Title

Transposable Elements and their Epigenetic Regulators Are Necessary to Orchestrate the Transcriptome of the Developing Germline

Permalink

Author

DiRusso, Jonathan Adam

Publication Date

2025

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Transposable Elements and their Epigenetic Regulators Are Necessary to Orchestrate the

Transcriptome of the Developing Germline

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of

Philosophy in Molecular Biology

by

Jonathan Adam DiRusso

2025

ABSTRACT OF THE DISSERTATION

Transposable Elements and their Epigenetic Regulators Are Necessary to Orchestrate the

Transcriptome of the Developing Germline

by

Jonathan Adam DiRusso

Doctor of Philosophy in Molecular Biology

University of California, Los Angeles, 2025

Professor Amander Therese Clark, Chair

In order to maintain organismal fitness, the germline must be able to transmit high-fidelity genomic information from one generation to the next. Transposable Elements (TEs), genomic elements which are capable of mobility within the genome, pose a risk to this process. Within a given species, only a small subset of TEs remain mobile. These mobile elements are often evolutionarily young compared to others in the genome and must be durably silenced during germline development so that their mobility does not impact fitness of the offspring. Conversely, evolutionarily older TEs lose their ability to mobilize due to deterioration or loss of pro-viral regions, which encode viral proteins. In tandem with their deterioration, TEs are often "endogenized" or "domesticated" and become cis-regulatory elements, for instance by harboring binding motifs of transcription factors, repressors, or insulators. As a result, a conserved property of these TEs is their capacity to embed themselves into the cis-regulatory network of tissues, especially in the early embryo.

Primordial Germ Cells (PGCs) are embryonic precursors to the adult germline. In mammals, PGCs are specified early in embryonic development (at embryonic day (E) E6.25 in

mice and ~E11-12 in humans). The transcriptional networks that drive and reinforce acquisition of PGC identity have been well interrogated using *in vitro* models of PGC specification in the mouse (m) and human (h) via generation of PGC-Like Cells (PGCLCs). Here we show that an evolutionarily young TE subfamily, LTR5Hs, is epigenetically remodeled during human PGC specification and bound by critical human PGC factors, supporting the hypothesis these elements act as enhancers. We go on to show that ectopic epigenetic repression of LTR5Hs results in inefficient human PGCLC induction, thus establishing enhancer activity of LTR5Hs as necessary for human PGC specification. These findings demonstrate that LTR5Hs is necessary for specification of the human germline.

After specification, PGCs will migrate while undergoing epigenetic reprogramming including DNA demethylation and imprint erasure. In the mouse, PGCs migrate into the developing genital ridge, undergo rapid mitotic proliferation and differentiate into either testicular germ cells, which become pro-spermatogonia, or into meiotic germ cells which will undergo meiosis and mature into oocytes. To ask how control of TEs impacts a later stages of PGC development, differentiation, we used an *in vivo* mouse model and employed PGC-specific conditional knockout of TRIM28. TRIM28 is an epigenetic scaffold necessary to repress many TEs, especially those which are evolutionarily young. We found that TRIM28 is regulated in a sex-specific manner and that TRIM28 loss in PGCs during differentiation results in upregulation genes typical of Zygotic Genome Activation, which in the mouse happens during the 2C stage of embryonic development. Upregulation of a 2C-like transcriptome results in reduced mitotic expansion, inefficient fate restriction to the adult germline, and a failure of PGCs to differentiate in males and to properly progress through meiosis in females. Thus, precise control of transposable elements is important not only for germline specification, but also to protect the germline transcriptional program as PGCs differentiate.

The dissertation of Jonathan Adam DiRusso is approved.

Patrick Allard

Hilary Ann Coller

Jeffrey Aaron Long

Margot Elizabeth Quinlan

Amander Therese Clark, Committee Chair

University of California, Los Angeles

2025

Dedication

To my family, husband, friends and chosen family, I could not have done without your love and

encouragement. To my past mentors, thank you for allowing me the room to try, to fail and to

learn. This would not be possible without your generosity, patience, tutorage and friendship. To

everyone who has helped me along this journey, thank you.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Throughout my PhD, I have been lucky to have been supported by a wonderful mentor, enthusiastic lab mates, and countless support staff. I would like to thank my thesis advisor Dr. Amander Clark for her guidance, support, and patience with a project that provided innumerable twists and turns while allowing me to explore some of my favorite corners of developmental biology. Dr. Clark's commitment to teaching, mentorship and supporting her students allowed me to thrive during my Ph. D., for which I will always be thankful. I would also like to thank my undergraduate mentor, Michele Markstein, for teaching me so much of what it means to work at the bench and instilling a love of experimental science and exploration. Without such incredible mentors past and present, I would not have achieved all that I have been able to.

I would also like to thank the Clark lab for their support, input and for making sure no day in lab was ever boring or uneventful. A special thank you to Matthew Lowe, who took me under his wing and patiently taught me mouse husbandry and embryonic dissections from a COVID-safe 6 feet away, and who provided countless hours of insight and support as I began this project. Thank you to Erica Pandolfi, who taught me how to culture human embryonic stem cells and *in vitro* production of hPGCLCs. Thank you to Timothy Hunt, who taught me how to make medias, culture mouse pluripotent stem cells and perform FACS sorts. Thank you to Fei-Man Hsu, who taught me how to analyze RNA-seq data and put me on a path to be able to perform many of the analyses in this thesis. Thank you to Sissy Wamaitha for teaching me immunofluorescence, countless suggestions on data analysis and many, many productive conversations on how to frame my research. A special thanks to Enrique for helping me get onboarded to the microscopy cores during COVID and to my desk mate Youngsun for all of his scientific and life advice.

To all my fellow graduate students, thank you so much for being such a wonderfully fun, creative and supporting group. A special thanks to Auriana Arabpour for keeping me smiling during long nights and early mornings, for countless coffee runs so I could find out "if this makes

home area director and department chair. To Margot Quinlan, for an amazingly fun and different rotation, and for your personal and scientific support during my Ph. D. To Patrick Allard for insightful comments, feedback and support on how to present and package my findings.

