PLOS GENETICS

RESEARCH ARTICLE

A hominoid-specific endogenous retrovirus may have rewired the gene regulatory network shared between primordial germ cells and naïve pluripotent cells

Jumpei Ito[1], **Yasunari Seita**[2,3], **Shohei Kojima**[4], **Nicholas F. Parrish**[4], **Kotaro Sasaki**[2,5☯*], **Kei Sato**[1,6,7,8,9☯*]

1 Division of Systems Virology, Department of Microbiology and Immunology, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan, **2** Department of Biomedical Sciences, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **3** Bell Research Center for Reproductive Health and Cancer, Nagoya, Aichi, Japan, **4** Genome Immunobiology RIKEN Hakubi Research Team, RIKEN Center for Integrative Medical Sciences and RIKEN Cluster for Pioneering Research, Yokohama, Japan, **5** Institute for Regenerative Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **6** Graduate School of Medicine, The University of Tokyo, Tokyo, Japan, **7** International Research Center for Infectious Diseases, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan, **8** International Vaccine Design Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan, **9** CREST, Japan Science and Technology Agency, Saitama, Japan

☯ These authors contributed equally to this work.
* keisato@g.ecc.u-tokyo.ac.jp (KS); ksasaki@vet.upenn.edu (KS)

OPEN ACCESS

Citation: Ito J, Seita Y, Kojima S, Parrish NF, Sasaki K, Sato K (2022) A hominoid-specific endogenous retrovirus may have rewired the gene regulatory network shared between primordial germ cells and naïve pluripotent cells. PLoS Genet 18(5): e1009846. https://doi.org/10.1371/journal.pgen.1009846

Editor: Cédric Feschotte, Cornell University, UNITED STATES

Received: September 30, 2021

Accepted: April 8, 2022

Published: May 12, 2022

Data Availability Statement: The authors confirm that all data underlying the findings are fully available without restriction. The RNA-seq data reported in this paper are available in GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE167570). The data produced in this study are available from the Mendeley Data Repository (http://dx.doi.org/10.17632/w5gfs9mdrr.1). The gene and TE annotation file is described in detail in and is available from the GitHub repository (https://github.com/TheSatoLab/TE_scRNA-Seq_analysis_

Abstract

Mammalian germ cells stem from primordial germ cells (PGCs). Although the gene regulatory network controlling the development of germ cells such as PGCs is critical for ensuring gamete integrity, substantial differences exist in this network among mammalian species, suggesting that this network has been modified during mammalian evolution. Here, we show that a hominoid-specific group of endogenous retroviruses, LTR5_Hs, discloses enhancer-like signatures in human *in vitro*-induced PGCs, PGC-like cells (PGCLCs). Human PGCLCs exhibit a transcriptome signature similar to that of naïve-state pluripotent cells. LTR5_Hs are epigenetically activated in both PGCLCs and naïve pluripotent cells, and the expression of genes in the vicinity of LTR5_Hs is coordinately upregulated in these cell types, contributing to the establishment of the transcriptome similarity between these cell types. LTR5_Hs are preferentially bound by transcription factors that are highly expressed in both PGCLCs and naïve pluripotent cells (*KLF4*, *TFAP2C*, *NANOG*, and *CBFA2T2*), suggesting that these transcription factors contribute to the epigenetic activation of LTR5_Hs in these cells. Comparative transcriptome analysis between humans and macaques suggests that the expression of many genes in PGCLCs and naïve pluripotent cells is upregulated by LTR5_Hs insertions in the hominoid lineage. Together, this study suggests that LTR5_Hs insertions may have finetuned the gene regulatory network shared between PGCLCs and naïve pluripotent cells and coordinately altered the gene expression in these cells during hominoid evolution.

Hwang_et_al/blob/master/CellRanger/input/hg38_TE_noAlt_unique.gtf.gz).

Funding: This study was supported in part by JSPS KAKENHI Grant-in-Aid for Early-Career Scientists JP20K15767 (to J. I); JSPS Research Fellow PD JP19J01713 (to J. I); AMED Research Program on Emerging and Re-emerging Infectious Diseases 20fk0108146 (to K. S), 19fk0108171 (to K. S), 20fk0108270 (to K. S) and 20fk0108413 (to K. S); AMED Research Program on HIV/AIDS 19fk0410019 (to K. S) and 20fk0410014 (to K. S); JST CREST (to K. S); JST J-RAPID JPMJJR2007 (to K. S); JST SICORP (e-ASIA) JPMJSC20U1 (to K. S); JSPS KAKENHI Grant-in-Aid for Scientific Research B 18H02662 (to K. S), JSPS KAKENHI Grant-in-Aid for Scientific Research on Innovative Areas 16H06429 (to K. S), 16K21723 (to K. S), 17H05823 (to K. S), 17H05813 (to K. S), and 19H04826 (to K. S); ONO Medical Research Foundation (to K. S); Ichiro Kanehara Foundation (to K. S); Mochida Memorial Foundation for Medical and Pharmaceutical Research (to K. S); Daiichi Sankyo Foundation of Life Science (to K. S); Sumitomo Foundation (to K. S); Uehara Foundation (to K. S); Takeda Science Foundation (to K. S); The Tokyo Biochemical Research Foundation (to K. S); International Joint Research Project of the Institute of Medical Science, the University of Tokyo 2020-K3003 (to K. Sas and K. S); Open Philanthropy fund from Silicon Valley Community Foundation 2019-197906 (to K. Sas) and a grant from the Pennsylvania Department of Health (to K. Sas). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

To ensure the health of the next generation and the continuation of a species, the development of germ cells, including primordial germ cells (PGCs), is strictly controlled by a complex gene regulatory network. Nevertheless, the gene regulatory network controlling the germ cell development has been substantially diversified during mammalian or even primate evolution. Here, our integrated analyses using multiomics and comparative genomics resources suggest that hominoid-specific insertions of endogenous retroviruses are epigenetically activated in both *in vitro*-induced PGCs and naïve pluripotent cells and may have coordinately altered the expression of the adjacent genes in these cells. This study provides evidence suggesting that the gene regulatory network shared between PGCs and naïve pluripotent cells may have been rewired by ERV insertions during hominoid evolution.

Introduction

Mammalian germ cells are first established as primordial germ cells (PGCs) from pluripotent cells, such as epiblasts, in postimplantation embryos [1–3]. Aberrations in germ cells lead to immediate infertility, genetic or epigenetic disorders in offspring, and genome integrity impairment. Therefore, the differentiation of germ cells, including PGCs, is strictly controlled by a complex gene regulatory network [1–3].

There is increasing interest in investigating the gene regulatory network in human germ cells. However, it is ethically difficult to routinely access human germ cells, particularly those from humans at early stages of development. Previous studies have established methodologies to artificially induce human germ cells such as PGCs from human induced pluripotent stem cells (iPSCs) [4,5]. *In vitro*-derived PGCs, referred to as PGC-like cells (PGCLCs), are considered to represent the premigratory stage of PGCs which present until 3 weeks postfertilization in humans [6]. This stage of human PGCs a are not accessible due to ethical and legal constraints and is difficult to investigate, while more later stages of human PGCs (e.g., migrating PGCs) are relatively accessible and have been investigated in previous studies [7–9]. Furthermore, recent studies have established methodologies to induce more differentiated stages of germ cells such as prospermatogonia cells from iPSCs [10,11]. These methods have enabled us to characterize the mechanisms of human germ cell development in detail. For example, previous studies using these methods have identified the critical transcription factors of human PGCLCs, such as *PRDM1*, *SOX17*, *TFAP2C*, and *TFAP2A* [4,5,9,12].

The gene regulatory network controlling the development of germ cells such as PGCs is critical for gamete integrity. However, substantial differences exist in this network among mammalian species. For example, various transcription factors (TFs) are differentially expressed between humans and mice [7]. In particular, *SOX17* is a critical transcription factor of PGCLC fate specification in humans but not in mice [4,5,12]. Additionally, a substantial number of genes are differentially expressed between human PGCLCs and PGCs of the crab-eating macaque (*Macaca fascicularis*), Old World monkey (OWM), although the expression patterns of the critical transcription factors of PGCs are conserved between the two species [6]. These observations suggest that the gene regulatory network controlling germ cell development has been finetuned during mammalian or even primate evolution.

Diversification of the gene regulatory networks is one of the molecular bases of evolution and is driven by the turnover of regulatory sequences such as enhancers [13,14]. A substantial

proportion of transposable elements (TEs) work as enhancers and play critical roles in gene regulatory networks and their evolution [15]. Endogenous retroviruses (ERVs) are a class of TEs originating from past retroviral infections. ERVs are particularly rich sources for creation of new enhancers since they contain many regulatory elements in their long terminal repeat (LTR) sequences, which originally functioned as viral promoters [16–18]. Notably, since ERV loci belonging to the same ERV group share the same set of regulatory elements, numerous inserted ERV loci can coordinately alter the expression patterns of multiple genes [18–20]. Furthermore, ERVs tend to possess regulatory elements that are activated in germline or early embryonic niches to proliferate in the germline genome [18,21]. For example, LTR5_Hs, the youngest human ERV subfamily expanded in the hominoid lineage (including humans, chimpanzees, gorillas, orangutans, and gibbons, but not OWMs), are transcriptionally and epigenetically activated in early embryonic cells such as the inner cell masses (ICM) of blastocysts [22–25]. Furthermore, previous studies demonstrated the enhancer activity of LTR5_Hs using the epigenetic perturbation by CRISPR activation (CRISPRa) and inhibition (CRISPRi) systems in embryonic carcinoma cells and embryonic stem cells (ESCs) [22,25]. Therefore, it is possible that ERVs are involved in the evolution of the gene regulatory network in germ or early embryonic cells [22,25–27].

Human PGCs and PGCLCs exhibit complex and mixed transcriptome signatures since various gene expression programs are initiated at this stage [8,12]. In particular, human PGCs and PGCLCs highly express genes associated with naïve pluripotency [7,9,28,29]. Pluripotency is classified into naïve and primed states, which represent the ground and more-differentiated states, respectively [30–32]. For example, ICM of blastocysts or preimplantation epiblasts show naïve pluripotency, while postimplantation epiblasts show primed pluripotency [30–32]. Several key TFs, including naïve pluripotency factors (e.g., *NANOG*, *KLF4*, and *TFCP2L1*) and some critical transcription factors of PGCLCs (e.g., *TFAP2C* and *PRDM1*), are commonly upregulated in human PGCLCs and naïve pluripotent cells [4,5,7,9,12,33–35]. These observations suggest that the core gene regulatory network, which is driven by the key TFs above, might be shared between PGCLCs and naïve pluripotent cells and play essential roles in establishing cellular identities in these cells. Indeed, a previous study showed that a substantial proportion of open chromatin regions are shared between human PGCLCs and naïve ESCs, and the shared open chromatin regions are frequently bound by TFAP2C [29]. Thus, the regulatory network shared between PGCLCs and naïve pluripotent cells has been recognized and investigated. However, the downstream genes regulated by this network and the functions of these genes have not been fully explored. Furthermore, the evolutionary origins of the *cis*-regulatory elements comprising this network and the evolution of this network have not been elucidated.

In the present study, we investigated the gene regulatory network shared between human PGCLCs and naïve pluripotent cells in detail. In this process, we found that several hundred loci of LTR5_Hs are epigenetically activated in PGCLCs in addition to naïve pluripotent cells [22–25]. This study provides evidence suggesting that LTR5_Hs insertions may have rewired the gene regulatory network shared between PGCLCs and naïve pluripotent cells during hominoid evolution and possibly accelerated germ cell evolution.

Results

Similarity of the gene expression signature of PGCLCs with that of naïve pluripotent cells

To characterize the similarity between PGCLCs and naïve pluripotent cells at the transcriptome level, we compared the transcriptomic signatures of PGCLCs and naïve ESCs. We

analyzed single-cell RNA sequencing (scRNA-Seq) datasets for *in vitro*-derived human male germ cells [Hwang et al. [10]] and for naïve and primed ESCs [Messmer et al. [36]] (**Fig 1A**). The Hwang et al. dataset contains information on germ cells that were sequentially differentiated from primed iPSCs: incipient mesoderm-like cells (iMeLCs), PGCLCs, multiplying prospermatogonia-like cells (MLCs), and mitotically quiescent T1 prospermatogonia-like cells (T1LCs), which are formed via transitional cells (TCs) (**Fig 1A**) [10]. Dimension reduction analysis suggested that the global transcriptome is highly similar between PGCLCs and naïve ESCs, consistent with previous reports (**Fig 1A**) [9,29].

To further assess the transcriptional similarity between PGCLCs and naïve ESCs, we first focused on the genes upregulated in both cell types. Accordingly, we assigned a PGCLC-specific expression score for each gene, which represents the similarity of the observed expression pattern to the defined "PGCLC-specific" expression pattern (**Fig 1B**; see **Definition of the PGCLC-specific expression score**). We confirmed that genes highly expressed in *in vivo* PGCs compared to the later stage of male germ cells tended to show a higher PGCLC-specific expression score (**S1 Fig**). According to this PGCLC-specific expression score and the log2-transformed fold change (log2 FC) of the expression score between naïve and primed ESCs, we classified the protein-coding genes into four categories: genes upregulated in both cell types, genes upregulated only in PGCLCs, genes upregulated only in naïve ESCs, and other genes (**Fig 1C and S1 Table**). As expected, the genes upregulated in PGCLCs substantially overlapped with those upregulated in naïve ESCs, supporting increased transcriptional similarity between these cell types (**Fig 1D**). Gene Ontology (GO) enrichment analysis showed that the three of the gene categories were enriched with distinct functional gene sets (**Fig 1E and S2 Table**). Notably, genes related to the "metabolism of carbohydrates" term were enriched among the genes upregulated in both PGCLCs and naïve ESCs (**Fig 1E**), suggesting that the mode of carbohydrate metabolism is similar between these cell types in humans, similar to the observation in mice [37,38].

To identify the candidates of TFs responsible for the transcriptional similarity between PGCLCs and naïve ESCs, we classified TFs according to their expression patterns (**Fig 1C and 1F**). Of the key transcription factors of PGCLCs (*TFAP2C*, *SOX17*, and *PRDM1*) [4,5,12], *TFAP2C* were upregulated in both PGCLCs and naïve ESCs, while *SOX17* was upregulated only in PGCLCs (**Figs 1C, 1F and S2A**). Although *PRDM1* was upregulated in naïve ESCs in addition to PGCLCs, the expression level of *PRDM1* in naïve ESCs was not so high (**S2A Fig**). Furthermore, key regulators of pluripotency (*NANOG*, *KLF4*, and *CBFA2T2*) were upregulated in both PGCLCs and naïve ESCs. Moreover, in addition to the native pluripotency-associated TFs (*KLF5*, *TFCP2L1*, and *ZNF42*) (**Figs 1C, 1F and S2A**), a substantial number of Krüppel-associated box (KRAB) domain zinc-finger protein (KZFP) family genes were upregulated only in naïve ESCs (**S3A and S3B Fig**), consistent with the findings of a previous study [22]. In contrast, the expression of KZFPs was generally low in PGCLCs but gradually increased during male germ cell development (**S3C Fig**).

In addition, we analyzed additional transcriptome datasets for PGCLCs [Kojima et al. [12] and the newly obtained data] and naïve ESCs [Takashima et al. [33] and Theunissen et al. [23]] and confirmed that the upregulation of the TFs mentioned above was observed across datasets (**S2B Fig**).

