnhancers become transcriptionally active during PGC specification both in vivo and in vitro with epigenetic reprogramming leading to increased chromatin accessibility, localized DNA demethylation, enrichment of H3K27ac, and occupation of key hPGC transcription factors. Inactivation of LTR5Hs TEEnhancers with KRAB mediated CRISPRi has a significant impact on germ cell specification. In summary, our data reveals the essential role of *Hominidae*-specific LTR5Hs TEEnhancers in human germ cell development.

<sup>1</sup> Zhejiang University-University of Edinburgh Institute (ZJU-UoE Institute), Zhejiang University School of Medicine, International Campus, Zhejiang University, 718 East Haizhou Rd., Haining 314400, China. <sup>2</sup> Department of Molecular Cell and Developmental Biology, University of California, Los Angeles, Los Angeles, CA 90095, USA. <sup>3</sup> Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90095, USA. <sup>4</sup> School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland. <sup>5</sup> Department of Orthopedic Surgery of the Second Affiliated Hospital of Zhejiang University School of Medicine, Zhejiang University, Hangzhou 310029, China. <sup>6</sup> Dr. Li Dak Sum & Yip Yio Chin Center for Stem Cell and Regenerative Medicine, Zhejiang University, Hangzhou, Zhejiang 310058, China. <sup>7</sup> Alibaba-Zhejiang University Joint Research Center of Future DigitalHealthcare, Zhejiang University, Hangzhou, Zhejiang 310058, China. <sup>8</sup> Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles, Los Angeles, CA 90095, USA. <sup>9</sup> Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, CA 90095, USA. <sup>10</sup> These authors contributed equally: Xinyu Xiang, Yu Tao. ✉email: [wanylulu@intl.zju.edu.cn](mailto:wanylulu@intl.zju.edu.cn); [clarka@ucla.edu](mailto:clarka@ucla.edu)

Proper formation of the adult germline is essential for the passage of genetic and epigenetic information from generation to generation. Primordial Germ Cells (PGCs) are specified during early embryonic development and constitute the founder germline cells that ultimately give rise to oocytes and sperm in the adult. As such, failure to specify PGCs leads to certain infertility in adulthood. Given the central importance of PGCs to reproduction, the developmental cues and regulatory milieu governing specification of PGCs has been broadly studied in various animal models<sup>1</sup>. While these models have proven instructive in PGC biology, constraints imposed by ethical and technical limitations have rendered the precise mechanisms governing human (h) PGC (hPGC) specification in vivo unclear.

Human PGCs originate from peri-implantation progenitors at day 11–12 (D11–12) post-fertilization just before gastrulation<sup>2</sup>, a time point at which clinical samples are prohibitively rare. Due to the inaccessibility of early hPGC development in vivo, an in vitro system for differentiating hPGC-Like Cells (hPGCLCs) from human pluripotent stem cells (hPSCs) has been established<sup>3,4</sup>. Using this system, both conserved and unique transcriptional networks regulating hPGC specification have been uncovered<sup>3–7</sup>. For instance, NANOG, PRDM1, TFAP2C, and PRDM14 are required for PGC specification and maintenance in both human and mouse embryos<sup>2,8–13</sup>. In contrast, SOX17 is crucial for hPGC specification<sup>3</sup>, but is dispensable in mouse; where instead SOX2 regulates the specification of mouse PGCs<sup>14,15</sup>. In addition to the transcription factors (TFs), differences can also be found in the gene regulatory elements required to specify PGCs, such as the utilization of a naïve enhancer at the *POU5F1* (*OCT4*) locus in hPGCs<sup>16</sup>, whereas this naïve enhancer sequence is not conserved in mouse<sup>17</sup>. Given this, we hypothesized that an additional source of variance in the regulatory networks governing PGC specification could be associated with transposable elements (TEs); repetitive elements which account for around half of the human genome.

Most of the TEs in the human genome are retrotransposons, which propagate through an RNA intermediate. Specifically, retrotransposon sequences are first transcribed as RNA, followed by reverse transcription to DNA before integration of a new copy into the genome<sup>18</sup>. Based on function and structure, retrotransposons are further classified as LINE- (long interspersed nuclear elements), SINE- (short interspersed nuclear elements), LTR- (long terminal repeats), or the *Hominidae*-specific SVA (SINE-VNTR-Alu)-elements<sup>18</sup>. Of particular interest when considering TE contribution to the regulatory landscape of the genome are Endogenous retroviruses (ERVs), a superfamily within the LTR retrotransposon class.

ERVs originate from ancient viruses that infected and integrated into the germline throughout evolution. Most Human ERVs appear to have entered the germline after the new world and old world monkey split<sup>19–23</sup>. Even though LTR retrotransposons occupy ~8% of the human genome, almost all LTR retrotransposon sequences have lost their transposition ability<sup>18</sup>. Nevertheless, recent studies suggest that LTR retrotransposons, especially ERVs, can serve as regulatory sequences that participate in gene regulation networks<sup>24</sup>. In humans, ERV sequences harbor binding sites for OCT4, NANOG, and p53<sup>25,26</sup>. Specifically, ChIP-seq analysis has shown that human ERV elements account for roughly 25% of all bound NANOG and OCT4<sup>25</sup> and nearly one-third of all p53 binding sites<sup>26</sup>, demonstrating a profound contribution by human ERVs to the human regulatory landscape.

The most recent expansion of human ERVs occurred over the last 5–20 million years in the HERVK (human mouse mammary tumor virus like-2, HML-2) group<sup>27</sup>. Even though HERVK(HML2) elements are also found in old world primates, distinct phylogenetic differences exist between those found in *Hominidae* relative to

*Hominoidea*. For example, HERVK(HML2) elements which are found in both monkey and human genomes have LTR5A and LTR5B regulatory sequences, while the most recent *Hominidae*-specific HERVK(HML2) elements harbor the LTR5Hs regulatory sequence<sup>27</sup>. In addition, some HERVK(HML2) TEs contain intact open reading frames that can code for viral proteins<sup>28</sup>, with LTR5Hs-regulated HERVK(HML2) provirus expression proposed to be a property of naïve human embryonic stem cells (hESCs)<sup>29</sup>. Grow and colleagues further hypothesize that expression of full-length LTR5Hs-regulated HERVK(HML2) proviruses may confer a critical immunoprotective effect in the human pre-implantation embryo by stimulating IFITM-1, a viral restriction factor, potentially protecting against HERVK(HML-2)-like retrovirus re-infection<sup>29</sup>.

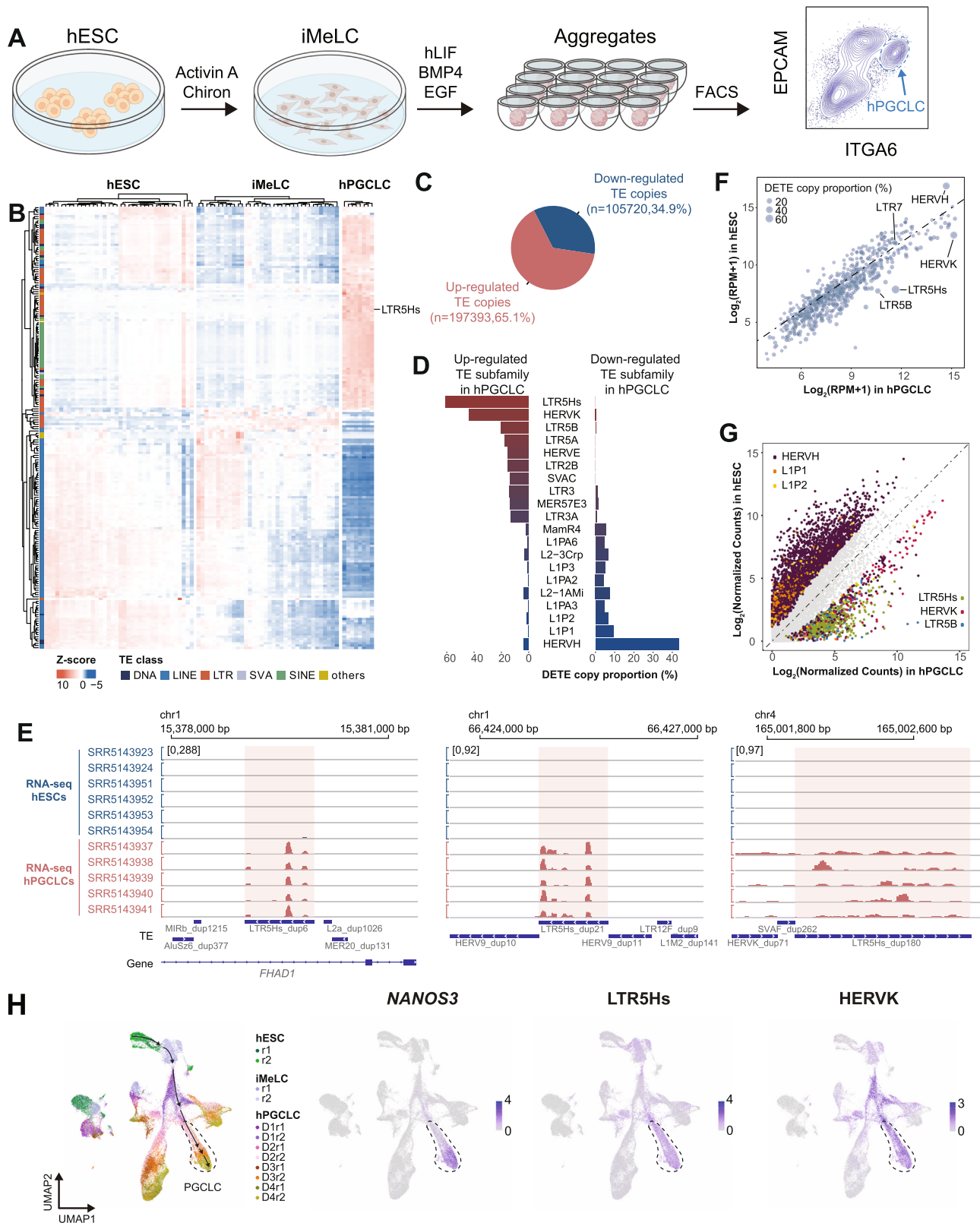
In addition to the production of viral particles, it is also known that many HERVK/LTR5Hs-, SVA-, and HERVH/LTR7- elements in the genome are accessible and marked by H3K27ac in human pluripotent cells, suggesting that they also serve a potential gene regulatory function associated with pluripotency<sup>30</sup>. Consistent with this, key pluripotency factors of the KLF family bind to and activate evolutionarily young TE Embedded eNancers (TEENhancers) found in LTR5Hs and SVA elements to facilitate human embryonic genome activation<sup>30</sup>. Thus, evolutionarily young *Hominidae*-specific TEs have extensively shaped the regulatory landscape of early embryonic development and this has likely fostered species divergence in the gene regulatory networks that regulate the development of cells in the reproduction system.

Here, we discovered that the *Hominidae*-specific LTR5Hs is expressed in hPGCs in vivo and hPGCLCs in vitro. Increased expression of LTR5Hs in hPGCLCs is associated with a remodeled epigenetic landscape leading to increased chromatin accessibility and localized DNA demethylation. Substantial binding of TFAP2C, NANOG, SOX17, SOX15, and enriched H3K27ac histone marks at LTR5Hs loci suggest a TEENhancer role for these TEs in hPGC specification. Inactivation of LTR5Hs TEENhancers compromises hPGCLC formation and de-regulates germline gene expression. In summary, our results reveal that LTR5Hs TEENhancers are involved in hPGC specification, and thus may cultivate the species specificity in human reproduction.

## Results

**Up-regulation of LTR5Hs transcript abundance in germline lineage.** In order to characterize dynamically expressed TE subfamilies during germ cell specification, we analyzed the RNA-seq data sets previously published from our lab<sup>16</sup>, including day 4 (D4) human PGC-like cells (hPGCLCs) differentiated from primed state hESCs through incipient mesoderm like cells (iMeLC) (Fig. 1A and Supplementary Data 1), an intermediate cell type between primed hESCs and hPGCLCs<sup>4</sup>. The D4 hPGCLCs are transcriptionally equivalent to early primate PGCs between D11–D21 post-fertilization<sup>2</sup>. To overview the expression pattern of TEs during hPGCLC induction, the top 200 TE subfamilies with the highest cross-sample variation were visualized (Fig. 1B). In general, we observed dynamic TE expression patterns during hPGCLC specification, with LTR5Hs being one of the top highly expressed TE subfamilies in hPGCLCs (Fig. 1B).