Last but not least I would like to thank my family and friends. To my husband Christopher for always supporting me through late nights and weekends, I would not have been able to do this without you. To my Mom and Dad, for their unconditional love and support throughout this while process. To Beth and my Mom, thank you for taking such good care of my Dad. You have made this possible in more ways than you will ever know.

It takes a village to raise a child, and the same is true for Ph. Ds. Thank you to everyone who has been on this journey with me.

VITA

Education:

University of Massachusetts, Amherst
B.S., Biochemistry and Molecular Biology

Amherst, MA
06/2017

University of California, Los Angeles
Ph.D., Molecular Biology

Los Angeles, CA
Expected 01/2025

Research Experience:

University of California, Los Angeles
Graduate Student Researcher, Dept. of Molecular, Cell and Dev. Biol.
Advisor: Amander T. Clark, Ph. D.

2019 – Present

Harvard Medical School
Visiting Post-Graduate Fellow in Genetics, Dept. of Genetics
Advisor: Norbert Perrimon, Ph. D.

2018-2019

University of Massachusetts, Amherst
Researcher, Dept. of Biology
Advisor: Michele Markstein, Ph. D.

2013-2018

Publications

DiRusso, J.A., Zhan, L., Tao, Y., Wang, A.L., Xiang, X., Robbins, A.C., Liu, W., Clark, A.T., TRIM28-mediated reprogramming of transposable elements safeguards PGC differentiation. *Under Review, Dev. Cell*, June 2024

Sousa, E., Mumu, S.K., Alvarado, C.C., Wu, Q.Y., Roberson, I., Espinoza, A., Hsu, F., Saito, K., Hunt, T.J., Faith, J., Lowe, M.G., **DiRusso, J.A.**, Clark, A.T. Reconstituted ovaries self-assemble without an ovarian surface epithelium. *Stem Cell Reports*, November 2023

DiRusso, J.A and Clark, A.T. Transposable elements in early human embryo development and embryo models. *Curr. Opinions in Genetics and Development.*, August 2023

Xiang, X.*, Tao, Y.*, **DiRusso, J.A.**, Hsu, F.M., Zhang, J., Xue, Z., Pontis, J., Trono, D., Liu, W., Clark, A.T. Human reproduction is regulated by retrotransposons derived from ancient Hominidae-specific viral infections. *Nature Comm.*, Jan. 2022 *co-first author

Selected Presentations:

DiRusso, JA., Zhan, L., Wang, A., Robbins, A., Tao, Y., Xiang, X., Liu, W., Clark, AT. "TRIM28 Regulation of LTR Elements Safeguards The PGC Epigenome During Sex Determination" (July 2024) Poster at the Society for the Study of Sexual Reproduction Annual Meeting, Dubin, Ireland

DiRusso, JA., Zhan, L., Wang, A., Robbins, A., Tao, Y., Xiang, X., Liu, W., Clark, AT. "TRIM28 Regulation of LTR Elements Safeguards The PGC Epigenome During Sex Determination" (July 2024) Poster at the International Society for Stem Cell Research Annual Meeting, Hamburg Germany

DiRusso, JA., Zhan, L., Wang, A., Robbins, A., Tao, Y., Xiang, X., Liu, W., Clark, AT. "Regulation of Endogenous Retroviruses by TRIM28 is Necessary to License Primordial Germ Cells for gametogenesis" (June 2023) Invited talk at the International Society for Stem Cell Research Annual Conference, Boston, Massachusetts, USA

DiRusso, JA., Wang, A., Robbins, A., Tao, Y., Xiang, X., Liu, W., Clark, AT. "Regulation of Endogenous Retroviruses by TRIM28 is Necessary to License Primordial Germ Cells for gametogenesis" (May, 2023) Invited seminar at the UCLA Molecular Biology Institute Cell and Developmental Biology Club

DiRusso, JA., Wang, A., Tao, Y., Xiang, X., Liu, W., Clark, AT. "Epigenetic Regulation Of Transposable Elements By TRIM28 Safeguards Primordial Germ Cell Differentiation" (September 2022) Poster at the Cold Spring Harbor Epigenetics Meeting

Selected Honors and Awards:

ISSCR Travel Award	July 2024
ISSCR Merit Abstract Award	July 2024
NIH Pre-Doctoral Fellowship	July 2023 – January 2025
F31 HD113346	
UCLA Molecular Biology Institute Whitcome Fellowship	July 2021 – April 2023
National Science Foundation Graduate Research Fellowship Program (NSF GRFP): Honorable Mention	March 2021
UCLA Dean's Scholar Award	September 2019

Chapter 1

Introduction

Introduction to Transposable Elements:

Transposable elements (TEs) are mobile DNA elements which are nearly ubiquitous across eukaryotes. Just a few species have been identified which completely lack transposable elements, all of which are intracellular parasites[1]. Despite their ubiquity, the exact function and the evolutionary interplay between TEs and their host species remains under active investigation. Broadly, TEs are categorized into two classes, Class I, RNA retrotransposons, and Class II, DNA transposons. Class I transposons transpose through an RNA intermediate which is reverse-transcribed and re-inserted in a new location by integrases into the host genome. Class II transposons are excised from the genome and re-inserted into a new location without duplication. As a result, Class I TEs move via a "copy and paste" mechanisms while Class II TEs move via a "cut and paste" mechanism, with a few notable exceptions such as the *Helitron* subclass[2].