Regulatory elements underlying the transcriptional similarity between PGCLCs and naïve ESCs

To identify the regulatory elements underlying the upregulation of genes in both PGCLCs and naïve ESCs, we investigated published datasets from an assay for transposase-accessible chromatin using sequencing (ATAC-Seq) obtained from PGCLCs and naïve/primed ESCs [22,29].

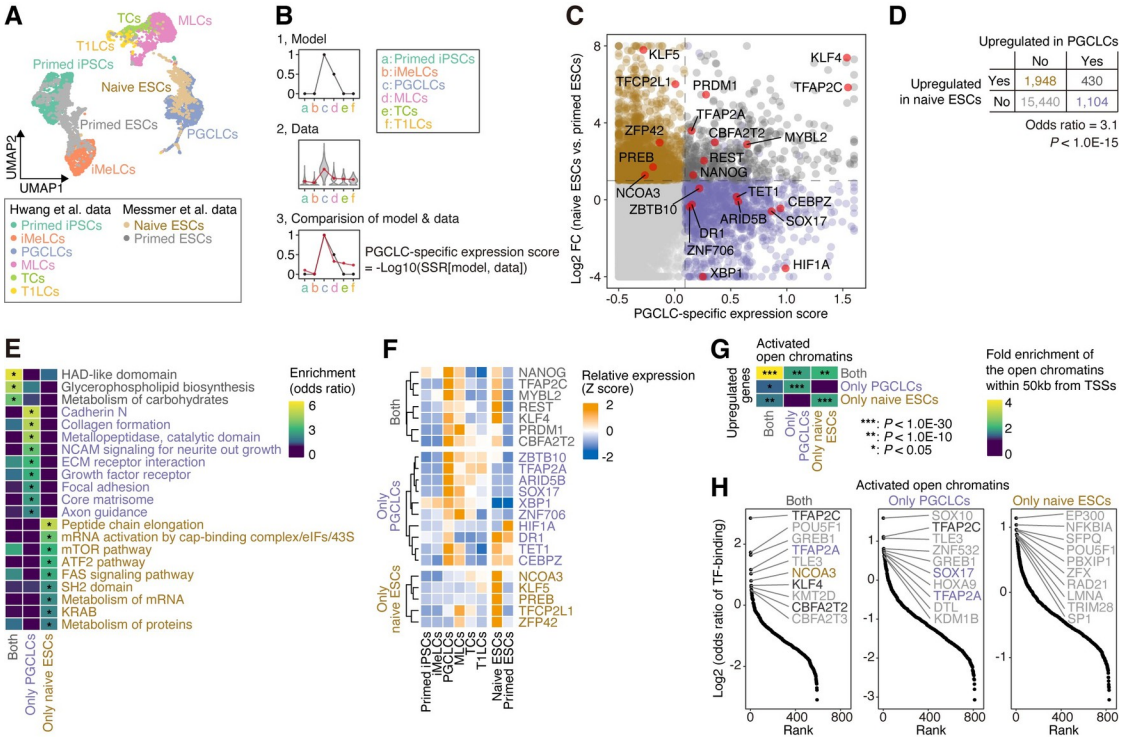

Fig 1. Characterization of the gene expression signature similarity between PGCLCs and naïve ESCs. (A) Dimension reduction analysis of scRNA-Seq data using UMAP [62]. Data for *in vitro*-derived human male germline development [Hwang et al. [10]] and for naïve and primed ESCs [Messmer et al. [36]] were integrated and subsequently used. The 3,000 protein-coding genes that were the most differentially expressed among cells were used. (B) Scheme for definition of the PGCLC-specific expression score. For each gene and TE, the sum of squared residuals (SSR) between the model (Panel 1) and the data (i.e., the normalized mean expression value for each cell type; Panel 2) was calculated (Panel 3). Subsequently, the SSR value was –log10-transformed (see **Definition of the PGCLC-specific expression score**). (C) Classification of protein-coding genes according to their expression patterns. The X-axis indicates the PGCLC-specific expression score, and the Y-axis indicates the log2 FC of the expression score in naïve ESCs vs. primed ESCs. The top 10% of genes with respect to the PGCLC-specific expression score were regarded as the genes upregulated in PGCLCs. Genes with log2 FC values > 1 and FDR values < 0.05 were regarded as upregulated in naïve ESCs. The genes were classified into four categories: genes upregulated in both cell types (dark gray), genes upregulated only in PGCLCs (purple), genes upregulated only in naïve ESCs (brown) and other genes (light gray). In addition, TFs (except for KZFPs) with elevated expression were annotated. The plot for KZFPs is shown in **S3A Fig**. (D) Association of the set of genes upregulated in PGCLCs with that in naïve ESCs. The *P* value was calculated with Fisher's exact test. (E) GO enrichment analysis results for the three gene categories (genes upregulated in both cell types, genes upregulated only in PGCLCs, and genes upregulated only in naïve ESCs). The gene sets that exhibited significant enrichment (odds ratio >2, FDR < 0.05; denoted by an asterisk) in any of the three gene categories are shown. (F) Expression patterns of the TFs annotated in (**C**). A violin plot is shown in **S2A Fig**. Although *TFAP2A* was first classified as a gene upregulated in both PGCLCs and naïve ESCs, we reclassified it as a gene upregulated only in PGCLCs since its expression in naïve ESCs was somewhat low (**Figs 1F and S2A**). (G) Enrichment of activated open chromatin regions in the vicinities of the upregulated genes. Three categories of open chromatin regions, namely those activated in both cell types, only PGCLCs, and only naïve ESCs compared to primed ESCs, were detected (log2 FC > 1; FDR < 0.05). Subsequently, for the three categories of open chromatin regions, the degrees of enrichment in the vicinity of (<50 kb from) the genes upregulated in both cell types, upregulated only in PGCLCs and upregulated only in naïve ESCs were calculated using the GREAT scheme [68] (see **Genomic Regions Enrichment of Annotations Tool (GREAT) enrichment analysis**). The *P* values were calculated with a binomial test. (H) Enrichment of TF-binding events in the open chromatin regions. A publicly available ChIP-Seq dataset provided by the GTRD [39] was used. For each TF, the enrichment (odds ratio) of the binding events in the respective categories of open chromatin regions compared to the other open chromatin regions was calculated. Statistical enrichment was calculated using Fisher's exact test. Of the TFs with FDR values <0.05, the top 10 TFs with respect to the odds ratio are annotated. The upregulated TFs shown in (**C**) and (**F**) are colored.

https://doi.org/10.1371/journal.pgen.1009846.g001

We first identified the open chromatin regions (i.e., ATAC-Seq peaks) that were activated in PGCLCs or naïve ESCs compared to primed ESCs and subsequently classified the open chromatin regions into three categories: those activated in both PGCLCs and naïve ESCs, those activated only in PGCLCs, and those activated only in naïve ESCs. Finally, we examined the

enrichment of the respective categories of open chromatin regions in the vicinity of (<50 kb from) the genes upregulated in both PGCLCs and naïve ESCs (**Fig 1G**). The open chromatin regions activated in both cell types were clearly enriched near the genes upregulated in both cell types, suggesting that the regulatory sequences activated in both cell types are important for controlling the upregulated genes common to these cell types (**Fig 1G**).

To identify the candidate TFs critical for controlling the regulatory elements identified above, we analyzed a publicly available chromatin immunoprecipitation sequencing (ChIP-Seq) dataset for 1,308 types of TFs provided by the Gene Transcription Regulation Database (GTRD) [39]. We used the data of 'merged' TF-binding sites, which were computed from ChIP-Seq data for a certain TF performed under various experimental conditions (e.g., cell line, treatment, and study). For the various TFs, we computed the enrichment of the binding events in each category of open chromatin regions compared to the other identified open chromatin regions (**Fig 1H**). The open chromatin regions activated in both PGCLCs and naïve ESCs were preferentially bound by TFs that were upregulated in both cell types (*TFAP2C*, *KLF4*, and *CBFA2T2*) or in one of these cell types (*TFAP2A* for PGCLCs and *NCOA3* for naïve ESCs). This result supports the importance of these TFs in regulating the genes upregulated in both cell types (**Fig 1H**).

TEs that are commonly upregulated in PGCLCs and naïve ESCs

To identify the TEs that are activated as enhancers during the human male germline developmental process, including PGCLCs, we analyzed the expression dynamics of TEs using the Hwang et al. scRNA-Seq dataset for *in vitro*-derived human male germ cells (**Fig 2**) [10]. We first used transcriptome data instead of epigenomic data since the transcriptional activity of TEs is known to reflect enhancer activity, similar to that of enhancer RNAs [40]. Pseudotime analysis [41] showed that the expression of TEs dynamically changed during *in vitro*-derived male germline development (**Fig 2A and 2B**). As described previously [10], the expression of most TEs (long interspersed nuclear elements [LINEs], short interspersed nuclear elements [SINEs], and SINE-VNTR-Alu [SVA] and DNA transposons) was gradually upregulated with the progression of development, presumably reflecting the gradual DNA demethylation that occurred during this process (**Fig 2B**) [7,35]. On the other hand, the expression of the various ERV subfamilies, including HERVH, LTR7, and LTR12C, was stage-specific (**Fig 2B**) [10]. In particular, the expression of some ERV subfamilies, such as HERVK, LTR5_Hs and HERVIP10FH, was specifically upregulated in PGCLCs and subsequently downregulated in cells at later stages (i.e., MLCs, TCs, and T1LCs) (**Fig 2B and 2C**). Notably, HERVK/LTR5_Hs (LTR5_Hs is a type of HERVK LTR sequence) was one of the top-ranked TEs with respect to the PGCLC-specific expression score (**Fig 2D, X-axis**). On the other hand, SVA transposons, a group of chimeric TEs originating partially from HERVK/LTR5_Hs [42], did not exhibit such a PGCLC-specific expression pattern (**Figs 2B, 2D and S4**).

Previous studies have shown that HERVK/LTR5_Hs is highly activated in naïve pluripotent cells, such as naïve ESCs and cells in ICM of blastocysts (**Fig 2C**) [22–25]. Indeed, our data showed that HERVK/LTR5_Hs was one of the top-ranked TEs upregulated in both PGCLCs and naïve ESCs (**Fig 2D**). Together, these results raise the possibility that LTR5_Hs may serve as enhancers shared between PGCLCs and naïve pluripotent cells and contribute to establishing the transcriptional similarity between these two cell types.

Increased enhancer-like signatures of LTR5_Hs in PGCLCs and naïve ESCs

To evaluate the enhancer potential of LTR5_Hs in PGCLCs and naïve ESCs, we investigated the chromatin accessibility and histone modification status of LTR5_Hs in these two cell types

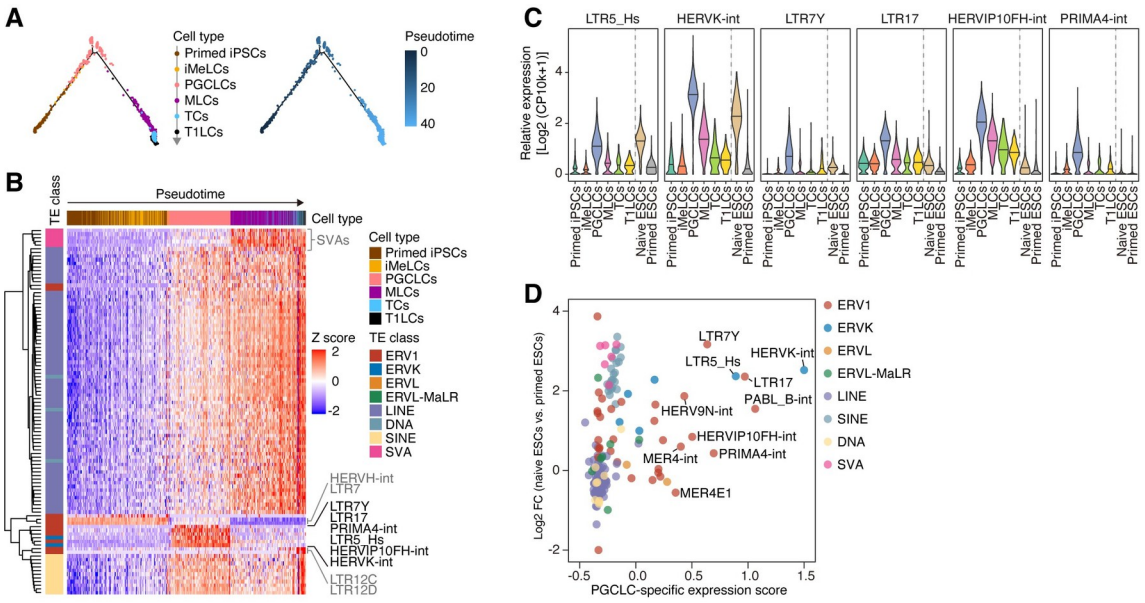

Fig 2. Specific expression of HERVK/LTR5_Hs in PGCLCs and naïve ESCs. (A) Pseudotime analysis [41] of scRNA-Seq data for *in vitro*-derived human male germline development [Hwang et al. [10]]. The 1,000 protein-coding genes that were the most differentially expressed throughout the development process were used. (B) Expression dynamics of TE subfamilies throughout male germline development. The cells are ordered according to the pseudotime shown in (**A**). The 100 TEs that were most differentially expressed among cell types are shown. (C) ERV subfamilies that were specifically expressed in PGCLCs [annotated in (**B**) in black]. In addition to the data for male germline development, data for naïve and primed ESCs [Messmer et al. [36]] are shown. (D) Identification of the TE subfamilies that were specifically upregulated in both PGCLCs and naïve ESCs. The X-axis indicates the PGCLC-specific expression score (defined in Fig 1B). The Y-axis indicates the log2 FC of the expression score between naïve ESCs vs. primed ESCs. The names of the top 10% TEs with respect to the PGCLC-specific expression score are annotated.

https://doi.org/10.1371/journal.pgen.1009846.g002

using ATAC-Seq and ChIP-Seq data targeting an active histone mark (i.e., H3K27ac), respectively (Fig 3). We examined the statistical enrichment of the two types of epigenetic signals on TEs in the various subfamilies (Fig 3A and 3B). In terms of both chromatin accessibility and active histone marks, LTR5_Hs was the top-ranked TE that was epigenetically activated in both PGCLCs and naïve ESCs. We next examined whether the chromatin accessibility of LTR5_Hs was greater in PGCLCs and naïve ESCs than in primed ESCs (Fig 3C). The open chromatin regions overlapping with LTR5_Hs tended to be activated in PGCLCs (Fig 3C, **upper panel**) and naïve ESCs (Fig 3C, **right panel**) compared to primed ESCs. Furthermore, LTR5_Hs was highly enriched in the open chromatin regions that were significantly activated in both PGCLCs and naïve ESCs (Fig 3C, **main panel and 3D**). Indeed, LTR5_Hs exhibited the strongest enrichment in these commonly activated open chromatin regions among all TEs (Fig 3E). Together, our findings show that LTR5_Hs exhibits enhancer-like epigenetic signatures in both PGCLCs and naïve ESCs.

We examined the enrichment of open chromatin regions in human PGCs on LTR5_Hs using a published ATAC-Seq dataset [29]. Although the developmental stage of PGCs is different from the stage represented by PGCLCs, we observed the strong enrichment of open chromatin regions on LTR5_Hs in PGCs, supporting the epigenetic activation of LTR5_Hs in PGCs (S5 Fig).