Since TEs are repetitive sequences in the genome, TE-derived RNA-seq reads are hard to quantify. To further identify TE subfamilies specific to hPGCLCs, we called differential expressed TE copies (DETE) in hPGCLCs compared with hESCs using a variety of methods recommended for TE quantification<sup>31–35</sup>. Briefly, RNA-seq reads were aligned to the reference genome with STAR or SQUIRE<sup>32,36</sup> (Supplementary Fig. 1A). Then, TE-derived RNA-seq reads over individual TE copies were quantified with featureCounts,



SQUIRE, Telescope, or Tetrascripts<sup>32–35</sup> followed by DETE calling with DESeq2<sup>37</sup> (Supplementary Fig. 1A and Supplementary Data 2–5). Using a four-fold difference and false discovery rate (FDR) <0.05 as a cut-off, we identified more up-regulated DETE copies in hPGCLCs compared to hESCs (65.1% for featureCounts; 66.1% for Telescope; 71.4% for Tetrascripts) except for SQUIRE

(48.7%) (Fig. 1C and Supplementary Fig. 1B). Since different TE subfamilies possess variable copy numbers, we reasoned large absolute DETE copy numbers may be due to the high total copy number for certain TE subfamilies. Therefore, to reveal TE subfamilies that are most dynamically expressed in hPGCLCs, we calculated the DETE copy numbers proportional to the total copy

**Fig. 1 Lineage-specific up-regulation of LTR5Hs in hPGCLC induction.** **A** Schematic illustration of hPGCLC in vitro differentiation procedure. **B** Heatmap for the top 200 TE subfamilies with the highest cross-sample variation in hESCs, iMeLCs, and hPGCLCs. The colored bar on the left indicates TE class. **C** Pie chart showing the proportion of up- or down-regulated DETE copies in hPGCLCs compared to hESCs using a cut-off of at least a 4-fold change and FDR < 0.05. **D** Top 10 up- or down-regulated TE subfamilies in hPGCLCs. X axis shows DETE copy numbers proportional to the total copy number of a specific TE subfamily. Only TE subfamilies with at least 80 copies and 8 DETE copies are kept for this analysis. **E** Screenshots showing representative hESC and hPGCLC RNA-seq tracks over LTR5Hs integrants. Red shaded rectangle region indicates individual LTR5Hs copies. **F** Scatterplot for aggregated expression level of each TE subfamily in hESCs and hPGCLCs. The size of each dot represents the proportion of DETE copy numbers relative to the total copy number of each TE subfamily. **G** Scatterplot of the expression of individual TE copies belonging to the top three up- or down-regulated DETE subfamilies. Gray dots represent TE copies which are not differentially expressed. **H** UMAP of scRNA-seq dataset for two replicates (r) of UCLA2 hESCs, iMeLCs, and D1 to D4 hPGCLCs (left), representative expression pattern for NANOS3, LTR5Hs, and HERVK. Differentiation trajectory of hPGCLCs is denoted by arrows, hPGCLC population is indicated by dashed line. DETE analysis for this figure is analyzed by the STAR + featureCounts + DESeq2 method. Source data underlying **B**, **D**, and **F** are provided as a Source Data file.

numbers for a specific TE subfamily and plotted the top 10 up- or down-regulated TE subfamilies in hPGCLCs (Fig. 1D and Supplementary Fig. 1C).

With different methods, we consistently observed primate- and *Hominidae*-specific TEs including LTR5Hs/HERVK as top up-regulated, and HERVH as top down-regulated TE subfamilies in hPGCLCs (Fig. 1D, E and Supplementary Fig. 1C). We next analyzed the aggregated transcript abundance for each TE subfamilies and obtained similar conclusions (Fig. 1F and Supplementary Fig. 2A). To better display the transcript abundance dynamics for DETE subfamilies, we also plotted the individual DETE copies for the top 3 up- or down-regulated DETE subfamilies, confirming the up-regulation of LTR5Hs/HERVK in hPGCLCs (Fig. 1G and Supplementary Fig. 2B).

LTR5Hs serves as the regulatory elements for HERVK, while LTR7 serves as the regulatory elements for HERVH<sup>29,38</sup>. We observed up-regulation of both LTR5Hs and HERVK with hPGCLC induction and down-regulation of HERVH, while LTR7 expression levels were unchanged with hPGCLC induction (Fig. 1D, F-G and Supplementary Fig. 1C and 2A-C). As recombination of ERVs leads to the formation of solo-LTRs in the genome<sup>39</sup>, we wanted to evaluate the transcript abundance of provirus-associated LTR5Hs or LTR7 compared to solo-LTR5Hs or solo-LTR7. To do so, we classified LTR5Hs and LTR7 further into HERVK-LTR5Hs, solo-LTR5Hs, HERVH-LTR7, and solo-LTR7. Transcript abundance analysis indicated significant up-regulation of both HERVK-LTR5Hs and solo-LTR5Hs in hPGCLCs (Supplementary Fig. 2D). However, expression levels of HERVH-LTR7 and solo-LTR7 showed no significant changes between hESCs and hPGCLCs (Supplementary Fig. 2D). Our observations suggested that the down-regulation of HERVH in hPGCLCs is uncoupled from expression changes at LTR7.

To investigate the expressions of TEs during hPGCLC induction with single-cell resolution, we re-analyzed the 10X Genomics single-cell RNA-seq data published by our lab<sup>2</sup>. Using NANOS3 as a marker for hPGCLCs, we clearly identified the up-regulation of LTR5Hs, HERVK and down-regulation of HERVH with differentiation of hPGCLCs, while the expression of other selective TE subfamilies were either at background levels or not specific to hPGCLCs (Fig. 1H and Supplementary Fig. 3). For additional interrogation of TEs expressed by hPGCs in vivo or hPGCLCs in vitro the following searchable website has been created and is freely available at <https://labw.org/germlineTE/>.

Human in vivo PGCs start to emerge around embryonic D11-D12<sup>2</sup>. To determine whether newly specified hPGCs in vivo express LTR5Hs, we re-analyzed the scRNA-seq (SMART-Seq) data from two Carnegie Stage 7 (CS7) embryos corresponding to embryonic D15 and D17 post-fertilization<sup>40</sup>. Seven hPGCs were annotated by Tyser et al. in this data set, and four sets of seven other randomly chosen cells were annotated as epiblast, primitive streak, emergent mesoderm, and advanced mesoderm were

included in our analysis of selected TEs (Supplementary Fig. 4A). Using this single-cell RNA-Seq data we showed that LTR5Hs and HERVK are up-regulated in hPGCs in vivo<sup>40</sup>.

We also investigated whether up-regulation of LTR5Hs was specific to hPGCLCs during in vitro somatic cell differentiation by examining the expression of LTR5Hs in RNA-seq datasets from in vitro multilineage differentiation from primed hESCs<sup>41</sup>. This analysis showed that LTR5Hs and HERVK were not enriched during the differentiation of hESCs into mesenchymal stem cell (MSC), neural progenitor cell (NPC), trophoblast-like cell (TBL), and mesendoderm (ME) (Supplementary Fig. 4B). In contrast, and consistent with previous findings, LTR7 and HERVH showed enriched expression in primed hESCs and ME relative to the other cell types<sup>41</sup> (Supplementary Fig. 4B).

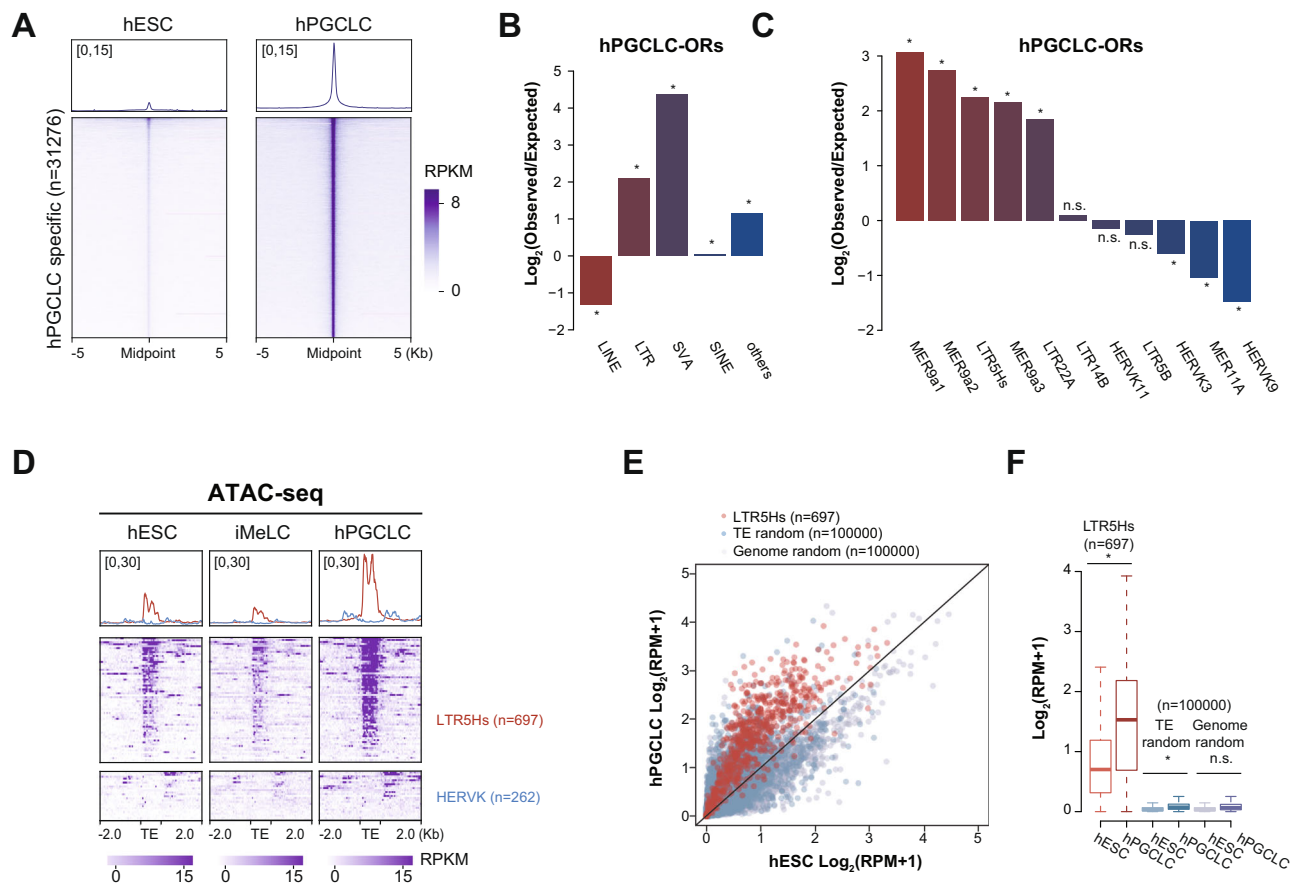
Overall, our in-depth analysis of RNA-seq and scRNA-seq data sets during hPGCLC induction from hESCs, scRNA-seq data from in vivo CS7 PGCs, and RNA-seq data sets from hESC multilineage differentiation collectively showed LTR5Hs is uniquely up-regulated with hPGCLC induction in vitro and is expressed by hPGCs in vivo.

### Increased chromatin accessibility of LTR5Hs in hPGCLCs.

Given the potential enhancer role of TEs in regulating gene expression<sup>30,42</sup>, we next evaluated changes in chromatin accessibilities during hPGCLC induction with our previously published Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) data<sup>16</sup>. Comparing ATAC-seq data of primed-state hESCs, iMeLCs, and hPGCLCs, we identified 31,276 and 90,201 ATAC-seq peaks that become specifically more accessible in hPGCLCs (referred as hPGCLC open regions, hPGCLC-ORs) and hESCs (referred as hESC open regions, hESC-ORs), respectively (Fig. 2A, Supplementary Fig. 5A, and Supplementary Data 6).

To uncover specific TE subfamilies enriched in the hPGCLC-ORs and hESC-ORs, we annotated the genomic distribution of those regions and investigated their enrichment over TE regions. As TEs are not randomly distributed across the genome, we generated randomly shuffled regions as controls by adjusting the relative proportion of genomic distribution comparable to hPGCLC-ORs or hESC-ORs<sup>43</sup> (Supplementary Fig. 5B). Compared with control regions, our analysis revealed that LTR- and SVA-classes were significantly enriched in both hPGCLC- and hESC-ORs (Fig. 2B and Supplementary Fig. 5C). Further analysis of LTR-class containing open regions showed that ERVK was the top enriched LTR family in both hPGCLC- and hESC-ORs (Supplementary Fig. 5D, E). Within the ERVK family, the enriched TE subfamilies diverged between hPGCLC-ORs and hESC-ORs. Interestingly, MER9a1, MER9a2, and LTR5Hs were ERVK subfamilies that were significantly enriched in hPGCLC-ORs, while LTR22A, MER11B, and LTR22C2 were significantly enriched in hESC-ORs (Fig. 2C and Supplementary Fig. 5F). In addition to the LTR family, we also detected enrichment of SVA





**Fig. 2** Increased chromatin accessibility over LTR5Hs with hPGCLC induction. **A** Heatmap and metplot for ATAC-seq signals over hPGCLC-ORs ( $n = 31276$ ). **B, C** Enrichment of TE classes (**B**), and TE subfamilies within the ERVK family (**C**) for hPGCLC-ORs over random shuffled regions with comparable genomic distributions (\* $p$ -value < 0.05, binomial test; n.s. = not significant). **D** Heatmap and metplot of ATAC-seq signals over all LTR5Hs ( $n = 697$ ) and HERVK copies ( $n = 262$ ) in hESCs, iMeLCs, and hPGCLCs. **E, F** Scatterplot (**E**) and boxplot (**F**) of ATAC-seq signals over LTR5Hs, randomly shuffled TE and genomic regions in hESCs and hPGCLCs (\* $p$ -value < 0.05, Welch Two Sample  $t$ -test; n.s. = not significant). In **F** the middle line represents the median; boxes represent the 25th (bottom) and 75th (top) percentiles; and whiskers represent the minimum and maximum points within 1.5× the interquartile range. Source data underlying **B, C**, and **F** are provided as a Source Data file.