Within the class *Mammalia*, the majority of TEs are Class I retrotransposons[3] (Figure 1-1). Broadly, retrotransposons can be broken into two subclasses, Long Terminal Repeat (LTR)- and non-LTR retrotransposons. Among the non-LTR retrotransposons are Long Interspersed Nuclear Element (LINE) and short interspersed nuclear element (SINE) superfamilies. LINE elements are autonomous and encode the machinery they need for transposition, while SINEs are non-autonomous and rely on LINE-encoded machinery for their transposition[2]. Full-length Long Terminal Repeat (LTR)-subclass elements are likewise autonomous. Included within the LTR subclass are Endogenous Retroviruses (ERVs), which are found in two configurations within the genome. Some ERVs are full-length elements, with flanking LTRs and the associated internal pro-virus, which comprises the viral genome[1,2]. Although some ERVs have remnants of a pro-viral genome, they often degrade or undergo recombination over evolutionary time, leading to loss of Open Reading Frames (ORFs) and therefore lose the capacity to produce viral proteins. ERVs are also found as solo-LTRs, having lost their pro-viral region through

recombination of the LTR sequences which flank full-length ERVs[3]. Solo-LTRs account for most

ERV elements in the genome, and similarly degrade over time.
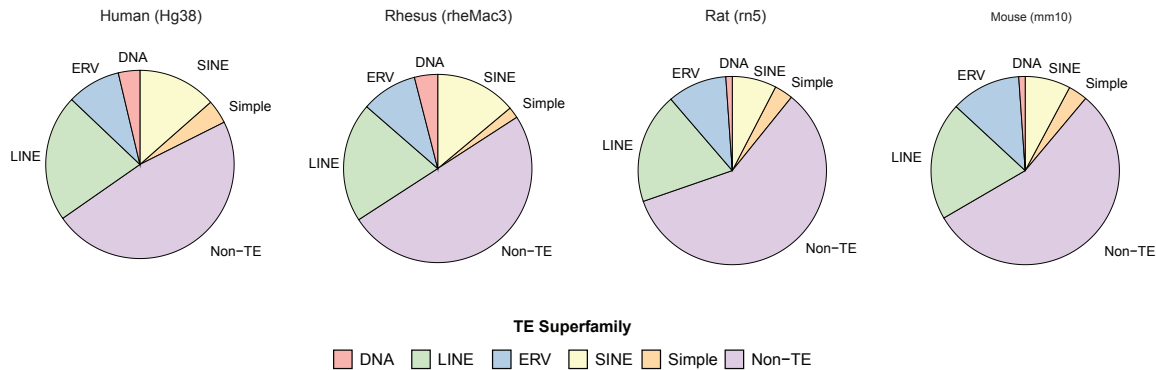
Figure 1-1: Relative abundance of transposable elements in commonly studied mammalian genomes. Genome assembly from which statistics were compiled in paratheses. Data from repeatmasker, accessed January 2025.

Within *Mammalia*, the divergence between *Monotremata* and *Theria*, the latter of which

includes placental mammals, is marked by a rapid expansion of ERV elements[3,4]. This is of

particular interest, as ERV-derived genes have been implicated in placental function, including

co-option of *syncytin* genes in mice and PEG-family genes more broadly across mammals[5,6].

The co-option of these ERV-derived genes suggests that TEs can also drive novel function

either by delivering novel genes or by acting as regulatory elements[7-13]. Therefore, despite the

capacity of TEs to be deleterious, they can also contribute positively to host fitness. As such,

they deserve careful consideration within the germline, a platform for heritable genomic novelty

which exists under exquisite evolutionary pressure. Broadly, this thesis investigates the

contributions of TEs and their regulators to organismal fitness. These are investigated using

both an *in vitro* model of human PGC specification as well as an *in vivo* mouse model allowing

study of PGC differentiation and germline maturation.

Regulation of Transposable Elements:

Given the capacity of transposable elements to disrupt genomic fidelity by insertion into

coding regions or by driving improper recombination during meiosis, control of TEs in the

germline is of critical importance[14,15]. As such, mammals have developed an expansive

repertoire of epigenetic machinery to silence transposable elements. Although this thesis will

focus on mechanisms which regulate TEs during embryonic and germline development, some

of these mechanisms are employed in somatic cells as well.

Both DNA methylation and TRIM28 (a.k.a KAP1)/KRAB-ZFP-mediated repression of TEs

have been well studied in Pluripotent Stem Cells (PSCs) and in Primordial Germ Cells (PGCs).

DNA methylation involves addition of a methyl group to the fifth prime carbon of cytosine, thus

generating 5-methylcytosine (5mC). The conversion of cytosine to 5mC is catalyzed by a family

of proteins called DNA Methyltransferases (DNMTs). In the mouse and human

Dnmt3a/*DNMT3A* and *Dnmt3b*/*DNMT3B* are *de novo* methyltransferases, which interact with a

non-catalytic DNMT, *Dnmt3l*/*DNMT3L*, to establish 5mC. A third *de novo* DNMT, *Dnmt3c,* has

been identified in the murine germline and specifically methylates transposable elements.

Collectively, the activity of the *de novo* DNMTs is required to re-establish imprints and to silence

transposable elements.[15-19] A third DNMT, *Dnmt1*/*DNMT1*, is a maintenance methyltransferase

which acts on hemimethylated DNA, which is created during DNA replication. The maintenance

DNA methylation activity of *Dnmt1*/*DNMT1* is facilitated by UHRF1, which binds hemimethylated

DNA, thus targeting DNMT1 to these sites. While DNMT1 does possess some weak *de novo*

methylation activity, it seems to be targeted only to ERVs and at loci which are likewise bound

by TRIM28 and enriched in H3K9me3.[20]

While *de novo* methylation by DNMT3A and DNMT3B is normally targeted to loci by

DNMT3L, DNMTs 1, 3A and 3B can be targeted to loci by *Trim28*/*TRIM28* as well. Interactions

between DNMTs and TRIM28 are especially important for imprinting genes during early

embryogenesis in the mouse, although whether that role is reprised in the human embryo

remains unknown in part due to the scarcity of human embryos consented for research [21-23]. It is

possible this between relationship TRIM28 and DNMTs exists in PGCs as well. In the mouse,

PGCs downregulate *Uhrf1* during DNA demethylation, which severely reduces *Dnmt1*

maintenance DNA methylation. Despite this, *Dnmt1* maintains DNA methylation at young ERVs and imprinted genes, as demonstrated by a PGC-specific knockout of DNMT1[18]. As *de novo* activity of DNMT1 has been shown to co-localize in the genome with *Trim28* binding[20], it is possible *Trim28* likewise directs *Dnmt1* to targeted elements in PGCs.