Potential regulators of LTR5_Hs in PGCLCs and naïve pluripotent cells

We next surveyed the TFs that bind to LTR5_Hs and control its activity in PGCLCs and naïve ESCs (Fig 4 **and** S3 Table). We analyzed the publicly available ChIP-Seq dataset for 1,308

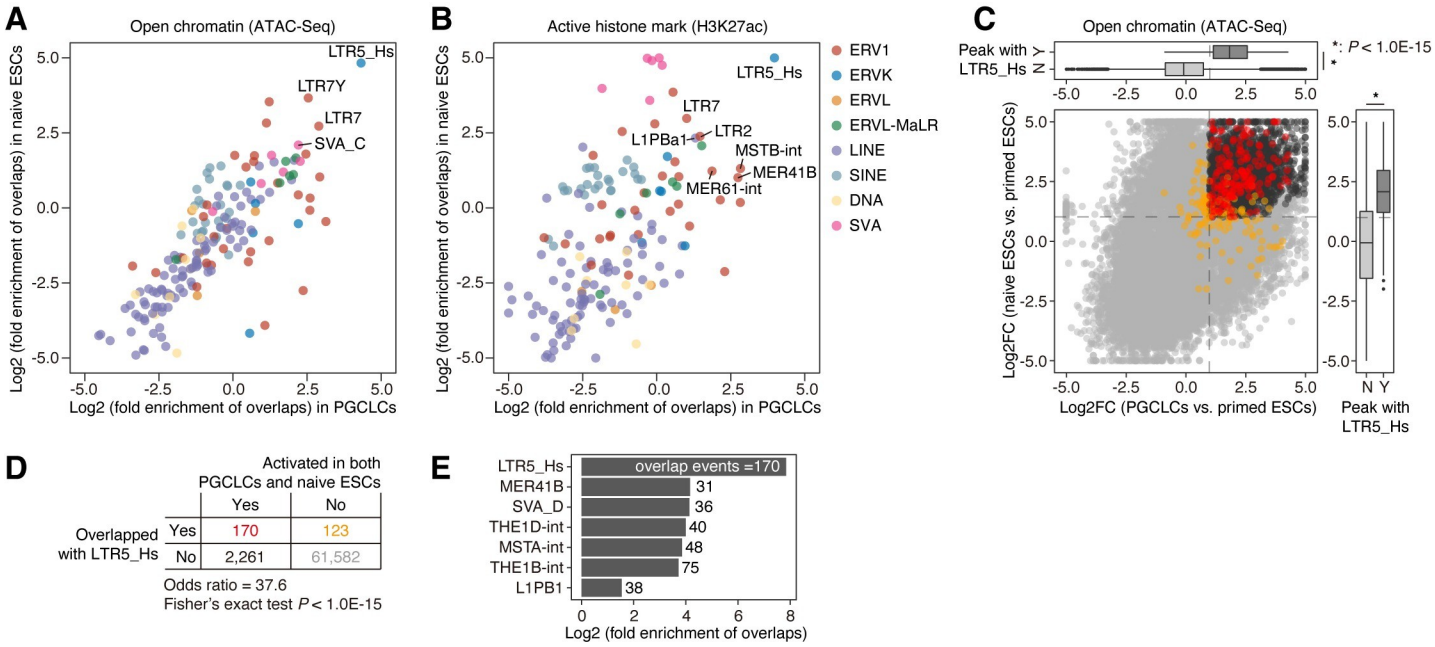

Fig 3. Potential enhancer activity of LTR5_Hs in PGCLCs and naïve ESCs. (A and B) Fold enrichment of the genomic overlap between TE loci and the peaks of ATAC-Seq (**A**) and ChIP-Seq targeting an active histone mark, H3K27ac (**B**). The fold enrichment value compared to the random expectation was calculated by the genomic permutation test. The X-axis and Y-axis indicate the log2-transformed fold enrichment values in PGCLCs and naïve ESCs, respectively. The ATAC-Seq and ChIP-Seq data originated from Pontis et al. [22] and Chen et al. [29]. (C) Upregulation of the chromatin accessibility of LTR5_Hs loci in PGCLCs and naïve ESCs compared to primed ESCs. For each ATAC-Seq peak (i.e., open chromatin region), the log2 FC scores of the chromatin accessibility in PGCLCs vs. primed ESCs (the X-axis) and naïve ESCs vs. primed ESCs (the Y-axis) are shown. In the main panel, the peaks overlapping with LTR5_Hs are colored red or orange. The peaks are colored red or black if they were upregulated in both PGCLCs and naïve ESCs (log2 FC > 1; FDR < 0.05). The color scheme is summarized in (**D**). In the upper and right panels, the marginal distributions for the X- and Y-axes, respectively, are shown (Y [Yes], overlapped with LTR5_Hs; N [No], not overlapped). An asterisk denotes $P < 1.0E\text{-}15$ based on the two-tailed Wilcoxon rank sum test. (D) The enrichment of LTR5_Hs in the ATAC-Seq peaks upregulated in both PGCLCs and naïve ESCs compared to primed ESCs. The *P* value was calculated with Fisher's exact test. (E) The enrichment of the various TE subfamilies in the ATAC-Seq peaks was upregulated in both PGCLCs and naïve ESCs. The fold enrichment value compared to the random expectation and the statistical significance were computed with the genomic permutation test. The number of overlapping events is shown on each bar. The results for TEs with significant enrichment (FDR < 0.05; log2 fold enrichment > 1; overlap events > 20) are shown.

https://doi.org/10.1371/journal.pgen.1009846.g003

types of TFs and identified TFs that preferentially bound to LTR5_Hs. Of these TFs, we extracted TFs that were expressed specifically in PGCLCs and naïve ESCs (**Fig 4A and 4B**). Of the TFs that preferentially bound to LTR5_Hs, *NANOG*, *TFAP2C*, *KLF4*, and *CBFA2T2* were upregulated in both PGCLCs and naïve ESCs (**Figs 1C, 4C and 4D**). Furthermore, *SOX17* and *TFAP2A* were specifically upregulated in PGCLCs, while *KLF5* was upregulated in naïve ESCs (**Figs 1C, 4C and 4D**). Notably, these TFs are known to play central roles in gene regulation in PGCLCs (i.e., *SOX17* and *TFAP2A*) [5,9], naïve pluripotent cells (i.e., *KLF5*) [43,44] or both cell types (i.e., *NANOG*, *TFAP2C*, *KLF4*, and *CBFA2T2*) (**Fig 1H**) [4,5,7,9,12,33,34,45]. Moreover, we confirmed that LTR5_Hs were preferentially bound by KLF4, NANOG, POU5F1, and TFAP2C in naïve ESCs, by TFAP2C in PGCLCs, and by SOX17 in a germ cell tumor cell line using publicly available ChIP-Seq datasets [29,46,47] (**S6 Fig**). Together, our data suggest that these TFs contribute to the epigenetic activation of LTR5_Hs in PGCLCs and naïve pluripotent cells.

Expression patterns of the genes adjacent to LTR5_Hs in PGCLCs and ESCs

To elucidate the roles of LTR5_Hs in gene regulation in PGCLCs and naïve pluripotent cells, we investigated the expression patterns of the genes adjacent to (<50 kb from) the LTR5_Hs

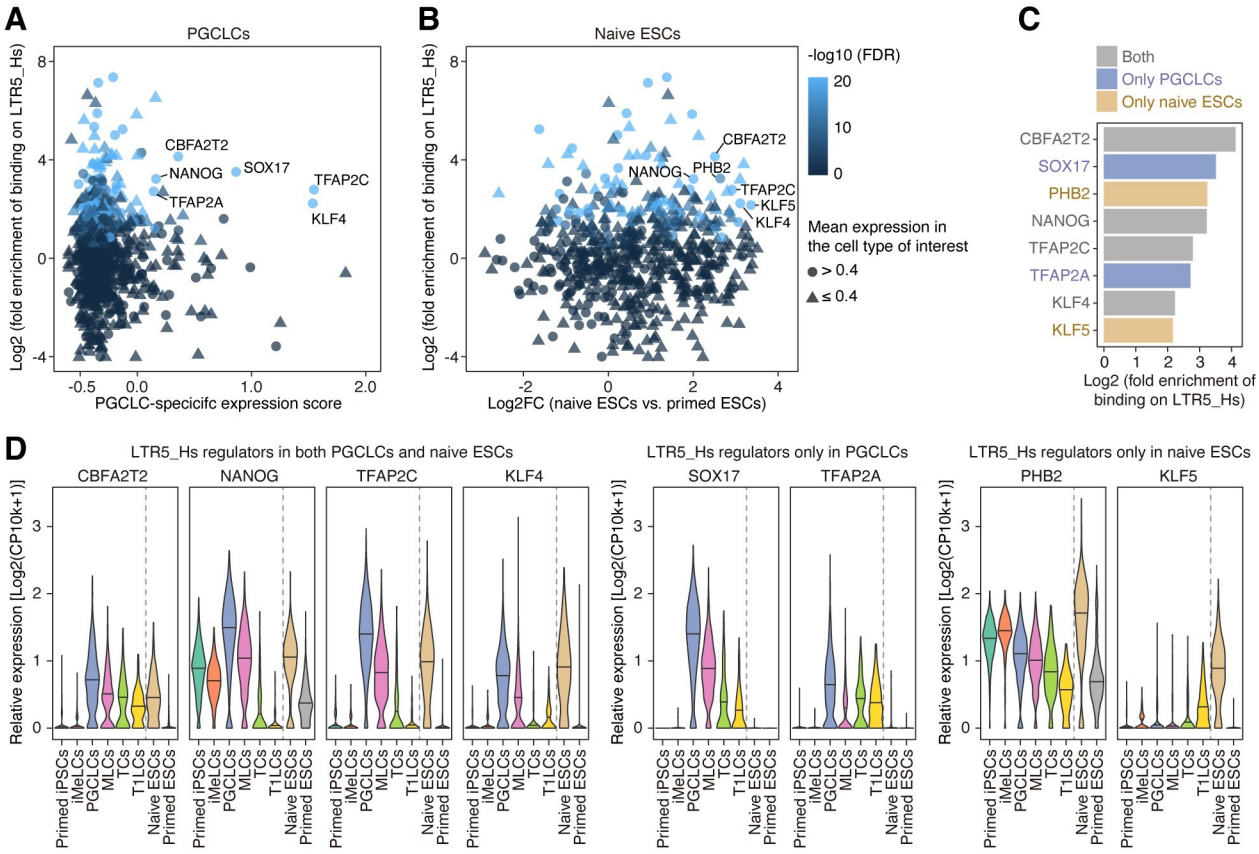

Fig 4. Identification of the potential regulators of LTR5_Hs in PGCLCs and naïve ESCs. (A and B) Identification of the TFs that bind to LTR5_Hs and are upregulated in PGCLCs (**A**) and naïve ESCs (**B**). For each TF, the statistical enrichment of the binding events on LTR5_Hs was calculated based on the publicly available ChIP-Seq dataset provided by the GTRD [39]. The Y-axis indicates the log2-transformed fold enrichment of the TF-binding events compared to the random expectation. The X-axis indicates the PGCLC-specific expression score (**A**) or the log2 FC of the expression score in naïve ESCs vs. primed ESCs (**B**). The symbols are colored according to the statistical significance of the TF-binding enrichment calculated by the genome permutation test. The symbol shape represents the mean expression level in PGCLCs (**A**) and naïve ESCs (**B**). The potential regulators of LTR5_Hs are annotated. The potential regulators were defined as the TFs satisfying the following criteria: (i) TFs that exhibited significant binding enrichment on LTR5_Hs (log2 fold enrichment > 2; FDR < 0.05; binding events > 20); (ii) for regulators in PGCLCs, TFs that were specifically upregulated in PGCLCs (the top 10% TFs with respect to the PGCLC-specific expression score; mean relative expression (log2[CP10k+1] > 0.4 in PGCLCs); and (iii) for regulators in naïve ESCs, TFs that were specifically upregulated in naïve ESCs (log2 FC > 2; FDR < 0.05; mean relative expression > 0.4 in naïve ESCs). (C) Classification of the potential LTR5_Hs regulators. The X-axis indicates the log2-transformed fold enrichment of the TF-binding events. (D) Expression patterns of TFs identified as potential LTR5_Hs regulators.

https://doi.org/10.1371/journal.pgen.1009846.g004

loci with transcriptomic or epigenetic activity (Fig 5 **and** S4 Table). The genes adjacent to LTR5_Hs tended to be specifically upregulated in both PGCLCs (Fig 5A**, upper panel**) and naïve ESCs (Fig 5A**, right panel**). Notably, the genes adjacent to LTR5_Hs were strikingly enriched with genes upregulated in both PGCLCs and naïve ESCs (Fig 5A**, main panel and 5B**). Indeed, of the genes commonly upregulated in PGCLCs and naïve ESCs, approximately 25% (107/430) were located in the vicinity of LTR5_Hs (Fig 5B). These results suggest that the epigenetic activation of LTR5_Hs is associated with the upregulation of adjacent gene expression in these cell types. GO enrichment analysis showed that genes associated with the "glucose metabolism" and "glycogen breakdown" terms were particularly enriched among the genes adjacent to LTR5_Hs and upregulated in both cell types (Fig 5C **and** S5 Table). These are child terms of the "metabolism of carbohydrates" term, which was significantly enriched for the genes upregulated in both PGCLCs and naïve ESCs (Fig 1E). Furthermore, glucose metabolism-related genes (i.e., *AGL*, *ENO2*, *PFKL*, *PHKA1*, and *PYGB*) were in the vicinity of

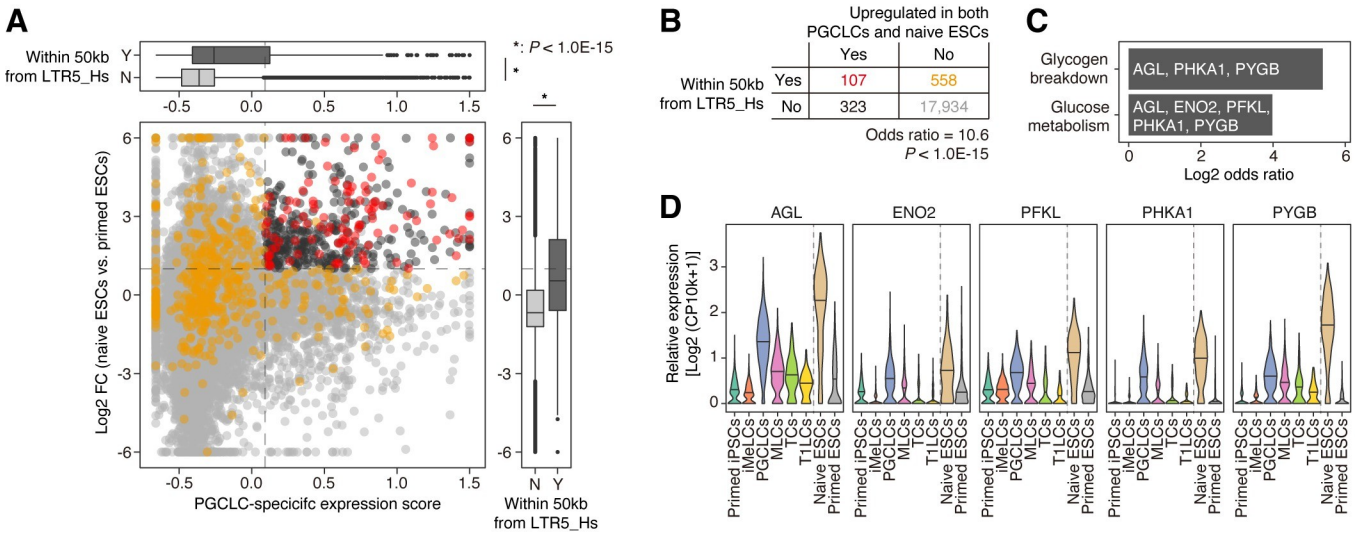

Fig 5. Expression patterns of the genes adjacent to LTR5_Hs in PGCLCs and naïve ESCs. (A) Association of the expression patterns of genes and their distance from LTR5_Hs in the genome. The X-axis indicates the PGCLC-specific expression score, and the Y-axis indicates the log2 FC of the expression score in naïve ESCs vs. primed ESCs. Genes were stratified according to whether they were present within 50 kb of LTR5_Hs with epigenetic or transcriptomic signals. In the main panel, the genes in the vicinity of LTR5_Hs are colored red or orange. The genes are colored red or black if they were upregulated in both PGCLCs (the top 10% of genes with respect to the PGCLC-specific expression score) and naïve ESCs (log2 FC > 1; FDR < 0.05). The color scheme is summarized in (**B**). In the top and right panels, the marginal distributions for the X- and Y-axes, respectively, are shown (Y [Yes], adjacent to LTR5_Hs; N [No], not adjacent). An asterisk denotes $P < 1.0E-15$ in the two-tailed Wilcoxon rank sum test. (B) Enrichment of the genes adjacent to LTR5_Hs among the genes upregulated in both PGCLCs and naïve ESCs. The *P* value was calculated with Fisher's exact test. (C) Results of the GO enrichment analysis. The gene sets with significant enrichment (FDR < 0.05) are shown. The names of the hit genes are shown on each bar. (D) Expression patterns of the genes present in the vicinity of LTR5_Hs and related to glucose metabolism.

https://doi.org/10.1371/journal.pgen.1009846.g005

LTR5_Hs and were highly upregulated in both PGCLCs and naïve ESCs (Fig 5D). The genes play central roles in energy generation via glycolysis (*PHKA1* and *PYGB*) and glycogenolysis (*AGL*, *ENO2*, and *PFKL*) (S7 Fig). These results suggest that the epigenetically activated LTR5_Hs may play a role in the regulation of glucose metabolism in both PGCLCs and naïve ESCs (see Discussion).