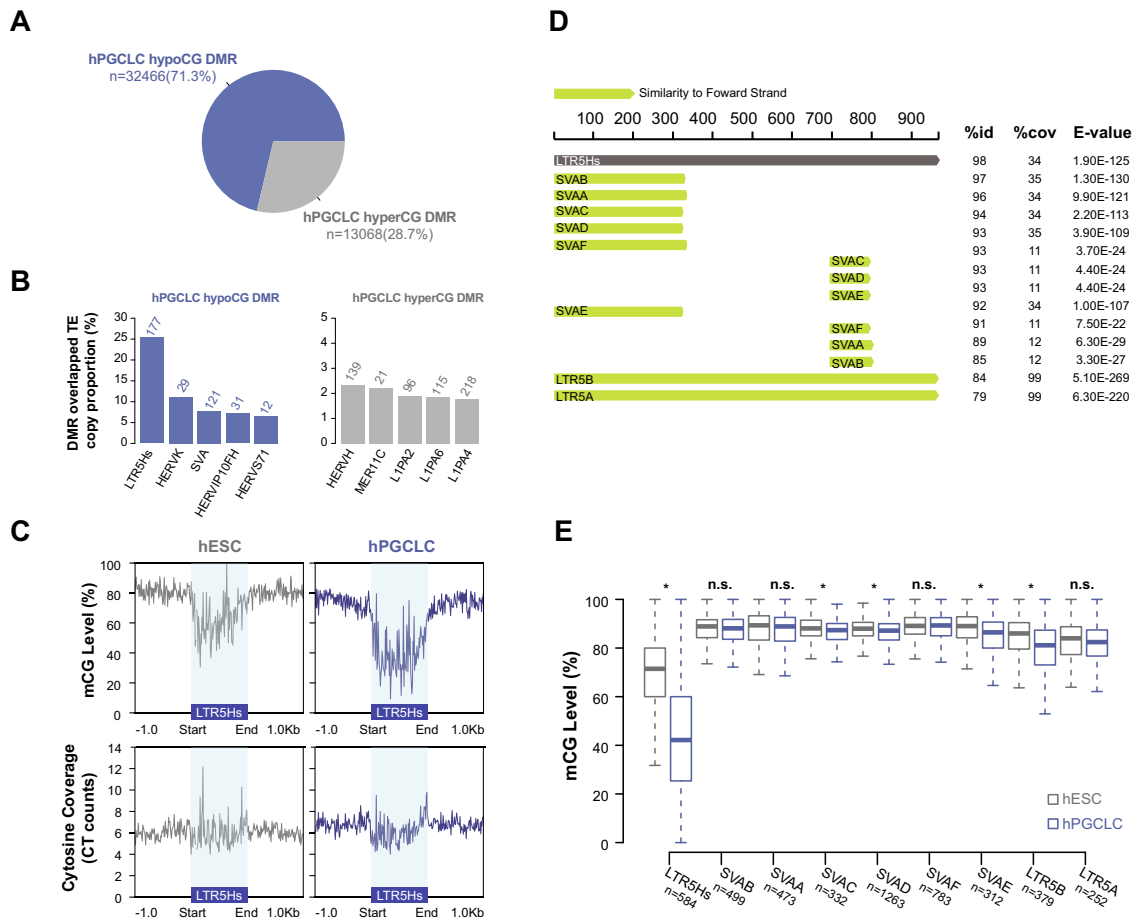
family members including SVAC and SVAD in hPGCLC-ORs, and enrichment of SVAB and SVAA in hESC-ORs (Supplementary Fig. 5G, H).

Analysis for ATAC-seq signals over LTR5Hs further confirmed the increased chromatin accessibility over LTR5Hs in hPGCLCs, while the chromatin landscape of HERVK was not accessible in hPGCLCs (Fig. 2D). As a contrast, LTR7 showed comparable chromatin accessibility in hESCs, iMeLCs, and hPGCLCs, while HERVH was not accessible in any of the cell types (Supplementary Fig. 5I). We next quantified the chromatin accessibility of LTR5Hs in hPGCLCs. By comparing to 100,000 randomly chosen TE copies or genomic regions, we observed that the majority of LTR5Hs loci became more open with hPGCLC induction, while we observed no dramatic changes for control regions (Fig. 2E, F).

**Localized hypomethylation over LTR5Hs in hPGCLCs.** Considering the chromatin accessibility changes over LTR5Hs, we next examined the DNA methylation landscape of LTR5Hs in hPGCLCs. Our previous study suggested there was no obvious genome-wide DNA demethylation in hPGCLCs compared to hESCs<sup>44</sup>. Consistent with our previous conclusion, re-analysis of our hESC and hPGCLC D4 whole Genome Bisulfite Sequencing (WGBS) data showed comparable average CG methylation in hESCs and hPGCLCs (Supplementary Fig. 6A). However,

differential methylated region (DMR) analysis identified 32466 hypomethylated CG (hypoCG, 71.3%) and 13068 hypermethylated CG (hyperCG, 28.7%) DMRs in hPGCLCs compared to primed hESCs (Fig. 3A and Supplementary Data 7). Among those DMRs, we observed LTR5Hs as the top TE subfamily that overlapped with hPGCLC hypoCG DMRs, and HERVH as the top TE subfamily that overlapped with the hPGCLC hyperCG DMRs (Fig. 3B).

Metaplots of CG methylation levels over LTR5Hs revealed CG demethylation across the whole LTR5Hs sequences (Fig. 3C). To rule out mappability issues in highly repetitive sequences, we also examined the cytosine coverages over LTR5Hs and detected comparable mappability within the LTR5Hs regions compared to the flanking genomic sequences (Fig. 3C). SVAD and LTR5Hs share common sequences, and both contribute to maintenance of the transcriptional regulatory network in naïve hESCs<sup>30</sup> (Fig. 3D and Supplementary Fig. 6B). To investigate whether SVAD was also demethylated in hPGCLCs, we plotted the CG methylation level over SVAD and detected modest demethylation close to the 3' end of SVAD (Supplementary Fig. 6B). We reasoned this modest demethylation on SVAD was likely due to sequence conservation between LTR5Hs and this region of the SVAD. To test this hypothesis, we next focused on LTR5Hs and related TE clades that share the most sequence similarities with LTR5Hs: SVAB, SVAA, SVAC, SVAD, SVAF, SVAE, LTR5B, and LTR5A,



**Fig. 3 Localized DNA demethylation over LTR5Hs in hPGCLCs.** **A** Percentage of hypoCG and hyperCG DMRs in hPGCLCs compared to hESCs. **B** Bar plot showing enrichment of TE subfamilies in hyperCG or hypoCG DMRs as a proportion of total TE copy number. **C** Metaplot of aggregate CG methylation level (top) and cytosine coverage (bottom) over LTR5Hs in hESCs and hPGCLCs. Blue shaded rectangle region indicates annotated LTR5Hs regions. **D** The consensus sequence similarity of LTR5Hs and related TE clades from Dfam<sup>45</sup>. Percent identity between the entry consensus sequences (%id), percent shared coverage (%cov) and match e-value (E-value) are displayed on the right. **E** Boxplot of CG methylation level over LTR5Hs and related TE clades in hESCs and hPGCLCs. Only TE subfamilies with a copy number >100 are included in the plot. \**p*-value < 0.05, Welch Two Sample *t*-test; n.s. represents not significant. The middle line represents the median; boxes represent the 25th (bottom) and 75th (top) percentiles; and whisker bars represent the minimum and maximum points within the 1.5× interquartile range. Source data underlying **E** is provided as a Source Data file and originates from *n* = 2 biological replicates (separate experiments) of hESCs and hPGCLCs generated from the UCLA2 hESC line.

obtained from the Dfam database<sup>45</sup> (Fig. 3D). Of this clade, only LTR5Hs displayed extreme CG demethylation during hPGCLC induction, while none of other related TE clades showed this trend (Fig. 3E and Supplementary Fig. 6D). This result suggested that the localized demethylation at LTR5Hs is specific to LTR5Hs itself, rather than to the LTR5Hs related sequences in SVA.

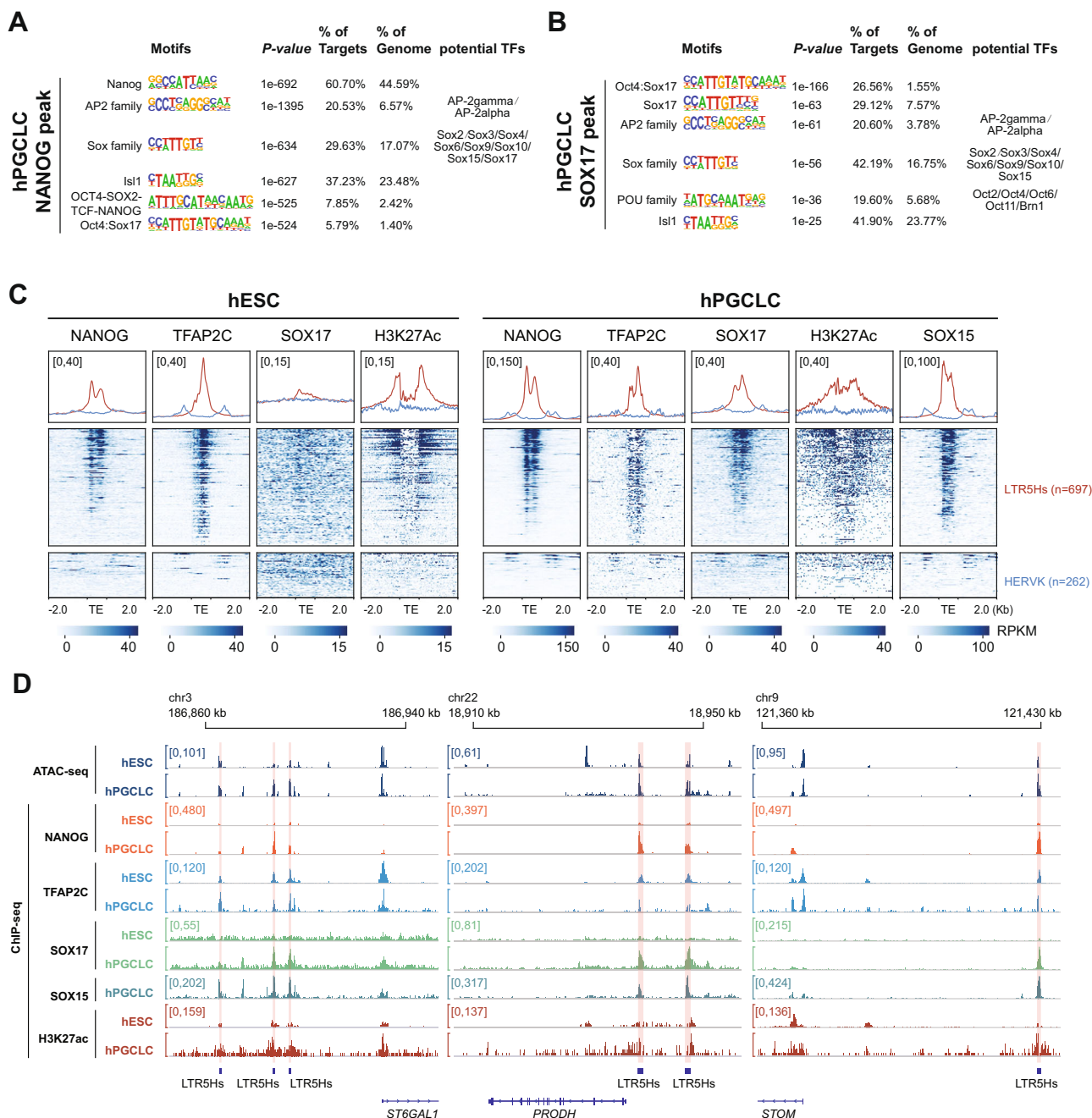
Based on our observations of RNA expression, chromatin accessibility, and localized demethylation of LTR5Hs in hPGCLCs, we thus hypothesized that the epigenetic activity of LTR5Hs might mediate a human-specific epigenetic landscape for hPGC specification.