Independent of DNMTs, *Trim28/TRIM28* also complexes with *Setdb1/SETDB1*, a lysine histone methyltransferase which catalyzes H3K9me3[24]. TRIM28 itself lacks a DNA-binding domain and is recruited to targeted loci by Kruppel-Associated Box Zinc Finger Proteins (KRAB-ZFPs) (Figure 1-2). KRAB-ZFPs are an extensive family of proteins found in all known tetrapods[25,26]. While some KRAB-ZFPs are highly conserved, most species have undergone extensive species-specific KRAB-ZFP expansion which regulate species-specific TEs and genomic elements[26]. While *Trim28/TRIM28* is essential for TE repression, the regulation of KRABs is equally important, and their expression changes rapidly during development of the embryo and germline, reflecting their central role in ensuring proper spaciotemporal silencing of transposable elmenets[27]. In chapter 4, we demonstrate that there are multiple programs controlling TE accessibility in the murine germline, and that *Trim28* maintains repression at young TEs, but also represses an aberrant transcriptional state capable of de-railing mouse PGC differentiation. In brief, these results suggest that TRIM28-mediated repression is necessary not only for control of TEs but also to maintain the PGC transcriptome at a time when it is exceptionally plastic.
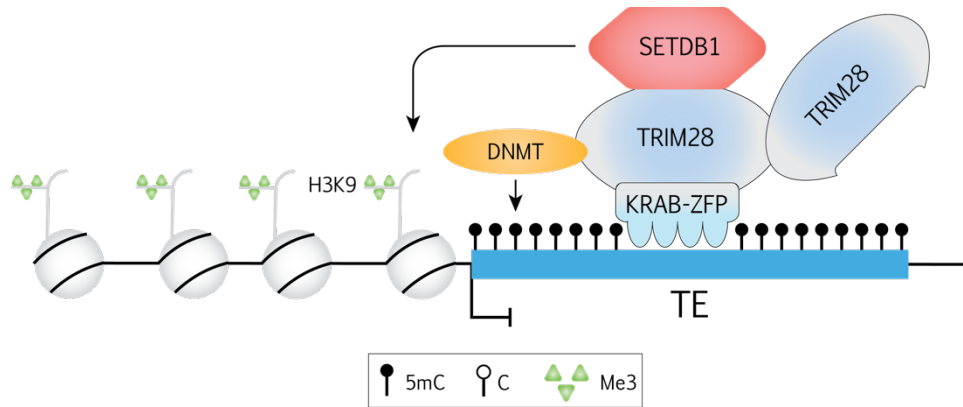
Figure 1-2: Schematic of a TRIM28 dimer interacting with a KRAB-ZFP, SETDB1 and DNMTs at the locus of a targeted TE.

Finally, the male and female germline likewise have mechanisms of TE control which are specific to the germline. In pro-spermatogonia, the piRNA system represses TEs and is essential to properly re-acquire genome-wide DNA methylation[28-30]. The contribution of the piRNA system is best explored in the mouse, as it is possible to collect ample embryonic and neo-natal germ cells as they undergo *de novo* methylation (beginning roughly E15.5). In mouse pro-spermatogonia, piRNAs are transcribed from piRNA clusters, after which they are processed by the Argonaut family proteins *Piwil2* (MILI) and *Piwil4* (MIWI2). After piRNA processing by MILI, piRNAs are loaded on to MIWI2, which translocates to the nucleus and binds the pre-processed mRNA of actively transcribed targets[31,32]. After binding to the nascent transcript, MIWI2 directs DNA methylation at the transcribed loci. Interestingly, knockout of either MIWI2 or MILI leads to the derepression of specific subfamilies of TEs, suggesting that they may have yet unresolved biases or interacting partners that produce these biases[29]. In the female, it has been hypothesized that TEs in the growing oocyte are targeted by an endo-siRNA mechanism. One such candidate for this activity is *Ago2* loss of which disrupts meiosis and results in spindle defects and TE derepression[33]. Still, whether this mechanism is alone sufficient for TE repression in the oocyte and whether *Ago2* has other functions and merely targets TEs indirectly is unresolved.

Transposable Elements as Regulatory Elements:

While much work on TEs, especially in the germline, is focused on their repression, it has also been shown that TEs have the capacity to contribute to cis-regulatory networks. In particular, ERVs are highly enriched for transcription factor binding motifs, and functional validation using gene editing or ectopic heterochromatic silencing have demonstrated that these elements are functional, bona fide cis-regulatory units[7,8,10,34]. Indeed, in both the mouse and human, a significant fraction of NANOG, OCT4 and CTCF are bound to transposable elements, with many bound to ERVs, in pluripotent stem cells (PSCs)[7]. Following this discovery, TEs which marked early embryonic states, including the ICM, pluripotent epiblast, and trophectoderm were identified in both mouse and human[10,11,35-38] (Figure 1-3). An important caveat is that these studies rely on fluctuations in TE expression across the genome and are often tested by silencing elements of entire subfamilies without regard for specific loci. As TE subfamilies often have hundreds or thousands of integrants across the genome, identifying which specific integrants function as cis-regulatory elements remains an open challenge.