Gene expression alterations likely driven by LTR5_Hs during primate evolution

LTR5_Hs proliferated in hominoid genomes after the divergence of hominoids and OWMs [18]. To elucidate the alterations in gene expression driven by LTR5_Hs insertions, we performed comparative transcriptome analysis between humans and an OWM, the crab-eating macaque, focusing on PGCLCs (for humans) or the premigratory stage of PGCs (for macaques) and naïve pluripotent cells (Fig 6). Similar to the findings in Fig 5A, the results revealed that genes adjacent to LTR5_Hs in the human genome tended to be upregulated in both PGCLCs and naïve ESCs compared to primed ESCs (Fig 6A and 6C). On the other hand, the macaque orthologs of the human genes adjacent to LTR5_Hs did not show such a clear trend (Fig 6B and 6D). Furthermore, the genes commonly upregulated in human PGCLCs and naïve ESCs did not highly overlap with those in macaque PGCs and pluripotent cells (12%, 61 out of 512 genes in humans), although the upregulation of key TFs, such as *KLF4*, *NANOG*, *TFAP2C*, *PRDM1*, and *CBFA2T2*, was conserved between the two species (Fig 6E and S6 Table). Moreover, of the genes that were upregulated in both human PGCLCs and naïve ESCs but not in macaque PGCs and naïve pluripotent cells, approximately 21% (95 out of 451 genes) were in the vicinity of LTR5_Hs (Fig 6E). We hereafter refer to these 95 genes as

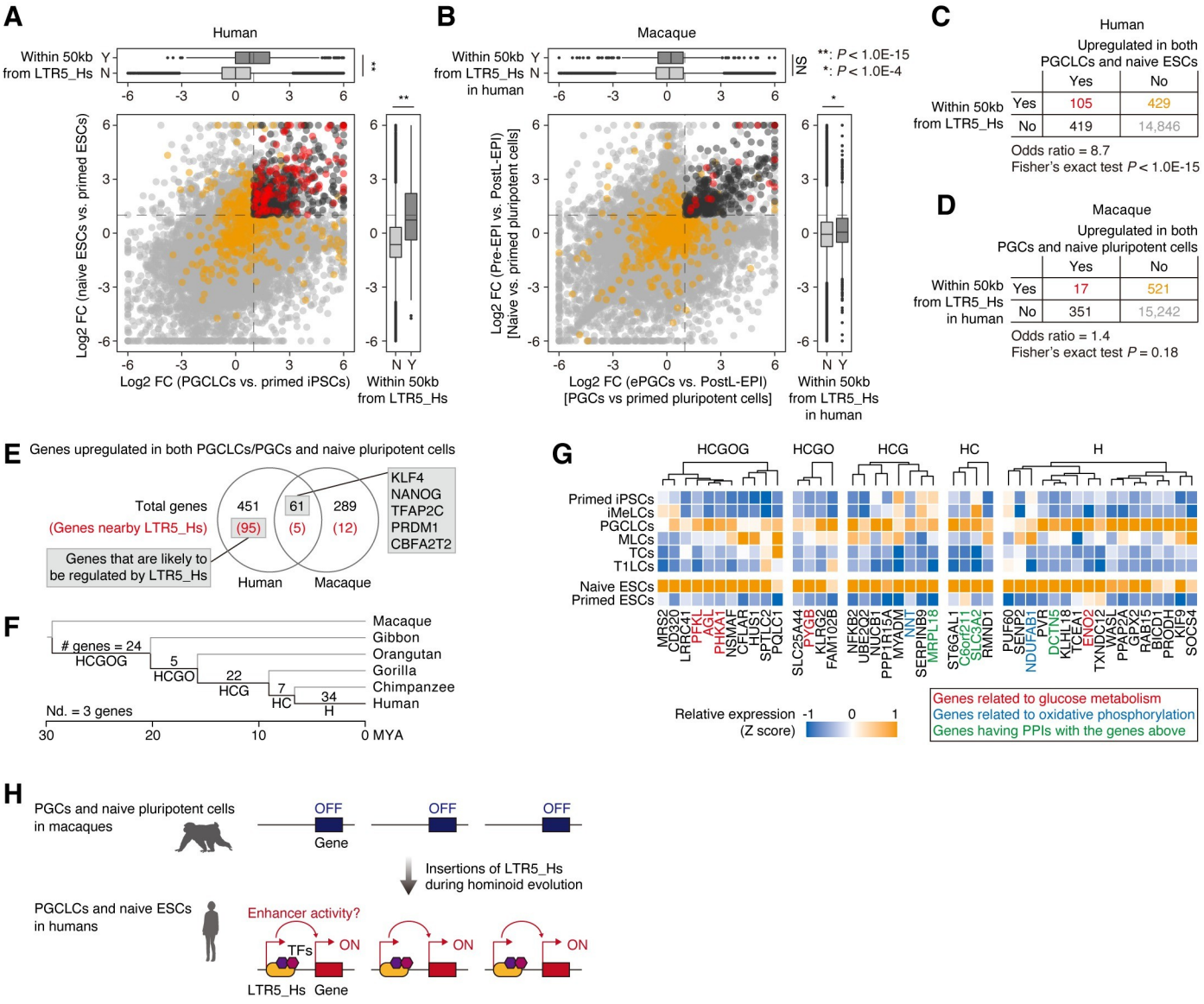

Fig 6. Comparative transcriptome analysis between humans and crab-eating macaques. (A and B) Comparative analysis of the gene expression patterns in PGCLCs/PGCs and naïve pluripotent cells between humans (**A**) and crab-eating macaques (**B**). (**A**) is similar to Fig 5A, but the X-axis indicates the log2 FC of the expression score in PGCLCs vs. primed iPSCs. In (**B**), the X-axis indicates the log2 FC of the expression score in early PGCs (ePGCs) vs. postimplantation late epiblasts (postL-EPIs; primed pluripotent cells), while the Y-axis indicates that in preimplantation epiblasts (pre-EPIs; naïve pluripotent cells) vs. postL-EPIs. In (**B**), the macaque genes are colored red or orange if their orthologs in humans are present within 50 kb of active LTR5_Hs. *, *P* value < 1.0E-4; **, *P* value < 1.0E-15; and NS, *P* value > 0.05. Human scRNA-Seq data [Messmer et al. [36] and Kojima et al. [12]] and macaque data [Sasaki et al. [6]] were used. (C and D) Enrichment of the human genes adjacent to LTR5_Hs (**C**) or their orthologs in macaques (**D**) among the genes upregulated in both PGCLCs/PGCs and naïve pluripotent cells. (E) Comparison of the genes upregulated in both PGCLCs/PGCs and naïve pluripotent cells between humans and macaques. The numbers in parentheses denote the numbers of genes adjacent to LTR5_Hs in the human genome or their orthologs in macaques. Only genes with ortholog information are included. The 95 genes (i) present in the vicinity of LTR5_Hs and (ii) that exhibited PGC- and naïve-specific expression patterns only in humans were defined as the genes likely to be regulated by LTR5_Hs. (F) Stratification of the genes that are likely to be regulated by LTR5_Hs according to the insertion date of the associated LTR5_Hs. On the various branches of the primate species tree, the numbers of the genes that are likely to be regulated by LTR5_Hs inserted in the corresponding branch are shown. The species tree was created with TimeTree [74]. Nd, not determined. (G) Expression patterns of the genes likely to be regulated by LTR5_Hs. Genes related to glucose metabolism, genes related to oxidative phosphorylation, and genes whose proteins engage in PPIs with the proteins encoded by the genes above (see S11B Fig) are annotated. Only genes exhibiting higher expression [mean expression (log2[CP10k+1] > 0.3] in both PGCLCs and ESCs are shown. (H) Summary of finings in the present study and the proposed model.

https://doi.org/10.1371/journal.pgen.1009846.g006

the genes likely to be regulated by LTR5_Hs (Fig 6E). Taken together, these results suggest that LTR5_Hs insertions may have altered the expression patterns of their adjacent genes to the PGCLC- and naïve-specific patterns in the hominoid lineage.

In a previous study (Fuentes et al.) [25], CRISPRa/i that target LTR5_Hs were established, and genes perturbed by these CRISPRa/i systems (i.e., genes regulated by LTR5_Hs) were identified in embryonic carcinoma cells. We compared the 95 genes that are likely to be regulated by LTR5_Hs defined in our analysis (shown in Fig 6E) with the genes perturbed by LTR5_Hs-targeting CRISPRa/i systems in that study [25] (**S8 Fig and S7 Table**). We found that these 95 genes significantly overlapped with the upregulated (56 out of 195) or downregulated (30 out of 73) genes by the LTR5_Hs-targeting CRISPR systems from the previous study [25], supporting the hypothesis that these genes are likely to be modulated by LTR5_Hs.

Finally, we examined the expression levels of i) HERVK/LTR5_Hs, ii) the glucose metabolic genes, and iii) genes that are likely to be regulated by LTR5_Hs (shown in Fig 6E) in human *in vivo* PGCs and early embryonic cells including naïve pluripotent cells such as ICM of blastocysts using publicly available scRNA-Seq datasets [8,48–50]. As described in the Introduction section, the developmental stages of the previously investigated PGCs (migrating PGCs) are more differentiated than the stage represented by PGCLCs (premigratory PGCs) [6]. Nevertheless, HERVK/LTR5_Hs and the genes mentioned above were more highly expressed in migrating PGCs than in the later stages of germ cells or somatic cells (S9 Fig). Similarly, HERVK/LTR5_Hs and the genes described above were highly expressed in naïve-like state cells (e.g., ICM of blastocysts) (S10 Fig). Together, these results suggest that our findings observed in *in vitro* cells (PGCLCs and naïve ESCs) can be recapitulated in their *in vivo* counterparts.

Gradual progression of LTR5_Hs-mediated gene expression alterations during hominoid evolution

The LTR5_Hs insertions started after hominoid-OWM divergence and continued even after human-chimpanzee divergence (S11A Fig) [18]. This result suggests that the gene expression alterations driven by LTR5_Hs may have proceeded gradually during hominoid evolution. To address this possibility, we first determined the insertion dates of LTR5_Hs loci (**S11A Fig and S8 Table**). Subsequently, the genes that are likely to be regulated by LTR5_Hs (Fig 6E) were classified according to the insertion dates of the associated LTR5_Hs loci (**Figs S11A and 6F**). As shown in Fig 6F, 24 out of 95 genes were associated with LTR5_Hs loci that were inserted in the common ancestor of the hominoid lineage (i.e., the branch "HCGOG" in Fig 6F). On the other hand, the majority of the genes (63 genes) were associated with LTR5_Hs loci that were inserted after the common ancestor of Homininae (human, chimpanzee, and gorilla) (Fig 6F). Of these, 34 genes were associated with human-specific LTR5_Hs loci (Fig 6F). Finally, we examined the insertion dates of LTR5_Hs loci that are likely to regulate genes related to the glucose metabolism pathway (shown in Fig 5C and 5D) and the genes encoding proteins that exhibit protein–protein interactions (PPIs) with the proteins encoded by the genes mentioned above (**Figs S11B and 6G**). Most of the core glucose metabolic genes (4 out of 5 genes) were associated with the LTR5_Hs loci inserted in the common ancestors of Hominoidea or Hominidae (humans, chimpanzees, gorillas, and orangutans) (**Figs S11B and 6G**). On the other hand, one of the core glucose metabolic genes (*ENO2*), the genes whose proteins have PPIs with the proteins of the core glucose metabolic genes described above, and the genes related to oxidative phosphorylation (i.e., *NDUFAB1* and *NNT*) were associated with the LTR5_Hs that were inserted more recently (**Figs S11B and 6G**).

LTR5_Hs insertions continued even after human speciation, and some LTR5_Hs loci are insertionally polymorphic in modern human populations [51]. To address whether the

LTR5_Hs that are likely important for the gene regulation in PGCLCs and naïve pluripotent cells are insertionally polymorphic, we identified LTR5_Hs loci that are present in the human reference genome (GRCh38) but are not fixed in 2,504 human genomes used as a global reference of human genome variation (**S9 Table**) [52]. Subsequently, we checked whether these polymorphic LTR5_Hs loci overlap with the LTR5_Hs loci that are likely to regulate gene expression (**S12 Fig**). Of the 11 polymorphic LTR5_Hs loci detected, two are in the vicinity of genes (*FOLR1* and *TNK1*) upregulated in both PGCLCs and naïve ESCs. This suggests the possibility that very recent insertions of LTR5_Hs have also contributed to alterations in gene expression in these cell types. Together, these results support that the gene expression alterations driven by LTR5_Hs in PGCs and naïve pluripotent cells may have proceeded gradually during hominoid evolution.

Discussion

Previous studies have suggested that there are similarities in gene expression between PGCLCs and naïve pluripotent cells. However, most of these studies have focused on several key TFs and have not characterized the similarity at the whole-transcriptome level in detail [4,5,7,9,12,29,33,34]. In the present study, we characterized the transcriptome signature and regulatory sequences shared between PGCLCs and naïve ESCs in detail and illuminated the presence of a shared gene regulatory network between these cell types (**Fig 1**).