**LTR5Hs may serve as germ cell-specific TE enhancers.** A previous study on evolutionary young TEs in human early embryogenesis suggested that LTR5Hs and SVAD elements may serve as TE enhancers, which are involved in species-specific transcriptional networks<sup>30</sup>. To explore whether LTR5Hs functions as TE enhancers in hPGCLC induction, we inspected the binding profiles of key TFs as well as the enhancer histone mark, H3K27ac at LTR5Hs.

Previous studies have shown that Transcription Factor AP-2 (Activating enhancer-binding Protein 2) Gamma (TFAP2C), a TF

from the AP2 family, is required for hPGCLC induction<sup>3,16</sup>. SRY-box (Sex-determining Region Y box) TFs SOX17, SOX15 and its downstream target ETV5 have also been reported as critical factors for hPGCLC induction and maintenance<sup>3,7,46</sup>. In addition, homeobox protein NANOG has been proposed as an indispensable pluripotency factor in PGC fate determination<sup>8,47</sup>. Motif analysis of hPGCLC-ORs which overlapped with LTR5Hs showed enrichment of known factors critical for PGC biology, including ets-ebox, AP2, and Sox family (Supplementary Fig. 7A). These results were consistent with reports that the Oct4:Sox17, AP-2 Gamma, and Sox15 motifs were highly enriched in hPGCLC-ORs<sup>16</sup>. To examine the binding of TFs at LTR5Hs in hESCs and hPGCLCs, we evaluated our previous published TFAP2C ChIP-seq (Chromatin Immunoprecipitation followed by sequencing)<sup>2</sup>, previously published H3K27ac ChIP-seq data<sup>2,48</sup>, as well as published SOX15 CUT&Tag-seq (Cleavage Under Targets and Tagmentation)<sup>7</sup>. In addition, we performed ChIP-seq of NANOG and SOX17 in hESCs and hPGCLCs.

Motif analysis for NANOG and SOX17 ChIP-seq peaks in hPGCLCs validated the quality of our ChIP-seq experiments, with the most enriched motif as Nanog and the Oct4:Sox17 fused motif, respectively (Fig. 4A, B). Interestingly, we detected the



**Fig. 4** LTR5Hs may serve as TEEnhancers in hPGCLCs. **A, B** Top enriched motifs within NANOG (**A**) and SOX17 (**B**) ChIP-seq peaks in hPGCLCs. Hypergeometric test is performed using Homer<sup>77</sup>. **C** Heatmaps and metaplots of NANOG, TFAP2C, SOX17, and H3K27ac ChIP-seq signals in hESCs and hPGCLCs, and SOX15 CUT&Tag-seq signals in hPGCLCs over all LTR5Hs ( $n = 697$ ) and HERVK ( $n = 262$ ). **D** Screenshots for ATAC-seq signals, NANOG, TFAP2C, SOX17, and H3K27ac ChIP-seq signals in hESCs and hPGCLCs, and CUT&Tag-seq signals for SOX15 over LTR5Hs TEEnhancers and their potential target genes (*ST6GAL1*, *PRODH*, and *STOM*).

enrichment of AP2 family, SOX family, and POU family motifs in both the NANOG and SOX17 ChIP-seq peaks in hPGCLCs (Fig. 4A, B). Therefore, our results implied the existence of an interconnected transcriptional regulatory network in hPGCLCs.

To address this, we next analyzed the binding profiles of key TFs and H3K27ac in hESCs and hPGCLCs over LTR5Hs and HERVK with LTR7 and HERVH used as controls (Fig. 4C, Supplementary Fig. 7B, and Supplementary Data 8). Overall, we observed extensive binding of NANOG (39.7%), TFAP2C (58.7%), and an enrichment of H3K27ac, but no binding of SOX17, over the majority of LTR5Hs copies in undifferentiated hESCs (Fig. 4C and Supplementary Fig. 7C). For LTR7, we

observed moderate binding of NANOG (31.0%) and TFAP2C (14.6%), and a slight enrichment of H3K27ac in hESCs (Supplementary Fig. 7B, C). In contrast, with differentiation of hPGCLCs we observed universal binding of NANOG (71.3%), TFAP2C (62.4%), SOX15 (60.4%), and SOX17 (24.0%) as well as the enrichment of H3K27ac at LTR5Hs (Fig. 4C and Supplementary Fig. 7D). The binding of key hPGC TFs, as well as enrichment of H3K27ac at LTR5Hs suggests an enhancer role for LTR5Hs with hPGCLC induction. For instance, a 40-kb distal LTR5Hs has been proposed to act as super-enhancer for naive pluripotency gene *ST6GAL1*<sup>30,49</sup>. We also observed the extensive binding of NANOG, TFAP2C, SOX17, and SOX15 over the

LTR5Hs nearby *ST6GAL1* in hPGCLCs (Fig. 4D). Similar binding patterns were observed for LTR5Hs near the hPGCLC up-regulated genes *PRODH* and *STOM* (Fig. 4D). For LTR7, we observed modest binding of NANOG (31.2%) in hPGCLCs with negligible binding of SOX15 (8.8%), TFAP2C (7.4%), SOX17 (0.5%), or H3K27ac enrichment (Supplementary Fig. 7B, D). No signs of NANOG, TFAP2C, SOX17, SOX15, or H3K27ac enrichment were detected over HERVK or HERVH (Fig. 4C and Supplementary Fig. 7B).

The substantial binding of key hPGCLC key TFs, along with the localized remodeling of the epigenetic landscape, led us to propose that LTR5Hs may serve as a hPGCLC-specific TEEnhancer to regulate hPGCLC induction.

**LTR5Hs TEEnhancers are essential for hPGCLC Induction.** To evaluate the functional relevance of LTR5Hs TEEnhancers in hPGCLC induction, we transduced UCLA2 hESCs lines with lentivirus encoding dCas9-KRAB fusion protein together with validated gRNAs targeting LTR5Hs (referred as CRISPRi-LTR5Hs)<sup>30</sup> (Fig. 5A). As control, hESCs were transduced with dCas9-KRAB with no gRNAs (referred as CRISPRi-empty). Then, CRISPRi-empty and CRISPRi-LTR5Hs hESC lines were induced to differentiate into hPGCLCs (Fig. 5A). By tethering KRAB protein to LTR5Hs loci with CRISPRi, the H3K9me3 repressive mark would be induced at targeted loci, thus inactivating LTR5Hs TEEnhancers<sup>50</sup>. At day 4 of hPGCLC induction, we quantified the percentage of hPGCLCs using Fluorescence-Activated Cell Sorting (FACS). In the CRISPRi-LTR5Hs lines, we consistently observed a significant reduction in the percentage of hPGCLCs compared to CRISPRi-empty controls (Fig. 5B, C). We further validated our results by repeating the experiments in the UCLA1 hESC line and obtained the same conclusion (Supplementary Fig. 8A). Collectively, our results suggested LTR5Hs TEEnhancers are involved in hPGCLC induction.

To uncover potential downstream LTR5Hs TEEnhancer-regulated genes involved in hPGCLC biology, we performed RNA-seq of CRISPRi-LTR5Hs and CRISPRi-empty sorted hPGCLCs. Analyzing the DE TE copies in CRISPRi-LTR5Hs compared with CRISPRi-empty, we detected 264 (85.7%) down-regulated and 44 (14.3%) up-regulated TE copies, with HERVK and LTR5Hs as TE subfamily with the most down-regulated DE TE copies and HERVH with the most up-regulated DE TE copies (Fig. 5D, E and Supplementary Data 9).

We then scanned the potential target sites for LTR5Hs gRNAs in the human genome by allowing a maximum of three mismatches with the LTR5Hs targeting guides. In total, we identified 6044 predicted target sites for the two gRNAs used to target LTR5Hs, among which 942 (15.59%) were located on 76.76% of all LTR5Hs copies (Supplementary Fig. 8B and Supplementary Data 10). Consistent with previous findings, SVA family members, especially SVAD, were also predicted to be targeted by the two gRNAs<sup>30</sup>, while few predicted sites targeted to genic regions (Supplementary Fig. 8B). Even though the gRNAs could be targeted to SVAD, we found no evidence for down-regulation of SVAD in CRISPRi-LTR5Hs hPGCLCs (Supplementary Fig. 8C). Using the same gRNAs, Pontis *et al.* observed significant repression of SVAD in CRISPRi-LTR5Hs naïve hESCs<sup>30</sup> (Supplementary Fig. 8C). This difference between naïve hESCs and hPGCLCs is likely due to the very low SVAD expression levels in hPGCLCs compared to naïve hESCs (Supplementary Fig. 8C). Additionally, as our hPGCLCs are differentiated from primed hESCs in which SVAD elements are not expressed (this study) or adorned with H3K27ac (Pontis *et al.*<sup>30</sup>), we would not expect interference in hPGCLC induction from off-target SVAD silencing. Overall, we detected significant

down-regulation of LTR5Hs, HERVK, and up-regulation of HERVH in CRISPRi-LTR5Hs hPGCLC while no changes in SVAD or LTR7 (Supplementary Fig. 8C, D).

We then analyzed the effect of CRISPRi-LTR5Hs on gene expression. Consistent with the DE TE pattern, we detected 124 (80%) of down-regulated DEGs (differential expressed genes) (using 1.5-fold change and FDR <0.05 as cut-off), while only 31 (20%) DEGs were up-regulated in CRISPRi-LTR5Hs compared to control (Fig. 5F and Supplementary Data 11). Considering the mild gene expression changes, we also included a MA plot to control for data normalization during DEG calling (Supplementary Fig. 8E).

To evaluate whether LTR5Hs was significantly associated with the DEGs in CRISPRi-LTR5Hs, we employed RAD (Region Associated DEG) analysis<sup>51</sup>. With this analysis, we discovered that down-regulated DEGs in CRISPRi-LTR5Hs were significantly enriched within 200 kb next to LTR5Hs copies (Fig. 5G and Supplementary Data 12). As a control, no significant association was found between randomly shuffled regions and CRISPRi-LTR5Hs DEGs (Fig. 5G and Supplementary Data 12). Correspondingly, RAD analysis for CRISPRi gRNAs predicted sites associated CRISPRi-LTR5Hs DEGs showed a similar pattern (Supplementary Fig. 8F and Supplementary Data 13).

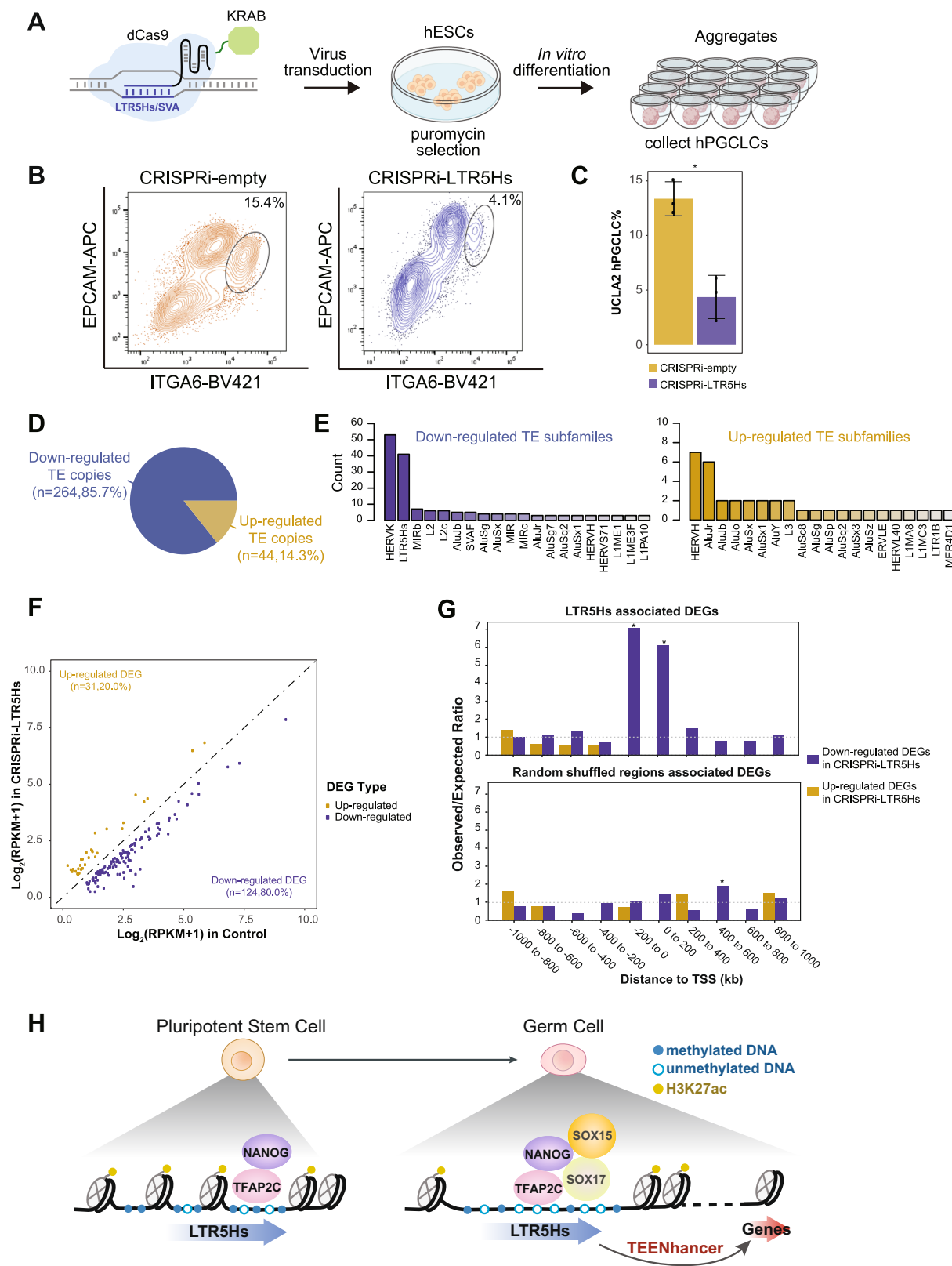
To rule out the possibility that the decrease in hPGCLC induction in the CRISPRi-LTR5Hs was derived from indirect effects, such as differentiation delay, or loss of pluripotency, we examined the expression of hPGC and pluripotency marker genes (Supplementary Fig. 9A). We detected no obvious changes in expression of these marker genes in CRISPRi-LTR5Hs and control samples, and thus conclude that the reduced induction of hPGCLCs was likely caused by a direct effect of interference with LTR5Hs accessibility and enhancer function (Supplementary Fig. 9A).