In mouse PSCs, RLTR9E and RLTR13D6 elements were shown to have enhancer-like properties, including H3K27ac enrichment, enrichment of NANOG, SOX2 and OCT4 binding motifs and the capacity to drive gene expression as measured by a synthetic luciferase assay[12,13]. Later, elegant experiments demonstrated that while a number of RLTR13D6 elements had enhancer-like properties, they were dispensable for self-renewal and silencing them resulted in little difference in gene expression in mPSCs[35]. Thus, it is also possible for TEs to be remodeled as if they were cis-regulatory elements, but to lack bona fide regulatory activity. While evidence of TEs acting as necessary enhancers during development is lacking in the mouse, numerous studies have shown the human embryo is highly reliant on TEs as cis-regulatory elements during embryogenesis.

In the human, cis-regulatory TEs are required beginning at Zygotic Genome Activation, which in humans occurs during the 8-cell (8C) stage of development. Experiments have shown

that SINE SVA elements are remodeled to lose H3K9me3 and gain accessibility as the embryo

enters the 8C stage. If this process is blocked, the embryo stalls and there is a failure to

develop, establishing TEs as cis-regulatory elements with bona-fide function in human

embryos[39]. In Chapter 3, we add yet another example of cis-regulatory TEs in the human

embryo, where we show that a *Hominidae* -specific ERV, LTR5Hs, is required for specification of

the germline. Indeed, a similar approach where ectopic H3K9me3 repression of LTR5Hs is

induced by CRISPRi caused reduced PGC-Like Cell specification, thus demonstrating LTR5Hs

is a bona fide enhancer and cis-regulatory element for PGC specification.
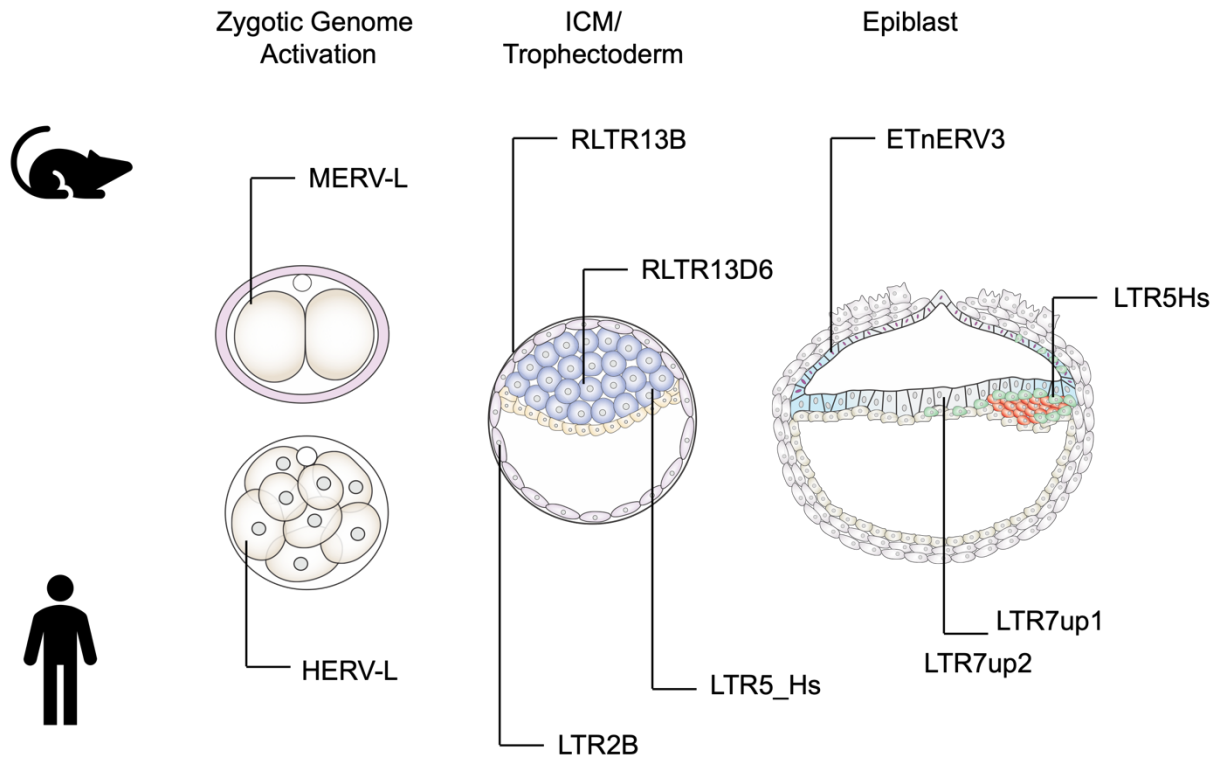
Figure 1-3: Examples of TEs which are highly expressed in embryonic tissues regardless of cis-regulatory function across early embryogenesis in mice (top) and humans (bottom).

Introduction to PGC Specification:

In both the human (h) and the mouse (m), PGC specification occurs early in embryonic

development. In the mouse, PGCs are specified in the post-implantation embryo beginning at

day E6.25, with the full foundry population of PGCs specified by E7.5[40,41]. In humans, the

earliest PGCs have been identified between ~E11 and E13 in cultured human embryos and in

histological analysis of rare human embryos[42,43]. Despite the conservation of much of the PGC

program, the signals and transcriptional networks that specify PGCs diverge between the

mouse and the human. In the mouse, *Bmp4* and *Bmp8a* are produced by the extra-embryonic

ectoderm (ExE) and *Bmp2* from the visceral endoderm[44-48]. Notably, primate embryos are not an

"egg cylinder" like the mouse but rather form a bilaminar disc which lack an extra-embryonic

ectoderm. While the inductive role of BMP4 in PGC specification is maintained in the human,

the source of BMP4 during PGC specification remains an open question, as scarcity of human

embryos and limited tools which can be employed makes understanding human PGC induction

in vivo difficult.