We showed that numerous LTR5_Hs loci are epigenetically activated in both PGCLCs and naïve ESCs (**Figs 3 and 5**). Although the enhancer-like signatures of LTR5_Hs in naïve pluripotent cells have been reported in previous studies [22–25], our data highlight the pleiotropic activation of LTR5_Hs in PGCLCs and naïve ESCs, which likely contributes to the establishment of transcriptome similarity between these cells. The results of our comparative transcriptome analysis between humans and macaques support the hypothesis that LTR5_Hs insertions have altered the expression patterns of their adjacent genes in PGC- and naïve pluripotent cell-specific manners during hominoid evolution (**Fig 6**). Despite the centrality of PGCs and naïve pluripotent cells to maintenance of the germline (and by extension the species), our results suggest that gene expression in these cells may vary between humans based on polymorphisms in specific LTR5_Hs loci. Together, our data suggest that LTR5_Hs insertions may have gradually rewired the core gene regulatory network shared between PGCLCs and naïve pluripotent cells during hominoid evolution (**Fig 6H**).

We found that genes related to the metabolism of carbohydrates, including glucose, were commonly upregulated in PGCLCs and naïve ESCs (**Fig 1E**). In mice, the manner of glucose metabolism is similar between PGCLCs and naïve pluripotent cells [37,38,53,54]: mouse PGCLCs and naïve pluripotent cells use both glycolysis and oxidative phosphorylation (i.e., both aerobic and anaerobic respiration, referred to as bivalent glucose metabolism), while primed pluripotent cells depend exclusively on glycolysis (i.e., anaerobic respiration). On the other hand, the manner of glucose metabolism in human PGCs and PGCLCs is still unclear, although naïve human ESCs use bivalent glucose metabolism similar to that in naïve mouse ESCs [37,53,55]. Together with the previous findings described above, our data suggest that human PGCs and PGCLCs may also exhibit glucose metabolism similar to that of naïve ESCs (i.e., bivalent glucose metabolism), consistent with the case in mice. Notably, the manner of glucose metabolism affects the cellular identities of PGCLCs and naïve ESCs in mice [38]. Therefore, future functional studies seeking to characterize glucose metabolism in human PGCLCs and PGCs are warranted.

Previous studies have demonstrated that naïve pluripotent cells in humans exhibit higher glycolytic activity than primed pluripotent cells, while naïve pluripotent cells in mice and

common marmosets (a New World Monkey) do not [55,56]. These findings suggest that glycolytic activity in naïve pluripotent cells was elevated in the hominoid or more ancestral lineages at least after human-marmoset divergence. The data obtained in the present study suggest that the expression of genes related to glucose metabolism is likely controlled by LTR5_Hs in PGCLCs and naïve pluripotent cells and was likely upregulated in these cells during hominoid evolution (**Figs 5C, 5D and 6G**). Together, these findings raise the possibility that LTR5_Hs insertions are associated with elevations in glycolytic activity in naïve pluripotent cells (and possibly in PGCs/PGCLCs) during hominoid evolution. Since the manner of glucose metabolism substantially affects the identities of these cells [38], the epigenetic activation of LTR5_Hs may affect the establishment or maintenance of these cells in humans by modulating glucose metabolism.

Our arguments in the present study are mainly based on the results of association analyses using publicly available and snapshot datasets. Therefore, although our results strongly support the hypothesis that the enhancers derived from LTR5_Hs play a pivotal role in the gene regulatory network shared between PGCLCs and naïve pluripotent cells, further experimental validation is needed to demonstrate this hypothesis. Furthermore, the biological significance of the upregulation of carbohydrate metabolism-related genes in PGCLCs and the association of these genes and LTR5_Hs should also be evaluated by experiments. A very recent study showed that epigenetic perturbation by the CRISPRi systems of LTR5_Hs led to a decreased induction level of PGCLCs from ESCs, suggesting that *cis*-regulatory elements derived from LTR5_Hs work in the gene regulatory network associated with PGCLC specification [57]. Further studies to elucidate the role of LTR5_Hs in the gene regulatory network and its evolution are needed.

In conclusion, our data suggest that the core gene regulatory network shared between PGCs/PGCLCs and naïve pluripotent cells may have been finetuned by LTR5_Hs insertions during hominoid evolution. This gene regulatory network modification may contribute to alterations in cellular characteristics, such as glucose metabolism, which are critical for the cellular identities of PGCs/PGCLCs and naïve pluripotent cells. The present study provides insights into germline evolution driven by selfish ERVs during hominoid evolution.

Materials and methods

Bulk RNA-Seq of PGCLCs

The iPSC (9A13 XY) line used in this study was established in a previous study [Hwang et al. [10]]. iPSCs were cultured on plates coated with recombinant laminin-511 E8 (BG iMatrix-511 Silk, Peprotech, Cranbury, NJ) and were maintained under feeder-free conditions in StemFit Basic04 medium (Ajinomoto, Tokyo, Japan) containing basic FGF (Peprotech) at 37˚C under an atmosphere of 5% CO_2 in air. For passaging or induction of differentiation, the cells were treated with a 1:1 mixture of TrypLE Select (Life Technologies, Waltham, MA) and 0.5 mM EDTA/PBS to enable their dissociation into single cells, and 10 mM ROCK inhibitor (Y-27632; Tocris, Abingdon, United Kingdom) was added.

PGCLCs were induced from iPSCs via iMeLCs as described previously [Sasaki et al. [4]] and purified using the surface markers EpCAM and INTEGRINα6. Total RNA was extracted from iPSCs and PGCLCs by using an RNeasy Micro Kit (Qiagen, Venlo, Netherlands) according to the manufacturer's instructions. cDNA was synthesized using 1 ng of purified total RNA, and cDNA libraries were constructed for RNA sequencing by using a SMART-Seq HT Kit (Takara, Shiga, Japan) and a Nextera XT DNA Library Preparation Kit (Illumina, San Diego, CA) according to the manufacturers' instructions. The libraries were sequenced using a single-end sequencing protocol on an Illumina NextSeq 500 instrument.

Single-cell and bulk RNA-Seq analyses of human data

In the present study, read count matrices containing both human gene expression and subfamily-level TE expression data were prepared. To generate the count matrices, the human reference genome sequence (GRCh38/hg38) without ALT contigs was used. In addition, the gene and TE transcript annotation file (i.e., GTF file) generated in a previous study [Hwang et al. [10]] was used. Briefly, this annotation file contains the gene transcript annotations for GRCh38/hg38 from GENCODE version 22 [58] and the TE annotations for GRCh38/hg38 from the RepeatMasker output file (15-Jan-2014). TE loci with low reliability scores (Smith-Waterman scores $< 2{,}500$) were excluded. The annotation file is described in detail in and is available from the GitHub repository (https://github.com/TheSatoLab/TE_scRNA-Seq_analysis_Hwang_et_al/blob/master/CellRanger/input/hg38_TE_noAlt_unique.gtf.gz).

Regarding the scRNA-Seq dataset for human early male germ cell development [Hwang et al. [10]], the read count matrix provided by Hwang et al. was used (https://github.com/TheSatoLab/TE_scRNA-Seq_analysis_Hwang_et_al/blob/master/count_matrix_data/vitro/data.merged.vitro.count.csv.gz). The read count matrix was generated using only reads that were uniquely mapped to the human reference genome.

A read count matrix was generated for the scRNA-Seq datasets for naïve and primed ESCs [Messmer et al. [36]], for PGCLCs and iPSCs [Kojima et al. [12]], and for *in vivo* male germ cells including PGCs [Li et al. [8]] and early embryonic cells [48–50]. The sequencing reads were downloaded and decrypted using the fastq-dump command in SRA Toolkit (https://ncbi.github.io/sra-tools/). If multiple FASTQ files were available for one single cell, the FASTQ files were concatenated. The sequencing reads were trimmed using Trimmomatic (version 0.39) [59] and subsequently mapped to the human reference genome using STAR (version 2.6.1c) [60] with the gene-TE transcript model described above. The read count matrix was constructed using featureCounts (version 1.6.3) [61]. In this process, only reads that were uniquely mapped to the human reference genome were used.

Bulk RNA-Seq data for naïve and primed ESCs [Takashima et al. [33] and Theunissen et al. [23]] and for PGCLCs (original data obtained in the present study) were analyzed according to the same pipeline described in the above paragraph.

The read abundance of each TE subfamily was calculated by summing the read counts of TE loci belonging to the TE subfamily using an in-house Python script (https://github.com/TheSatoLab/TE_scRNA-Seq_analysis_Hwang_et_al/blob/master/make_count_matrix/script/sum_TE_count.subfamily.py). The counts per 10,000 (CP10k) value was calculated as the relative expression level, and the log2-transformed CP10k with a pseudocount of one (log2[CP10k +1]) value was subsequently computed.

Information on the RNA-Seq datasets analyzed in the present study is summarized in **S10 Table**.

Pseudotime analysis

Pseudotime analysis of the scRNA-Seq data for *in vitro*-derived human male germ cell development [Hwang et al. [10]] was performed using Monocle 2 [41] according to the procedures in the official tutorial (http://cole-trapnell-lab.github.io/monocle-release/docs/). The expression read count data were normalized under the negative binomial distribution assumption. In the pseudotime analysis, the 1,000 protein-coding genes that were the most differentially expressed during human male germ cell development were used. DDRTree was selected for the dimension reduction method.

Data integration and dimension reduction analysis of scRNA-Seq data

Data integration between the Hwang et al. [10] and Messmer et al. [36] datasets followed by dimension reduction analysis was performed using Seurat 3 (version 3.2.2) [62] according to

the scheme described in the Seurat tutorial (https://satijalab.org/seurat/vignettes.html). For each scRNA-Seq dataset, the expression data were normalized using SCTransform [63] by regressing out the total expression levels of mitochondrial genes. Subsequently, the datasets were integrated using the Seurat "anchoring" framework [62]. In the data integration, the 3,000 most differentially expressed protein-coding genes in both datasets were used. The dimension reduction analysis was performed via uniform manifold approximation and projection (UMAP) [62] based on the integrated expression data. In the UMAP analysis, the first 30 principal components were used.

Definition of the PGCLC-specific expression score

In this analysis, scRNA-Seq data for *in vitro*-derived human male germ cell development [Hwang et al. [10]] were used. The dataset includes data for a series of cells that were sequentially differentiated from iPSCs (iPSCs, iMeLCs, PGCLCs, MLCs, TCs, and T1LCs). As shown in the upper panel of **Fig 1B**, the model representing the PGCLC-specific expression pattern was defined by a iPSC:iMeLC:PGCLC:MLC:TC:T1LC ratio of 0:0:1:0.5:0:0 (referred to as the model). In this model, the expression value of MLCs was set to 0.5 since it is known that the critical TFs of PGCs (e.g., *TFAP2A*, *TFAP2C*, *SOX17*, and *NANOG*) remain weakly expressed in MLCs (and in multiplying prospermatogonia cells, the *in vivo* counterparts of MLCs) (**Fig 4D**) [10]. As shown in the middle panel of **Fig 1B**, for each gene and TE subfamily, the data representing the expression pattern were defined. Briefly, the relative expression (log2[CP10k +1]) values in the various cells were normalized as Z scores. Next, the mean expression values in the different cell types were calculated according to the Z scores above, and these mean expression values were rescaled to fit between 0 and 1. Here, a series of rescaled mean expression values is referred to as the data. Finally, as shown in the lower panel of **Fig 1B**, the sum of squared residuals (SSR) between the model and the data was calculated, and the SSR value was subsequently −log10-transformed. This −log10(SSR) value was defined as the PGCLC-specific expression score. This analysis was performed using an in-house script ("calc_PGC_specific_expression_score.R") available from the GitHub repository (https://github.com/TheSatoLab/LTR5_Hs_PGC_Naive_enhancer).

Differential gene expression analysis

Differential gene expression analysis was performed using DESeq2 (version 1.26.0) [64]. Only protein-coding genes were included in this analysis. Genes with relatively low expression levels (i.e., those with a 90^{th} percentile of reads per million value < 0.2) were excluded from the analysis. The statistical significance was calculated with the Wald test. The false discovery rate (FDR) value was calculated by the Benjamini-Hochberg (BH) method.

Classification of protein-coding genes and TFs according to their expression patterns

In this analysis, the protein-coding genes that were expressed in the dataset of either Hwang et al. [10] or Messmer et al. [36] were used. Genes upregulated in PGCLCs were defined as the top 10% of genes with respect to the PGCLC-specific expression score among the genes expressed in the Hwang et al. dataset. Genes upregulated in naïve ESCs were defined as the genes with log2 FC values > 1 and FDR values < 0.05 in the differential gene expression analysis between naïve ESCs vs. primed ESCs using DESeq2. According to the above definitions, the genes were classified as genes upregulated in both cell types, genes upregulated only in PGCLCs, genes upregulated only in naïve ESCs, and other genes.

The TFs shown in **Fig 1C and 1F** were selected according to the following scheme. Briefly, a list of human TFs was downloaded from The Human Transcription Factors database (version 1.01; http://humantfs.ccbr.utoronto.ca/index.php) [39]. *CBFA2T2* was manually added to the list of TFs. The listed TFs were classified as TFs upregulated in both cell types, TFs upregulated only in PGCLCs, TFs upregulated only in naïve ESCs, and other TFs according to the scheme described in the above paragraph. Of the TFs upregulated only in PGCLCs or only in naïve ESCs, the TFs with a mean log2(CP10k+1) value>0.6 in the corresponding cell type were selected. Of the TFs upregulated in both cell types, the TFs with a mean log2(CP10k+1) value >0.6 in either PGCLCs or naïve ESCs and with a mean log2(CP10k+1) value >0.3 in the other cell type were selected. In addition, *TFAP2A* was manually added to the list of the shown TFs. Information on the gene classification is summarized in **S1 Table**.

GO enrichment analysis

A gene-gene set association file including Molecular Signatures Database (MSigDB) canonical pathways and InterPro entries was used. The MSigDB canonical pathways were downloaded from MSigDB (http://software.broadinstitute.org/gsea/msigdb; version 6.1). InterPro entries were obtained from BioMart on the Ensembl website (www.ensembl.org; accessed on 13th February 2018).

The statistical significance values of the overlaps between the list of genes of interest and the predefined gene sets were calculated by one-tailed Fisher's exact test. FDR values were calculated using BH method. As a universal (or background) set of genes, the protein-coding genes satisfying the following criteria were used: 1) genes included in the gene-gene set association file above and 2) genes whose expression was detected in either of the scRNA-Seq datasets [Hwang et al. [10] or Messmer et al. [36]].

In the GO enrichment analysis shown in **Fig 1E**, the redundant gene sets whose members highly overlapped with each other were removed from the results. First, the gene sets with significant enrichment (FDR < 0.05) were ranked according to the odds ratio values. Second, if the gene members of a certain gene set highly overlapped with those of the upper-ranked gene sets, the gene set was removed from the results. Two gene sets were regarded as highly overlapping if the Jaccard index was greater than 0.5. This gene set filtering was performed with an in-house script ("rmRedundantGS_based_on_OR.py") available from the GitHub repository (https://github.com/TheSatoLab/LTR5_Hs_PGC_Naive_enhancer).

ATAC-Seq and ChIP-Seq analyses

Sequencing reads obtained from ATAC-Seq or ChIP-Seq were mapped to the human reference genome (GRCh38/hg38) using the BWA-MEM algorithm (version 0.7.17) [65]. Reads mapped to the mitochondrial genome or with low mapping scores (mapping quality, MAPQ < 10) were removed using SAMtools (version 1.10) [66]. In addition, PCR-duplicated reads were removed using Picard MarkDuplicates (version 2.18.16) (http://broadinstitute.github.io/picard/). Peak calling was performed using MACS2 callpeak (version 2.2.6) (https://pypi.org/project/MACS2/) with the threshold FDR < 0.05. For ChIP-Seq, the input control files were used in the peak calling step if the files were available. If >50,000 peaks were detected in one dataset, only the top 50,000 peaks with respect to statistical significance were used in the downstream analyses. Information on the analyzed data is summarized in **S10 Table**.