Even though significantly fewer hPGCLCs were induced with CRISPRi-LTR5Hs differentiation, no canonical hPGCLC marker was repressed in these cells (Supplementary Fig. 9A). Therefore, to identify potential new LTR5Hs TEEnhancer-regulated genes in hPGCLCs, we further analyzed the DEGs up-regulated with hPGCLC differentiation from hESCs and the DEGs down-regulated in CRISPRi-LTR5Hs hPGCLCs compared with CRISPRi-empty hPGCLCs (Supplementary Fig. 9B and Supplementary Data 14). We identified significant overlap (95/124, 76.6% for CRISPRi-LTR5Hs down-regulated DEGs) between down-regulated DEGs in CRISPRi-LTR5Hs hPGCLCs and hPGCLC-specific up-regulated DEGs (Supplementary Fig. 9B and Supplementary Data 14). We thus reasoned that those 95 genes were likely to be the direct targets of LTR5Hs TEEnhancers and we predicted that these genes might have a role in hPGCLC biology. For instance, *PRODH*, *ST6GAL1*, and *STOM* were candidate genes specifically expressed in hPGCLCs, repressed in CRISPRi-LTR5Hs hPGCLCs, and showed hPGCLC specificity relative to somatic cells at single-cell resolution. In addition, these genes were potentially regulated by LTR5Hs TEEnhancers (Fig. 4D and Supplementary Fig. 9C, D).

## Discussion

Despite having been initially coined “controlling elements” by Barbara McClintock, TEs were long discarded as parasitic genetic elements. Within the last decade it has become apparent that TEs contribute profoundly to the regulatory landscape of the human genome. Although many TEs are epigenetically silenced by defensive mechanisms, some TE sequences are domesticated by the host during evolution and are therefore kept under selective pressure<sup>24</sup>. *Hominidae*-specific TEs are relatively recent and do not exist in new world-monkeys or even old-world monkeys such





as the rhesus or cynomolgous macaque. These relatively new *Hominidae* TEs have evolved species-specific functions which are unique to apes, and in some cases are unique to humans.

Here we have shown that one of these *Hominidae*-specific elements, LTR5Hs, is detected at the RNA level both *in vitro* in hPGCLCs and *in vivo* in hPGCs using single-cell RNA-seq data

from a CS7 human embryo. Using our *in vitro* model, we likewise show that LTR5Hs elements acquire an open chromatin state, are hypomethylated and bound by key PGC TFs including NANOG, TFAP2C, SOX17, and SOX15 after hPGCLC induction. Further supporting the role of LTR5Hs as enhancers necessary for germ cell specification, we found that LTR5Hs elements are decorated

**Fig. 5 Inactivation of LTR5Hs TEENhancers leads to less hPGCLC formation.** **A** Schematic illustration depicting transduction of dCas9-KRAB-gRNAs targeting LTR5Hs in hESCs followed by hESCs differentiation to hPGCLCs. **B** Representative flow cytometry plots for hPGCLC differentiating from CRISPRi-LTR5Hs and CRISPRi-empty in UCLA2 hESC lines. The black circles denote the hPGCLC population as defined by ITGA6/EPCAM double-positive cells. **C** Barplot showing the percentage of hPGCLCs in CRISPRi-empty relative to CRISPRi-LTR5Hs groups in UCLA2 (biological replicates  $n = 3$ ; \* $p$ -value = 0.0042; error bars showing mean  $\pm$  SEM). **D** Pie chart showing up- or down-regulated DETE copies in CRISPRi-LTR5Hs compared with CRISPRi-empty controls, as defined by at least a 4-fold change in expression and FDR  $< 0.05$ . **E** Barplot of the TE subfamilies with the most up- or down-regulated DETE copies. **F** Scatterplot of the expression level for identified up- or down-regulated DEGs in CRISPRi-LTR5Hs compared with CRISPRi-empty control, using a cut-off of at least 1.5-fold change in expression and a FDR of  $< 0.05$ . **G** RAD analysis for the association between LTR5Hs (upper panel) or random shuffled regions (lower panel) with CRISPRi-LTR5Hs DEGs. \* $p$ -value  $< 0.05$ , two-sided Welch Two Sample  $t$ -test. **H** Proposed model for the role of LTR5Hs as TEENhancers in PGC specification. Source data underlying **C** and **G** are provided as a Source Data file.

with H3K27ac in hPGCLCs, and that silencing of LTR5Hs using CRISPRi reduces the efficiency of hPGCLC induction, in part due to loss of LTR5Hs enhancer function. Together these results show that *Hominidae*-specific LTR5Hs could serve as TEENhancers necessary for hPGC specification (Fig. 5H), and therefore could be considered essential to successful human reproduction.

Both HERVK-associated and solo-LTR5Hs integrants function as TEENhancers necessary for the maintenance of naïve pluripotency<sup>29,30</sup>. In the naïve state, LTR5Hs copies are hypomethylated, marked with H3K27ac and are synergistically bound by OCT4, p300, and key KLF-family members, most notably KLF4 and KLF17<sup>29,30</sup>. While OCT4 is expressed in both primed and naïve pluripotency, KLF17 and KLF4 are naïve-specific TFs, suggesting that binding by KLF4-KLF17 is necessary for LTR5Hs TEENhancer function in naïve hESCs. Further supporting evidence for this conclusion is the observation that over expression of KLF4-KLF17 in the primed state of pluripotency is sufficient to open the LTR5Hs TEENhancers and regulate neighboring gene expression<sup>30</sup>. Our recent study suggests that KLF17 is not expressed in hPGCLCs, whereas KLF4 is up-regulated upon hPGCLC induction but is not functionally required for the induction or proliferation of hPGCLCs<sup>52</sup>. Thus, we propose that unlike naïve pluripotent stem cells where KLF4 and KLF17 bind to TEENhancers, the LTR5Hs TEENhancers in hPGCLCs utilize SOX17, SOX15, TFAP2C, NANOG, and ETV5. These data collectively argue that while LTR5Hs copies function as TEENhancers in both naïve hESCs and hPGCLCs, the TF networks that reinforce LTR5Hs TEENhancer function in each cell state are distinct.

Despite lack of KLF4 activity and KLF17 expression in germ cell specification, hPGCLCs in vitro and hPGCs in vivo exhibit a naïve-like pluripotent molecular program that has similarities to the naïve state in pre-implantation human embryos. This includes two active X chromosomes in females, genome-wide DNA demethylation and expression of naïve pluripotent TFs including KLF4, TFPCP2L1, and TFAP2C<sup>9,16,52-54</sup>. 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<sup>17,30</sup>. Our results imply that the commissioning of LTR5Hs TEENhancers during hPGCLC induction is driven by the marking of these sites in primed hESCs by a basic network of TFAP2C and NANOG, which is then reinforced with the recruitment of SOX17 and SOX15 during hPGCLC induction. Further supporting this interpretation, time-resolved ATAC-seq during hPGCLC induction from Wang *et al.* shows Sox15 and Oct4:Sox17 motifs become preferentially open during the second day of the four-day hPGCLC differentiation protocol<sup>7</sup>, roughly concomitant with enrichment of naïve-state gene profiles by hPGCLCs<sup>2</sup>. Thus, proper LTR5Hs TEENhancer activity may be necessary for acquisition of a naïve-like transcriptome during hPGCLC induction.

Interestingly, we also observed strong enrichment of ets-ebox binding motifs in hPGCLC-ORs (Supplementary Fig. 7A), which

are bound by ETS-family TFs, including ETV4 and ETV5. Recently, it has been proposed that ETV5 is necessary for hPGCLC maintenance, functioning downstream of SOX15. In the absence of SOX15, ETV5 expression is reduced and, reciprocally, efficiency of hPGCLC induction is reduced in the absence of ETV5<sup>7</sup>. These data lead us to hypothesize that ETV5 may also bind LTR5Hs elements during or after hPGCLC induction, possibly following SOX15-mediated commissioning of LTR5Hs enhancers.

We have found that TFAP2C and NANOG are bound to LTR5Hs in undifferentiated hESCs. Our data established a model whereby TFAP2C and NANOG license LTR5Hs in hESCs, and following entry into hPGCLC differentiation, SOX17 and SOX15 cooperate with TFAP2C and NANOG to recruit chromatin remodeling complexes to open chromatin and promote DNA demethylation at LTR5Hs, thus enabling their activity as enhancers. Our results also suggest that localized DNA demethylation over LTR5Hs precedes the global DNA demethylation in the germline, which is a hallmark of hPGC development in the embryo<sup>9,53,54</sup>, further implicating proper commissioning of LTR5Hs enhancer elements as an essential step in, and not a consequence of, hPGCLC induction.

Curiously, despite strong sequence conservation between SVAD elements and the 3' end of LTR5Hs elements (Fig. 3D), we observe distinct differences in the epigenetic state of these subfamilies after hPGCLC induction. SVAD integrants show less extensive DNA demethylation and accessibility in hPGCLCs with SVAD transcripts being expressed at low levels in hPGCLCs. It has become increasingly appreciated that enhancer elements are often produced by bi-directional, unspliced and often non-polyadenylated RNA Polymerase II-transcribed RNAs, termed enhancer (e) RNAs<sup>55-57</sup>, and that eRNA transcription levels are often positively correlated with the transcriptional levels of nearby genes<sup>55</sup>. Although the function of eRNAs remains enigmatic, production of eRNA has become a hallmark of strong enhancer function. Still, non-transcribed enhancers may act as weak enhancers<sup>57</sup>. Given that SVAD is modestly demethylated and has weak enhancement of chromatin accessibility, it remains possible that SVAD may have some weak enhancer activity in hPGCLCs. In contrast, robust LTR5Hs transcript detection, dramatic DNA demethylation, and chromatin accessibility suggest that LTR5Hs elements act as strong enhancers in hPGCLCs.

While hPGCs acquire a naïve-like transcriptome, they do not fully exit primed state, and demonstrate characteristics of both states<sup>2</sup>. While LTR5Hs/HERVK expression has been linked to a naïve state, enrichment of LTR7/HERVH expression has likewise been associated with the primed pluripotent state<sup>58</sup>, although some LTR7 elements show hallmarks of enhancer function in naïve hESCs<sup>30</sup>. Interestingly, while we detected an up-regulation of LTR5Hs expression, we did not observe any changes in LTR7 expression with hPGCLC induction. Despite no change to LTR7 expression, we did observe a modest decrease in NANOG binding and a decrease in H3K27ac enrichment at certain LTR7 copies

upon hPGCLC induction. Thus, while LTR7 expression remains unchanged between primed state hESCs and hPGCLCs, LTR7 enhancer function seems to be decommissioned during hPGCLC induction. This suggests that, in some contexts, LTR7 enhancer function might be uncoupled from RNA production at LTR7 loci.

In addition to gene regulation at enhancers and promoters, human ERVs 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<sup>59</sup>. Given that HERVH is down-regulated and LTR5Hs sequences are up-regulated during hPGCLC induction, it is possible that LTR5Hs may also be required for the assembly of genome 3-D architecture in hPGCLCs, and therefore the failure to fully repress HERVH in the CRISPRi-LTR5Hs hPGCLCs.