   To overcome these challenges posed by the rare nature of embryonic PGCs in the

human, *in vitro* approaches have been developed to model PGC specification *in vitro* in both

mouse (m) and human (h) models. As the founder population of PGCs in both the mouse and

human is exceptionally limited, *in vitro* models have been imperative in being able to better

interrogate the role of inductive cues and their effects on the transcriptional networks of early

PGCs. In humans, *BMP4* or *BMP2* alone is sufficient to induce PGC using two inductive

approaches, while in the mouse loss of *BMP2* results in a significant reduction of PGCs *in vivo*

but is not required *in vitro*[49-51]. PGCLCs of both mice and humans show high transcriptional

fidelity with *in vivo* PGCs[9,51]. As such, they have provided a reliable model to examine

transcription factor and genome-wide epigenetic dynamics during specification.

   In addition to differences in the response to inductive cues, mPGCs and hPGCs also

differ in their transcriptional networks (Figure 1-4). In the mouse, *Blimp1* is expressed at E6.25

marking specified PGCs and promotes *Prdm14* expression, which in turn stimulates *Tfap2c*,

Sox2, *Nanog* and *Oct4*[41,52]. Thus, in mPGCs *Blimp1* is a master regulator and upstream of most

of the PGC program. In the human, *SOX2* is not expressed in PGCs, while *SOX17* and *SOX15*

are[50,53]. *SOX17*, which promotes endoderm differentiation, is upregulated in hPGCLCs in

tandem with *PRDM1*, which suppresses the transcription of endodermal genes and promotes

the hPGC program. Other key hPGC genes, such as *TFAP2C* and *PRDM14,* are detected after

SOX17[50], suggesting that SOX17 in hPGCs acts more upstream than SOX2 in mPGCs, making

it unlikely SOX17 is merely replacing SOX2 without novel function. As in the mouse, *PRDM14*

re-enforces the hPGC program by upregulating or enhancing expression of PGC genes,

including *TFAP2C*, *SOX15* and *SOX17*[54]. In Chapter 3, we find that part of hPGC specification

in the human requires epigenetic reprogramming of a *Hominidae*-specific TE, LTR5Hs, which is

bound by SOX17, SOX15, TFAP2C and NANOG. Additionally, we show without remodeling of

LTR5Hs elements hPGC specification is reduced. Thus, in addition to divergences in the

transcriptional program, we also show that the cis-regulatory network employed to re-enforce

and stabilize hPGC identity is dependent on TEs which are not retained in the mouse,

exemplary of the evolutionary novelty which can drive conserved aspects of germline formation.
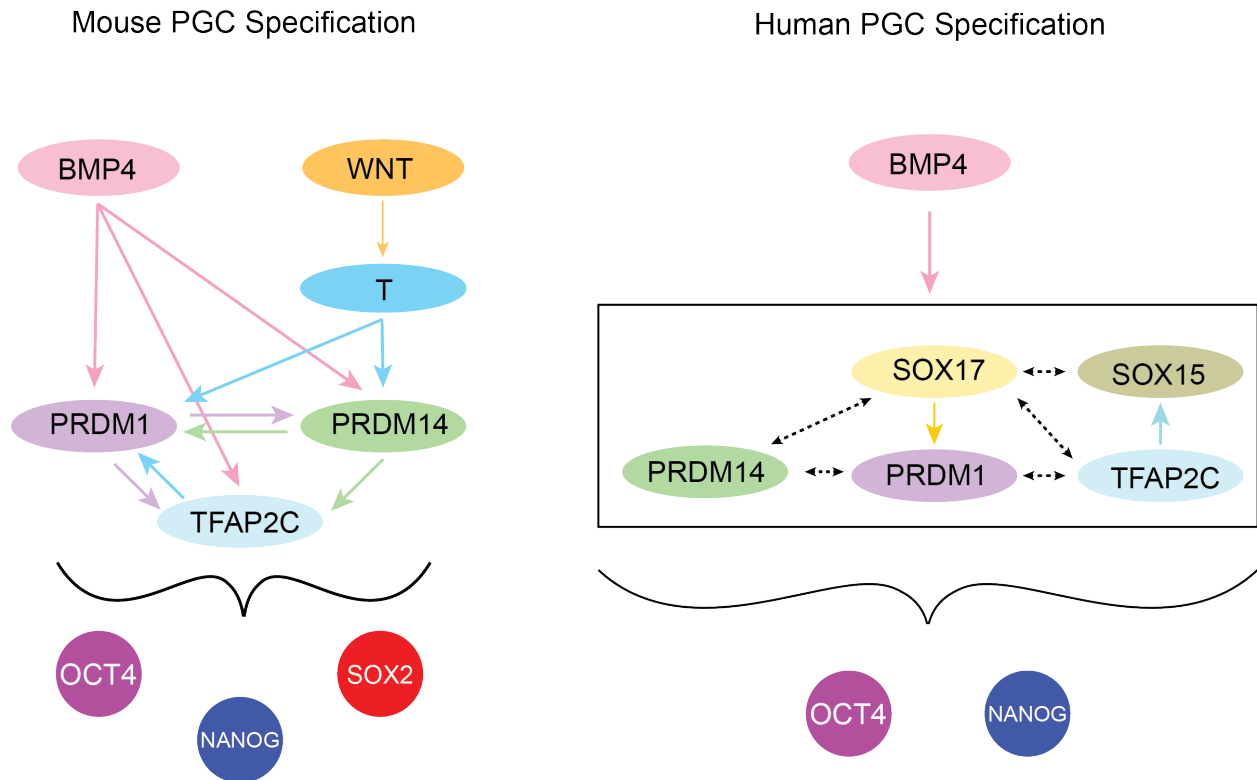
Figure 1-4: A schematic showing interactions in the mouse (left) and human (right) PGC specification networks. Unknown interactions are highlighted with black dashed arrows. Direct interactions are marked with colored solid arrows.