Identification of the open chromatin regions activated in PGCLCs or naïve ESCs compared to primed ESCs

First, the union (or merged) set of ATAC-Seq peaks between the two compared conditions (e.g., naïve ESCs vs. primed ESCs) was defined using the bedtools merge function (version

v2.27.0) [67]. Second, from the sequencing read alignment (BAM) file of each ATAC-Seq run, the reads that were assigned to the various merged peaks were counted using featureCounts (version 1.6.3) [61]. Finally, the peaks (i.e., open chromatin regions) that were activated (log2 FC $>$ 1; FDR $<$ 0.05) in PGCLCs or naïve ESCs compared to primed ESCs were identified using DESeq2 (version 1.26.0) [64]. Subsequently, the open chromatin regions were classified into those upregulated in both cell types, those upregulated only in PGCLCs, those upregulated only in naïve ESCs, and others.

Genomic Regions Enrichment of Annotations Tool (GREAT) enrichment analysis

As shown in **Fig 1G**, the enrichment of the open chromatin regions of interest (the open chromatin regions activated in both cell types, only PGCLCs, and only naïve ESCs) in the vicinity of the genes of interest (the genes upregulated in both cell types, only PGCLCs, and only naïve ESCs) was calculated according to the GREAT scheme [68]. This method is explained in detail elsewhere [69]. Briefly, regions of interest were defined as the regions within 50 kb of the transcription start sites (TSSs) of the genes of interest. Background regions were defined as the regions within 50 kb of the TSSs of all protein-coding genes. The lengths of the regions of interest and the background regions were calculated and referred to as Li and Lb, respectively. In the regions of interest and the background regions, the open chromatin regions were counted (referred to as counts of interest [Ci] and background counts [Cb], respectively). The fold enrichment value was calculated by dividing Ci/Cb by Li/Lb, and the statistical significance was evaluated using a binomial test. This analysis was performed using an in-house script ("great_pairwise.py") available from the GitHub repository (https://github.com/TheSatoLab/LTR5_Hs_PGC_Naive_enhancer).

Enrichment analysis of TF binding sites on the set of open chromatin regions of interest

A public ChIP-Seq dataset for 1,308 types of TFs provided by the GTRD (version 19.10) [39] was used. The ChIP-Seq peak data file "Homo sapiens_macs2_clusters.interval.gz" was downloaded from the database above (http://gtrd19-10.biouml.org/) on 20th May 2020. This file contains the single set of peaks (i.e., clustered peaks) for each TF. In this file, the peaks that had been computed for the same TF under the different experimental conditions (e.g., cell line, treatment, and study) were joined into clusters. For the various TFs, we detected overlaps between the TF binding sites and the open chromatin regions. Next, we classified the open chromatin regions according to (i) whether the open chromatin regions overlapped with the TF binding sites and (ii) whether the open chromatin regions belonged to a set of open chromatin regions of interest (i.e., those activated in both cell types, only PGCLCs, and only naïve ESCs). Subsequently, the odds ratios and *P* values were calculated with Fisher's exact test. The FDR values were calculated with the BH method.

Genomic permutation test

To calculate the fold enrichment of the overlaps between TE loci and a set of genomic regions of interest (e.g., ATAC-Seq peaks), randomization-based enrichment analysis (i.e., a genomic permutation test) was performed. The genomic regions of interest were randomized using the bedtools shuffle function [67]; subsequently, the genomic regions of interest on TE loci in the randomized data were counted. This process was repeated 100 times, and the mean value of the counts in the randomized datasets was regarded as the random expectation value. The fold enrichment was calculated by dividing the observed count by the random expectation value. The P value was calculated according to the assumption of a Poisson distribution. The random

expectation value was used as the lambda parameter of the Poisson distribution. This analysis was performed using an in-house script (calc_enrichment_randomized.great.py) available from the GitHub repository (https://github.com/TheSatoLab/LTR5_Hs_PGC_Naive_enhancer).

Identification of the potential regulators of LTR5_Hs in PGCs and naïve pluripotent cells

We first identified the TFs that preferentially bind to LTR5_Hs using public ChIP-Seq data for 1,308 types of TFs provided by the GTRD (version 19.10) [39].

For the various TFs, we calculated the fold enrichment of the TF-binding events on LTR5_Hs over the random expectation as well as the statistical significance using the genomic permutation test described in the above section. Next, we integrated the TF binding enrichment data with the expression pattern data of these TFs. To identify the potential regulators of LTR5_Hs in PGCs, the PGCLC-specific expression score defined in the above section was used. To identify the regulators in naïve pluripotent cells, the log2 FC values of the expression levels between naïve ESCs vs. primed ESCs computed using DESeq2 [64] were used. The potential regulators of LTR5_Hs were defined as the TFs satisfying the following criteria: (i) TFs that exhibited significant binding enrichment on LTR5_Hs (log2-fold enrichment > 2; FDR < 0.05; binding events > 20); (ii) for regulators in PGCLCs, TFs that were specifically upregulated in PGCLCs (in the top 10% with respect to the PGCLC-specific expression score; mean relative expression (log2[CP10k+1] > 0.4 in PGCLCs); and (iii) for regulators in naïve pluripotent cells, TFs that were specifically upregulated in naïve ESCs (log2[FC] > 2; FDR < 0.05; mean relative expression > 0.4 in naïve ESCs).

Definition of the genes in the vicinity of active LTR5_Hs

The "active" LTR5_Hs loci, namely, the LTR5_Hs loci with transcriptomic or epigenetic signals, were defined. Specifically, LTR5_Hs loci with transcriptomic signals were defined as loci whose expression was detected in >0.5% of the cell population in any of the following scRNA-Seq datasets: (i) the PGCLC dataset of Hwang et al. [10], (ii) the PGCLC dataset of Kojima et al. [12], and (iii) the naïve ESC dataset of Messmer et al. [36]. The LTR5_Hs loci with epigenetic signals were defined as loci that overlapped with the epigenetic signal peaks in any of the following ATAC-Seq or ChIP-Seq (targeting H3K27ac) datasets: (i) the PGCLC ATAC-Seq or ChIP-Seq dataset of Chen et al. [29] and (ii) the datasets of naïve ESCs in Pontis et al. [22]. Information on the active LTR5_Hs loci is summarized in **S8 Table**.

The genes in the vicinity of the active LTR5_Hs were also defined. The TSSs of the various transcripts for each protein-coding gene were extracted from the GENCODE gene annotation model (version 22) [58]. The distance from the TSS of each gene to the closest LTR5_Hs copy was computed using the bedtools closest function [67]. Subsequently, for each gene, the minimum distance from the TSS to the active LTR5_Hs copy was calculated. A gene in the vicinity of the active LTR5_Hs was defined as a gene within 50 kb of the minimum distance defined above. We performed a sensitivity analysis on the distance parameter for proximity definition and confirmed that our conclusions are relatively robust at various proximity definitions (e.g., within 20kb, 50kb, 100kb, 200kb, and 500kb) (**S13 Fig**).

scRNA-Seq analysis of crab-eating macaque data and comparative transcriptome analysis between humans and macaques

For analysis of crab-eating macaque data, the reference genome (macFas5.fa), gene transcriptome annotation (genes/macFas5.ensGene.gtf; corresponding to the Ensembl 99 gene

transcriptome annotation), and RepeatMasker output files (macFas5.fa.out) were downloaded from the University of California, Santa Cruz (UCSC) Genome Browser (http://hgdownload.soe.ucsc.edu/goldenPath/macFas5/bigZips/) on 23rd March 2020. The gene-TE transcript model for crab-eating macaques was constructed according to the same procedure used for humans. For the gene model, transcripts with the flag "retained intron" were excluded. For the TE model, TE loci with low reliability scores (i.e., Smith-Waterman scores $< 2{,}500$) were excluded. Additionally, the regions of TE loci overlapping with the gene transcripts were also excluded. The gene-TE transcript model was generated by concatenating the gene and TE models.

The scRNA-Seq dataset of early embryos and germ cells from crab-eating macaques [Sasaki et al. [6]] was analyzed. Briefly, the sequencing reads were trimmed using Trimmomatic (version 0.39) [59] and subsequently mapped to the reference genome using STAR (version 2.6.1c) [60] with the gene-TE transcript model above. The read count matrix was constructed using featureCounts (version 1.6.3) [61].

Gene ortholog information between humans and crab-eating macaques was downloaded from the Ensembl database (version 99) via BioMart (https://www.ensembl.org) on 23rd March 2020.

Phylogenetic analysis of the LTR5 family

LTR5A, LTR5B, and LTR5_Hs loci with Smith-Waterman scores $\geq 2{,}500$ were extracted from the RepeatMasker output file (15-Jan-2014; for GRCh38/hg38). Subsequently, the sequences of these LTR5 loci were extracted from the human reference genome (GRCh38/hg38) using the bedtools getfasta function [67]. A multiple sequence alignment (MSA) of these LTR5 loci was constructed using MAFFT with the FFT-NS-i algorithm (version 7.407) [70]. In the MSA, the alignment sites with <85% site coverage were eliminated using the in-house script "select_alignment_site.py" available from the GitHub repository (https://github.com/TheSatoLab/primate_A3_repertoire_and_evolution/blob/main/Trees/script). Subsequently, the sequences that had gaps in >15% of alignment sites were eliminated using the script above. In addition, tree-based filtering of the underlying dataset was performed prior to construction of a final tree. A preliminary tree was constructed, and phylogenetic outlier sequences, which have extremely long external branches (i.e., standardized external branch lengths > 3), were subsequently detected and discarded from the MSA used for final tree construction. The phylogenetic tree of LTR5 loci was reconstructed using RAxML (version 8.2.11) [71] with the GTRCAT model.

Investigation of the distribution of orthologs of human LTR5 loci across Simiiformes

LiftOver chain files were downloaded from the UCSC Genome Browser (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/liftOver/) (**S11 Table**). Using the LiftOver program (http://genome.ucsc.edu/cgi-bin/hgLiftOver) and the LiftOver chain files, the genomic coordinates of LTR5 loci in the human reference genome were converted to those in another species with the option "Minmatch = 0.5". If the conversion was successful, we inferred that the orthologs of the LTR5 loci were likely present in the corresponding genome.

Estimate of the insertion dates of LTR5_Hs loci and stratification of the genes likely to be regulated by LTR5_Hs according to the insertion dates

The insertion dates of the various LTR5_Hs loci were estimated according to information on both (i) the distributions of orthologous insertions across primates and (ii) the positions of

LTR5 loci in the phylogenetic tree. Since there were a substantial number of missing values in the ortholog distribution information, we used phylogenetic information in addition to ortholog information to robustly estimate the LTR5_Hs insertion dates. First, LTR5_Hs loci were ordered according to the phylogenetic relationship (from older to younger). Second, using the framework of a sliding window analysis, the final positions of LTR5_Hs loci where more than three out of ten LTR5_Hs loci had orthologous insertions were determined for each primate of interest (chimpanzee, gorilla, orangutan, gibbon, macaque, and marmoset). For each species, LTR5_Hs loci that were older than the final LTR5_Hs copy were regarded as LTR5_Hs loci that were inserted before the divergence between humans and the corresponding species. Information on the estimated insertion dates is summarized in **S8 Table**.

The genes that are likely to be regulated by LTR5_Hs were stratified according to the insertion dates of the associated LTR5_Hs loci. If the associated LTR5_Hs of one gene was not included in the phylogenetic tree of LTR5 loci, the gene was categorized as "not determined". In addition, if multiple LTR5_Hs loci with distinct insertion dates were associated with one gene, the gene was also categorized as "not determined".

PPI network analysis

PPI network information for humans was downloaded from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database (version 11.0; "9606.protein.links.v11.0.txt.gz") [72]. The PPI links with confidence scores >400 were used for the analysis. The number of interacting partners of each gene was computed with the igraph package implemented in R (https://igraph.org/).

Detection of LTR5_Hs insertions that are present in the human reference genome but not fixed in the human population

High-coverage whole genome sequencing (WGS) datasets in 1000 Genome Project [52] were downloaded from the following URL: 'ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/'. We searched WGS data for reads spanning the insertion site of of LTR5_Hs loci as follows. We first detected/annotated LTR5_Hs from GRCh38 using RepeatMasker with repeat sequence library provided from RepBase (version 24.01). We used the '-s -no_is' options to sensitively detect LTR5_Hs. Next, we searched for reads skipping annotated LTR5_Hs, that is, reads mapping to the genomic regions flanking the LTR5_Hs insertion site (i.e. the predicted state/sequence of this locus prior to LTR5_Hs integration). We screened reads in the WGS datasets and extracted soft-clipped reads with 'SA:Z' tag. During this step, supplementary reads were excluded from analysis. We checked the mapped positions of the clipped and non-clipped regions on GRCh38. Here after, we refer to the clipped and non-clipped regions as to clipped_seq and non_clipped_seq, respectively. We next filtered out reads of which clipped_seq and non_clipped_seq are mapping to different chromosomes. Then we checked whether the clipped_seq and non_clipped_seq are mapping to flanking regions of an annotated LTR5_Hs locus. In this step, we considered that a read is a skipping read if both the clipped_seq and non_clipped_seq map to 25-nt from the ends of an annotated LTR5_Hs locus. We found 11 LTR5_Hs loci that are likely absent in at least one datasets. The mean count of skipping reads per LTR5_Hs locus in a single dataset ranged from 3.4 to 12.8. To exclude potential false positives due to any technical reasons, such as index hopping, we considered that an individual lacks at least one allele of a LTR5_Hs copy if two or more skipping reads were found at the LTR5_Hs locus.

Data visualization

All data visualizations were performed in R (version 3.6.3). Heatmaps were drawn using ComplexHeatmap [73]. The phylogenetic tree was visualized with ggtree (http://bioconductor.org/packages/release/bioc/html/ggtree.html). The PPI network was visualized using ggnet2 (https://briatte.github.io/ggnet/). The other data were visualized with ggplot2 (https://ggplot2.tidyverse.org/).

Statistical analysis

Statistical analysis was performed in R (version 3.6.3). Statistical significance was evaluated by the two-tailed Wilcoxon rank sum test unless otherwise noted. FDR values were calculated by BH method.