Finally, here we have identified three potential LTR5Hs-regulated genes which may be important to hPGC biology based on their selective expression in hPGCLCs and their down-regulation following CRISPRi-LTR5Hs treatment. Of particular interest is *ST6GAL1*, a sialyltransferase<sup>60</sup> that produces CD75, a cell-surface glycoprotein that serves as a marker of naïve hESCs<sup>49</sup>. *ST6GAL1* is likewise regulated by LTR5Hs in naïve hESCs<sup>30</sup>, offering further support to our hypothesis that LTR5Hs TE enhancers act to reinforce elements of the naïve transcriptome during hPGCLC/hPGC maintenance. While the role of both *ST6GAL1* and CD75 remains enigmatic, knockdown of *ST6GAL1* during reprogramming of human dermal fibroblast (HDF) impedes reprogramming and causes a delay in the expression of *NANOG*, *OCT4*, and *SOX2* RNA<sup>61</sup>. Conversely, knockdown of *ST6GAL1* in primed hESC had a modest effect on the transcriptome, causing an up-regulation of genes associated with organogenesis<sup>61</sup>. Recent work by Liu et al.<sup>62</sup> has produced a high-resolution roadmap of the transcriptome during HDF reprogramming, which uncovered an intermediate state immediately prior to a lineage bifurcation between primed and naïve transcriptome acquisition. It is tempting to speculate that *ST6GAL1* may be necessary to efficiently pass through this intermediate state and that during hPGC specification or hPGCLC induction *ST6GAL1* has a similar role as latent pluripotency is re-established following specification or induction, respectively.

Modern and archaic humans began to diverge ~500,000 years ago with modern humans becoming the dominant surviving human species ~50,000 years ago<sup>63,64</sup>. 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 LTR5Hs became beneficial to germ cell specification and consequently improved human reproductive fitness. As we have identified multiple TF networks that converge on LTR5Hs, it is also possible that other factors not profiled in this work contribute to the specification and reinforcement of PGC fate. Likewise, advances in recent techniques to model the early embryo could provide additional platforms to dissect the networks which delineate the naïve state networks in the pre-implantation embryo.

## Methods

**Ethics statement.** The UCLA2 and UCLA1 hESC lines were derived at UCLA by the UCLA Pluripotent Stem Cell Core Facility following Institutional Review Board (IRB) and UCLA Embryonic Stem Cell Research Oversight (ESCRO) Committee Approvals. Informed consent was obtained after the embryo donors contacted the UCLA Broad Stem Cell Research Center to inquire about donating surplus embryos following in vitro fertilization. Embryo donors were not paid and were able to freely withdraw consent to use the embryos for stem cell research up to the point of hESC derivation when the embryo is destroyed. Informed consent was obtained from all embryo donors prior to sending frozen donated embryos to UCLA. Once derived, the hESC lines were authenticated using Affymetrix

Genome-wide Human SNP Array 6.0 to detect Single Nucleotide Polymorphisms and Copy Number Variant (SNP/CNV) prior to distribution. The UCLA1 and UCLA2 hESC lines are provided to researchers de-identified, with all links and identifiers broken prior to distribution. All de-identified 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 using the de-identified hESC lines were approved by the UCLA Embryonic Stem Cell Research Oversight Committee.

**Cell culture.** UCLA2 and UCLA1 hESC are cultured as previously described<sup>16</sup>, 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), 55  $\mu$ M 2-Mercaptoethanol (Life Technologies, 21985-023), 10 ng/mL recombinant human FGF basic (Proteintech, HZ1285), and 50 ng/mL primocin (InvivoGen, ant-pm-2) in DMEM/F12 media (GIBCO, 11330-032). The primed hESCs were split by 1 mg/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). HEK293 cells were acquired from ATCC (Cat# CRL-3216). No lines used in this study belong to the International Cell Line Authentication Committee register of misidentified cell lines.

**hPGCLC differentiation.** Using the UCLA2 hESC line, the differentiation of hPGCLCs in vitro was performed as previously described<sup>4,5</sup>. 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 that had previously been coated with human plasma fibronectin (Life Technologies, 33016-015) for at least 1 hour (h). 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), 55  $\mu$ M 2-Mercaptoethanol (Life Technologies, 21985-023), 1x Penicillin/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 h, 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 3000 cells/well in 200  $\mu$ L/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), 55  $\mu$ M 2-Mercaptoethanol (Life Technologies, 21985-023), 1x 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 (Thermo Fisher Scientific, 25300120) for 10 minutes (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. For SOX17 ChIP-seq, all hPGCLCs were collected and sorted by BD FACSDiva v8.0.2. For NANOG ChIP-seq, 96 aggregates were sampled via FACS, while the remaining aggregates were dissociated in parallel before being fixed and flash frozen (see TF Chromatin Immunoprecipitation). The antibodies used in this study are: BV421 conjugated anti-human/mouse CD49f (ITGA6) (BioLegend; Cat#313624; RRID: AB\_2562244; Lot#B274314) at 1/80; APC-conjugated anti-human CD326 (EPCAM) (BioLegend; Cat#324208; RRID: AB\_756082; Lot#B284158) at 1/80.

**ChIP-seq protocol.** The ChIP-seq was performed as previously described<sup>17</sup>. Isolated hPGCLCs (SOX17) or whole hPGCLC aggregates (NANOG) were fixed using 1% formaldehyde (Thermo Fisher Scientific, Waltham MA) rotating at room temperature for 10 min. Fixation was quenched using 0.14 M Glycine (Sigma Aldrich, St. Louis MO), cells were pelleted by centrifugation at 3000 RPM for 5 min. Resulting cell pellets were flash frozen in liquid nitrogen and stored at -80 °C prior to immunoprecipitation.

Pellets were thawed on ice and resuspended in lysis buffer (10 mM Tris HCl pH 8, 0.25% Triton-X 100, 10 mM EDTA, 0.5 mM EGTA, supplemented with Halt Protease Inhibitor cocktail (Thermo Fisher Scientific, Waltham, MA)) at room temperature, rotating, for 15 min. The resulting lysate was pelleted by 5 min of



centrifugation at 4000 RPM. Pellet was resuspended in Nuclei isolation buffer (10 mM Tris-HCl pH 8.0, 200 mM NaCl, 10 mM EDTA, 0.5 mM EGTA supplemented with Halt Protease Inhibitor Cocktail (Thermo Fisher Scientific, Waltham MA)) at 4 °C, rotating for 10 min followed by 5 min of centrifugation at 4000 RPM. The resulting pellet was resuspended in 10 mM Tris HCl pH 8, 10 mM EDTA, 0.5 mM EGTA with protease inhibitors. Samples were sonicated using a Covaris (Woburn, MA) S220 (Intensity of 5, 200 cycles per burst, 5% duty cycle) in 8 cycles of 30 seconds on, 30 seconds off for an effective 4 min of sonication. Insoluble material was removed by centrifugation at 14,000 RPM for 10 min at 4 °C. In all, 10% of resulting soluble supernatant was saved as an input sample. To pre-clear, Protein A beads (30 µL/ sample) were washed in dilution buffer (16.7 mM Tris-HCl pH 8, 0.01% SDS, 1.1% TritonX-100, 1.2 mM EDTA, 167 mM NaCl) for three times. Protein A beads were resuspended in dilution buffer and added to samples so that 30 µL of Protein A Dynabeads were suspended in a volume of dilution buffer equal to the volume of soluble material. Chromatin was pre-cleared by incubation with Protein A Dynabeads (Thermo Fisher, Waltham MA) for 2 h, rotating at 4 °C. Beads were removed and 1.6 µg of anti-SOX17 antibody (Cat#AF1924, R and D Systems) or 1.2 µg of anti-NANOG antibody (Cat#AF1997, R and D Systems) were added and allowed to incubate rotating at 4 °C overnight. Antibodies bound using 60 µL of Protein G Dynabeads by incubation at 4 °C, rotating for 2 h. Antibody-bound beads were washed 2 × 4 min with 50 mM HEPES pH 7.9, 1% TritonX-100, 0.1% Deoxycholate, 1 mM EDTA, 140 mM NaCl at room temperature, followed by 2 washes with 50 mM HEPES pH 7.9, 0.1% SDS, 1% TritonX-100, 1 mM EDTA, 500 mM NaCl. Beads were subsequently washed twice with 10 mM Tris HCl, pH 8, 1 mM EDTA. Chromatin was eluted from beads by heating 65 °C, rotating at 1400 RPM in 50 mM Tris HCl pH 8, 1 mM EDTA, 1% SDS twice. To facilitate crosslinking reversal, eluate was left to incubate at 65 °C overnight. Eluate was treated with 15 µg RNase A at 37 °C for 30 min followed by treatment with 100 µg of Proteinase K at 56 °C. DNA was purified using Qiagen PCR purification kit according to manufacturer's instructions.

Eluted DNA was used to generate libraries for ChIP-seq using Tecan Genomics Ovation UltraLow V2 DNA-seq (0344NB, Redwood City, CA) according to the manufacturer's instructions. All ChIP-seq libraries were sequenced using a NovaSeq 6000 (Illumina, San Diego) on an NovaSeq SP lane using paired-end 100 base pair reads.

**CRISPR/dCas9-kRAB assay.** Two gRNAs (gRNA55 and gRNA57) that targeted LTR5Hs were designed by Pontis and colleagues<sup>30</sup>. The two gRNAs were cloned into pLV-dCas9-KRAB-T2a-Puro (Addgene 71236), and the plasmid with no LTR5Hs gRNA was used as a control. Using a second-generation lentiviral system we generated dCas9-KRAB-gRNA55, dCas9-KRAB-gRNA57, and dCas9-KRAB-empty virus in HEK293T cells (ATCC, Manassas, Virginia, Cat# CRL-3216). Supernatants that contain lentivirus were then collected and ultracentrifuged. Confluent hESC were trypsinized with 0.05% trypsin at 37 °C for 5 min, 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, 1 µg/mL puromycin was used to screen for positive cells for at least 5 days. Surviving cells were then used to perform downstream assays.

**RNA-sequencing.** RNA-seq was performed as previously described<sup>65</sup>. Briefly the hPGCLCs were directly sorted into 350 µL 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 Illumina NovaSeq 6000 sequencer.

**Statistics and reproducibility.** No statistical method was used to predetermine sample size and no data were excluded from the analyses. For CRISPRi experiments, hESCs from within a given cell line were pooled and randomly allocated to either CRISPRi-virus or control-virus conditions. ChIP-seq experiments were not randomized. Authors were not blinded to allocation during experiments and outcome assessment. All statistics were calculated using GraphPad Prism v9.2.0 (283) or R<sup>66</sup> v3.5.1 unless otherwise mentioned in the figure legend. The Statistical test methods used were provided in Source Data file.

### Bioinformatics analysis

**Reference genome.** Human reference genome GRCh38.97 from Ensembl<sup>67</sup> was used for STAR<sup>36</sup> v2.7.0e and BSMAP<sup>68</sup> v2.74 alignment, while human reference genome hg38 from UCSC<sup>69</sup> was used for SQUIRE<sup>32</sup> v0.9.9.92 for alignment. TE annotation file from repeatmasker (<http://repeatmasker.org/>) GRCh38 and gene annotation file from Ensembl<sup>67</sup> GRCh38.97 was utilized for all genomics analysis.

**TE quantification methods comparison.** Four methods for TE quantification were applied to call DETEs in hPGCLCs compared with hESCs, to identify TE sub-families specific to hPGCLCs more precisely. RNA-seq data of hESC to hPGCLC differentiation was from previous publication<sup>16</sup> GSE93126 (Supplementary Data 1).

Quality control for raw RNA-seq sequences was performed by FastQC<sup>70</sup> v0.11.8. Then the raw reads were aligned by STAR<sup>36</sup> v2.7.0e or SQUIRE<sup>32</sup> v0.9.9.92. For STAR alignment, maximal 1000 multiple mapped reads were allowed, and the best hit was kept (--outFilterMultimapNmax 1000 --outSAMmultNmax 1). SQUIRE Map function with default parameters was applied for alignment. The output bam format files were sorted and indexed by SAMtools<sup>71</sup> v1.9 for downstream analysis. Bigwig tracks were generated using deepTools<sup>72</sup> v3.4.3 by normalizing to RPKM (Reads Per Kilobase per Million mapped reads) using bin size of 10 bp.