Introduction to PGC Differentiation:

In the mouse, PGCs undergo two waves of global erasure of DNA methylation and global re-arrangement of their epigenome. The first stage of DNA demethylation is largely passive and relies on downregulation of *Dnmt3a*, *Dnmt3b* and *Uhrf1*, thus blocking *de novo* methylation and DNA replication-coupled maintenance of DNA methylation[55-57]. Around ~E10.5, DNA demethylation begins as PGCs begin to replicate. Mitotic expansion of PGCs allows for passive replication-coupled DNA demethylation. A second wave of DNA demethylation involves passive demethylation coupled with active locus-specific demethylation driven by *Tet1*, which produces a hypomethylated PGC genome by E13.5[55-57]. Only a few precise loci maintain DNA methylation to this point, including young TEs, especially IAPs[57].

During this window PGCs also undergo global reduction in the H3K9me2, gain of H3K27me3 (which rapidly changes sub-nuclear localization between E11.5 and E12.5) and gain

in H2A/H4R3me2[58-60]. This is of particular interest, as the window between E11.5 and E12.5 is marked by epigenetic and cell-identity changes to PGCs. The first of these is an exceptionally transient event between E11.5 and E12.5 where CAF-1 is excluded from the nucleus and HIRA becomes highly expressed[58]. CAF-1 is a H3.1/H3.2 chaperone complex which incorporates new H3/H4 into replicating DNA, while HIRA is a H3.3 chaperone which incorporates H3.3 into chromatin without regard for cell cycle[61,62]. This finding suggests there may be global changes to the histone composition landscape of PGCs which has yet to be identified. As H3.3 (and therefore HIRA) localizes to promoters, it is likewise possible that this timepoint marks a large change in the gene-regulatory program of PGCs. One such change during this window is the expression of *Dazl*, an RNA-binding protein which is required for meiotic entry and exit from the PGC state[63-65]. Without timely *Dazl* expression, mPGCs fail to silence *Nanog* and *Sox2* and are not competent to enter meiosis. For this reason, *Dazl* is thought to commit PGCs to the germline, a conclusion in part reached by the inability of PGCs to be reverted back to an Embryonic Stem Cell – like state (termed an Embryonic Germ Cell) after they upregulate *Dazl*[65,66]. At E12.5, at which point DAZL is readily detected, PGCs are highly sensitive to epigenetic perturbation. This may be because DAZL, which is required for meiosis, is upregulated only in PGCs which have undergone proper reprogramming of their epigenome and are therefore likely to be competent to give rise to fit offspring.

Like expression of DAZL and therefore meiotic competence, the timing of mPGC differentiation is also controlled by *Dnmt1*, PRC1 and PRC2, which work in concert to time meiotic entry [18,67,68]. Loss of any one of these components is sufficient to mistime PGC differentiation, especially entry into meiosis which is tightly controlled by the PGC epigenome. It is notable that lack of specific PRC2 components within PGCs can lead to unique phenotypes. For example, knockout of PRC2 component *Eed* in ovarian PGCs causes a more severe phenotype than knockout of another PRC2 component, *Ezh2*. Conversely, loss of *Ezh2* but not *Eed* drives de-repression of young ERV elements in ovarian PGCs only[68].

12

Following E12.5, PGCs in an ovarian environment will become exposed to retinoic acid (RA), which will stimulate *Stra8* and initiate meiosis[69]. In a testicular environment, Sertoli cells begin to form cords around PGCs, protecting them from RA signaling via *Cyp26b1*[70,71]. In line with these changes to the soma, PGCs begin to acquire sex-specific developmental trajectories and transcriptomes at E13.5, and by E14.5 the chromatin accessibility landscape of male or female PGCs is fully distinct [72] (Figure 1-5). In Chapter 4, we show that *Trim28* is regulated in a sex-specific manner between testicular and ovarian PGCs beginning at E12.5, with ovarian PGCs downregulating *Trim28* after E12.5. Moreover, we show that once PGCs begin differentiation they become differently sensitive to knockout of *Trim28*, with *Trim28* required for differentiation into pro-spermatogonia, while in ovarian PGCs loss of *Trim28* results in inefficient entry into meiosis in some PGCs which ultimately fail to progress through meiosis and therefore a failure to produce functional gametes in the adult. Thus, here we have identified a new aspect of sex-specific differentiation in testicular and ovarian PGCs and find that in part the role of TRIM28 during this transition to is protect the PGC program from derailment.
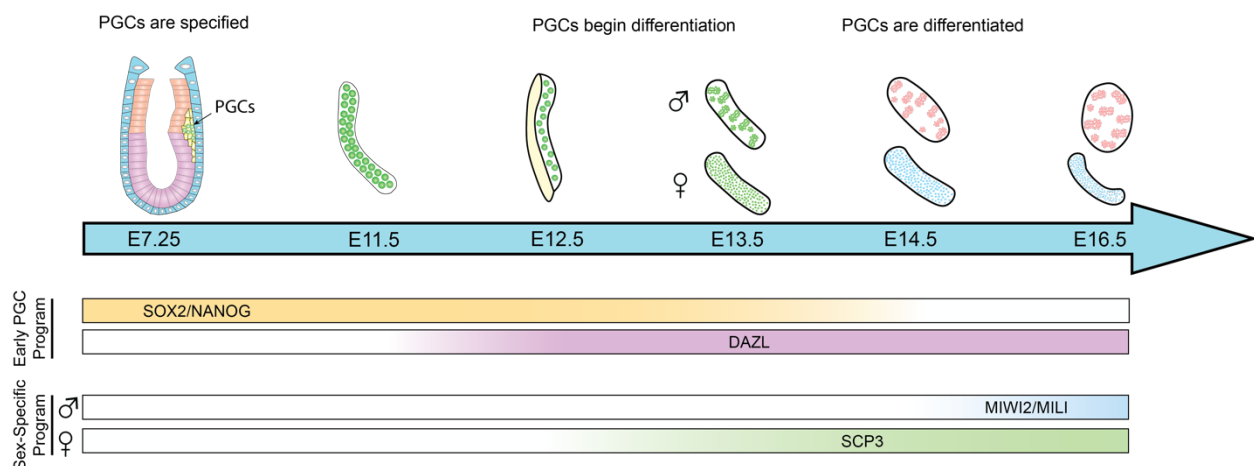
Figure 1-5: Schematic showing the timeline for PGC development in mouse, including downregulation of early PGC markers (SOX2/NANOG), upregulation of DAZL, and acquisition of sex-specific markers of differentiation.