Supporting information

S1 Fig. PGCLC-specific score of genes that are upregulated in PGCs compared to the later stages of male germ cells. The scRNA-Seq data from Li et al. [8] was analyzed. As controls, the scores of genes that are downregulated in PGCs and those of genes that did not significantly change between PGCs and the later stages are shown.
(TIF)

S2 Fig. TFs upregulated in both cell types, only PGCLCs, and only naïve ESCs. (A) Expression levels in various cell types from scRNA-Seq data for male germline development [Hwang et al. [10]] and for naïve and primed ESCs [Messmer et al. [36]]. The results for the TFs annotated in **Fig 1C** are shown. (B) Upregulation of TFs in PGCLCs and naïve ESCs observed across datasets. For the various datasets, the log2 FC values of the expression scores in PGCLCs vs. primed iPSCs or naïve ESCs vs. primed ESCs are shown. An asterisk denotes significant upregulation (FDR < 0.05; log2 FC > 1). A gray asterisk indicates that the expression level of the gene was not high (the mean expression level of the gene was below the 50th percentile for all expressed genes) even though significant upregulation was observed. For PGCLCs, the data of Hwang et al. [10] and Kojima et al. [12] were analyzed in addition to the original data in the present study. For naïve ESCs, the data of Messmer et al. [36], Takashima et al. [33], and Theunissen et al. [23] were analyzed.
(TIF)

S3 Fig. Expression patterns of KZFPs. (A) Classification of KZFPs according to their expression patterns. Highly expressed KZFPs in PGCLCs or naïve ESCs are annotated. The results for TFs other than KZFPs are shown in **Fig 1C**. (B) Distributions of the log2 FC values of the expression scores of KZFPs in naïve ESCs vs. primed ESCs. The dot color denotes the statistical significance of the gene expression change. (C) Expression patterns of KZFPs during *in vitro*-derived human male germline development. The heatmap shows the relative mean expression values in the various cell types. The upper panel shows the transitions of the individual (gray) and mean (red) expression values.
(TIF)

S4 Fig. Expression patterns of SVA transposons. The results for the SVA transposons included in the heatmap in **Fig 2B** are shown.
(TIF)

S5 Fig. The enrichment of open chromatin regions in human PGCs on LTR5_Hs. The ATAC-Seq data from Chen et al. [29] was analyzed. The log2 fold enrichment and statistical

significance values are shown as a volcano plot.
(TIF)

S6 Fig. The enrichment of TF binding sites on LTR5_Hs in cell types of interest. The enrichment for KLF4, NANOG, POU5F1, TFAP2C in naïve human ESCs and that for TFAP2C in PGCLCs are shown. In addition, the enrichment for SOX17 in a seminoma cell line (TCam-2 cells) is shown. The log2 fold enrichment and statistical significance scores are shown as a volcano plot. As controls, the enrichment scores for Tes other than LTR5_Hs are shown.
(TIF)

S7 Fig. Pathway maps of glycolysis and glycogen breakdown. Pathway maps of glycolysis (**A**) and glycogen breakdown (**B**). Genes that are likely to be regulated by LTR5_Hs (i.e., *AGL*, *ENO2*, *PFKL*, *PHKA1*, and *PYGB*) are highlighted in orange. The pathway maps originated from the Reactome pathway database (https://reactome.org/) (*65*).
(TIF)

S8 Fig. Comparison of genes regulated by LTR5_Hs defined in this study and a previous study (Fuentes et al.) [25]. The 95 genes that are likely to be regulated by LTR5_Hs (shown in Fig 6E) were compared with genes that were up- or downregulated by CRISPRa or CRISPRi systems in embryonic carcinoma cells in a previous study (Fuentes et al.) [25]. To perform a fair comparison, only genes contained in both datasets and located within 50 kb from LTR5_Hs insertions (447 genes) were included in this analysis. These genes were stratified according to (i) whether the genes were included in the 95 genes (shown in Fig 6E) and (ii) whether the genes were perturbed by CRISPR systems in Fuentes et al. [25] (adjusted p value < 0.05, log2 FC value > 1 for upregulation, log2 FC value < -1 for downregulation). Subsequently, the degree of overlap between the stratified gene sets was evaluated. The *P* value was calculated with Fisher's exact test. Information on the genes is summarized in S7 Table.
(TIF)

S9 Fig. Expression pattern of HERVK/LTR5_Hs and genes that are likely regulated by LTR5_Hs in human PGCs. The scRNA-Seq data from Li et al. [8], which includes male germ cells (migrating PGCs and multiplying and mitotically quiescent prospermatogonia) and somatic cells at 4–25 weeks post-fertilization, was analyzed. Please note that migrating PGCs are more differentiated than the stage represented by PGCLCs (pre-migratory stage, $\leq$3 weeks post-fertilization). A) Violin plot showing the expression of LTR5_Hs, HERVK, and the glucose metabolism-related genes. B) Heatmap showing the normalized mean expression of genes that are likely regulated by LTR5_Hs (defined in Fig 6E).
(TIF)

S10 Fig. Expression pattern of HERVK/LTR5_Hs and genes that are likely regulated by LTR5_Hs in human early embryonic cells including naïve ICM of blastocysts. The scRNA-Seq datasets from various studies [48–50], which include from Pronucleus cells to blastocysts at embryonic day 7, were merged and analyzed. A) Violin plot showing the expression of LTR5_Hs, HERVK, and the glucose metabolism-related genes. In blastocysts, dots for ICM (red), trophectoderm (blue), and pre-lineage (yellow) are colored. In addition to data for the embryonic cells described above, data for naïve and primed ESCs [36] are shown. B) Heatmap showing the normalized mean expression of genes that are likely regulated by LTR5_Hs (defined in Fig 6E).
(TIF)

S11 Fig. Stratification of the genes likely to be regulated by LTR5_Hs according to the insertion date of the associated LTR5_Hs. (A) Stratification of LTR5_Hs loci in the human genome according to their insertion dates. (i) Phylogenetic tree of the LTR5 family (including LTR5_Hs and related subfamilies [i.e., LTR5A and LTR5B]). (ii) Information on the distribution of orthologous insertions of LTR5 loci among primate genomes. According to the ortholog distribution and phylogeny, LTR5_Hs loci were stratified into five categories (HCGOG, HCGO, HCG, HC, and H). (iii) Epigenetic and transcriptomic statuses of various LTR5_Hs loci. (iv) LTR5_Hs loci that are likely to be associated with gene regulation. (B) PPI network for the genes likely to be regulated by LTR5_Hs. Only PPI links among the proteins encoded by the displayed genes are shown. The node color denotes the insertion date of the associated LTR5_Hs of the gene. The node size is proportional to the number of interacting partners in the whole PPI network. The glucose metabolism-related network is circled in orange. The PPI information originated from the STRING database [72].
(TIF)

S12 Fig. Potential roles of polymorphic LTR5_Hs insertions on the gene expression in PGCLCs and naïve ESCs. LTR5_Hs loci that are present in the human reference genome but not fixed in the human population (referred to as polymorphic LTR5_Hs loci) were identified using 1000 Genome Project datasets [52]. Information on the polymorphic LTR5_Hs loci is summarized in **S9 Table**. (A) Comparison of the polymorphic LTR5_Hs loci and the LTR5_Hs loci that are likely to regulate the gene expression in PGCLCs and naïve ESCs. The names of the overlapping LTR5_Hs loci are shown ("LTR5_Hs|chr11:72164373–72165341|+" and "LTR5_Hs|chr3:195927524–195928492|-"). (B) Geographical prevalence of the polymorphic LTR5_Hs loci in different human populations. Proportions of individuals with allele(s) lacking the LTR5_Hs insertion in different populations are shown. AFR, African; AMR, Ad Mixed American; EAS, East Asian; EUR, European; and SAS, South Asian. The map was generated using R maps (https://cran.r-project.org/web/packages/maps/index.html). (C) Expression levels of the genes associated with polymorphic LTR5_Hs in various cell types.
(TIF)

S13 Fig. A sensitivity analysis on the distance parameter for proximity definition. Protein-coding genes were classified into genes adjacent to LTR5_Hs or not according to the various distance thresholds for proximity definition. Subsequently, Log2 FC value (naïve ESCs vs. primed ESCs) and PGCLC-specific expression score were compared between the two gene categories.
(TIF)

S1 Table. Classification of protein-coding genes according to their expression patterns (related to Fig 1C).
(XLSX)

S2 Table. GO enrichment analysis results for the three gene categories (genes upregulated in both cell types, genes upregulated only in PGCLCs, and genes upregulated only in naïve ESCs) (related to Fig 1E).
(XLSX)

S3 Table. Identification of the potential regulators of LTR5_Hs in PGCLCs and naïve ESCs (related to Fig 4).
(XLSX)

S4 Table. Association of the expression patterns of genes and their distance from LTR5_Hs in the genome (related to Fig 5A).
(XLSX)

S5 Table. Results of GO enrichment analysis using the genes that are present nearby LTR5_Hs and upregulated in both PGCLCs and naïve ESCs (related to Fig 5C).
(XLSX)

S6 Table. Comparison of the genes upregulated in both PGCLCs/PGCs and naïve pluripotent cells between humans and macaques (related to Fig 6E).
(XLSX)

S7 Table. Comparison of genes regulated by LTR5_Hs defined in this study and a previous study (Fuentes et al.) [25].
(XLSX)

S8 Table. Information on respective LTR5 loci.
(XLSX)

S9 Table. LTR5_Hs loci that are present in the human reference genome (GRCh38) but not fixed in the human population.
(XLSX)

S10 Table. Sequencing dataset analyzed in the present study.
(XLSX)

S11 Table. LiftOver chain files used in the present study.
(XLSX)

Acknowledgments

We would like to thank Mai Suganami (Institute of Medical Science, The University of Tokyo, Japan) for technical supports; Junna Kawasaki (Institute for Frontier Life and Medical Sciences, Kyoto University, Japan) for thoughtful comments. The super-computing resource, SHIROKANE, was provided by Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan. The results shown here are in part based on data generated by the 1000 Genome Project (https://www.internationalgenome.org/about/).

Author Contributions

Conceptualization: Jumpei Ito, Kotaro Sasaki.

Data curation: Jumpei Ito, Yasunari Seita.

Formal analysis: Jumpei Ito, Shohei Kojima.

Funding acquisition: Jumpei Ito, Kotaro Sasaki, Kei Sato.

Investigation: Jumpei Ito, Yasunari Seita, Kotaro Sasaki.

Methodology: Jumpei Ito.

Resources: Yasunari Seita, Kotaro Sasaki.

Software: Jumpei Ito, Shohei Kojima.

Supervision: Nicholas F. Parrish, Kotaro Sasaki, Kei Sato.

Validation: Jumpei Ito, Yasunari Seita, Kotaro Sasaki, Kei Sato.

Visualization: Jumpei Ito, Yasunari Seita, Kei Sato.

Writing – original draft: Jumpei Ito.

Writing – review & editing: Jumpei Ito, Yasunari Seita, Shohei Kojima, Nicholas F. Parrish, Kotaro Sasaki, Kei Sato.

References

1. Saitou M, Yamaji M. Primordial germ cells in mice. Cold Spring Harb Perspect Biol. 2012; 4(11). Epub 2012/11/06. https://doi.org/10.1101/cshperspect.a008375 PMID: 23125014; PubMed Central PMCID: PMC3536339.
2. Tang WW, Kobayashi T, Irie N, Dietmann S, Surani MA. Specification and epigenetic programming of the human germ line. Nat Rev Genet. 2016; 17(10):585–600. Epub 2016/08/31. https://doi.org/10.1038/nrg.2016.88 PMID: 27573372.
3. Kobayashi T, Surani MA. On the origin of the human germline. Development. 2018; 145(16). Epub 2018/07/25. https://doi.org/10.1242/dev.150433 PMID: 30037844.
4. Sasaki K, Yokobayashi S, Nakamura T, Okamoto I, Yabuta Y, Kurimoto K, et al. Robust In Vitro Induction of Human Germ Cell Fate from Pluripotent Stem Cells. Cell Stem Cell. 2015; 17(2):178–94. Epub 2015/07/21. https://doi.org/10.1016/j.stem.2015.06.014 PMID: 26189426.
5. Irie N, Weinberger L, Tang WW, Kobayashi T, Viukov S, Manor YS, et al. SOX17 is a critical specifier of human primordial germ cell fate. Cell. 2015; 160(1–2):253–68. Epub 2014/12/30. https://doi.org/10.1016/j.cell.2014.12.013 PMID: 25543152; PubMed Central PMCID: PMC4310934.
6. Sasaki K, Nakamura T, Okamoto I, Yabuta Y, Iwatani C, Tsuchiya H, et al. The Germ Cell Fate of Cynomolgus Monkeys Is Specified in the Nascent Amnion. Dev Cell. 2016; 39(2):169–85. Epub 2016/10/26. https://doi.org/10.1016/j.devcel.2016.09.007 PMID: 27720607.
7. Tang WW, Dietmann S, Irie N, Leitch HG, Floros VI, Bradshaw CR, et al. A Unique Gene Regulatory Network Resets the Human Germline Epigenome for Development. Cell. 2015; 161(6):1453–67. Epub 2015/06/06. https://doi.org/10.1016/j.cell.2015.04.053 PMID: 26046444; PubMed Central PMCID: PMC4459712.
8. Li L, Dong J, Yan L, Yong J, Liu X, Hu Y, et al. Single-Cell RNA-Seq Analysis Maps Development of Human Germline Cells and Gonadal Niche Interactions. Cell Stem Cell. 2017; 20(6):858–73.e4. Epub 2017/05/02. https://doi.org/10.1016/j.stem.2017.03.007 PMID: 28457750.
9. Chen D, Sun N, Hou L, Kim R, Faith J, Aslanyan M, et al. Human Primordial Germ Cells Are Specified from Lineage-Primed Progenitors. Cell Rep. 2019; 29(13):4568–82.e5. Epub 2019/12/26. https://doi.org/10.1016/j.celrep.2019.11.083 PMID: 31875561; PubMed Central PMCID: PMC6939677.
10. Hwang YS, Suzuki S, Seita Y, Ito J, Sakata Y, Aso H, et al. Reconstitution of prospermatogonial specification in vitro from human induced pluripotent stem cells. Nat Commun. 2020; 11(1):5656. Epub 2020/11/11. https://doi.org/10.1038/s41467-020-19350-3 PMID: 33168808; PubMed Central PMCID: PMC7653920.
11. Yamashiro C, Sasaki K, Yabuta Y, Kojima Y, Nakamura T, Okamoto I, et al. Generation of human oogonia from induced pluripotent stem cells in vitro. Science. 2018; 362(6412):356–60. Epub 2018/09/22. https://doi.org/10.1126/science.aat1674 PMID: 30237246.
12. Kojima Y, Sasaki K, Yokobayashi S, Sakai Y, Nakamura T, Yabuta Y, et al. Evolutionarily Distinctive Transcriptional and Signaling Programs Drive Human Germ Cell Lineage Specification from Pluripotent Stem Cells. Cell Stem Cell. 2017; 21(4):517–32.e5. Epub 2017/10/07. https://doi.org/10.1016/j.stem.2017.09.005 PMID: 28985527.
13. Peter IS, Davidson EH. Evolution of gene regulatory networks controlling body plan development. Cell. 2011; 144(6):970–85. Epub 2011/03/19. https://doi.org/10.1016/j.cell.2011.02.017 PMID: 21414487; PubMed Central PMCID: PMC3076009.
14. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, et al. Enhancer evolution across 20 mammalian species. Cell. 2015; 160(3):554–66. Epub 2015/01/31. https://doi.org/10.1016/j.cell.2015.01.006 PMID: 25635462; PubMed Central PMCID: PMC4313353.
15. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 2017; 18(2):71–86. Epub 2016/11/22. https://doi.org/10.1038/nrg.2016.139 PMID: 27867194; PubMed Central PMCID: PMC5498291.

16. Jacques P, Jeyakani J, Bourque G. The majority of primate-specific regulatory sequences are derived from transposable elements. PLoS Genet. 2013; 9(5):e1003504. Epub 2013/05/16. https://doi.org/10.1371/journal.pgen.1003504 PMID: 23675311; PubMed Central PMCID: PMC3649963.

17. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. 2014; 24(12):1963–76. Epub 2014/10/17. https://doi.org/10.1101/gr.168872.113 PMID: 25319995; PubMed Central PMCID: PMC4248313.

18. Ito J, Sugimoto R, Nakaoka H, Yamada S, Kimura T, Hayano T, et al. Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. PLoS Genet. 2017; 13(7):e1006883. Epub 2017/07/13. https://doi.org/10.1371/journal.pgen.1006883 PMID: 28700586; PubMed Central PMCID: PMC5529029.

19. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 2008; 9(5):397–405. Epub 2008/03/28. https://doi.org/10.1038/nrg2337 PMID: 18368054; PubMed Central PMCID: PMC2596197.