Read quantification for individual TE copies were calculated using featureCounts<sup>35</sup> v2.0.0, SQUIRE<sup>32</sup> v0.9.9.92, Telescope<sup>34</sup> v2.0.0, or TETranscripts<sup>33</sup> v2.2.1. FeatureCounts, Telescope and TETranscripts used the sorted bam file from STAR, while SQUIRE used its own sorted bam file. Multiple mapped reads were included for TE quantification (featureCounts -M, TETranscripts --mode multi, SQUIRE, and Telescope using default parameters). Differentially expressed TEs (DETEs) were processed using R package DESeq2<sup>37</sup> v1.26.0 for the count matrices from TE quantification. Only TE with RPKM mean in either control or treatment group >1 were kept for further analysis. DETEs were obtained with at least 4-fold change and FDR < 0.05.

**RNA-seq analysis.** Other than methods used for the TE quantification, "STAR + featureCounts + DESeq2" method was applied for both TE and gene quantification and DETE/DEG calling in the article. Besides GSE93126 RNA-seq data<sup>16</sup> of hESC to hPGCLC differentiation (Supplementary Data 1), other RNA-seq datasets used in this article including RNA-seq of CRISPRi in hPGCLCs generated from this paper, RNA-seq data of hESC multilineage differentiation from previous publication<sup>41</sup> GSE16256 (Supplementary Data 1), and RNA-seq data of CRISPRi in naïve hESCs from previous publication<sup>30</sup> GSE117395 (Supplementary Data 1).

For RNA-seq data quality control, alignment and track generation, FastQC<sup>70</sup> v0.11.8, STAR<sup>36</sup> v2.7.0e and deepTools<sup>72</sup> v3.4.3 were applied as described in "TE quantification methods comparison" section.

Both gene and TE were quantified using FeatureCounts<sup>35</sup> v2.0.0, with "-M" option allowing the quantification for multiple mapped reads. For DETEs and DEGs calling by DESeq2<sup>37</sup> v1.26.0, only TE or gene with RPKM mean in either control or treatment group >1 were kept for further analysis. DETEs were obtained with at least 4-fold change and FDR < 0.05 while DEGs were obtained with at least 1.5-fold change and FDR < 0.05.

To visualize the top 200 TE subfamilies that are most dynamically expressed in hESCs, iMeLCs, and hPGCLCs, top 200 TE subfamilies with the largest variance for the normalized counts across the three cell types were kept. Z-score of the RPM (Reads Per Million mapped reads) for each TE subfamily was used for data visualization.

To analyze the expression level in hESCs and hPGCLCs over HERV associated LTR or solo LTRs, we classified solo LTR5Hs and solo LTR7 as following. The distance between LTR5Hs (or LTR7) individual copy to the nearest HERVK (or HERVH) was first calculated using bedtools<sup>73</sup> v2.29.2 closest function. The distance distribution was then summarized in R<sup>66</sup> v3.5.1. LTR5Hs within 100 bp distance to nearest HERVK were classified as HERVK-LTR5Hs, while others were defined as solo LTR5Hs. LTR7 within 10 bp distance to nearest HERVH were classified as HERVH-LTR7, while others were defined as solo LTR7.

**scRNA-seq analysis.** Two biological replicates for the scRNA-seq data of hESC to hPGCLC differentiation (UCLA2 line) was downloaded from previous publication<sup>2</sup> GSE140021 (Supplementary Data 1). The reads were quantified by 10x Genomics Cell Ranger<sup>74</sup> v3.1.0 to both gene and TE reference genome with default parameters. The generated cell-by-gene/TE unique molecular identifier (UMI) count matrix was analyzed in Seurat<sup>75</sup> R package v3.2.2. Due to limited coverage in scRNA-seq data, we aggregated reads from individual TE copies to TE subfamilies for downstream analysis.

Cells expressing 1000–7000 gene features and <20% mitochondrial genes were kept. The UMI counts were then normalized and log-transformed followed with identifying top 2000 variable features and scaling for both gene and TE UMI count matrix with default parameter. For batch correction between two replicates, we used Seurat's IntegrateData function with default parameter, which were used further for clustering and UMAP visualization. The scaled integrated data with variable genes was used to perform principal component analysis (PCA). UMAPs were calculated by RunUMAP function using top 50 principal components and resolution 1.

Raw data for scRNA-seq of two Carnegie Stage 7 human gastrula embryos was kindly shared by the authors from previous publication<sup>40</sup> (Supplementary Data 1). Single-cell RNA-seq data of seven PGCs as well as other randomly selected cells from annotated cell types (epiblast, primitive streak, emergent mesoderm and advanced mesoderm, annotated by Tyser et al.), were re-analyzed for gene and TE expression same as "STAR + FeatureCounts" RNA-seq analysis method. In brief, FastQC<sup>70</sup> v0.11.8 was used for quality control, STAR<sup>36</sup> v2.7.0e was used for alignment, both gene and TE were quantified by FeatureCounts<sup>35</sup> v2.0.0, with "-M" allowing multiple mapping for TEs.



**ATAC-seq analysis.** Raw ATAC-seq data from previous publication<sup>16</sup> GSE120648 (Supplementary Data 1) were downloaded followed by quality control with FastQC<sup>70</sup> v0.11.8. Then raw reads were aligned by STAR<sup>36</sup> v2.7.0e allowing maximal 1000 multiple mapped reads with no more than three mismatches and the best hit was kept (--outFilterMultimapNmax 1000 --outFilterMismatchNmax 3 --outSAMmultNmax 1). Splice junction was neglected by building STAR index without general feature format file and not allowing intron length (--alignIntronMax 1). PCR duplicates were removed using SAMtools<sup>71</sup> v1.9 rmdup function. SAMtools<sup>71</sup> v1.9 merge function was used to merge aligned reads in bam format for replicates in each cell type for downstream analysis to increase coverage.

ATAC-seq peaks were defined using the MACS2<sup>76</sup> v2.2.7.1 callpeaks function. Here we only kept peaks with a fold change enrichment >4 from the MACS2 output. In order to identify hPGCLC- or hESC-ORs, we used bedtools<sup>73</sup> v2.29.2 multiinter function with Ryan Layers's clustering, and the regions <100 bp were discarded. Bigwig tracks were generated using deeptools<sup>72</sup> v3.4.3 by normalizing to RPKM using binsize of 10 bp. ATAC-seq signal over hPGCLC- or hESC-ORs or TE regions were visualized using deeptools<sup>72</sup> v3.4.3.

To quantify ATAC-seq signals over LTR5Hs as well as random shuffled regions, we first generated 100,000 random shuffled TE and genomic regions. A hundred thousand ( $n = 100,000$ ) TE random regions were randomly selected 100,000 TE individual copies from all TE copies in human reference genome. A hundred thousand ( $n = 100,000$ ) genome random regions were randomly shuffled genomic regions with the same length as 100,000 TE individual copy regions generated by bedtools<sup>73</sup> v2.29.2 shuffle function. The ATAC-seq read counts over LTR5Hs, 100,000 random shuffled TE and genomic regions were calculated with bedtools<sup>73</sup> v2.29.2 multicov function. Then, read counts were normalized to total reads aligned in each sample using RPM and visualized in R<sup>66</sup> v3.5.1.

**TE enrichment analysis over hPGCLC- or hESC-ORs.** TE annotation for hPGCLC- or hESC-ORs was conducted by Homer<sup>77</sup> v4.7 annotatePeaks.pl function using GRCh38 TE annotation file.

To generated randomly shuffled regions with comparable genomic distribution to TEs, random shuffled regions for ATAC-seq hPGCLC- or hESC-ORs were adjusted by the relative proportion of genomic regions (promoter, exon, intron, TTS, 10 kb gene proximal region, 10–100 kb distal region or >100 kb intergenic region), according to Chuong et al.<sup>43</sup>. To be specific, the midpoints of hPGCLC/hESC-ORs were annotated to genomic regions (promoter, exon, intron, TTS, intergenic region) by Homer<sup>77</sup> v4.7 annotatePeaks.pl function. Then intergenic region was further divided into 10 kb gene proximal region, 10–100 kb distal region or >100 kb intergenic region by their distance to the nearest gene. Then, the entire human genome was divided into promoter, exon, intron, TTS, intergenic region (10 kb gene proximal region, 10–100 kb distal region or >100 kb intergenic region). The annotated midpoints of hPGCLC/hESC-ORs in each kind of genomic region were shuffled 10,000 times within the corresponding genomic region with bedtools<sup>73</sup> v2.29.2 shuffle function (-seed 1 to 10,000) by keeping the shuffled regions on the same chromosome (-chrom). Then, the shuffled regions with same seed number were merged to create shuffled hPGCLC/hESC-ORs maintaining the same genomic distribution as the original hPGCLC/hESC-ORs.

The expected TE occurrence was calculated by the average number of TE copies which were intersected with the 10,000 combined random shuffled hPGCLC/hESC-ORs. If the expected TE copy occurrence for certain TE subfamily was smaller than 1, it was rounded to 1. The observed TE occurrence was counted based on the TE annotation for hPGCLC/hESC-ORs. The value of  $\text{Log}_2$  transformed "observed TE occurrence/expected TE occurrence" was used as enrichment score, with one-sided exact binomial test for statistical test.

**Transcription factor motif enrichment analysis.** Motif file for over 400 transcriptional factors were collected from Homer<sup>77</sup> v4.7 and the position of each motif in GRCh38 genome were calculated using Homer scanMotifGenome.pl function. Next, hPGCLC-ORs overlapped with LTR5Hs were identified by bedtools<sup>73</sup> v2.29.2 intersect function. As control, those LTR5Hs overlapped hPGCLC-ORs were randomly distributed using bedtools<sup>73</sup> v2.29.2 shuffle function while keeping on the same chromosome (-chrom).

To analyze the enrichment of TF motifs over chromatin opened LTR5Hs, the frequency of occurrences for TF motifs in hPGCLC-ORs overlapped LTR5Hs and shuffled control were processed using bedtools<sup>73</sup> v2.29.2 intersect function. Top 50 TF motifs with highest enrichment ratios were plotted.

**WGBS analysis.** Raw WGBS data were downloaded from previous publication<sup>44</sup> GSE139115 (Supplementary Data 1). Reads were aligned with BSMAP<sup>68</sup> v2.74 by mapping reads to all four strands (-n 1), allowing maximum one equal best hits and less than two mismatches per read (-w 1, -v 2). Aligned reads in bam format for biological replicates of hESC and hPGCLCs were merged to increase the coverage using SAMtools<sup>71</sup> v1.9 merge function. Methratio.py script built in BSMAP were used to calculate cytosine counts only keeping unique mappings (-u), non-duplicated reads (-r), and reporting loci with zero methylation ratios (-z). Methylation level at CG sites was then calculated by  $\#C/(\#C + \#T)$ .

To visualize the CG methylation level and cytosine coverage over LTR5Hs and SVAD, only CG sites with  $\geq 3$  covered reads were retained. Wiggle tracks were

generated with customized perl script and converted to bigwig with wigToBigWig<sup>78</sup> v4 followed by data visualization with deeptools<sup>72</sup> v3.4.3.

To analyze CG methylation level over LTR5Hs and its related TE clades obtained from the Dfam<sup>45</sup> database that shared the most sequence similarity, #C and #C + #T count over each individual TE copy were extracted with customized python script and plotted in R.

DMR were defined using R package DMRcaller<sup>79</sup> v1.14.2 over GRCh38 whole genome using 200 bp as DMR bin size. Only bins with at least four CG sites and each CG sites should be covered by at least three reads were kept for further analysis. Minimal CG methylation difference of 0.2 and FDR less than 0.05 were applied to define DMRs. Bins defined as DMR and within 100 bp gap were merged.

**ChIP-seq analysis.** NANOG and SOX17 ChIP-seq data in hESCs and hPGCLCs were generated in this paper. TFAP2C ChIP-seq in hESCs and hPGCLCs, H3K27ac ChIP-seq in hPGCLCs were from previous publication<sup>2</sup> GSE140021 (Supplementary Data 1). H3K27ac ChIP-seq in hESCs was from previous publication<sup>48</sup> GSE69646 (Supplementary Data 1). SOX15 CUT&Tag-seq was from previous publication<sup>7</sup> GSE143345 (Supplementary Data 1).

For all ChIP-seq data, quality control was performed by FastQC<sup>70</sup> v0.11.8. Then reads were aligned by STAR<sup>36</sup> v2.7.0e allowing maximal 1000 multiple mapped reads with no more than three mismatches and the best hit was kept (--outFilterMultimapNmax 1000 --outFilterMismatchNmax 3 --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<sup>71</sup> v1.9 rmdup function.

Representative replicate for each condition was used for downstream analysis. ChIP-seq peaks were defined using the MACS2<sup>76</sup> v2.2.7.1 callpeaks function by setting ChIP file as treatment and input file as control. Bigwig tracks were generated using deeptools<sup>72</sup> v3.4.3 by normalizing to RPKM using binsize of 10 bp. ChIP-seq signal over TE regions were visualized using deeptools<sup>72</sup> v3.4.3. Motif annotation over ChIP-seq peak summits used Homer<sup>77</sup> v4.7 findMotifsGenome.pl function with fragment size 200 and masking repeats (-size 200 -mask). TF-bound LTR5Hs/LTR7 copies were identified by bedtools<sup>73</sup> v2.29.2 intersect function.