20. Lynch VJ, Nnamani MC, Kapusta A, Brayer K, Plaza SL, Mazur EC, et al. Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. Cell Rep. 2015; 10(4):551–61. Epub 2015/02/03. https://doi.org/10.1016/j.celrep.2014.12.052 PMID: 25640180; PubMed Central PMCID: PMC4447085.

21. Carter T, Singh M, Dumbovic G, Chobirko JD, Rinn JL, Feschotte C. Mosaic cis-regulatory evolution drives transcriptional partitioning of HERVH endogenous retrovirus in the human embryo. Elife. 2022; 11. Epub 2022/02/19. https://doi.org/10.7554/eLife.76257 PMID: 35179489.

22. Pontis J, Planet E, Offner S, Turelli P, Duc J, Coudray A, et al. Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. Cell Stem Cell. 2019; 24(5):724–35.e5. Epub 2019/04/23. https://doi.org/10.1016/j.stem.2019.03.012 PMID: 31006620; PubMed Central PMCID: PMC6509360.

23. Theunissen TW, Friedli M, He Y, Planet E, O'Neil RC, Markoulaki S, et al. Molecular Criteria for Defining the Naive Human Pluripotent State. Cell Stem Cell. 2016; 19(4):502–15. Epub 2016/07/19. https://doi.org/10.1016/j.stem.2016.06.011 PMID: 27424783; PubMed Central PMCID: PMC5065525.

24. Grow EJ, Flynn RA, Chavez SL, Bayless NL, Wossidlo M, Wesche DJ, et al. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature. 2015; 522(7555):221–5. Epub 2015/04/22. https://doi.org/10.1038/nature14308 PMID: 25896322; PubMed Central PMCID: PMC4503379.

25. Fuentes DR, Swigut T, Wysocka J. Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. Elife. 2018;7. Epub 2018/08/03. https://doi.org/10.7554/eLife.35989 PMID: 30070637; PubMed Central PMCID: PMC6158008.

26. Sakashita A, Maezawa S, Takahashi K, Alavattam KG, Yukawa M, Hu YC, et al. Endogenous retroviruses drive species-specific germline transcriptomes in mammals. Nat Struct Mol Biol. 2020; 27(10):967–77. Epub 2020/09/09. https://doi.org/10.1038/s41594-020-0487-4 PMID: 32895553.

27. Percharde M, Lin CJ, Yin Y, Guan J, Peixoto GA, Bulut-Karslioglu A, et al. A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. Cell. 2018; 174(2):391–405.e19. Epub 2018/06/26. https://doi.org/10.1016/j.cell.2018.05.043 PMID: 29937225; PubMed Central PMCID: PMC6046266.

28. Leitch HG, Smith A. The mammalian germline as a pluripotency cycle. Development. 2013; 140(12):2495–501. Epub 2013/05/30. https://doi.org/10.1242/dev.091603 PMID: 23715543.

29. Chen D, Liu W, Zimmerman J, Pastor WA, Kim R, Hosohama L, et al. The TFAP2C-Regulated OCT4 Naive Enhancer Is Involved in Human Germline Formation. Cell Rep. 2018; 25(13):3591–602.e5. Epub 2018/12/28. https://doi.org/10.1016/j.celrep.2018.12.011 PMID: 30590035; PubMed Central PMCID: PMC6342560.

30. De Los Angeles A, Ferrari F, Xi R, Fujiwara Y, Benvenisty N, Deng H, et al. Hallmarks of pluripotency. Nature. 2015; 525(7570):469–78. Epub 2015/09/25. https://doi.org/10.1038/nature15515 PMID: 26399828.

31. Weinberger L, Ayyash M, Novershtern N, Hanna JH. Dynamic stem cell states: naive to primed pluripotency in rodents and humans. Nat Rev Mol Cell Biol. 2016; 17(3):155–69. Epub 2016/02/11. https://doi.org/10.1038/nrm.2015.28 PMID: 26860365.

32. Nichols J, Smith A. Naive and primed pluripotent states. Cell Stem Cell. 2009; 4(6):487–92. Epub 2009/06/06. https://doi.org/10.1016/j.stem.2009.05.015 PMID: 19497275.

33. Takashima Y, Guo G, Loos R, Nichols J, Ficz G, Krueger F, et al. Resetting transcription factor control circuitry toward ground-state pluripotency in human. Cell. 2014; 158(6):1254–69. Epub 2014/09/13. https://doi.org/10.1016/j.cell.2014.08.029 PMID: 25215486; PubMed Central PMCID: PMC4162745.

34. Pastor WA, Liu W, Chen D, Ho J, Kim R, Hunt TJ, et al. TFAP2C regulates transcription in human naive pluripotency by opening enhancers. Nat Cell Biol. 2018; 20(5):553–64. Epub 2018/04/27. https://doi.org/10.1038/s41556-018-0089-0 PMID: 29695788; PubMed Central PMCID: PMC5926822.

35. von Meyenn F, Berrens RV, Andrews S, Santos F, Collier AJ, Krueger F, et al. Comparative Principles of DNA Methylation Reprogramming during Human and Mouse In Vitro Primordial Germ Cell Specification. Dev Cell. 2016; 39(1):104–15. Epub 2016/10/12. https://doi.org/10.1016/j.devcel.2016.09.015 PMID: 27728778; PubMed Central PMCID: PMC5064768.

36. Messmer T, von Meyenn F, Savino A, Santos F, Mohammed H, Lun ATL, et al. Transcriptional Heterogeneity in Naive and Primed Human Pluripotent Stem Cells at Single-Cell Resolution. Cell Rep. 2019; 26(4):815–24.e4. Epub 2019/01/24. https://doi.org/10.1016/j.celrep.2018.12.099 PMID: 30673604; PubMed Central PMCID: PMC6344340.

37. Hayashi Y, Otsuka K, Ebina M, Igarashi K, Takehara A, Matsumoto M, et al. Distinct requirements for energy metabolism in mouse primordial germ cells and their reprogramming to embryonic germ cells. Proc Natl Acad Sci U S A. 2017; 114(31):8289–94. Epub 2017/07/19. https://doi.org/10.1073/pnas.1620915114 PMID: 28716939; PubMed Central PMCID: PMC5547595.

38. Tischler J, Gruhn WH, Reid J, Allgeyer E, Buettner F, Marr C, et al. Metabolic regulation of pluripotency and germ cell fate through α-ketoglutarate. Embo j. 2019; 38(1). Epub 2018/09/28. https://doi.org/10.15252/embj.201899518 PMID: 30257965; PubMed Central PMCID: PMC6315289.

39. Yevshin I, Sharipov R, Kolmykov S, Kondrakhin Y, Kolpakov F. GTRD: a database on gene transcription regulation-2019 update. Nucleic Acids Res. 2019; 47(D1):D100–d5. Epub 2018/11/18. https://doi.org/10.1093/nar/gky1128 PMID: 30445619; PubMed Central PMCID: PMC6323985.

40. Kim TK, Hemberg M, Gray JM. Enhancer RNAs: a class of long noncoding RNAs synthesized at enhancers. Cold Spring Harb Perspect Biol. 2015; 7(1):a018622. Epub 2015/01/07. https://doi.org/10.1101/cshperspect.a018622 PMID: 25561718; PubMed Central PMCID: PMC4292161.

41. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods. 2017; 14(10):979–82. Epub 2017/08/22. https://doi.org/10.1038/nmeth.4402 PMID: 28825705; PubMed Central PMCID: PMC5764547.

42. Hancks DC, Kazazian HH Jr. SVA retrotransposons: Evolution and genetic instability. Semin Cancer Biol. 2010; 20(4):234–45. Epub 2010/04/27. https://doi.org/10.1016/j.semcancer.2010.04.001 PMID: 20416380; PubMed Central PMCID: PMC2945828.

43. Ema M, Mori D, Niwa H, Hasegawa Y, Yamanaka Y, Hitoshi S, et al. Krüppel-like factor 5 is essential for blastocyst development and the normal self-renewal of mouse ESCs. Cell Stem Cell. 2008; 3(5):555–67. Epub 2008/11/06. https://doi.org/10.1016/j.stem.2008.09.003 PMID: 18983969.

44. Yamane M, Ohtsuka S, Matsuura K, Nakamura A, Niwa H. Overlapping functions of Krüppel-like factor family members: targeting multiple transcription factors to maintain the naïve pluripotency of mouse embryonic stem cells. Development. 2018;145(10). Epub 2018/05/10. https://doi.org/10.1242/dev.162404 PMID: 29739838.

45. Tu S, Narendra V, Yamaji M, Vidal SE, Rojas LA, Wang X, et al. Co-repressor CBFA2T2 regulates pluripotency and germline development. Nature. 2016; 534(7607):387–90. Epub 2016/06/10. https://doi.org/10.1038/nature18004 PMID: 27281218; PubMed Central PMCID: PMC4911307.

46. Bayerl J, Ayyash M, Shani T, Manor YS, Gafni O, Massarwa R, et al. Principles of signaling pathway modulation for enhancing human naive pluripotency induction. Cell Stem Cell. 2021; 28(9):1549–65.e12. Epub 2021/04/30. https://doi.org/10.1016/j.stem.2021.04.001 PMID: 33915080; PubMed Central PMCID: PMC8423434.

47. Jostes SV, Fellermeyer M, Arévalo L, Merges GE, Kristiansen G, Nettersheim D, et al. Unique and redundant roles of SOX2 and SOX17 in regulating the germ cell tumor fate. Int J Cancer. 2020; 146(6):1592–605. Epub 2019/10/05. https://doi.org/10.1002/ijc.32714 PMID: 31583686.

48. Yan L, Yang M, Guo H, Yang L, Wu J, Li R, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol. 2013; 20(9):1131–9. Epub 2013/08/13. https://doi.org/10.1038/nsmb.2660 PMID: 23934149.

49. Xue Z, Huang K, Cai C, Cai L, Jiang CY, Feng Y, et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature. 2013; 500(7464):593–7. Epub 2013/07/31. https://doi.org/10.1038/nature12364 PMID: 23892778; PubMed Central PMCID: PMC4950944.

50. Petropoulos S, Edsgärd D, Reinius B, Deng Q, Panula SP, Codeluppi S, et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. Cell. 2016; 165(4):1012–26. Epub 2016/04/12. https://doi.org/10.1016/j.cell.2016.03.023 PMID: 27062923; PubMed Central PMCID: PMC4868821.

51. Wildschutte JH, Williams ZH, Montesion M, Subramanian RP, Kidd JM, Coffin JM. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proc Natl Acad Sci U S A. 2016; 113

(16):E2326–34. Epub 2016/03/24. https://doi.org/10.1073/pnas.1602336113 PMID: 27001843; PubMed Central PMCID: PMC4843416.

52. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015; 526(7571):68–74. Epub 2015/10/04. https://doi.org/10.1038/nature15393 PMID: 26432245; PubMed Central PMCID: PMC4750478.

53. Mathieu J, Ruohola-Baker H. Metabolic remodeling during the loss and acquisition of pluripotency. Development. 2017; 144(4):541–51. Epub 2017/02/16. https://doi.org/10.1242/dev.128389 PMID: 28196802; PubMed Central PMCID: PMC5312031.

54. Zhou W, Choi M, Margineantu D, Margaretha L, Hesson J, Cavanaugh C, et al. HIF1α induced switch from bivalent to exclusively glycolytic metabolism during ESC-to-EpiSC/hESC transition. Embo j. 2012; 31(9):2103–16. Epub 2012/03/27. https://doi.org/10.1038/emboj.2012.71 PMID: 22446391; PubMed Central PMCID: PMC3343469.

55. Gu W, Gaeta X, Sahakyan A, Chan AB, Hong CS, Kim R, et al. Glycolytic Metabolism Plays a Functional Role in Regulating Human Pluripotent Stem Cell State. Cell Stem Cell. 2016; 19(4):476–90. Epub 2016/09/13. https://doi.org/10.1016/j.stem.2016.08.008 PMID: 27618217; PubMed Central PMCID: PMC5055460.

56. Shiozawa S, Nakajima M, Okahara J, Kuortaki Y, Kisa F, Yoshimatsu S, et al. Primed to Naive-Like Conversion of the Common Marmoset Embryonic Stem Cells. Stem Cells Dev. 2020; 29(12):761–73. Epub 2020/03/20. https://doi.org/10.1089/scd.2019.0259 PMID: 32188344.

57. Xiang X, Tao Y, DiRusso J, Hsu FM, Zhang J, Xue Z, et al. Human reproduction is regulated by retrotransposons derived from ancient Hominidae-specific viral infections. Nat Commun. 2022; 13(1):463. Epub 2022/01/26. https://doi.org/10.1038/s41467-022-28105-1 PMID: 35075135; PubMed Central PMCID: PMC8786967.

58. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019; 47(D1):D766–d73. Epub 2018/10/26. https://doi.org/10.1093/nar/gky955 PMID: 30357393; PubMed Central PMCID: PMC6323946.

59. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30(15):2114–20. Epub 2014/04/04. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404; PubMed Central PMCID: PMC4103590.

60. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29(1):15–21. Epub 2012/10/30. https://doi.org/10.1093/bioinformatics/bts635 PMID: 23104886; PubMed Central PMCID: PMC3530905.

61. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014; 30(7):923–30. Epub 2013/11/15. https://doi.org/10.1093/bioinformatics/btt656 PMID: 24227677.

62. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019; 177(7):1888–902.e21. Epub 2019/06/11. https://doi.org/10.1016/j.cell.2019.05.031 PMID: 31178118; PubMed Central PMCID: PMC6687398.

63. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019; 20(1):296. Epub 2019/12/25. https://doi.org/10.1186/s13059-019-1874-1 PMID: 31870423; PubMed Central PMCID: PMC6927181.

64. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15(12):550. Epub 2014/12/18. https://doi.org/10.1186/s13059-014-0550-8 PMID: 25516281; PubMed Central PMCID: PMC4302049.

65. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14):1754–60. Epub 2009/05/20. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168; PubMed Central PMCID: PMC2705234.

66. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078–9. Epub 2009/06/10. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943; PubMed Central PMCID: PMC2723002.

67. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26(6):841–2. Epub 2010/01/30. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278; PubMed Central PMCID: PMC2832824.

68. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010; 28(5):495–501. Epub 2010/05/04. https://doi.org/10.1038/nbt.1630 PMID: 20436461; PubMed Central PMCID: PMC4840234.

69. Ito J, Kimura I, Soper A, Coudray A, Koyanagi Y, Nakaoka H, et al. Endogenous retroviruses drive KRAB zinc-finger protein family expression for tumor suppression. Sci Adv. 2020; 6(43). Epub 2020/10/23. https://doi.org/10.1126/sciadv.abc3020 PMID: 33087347; PubMed Central PMCID: PMC7577720.

70. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30(4):772–80. Epub 2013/01/19. https://doi.org/10.1093/molbev/mst010 PMID: 23329690; PubMed Central PMCID: PMC3603318.

71. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30(9):1312–3. Epub 2014/01/24. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623; PubMed Central PMCID: PMC3998144.

72. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019; 47(D1):D607–d13. Epub 2018/11/27. https://doi.org/10.1093/nar/gky1131 PMID: 30476243; PubMed Central PMCID: PMC6323986.

73. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016; 32(18):2847–9. Epub 2016/05/22. https://doi.org/10.1093/bioinformatics/btw313 PMID: 27207943.

74. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol. 2017; 34(7):1812–9. Epub 2017/04/08. https://doi.org/10.1093/molbev/msx116 PMID: 28387841.