**CRISPRi gRNA target sites prediction.** To search the predicted sites of LTR5Hs targeting gRNA, we used the Homer<sup>77</sup> v4.7 to generate motif file for gRNA plus PAM NGG sequence (CTCCCTAATCTCAAGTACCNGG, TGTTTCAGAGAGACGGGGTNGG) using seq2profile.pl and searched the targeting sites using scanMotifGenomeWide.pl with <3 mismatches. The target sites were annotated using gene and TE annotation and then categorized into either promoter, exonic, TE, intronic, or intergenic sites. If one target site was annotated with multiple categories, only one category would be retained with priority order of promoter, exon, TE, intron, and intergenic sites.

**RAD analysis.** For RAD analysis, we used the website application from Guo et al.<sup>51</sup>. For RAD analysis of LTR5Hs associated DEGs, up- and down-regulated DEGs in CRISPRi-LTR5Hs were input as DEGs lists; LTR5Hs bed file or randomly shuffled LTR5Hs bed file by bedtools<sup>73</sup> v2.29.2 shuffle function were input as Genomic Regions of Interest (gROI) file. For RAD analysis of CRISPRi gRNAs predicted sites associated DEGs, up- and down-regulated DEGs in CRISPRi-LTR5Hs were input as DEGs lists; bed file of CRISPRi gRNAs predicted sites was input as gROI file. For submit options, "GRCh38" was chose for reference genome, "1000, 800, 600, 400, 200, and 0 kb" was input as customized peak extend distance correspondingly, "hypergeometric test" was choosing for statistical test.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

All high-throughput sequencing data generated are accessible at NCBI's Gene Expression Omnibus (GEO) via GEO Series accession number GSE182218. Source data are provided with this paper.

## Code availability

Customized code/scripts used in this study are available from the corresponding author upon request.

Received: 19 September 2020; Accepted: 16 December 2021;

Published online: 24 January 2022

## References

- Hancock, G. V., Wamaitha, S. E., Peretz, L. & Clark, A. T. Mammalian primordial germ cell specification. *Development* **148**, dev189217 (2021).

2. Chen, D. et al. Human primordial germ cells are specified from lineage-primed progenitors. *Cell Rep.* **29**, 4568–4582.e5 (2019).
3. Irie, N. et al. SOX17 is a critical specifier of human primordial germ cell fate. *Cell* **160**, 253–268 (2015).
4. Sasaki, K. et al. Robust in vitro induction of human germ cell fate from pluripotent stem cells. *Cell Stem Cell* **17**, 178–194 (2015).
5. Chen, D. et al. Germline competency of human embryonic stem cells depends on eomesodermin. *Biol. Reprod.* **97**, 850–861 (2017).
6. Kojima, Y. et al. Evolutionarily distinctive transcriptional and signaling programs drive human germ cell lineage specification from pluripotent stem cells. *Cell Stem Cell* **21**, 517–532.e5 (2017).
7. Wang, X. et al. The chromatin accessibility landscape reveals distinct transcriptional regulation in the induction of human primordial germ cell-like cells from pluripotent stem cells. *Stem Cell Rep.* **16**, 1245–1261 (2021).
8. Yamaguchi, S. et al. Conditional knockdown of *Nanog* induces apoptotic cell death in mouse migrating primordial germ cells. *Development* **136**, 4011–4020 (2009).
9. Guo, F. et al. The transcriptome and DNA methylome landscapes of human primordial germ cells. *Cell* **161**, 1437–1452 (2015).
10. Ohinata, Y. et al. Blimp1 is a critical determinant of the germ cell lineage in mice. *Nature* **436**, 207–213 (2005).
11. Weber, S. et al. Critical function of AP-2gamma/TCFAP2C in mouse embryonic germ cell maintenance. *Biol. Reprod.* **82**, 214–223 (2010).
12. Yamaji, M. et al. Critical function of Prdm14 for the establishment of the germ cell lineage in mice. *Nat. Genet.* **40**, 1016–1022 (2008).
13. Sybirna, A. et al. A critical role of PRDM14 in human primordial germ cell fate revealed by inducible degrons. *Nat. Commun.* **11**, 1282 (2020).
14. Hara, K. et al. Evidence for crucial role of hindgut expansion in directing proper migration of primordial germ cells in mouse early embryogenesis. *Dev. Biol.* **330**, 427–439 (2009).
15. Kanai-Azuma, M. et al. Depletion of definitive gut endoderm in Sox17-null mutant mice. *Development* **129**, 2367–2379 (2002).
16. Chen, D. et al. The TFAP2C-regulated OCT4 naive enhancer is involved in human germline formation. *Cell Rep.* **25**, 3591–3602.e5 (2018).
17. Pastor, W. A. et al. TFAP2C regulates transcription in human naive pluripotency by opening enhancers. *Nat. Cell Biol.* **20**, 553–564 (2018).
18. Friedli, M. & Trono, D. The Developmental Control of Transposable Elements and the Evolution of Higher Species. *Annu. Rev. Cell Dev. Biol.* **31**, 429–451 (2015).
19. Turner, G. et al. Insertional polymorphisms of full-length endogenous retroviruses in humans. *Curr. Biol.* **11**, 1531–1535 (2001).
20. Reus, K. et al. HERV-K(OLD): ancestor sequences of the human endogenous retrovirus family HERV-K(HML-2). *J. Virol.* **75**, 8917–8926 (2001).
21. Barbulescu, M. et al. Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. *Curr. Biol.* **9**, 861–S1 (1999).
22. Jha, A. R. et al. Human endogenous retrovirus K106 (HERV-K106) was infectious after the emergence of anatomically modern humans. *PLoS ONE* **6**, e20234–e20234 (2011).
23. Medstrand, P. & Mager, D. L. Human-specific integrations of the HERV-K endogenous retrovirus family. *J. Virol.* **72**, 9782–9787 (1998).
24. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).
25. Kurnarso, G. et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42**, 631–634 (2010).
26. Wang, T. et al. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl Acad. Sci. USA* **104**, 18613–18618 (2007).
27. Subramanian, R. P., Wildschutte, J. H., Russo, C. & Coffin, J. M. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology* **8**, 90 (2011).
28. Fuchs, N. V. et al. Human endogenous retrovirus K (HML-2) RNA and protein expression is a marker for human embryonic and induced pluripotent stem cells. *Retrovirology* **10**, 115–115 (2013).
29. Grow, E. J. et al. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**, 221–225 (2015).
30. Pontis, J. et al. Hominoid-specific transposable elements and KZFPs facilitate human embryonic genome activation and control transcription in naive human ESCs. *Cell Stem Cell* **24**, 724–735.e5 (2019).
31. Teissandier, A., Servant, N., Barillot, E. & Bourc'his, D. Tools and best practices for retrotransposon analysis using high-throughput sequencing data. *Mobile DNA* **10**, 52 (2019).
32. Yang, W. R., Ardeltan, D., Pacyna, C. N., Payer, L. M. & Burns, K. H. SQUIRE reveals locus-specific regulation of interspersed repeat expression. *Nucleic Acids Res.* **47**, e27–e27 (2019).
33. Jin, Y., Tam, O. H., Paniagua, E. & Hammell, M. Tetrascripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593–3599 (2015).
34. Bendall, M. L. et al. Telescope: characterization of the retrotranscriptome by accurate estimation of transposable element expression. *PLoS Comput. Biol.* **15**, e1006453 (2019).
35. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
36. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
37. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
38. Kelley, D. & Rinn, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.* **13**, R107 (2012).
39. Thompson, P. J., Macfarlan, T. S. & Lorincz, M. C. Long terminal repeats: from parasitic elements to building blocks of the transcriptional regulatory repertoire. *Mol. Cell* **62**, 766–776 (2016).
40. Tyser, R. C. V. et al. Single-cell transcriptomic characterization of a gastrulating human embryo. *Nature* **600**, 285–289 (2021).
41. Xie, W. et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).
42. Fuentes, D. R., Swigut, T. & Wysocka, J. Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. *eLife* **7**, e35989 (2018).
43. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).
44. Gell, J. J. et al. An extended culture system that supports human primordial germ cell-like cell survival and initiation of DNA methylation erasure. *Stem Cell Rep.* **14**, 433–446 (2020).
45. Hubley, R. et al. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–D89 (2016).
46. Smela, M. P., Sybirna, A., Wong, F. C. K. & Azim Surani, M. Testing the role of sox15 in human primordial germ cell fate [version 2; peer review: 2 approved]. *Wellcome Open Res.* **4**, 122 (2019).
47. Chambers, I. et al. Nanog safeguards pluripotency and mediates germline development. *Nature* **450**, 1230–1234 (2007).
48. Ji, X. et al. 3D chromosome regulatory landscape of human pluripotent cells. *Cell Stem Cell* **18**, 262–275 (2016).
49. Collier, A. J. et al. Comprehensive cell surface protein profiling identifies specific markers of human naive and primed pluripotent states. *Cell Stem Cell* **20**, 874–890.e7 (2017).
50. Thakore, P. I. et al. Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nat. Methods* **12**, 1143–1149 (2015).
51. Guo, Y. et al. RAD: a web application to identify region associated differentially expressed genes. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btab075> (2021).
52. Hancock, G. V. et al. Divergent roles for KLF4 and TFCP2L1 in naive ground state pluripotency and human primordial germ cell development. *Stem Cell Res.* **55**, 102493 (2021).
53. Gkoutela, S. et al. DNA demethylation dynamics in the human prenatal germline. *Cell* **161**, 1425–1436 (2015).
54. Tang, W. W. C. et al. A unique gene regulatory network resets the human germline epigenome for development. *Cell* **161**, 1453–1467 (2015).
55. Kim, T.-K. et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
56. De Santa, F. et al. A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. *PLoS Biol* **8**, e1000384 (2010).
57. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
58. Theunissen, T. W. et al. Molecular criteria for defining the naive human pluripotent state. *Cell Stem Cell* **19**, 502–515 (2016).
59. Zhang, Y. et al. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat. Genet.* **51**, 1380–1388 (2019).
60. Hedlund, M., Ng, E., Varki, A. & Varki, N. M.  $\alpha$ -2-6-linked sialic acids on N-glycans modulate carcinoma differentiation in vivo. *Cancer Res.* **68**, 388–394 (2008).
61. Wang, Y.-C. et al. Glycosyltransferase ST6GAL1 contributes to the regulation of pluripotency in human pluripotent stem cells. *Sci. Rep.* **5**, 13317 (2015).
62. Liu, X. et al. Reprogramming roadmap reveals route to human induced trophoblast stem cells. *Nature* **586**, 101–107 (2020).
63. Sankararaman, S., Mallick, S., Patterson, N. & Reich, D. The combined landscape of denisovan and neanderthal ancestry in present-day humans. *Curr. Biol.* **26**, 1241–1247 (2016).
64. Wolf, A. B. & Akey, J. M. Outstanding questions in the study of archaic hominin admixture. *PLoS Genet.* **14**, e1007349 (2018).
65. Tao, Y. et al. TRIM28-regulated transposon repression is required for human germline competency and not primed or naive human pluripotency. *Stem Cell Rep.* **10**, 243–256 (2018).

66. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2019).
67. Howe, K. L. et al. Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
68. Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPPING program. *BMC Bioinform.* **10**, 232 (2009).
69. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
70. Andrews, S. A quality control tool for high throughput sequence data. *BibSonomy* (2010).
71. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
72. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. DeepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
73. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
74. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017).
75. Stuart, T. et al. Comprehensive Integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
76. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
77. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
78. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).
79. Catoni, M., Tsang, J. M., Greco, A. P. & Zabet, N. R. DMRcaller: a versatile R/Bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts. *Nucleic Acids Res.* **46**, e114 (2018).

## Acknowledgements

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, culturing, and distributing the UCLA hESC lines from the BSCRC Pluripotent Stem Cell Core Facility, Suhua Feng in assistance with high throughput sequencing. We gratefully acknowledge Shankar Srinivas (University of Oxford, Oxford) and Antonio Scialdone (Helmholtz Zentrum München – German Research Center for Environmental Health), for sharing the single-cell RNA-Seq data of a CS7 human embryo. This study was supported by the NIH/NICHD R01HD079546 (to A.C.), the Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research

at UCLA Research Award Program (to A.C.), the National Natural Science Foundation of China 32170551 (to W.L.), the Zhejiang Provincial Natural Science Foundation of China LQ20C060004 (to W.L.), the Fundamental Research Funds for the Central Universities 2021QN81016 (to W.L.), and Alibaba Cloud (to W.L.).

## Author contributions

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

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-28105-1>.

**Correspondence** and requests for materials should be addressed to Wanlu Liu or Amander T. Clark.

**Peer review information** *Nature Communications* thanks Guillaume Bourque and the other anonymous reviewers for their contribution to the peer review of this work.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022