References:

1   Wells, J. N. & Feschotte, C. A Field Guide to Eukaryotic Transposable Elements. *Annu Rev Genet* **54**, 539-561 (2020). https://doi.org:10.1146/annurev-genet-040620-022145

2   Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**, 973-982 (2007). https://doi.org:10.1038/nrg2165

3   Senft, A. D. & Macfarlan, T. S. Transposable elements shape the evolution of mammalian development. *Nat Rev Genet* **22**, 691-711 (2021). https://doi.org:10.1038/s41576-021-00385-1

4   Sotero-Caio, C. G., Platt, R. N., 2nd, Suh, A. & Ray, D. A. Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biol Evol* **9**, 161-177 (2017). https://doi.org:10.1093/gbe/evw264

5   Dupressoir, A. *et al.* A pair of co-opted retroviral envelope syncytin genes is required for formation of the two-layered murine placental syncytiotrophoblast. *Proc Natl Acad Sci U S A* **108**, E1164-1173 (2011). https://doi.org:10.1073/pnas.1112304108

6   Dupressoir, A. *et al.* Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. *Proc Natl Acad Sci U S A* **106**, 12127-12132 (2009). https://doi.org:10.1073/pnas.0902925106

7   Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**, 631-634 (2010). https://doi.org:10.1038/ng.600

8   Pontis, J. *et al.* Primate-specific transposable elements shape transcriptional networks during human development. *Nat Commun* **13**, 7178 (2022). https://doi.org:10.1038/s41467-022-34800-w

9   Xiang, X. *et al.* Human reproduction is regulated by retrotransposons derived from ancient Hominidae-specific viral infections. *Nat Commun* **13**, 463 (2022). https://doi.org:10.1038/s41467-022-28105-1

10   Pontis, J. *et al.* Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. *Cell Stem Cell* **24**, 724-735 e725 (2019). https://doi.org:10.1016/j.stem.2019.03.012

11   Theunissen, T. W. *et al.* Molecular Criteria for Defining the Naive Human Pluripotent State. *Cell Stem Cell* **19**, 502-515 (2016). https://doi.org:10.1016/j.stem.2016.06.011

12   Sundaram, V. *et al.* Functional cis-regulatory modules encoded by mouse-specific endogenous retrovirus. *Nat Commun* **8**, 14550 (2017). https://doi.org:10.1038/ncomms14550

13   He, J. *et al.* Transposable elements are regulated by context-specific patterns of chromatin marks in mouse embryonic stem cells. *Nat Commun* **10**, 34 (2019). https://doi.org:10.1038/s41467-018-08006-y

14   Zamudio, N. *et al.* DNA methylation restrains transposons from adopting a chromatin signature permissive for meiotic recombination. *Genes Dev* **29**, 1256-1270 (2015). https://doi.org:10.1101/gad.257840.114

15   Bourc'his, D. & Bestor, T. H. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431**, 96-99 (2004). https://doi.org:10.1038/nature02886

16   Jia, D., Jurkowska, R. Z., Zhang, X., Jeltsch, A. & Cheng, X. Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature* **449**, 248-251 (2007). https://doi.org:10.1038/nature06146

17   Hata, K., Okano, M., Lei, H. & Li, E. Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development* **129**, 1983-1993 (2002). https://doi.org:10.1242/dev.129.8.1983

18   Hargan-Calvopina, J. *et al.* Stage-Specific Demethylation in Primordial Germ Cells Safeguards against Precocious Differentiation. *Dev Cell* **39**, 75-86 (2016). https://doi.org:10.1016/j.devcel.2016.07.019

19   Barau, J. *et al.* The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. *Science* **354**, 909-912 (2016). https://doi.org:10.1126/science.aah5143

20   Haggerty, C. *et al.* Dnmt1 has de novo activity targeted to transposable elements. *Nat Struct Mol Biol* **28**, 594-603 (2021). https://doi.org:10.1038/s41594-021-00603-8

21   Quenneville, S. *et al.* In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Mol Cell* **44**, 361-372 (2011). https://doi.org:10.1016/j.molcel.2011.08.032

22   Alexander, K. A., Wang, X., Shibata, M., Clark, A. G. & Garcia-Garcia, M. J. TRIM28 Controls Genomic Imprinting through Distinct Mechanisms during and after Early Genome-wide Reprogramming. *Cell Rep* **13**, 1194-1205 (2015). https://doi.org:10.1016/j.celrep.2015.09.078

23   Messerschmidt, D. M. *et al.* Trim28 is required for epigenetic stability during mouse oocyte to embryo transition. *Science* **335**, 1499-1502 (2012). https://doi.org:10.1126/science.1216154

24   Schultz, D. C., Ayyanathan, K., Negorev, D., Maul, G. G. & Rauscher, F. J., 3rd. SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev* **16**, 919-932 (2002). https://doi.org:10.1101/gad.973302

25   Rowe, H. M. *et al.* KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature* **463**, 237-240 (2010). https://doi.org:10.1038/nature08674

26   Imbeault, M., Helleboid, P. Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550-554 (2017). https://doi.org:10.1038/nature21683

27   Ecco, G. *et al.* Transposable Elements and Their KRAB-ZFP Controllers Regulate Gene Expression in Adult Tissues. *Dev Cell* **36**, 611-623 (2016). https://doi.org:10.1016/j.devcel.2016.02.024

28   Aravin, A. A. *et al.* A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol Cell* **31**, 785-799 (2008). https://doi.org:10.1016/j.molcel.2008.09.003

29   Manakov, S. A. *et al.* MIWI2 and MILI Have Differential Effects on piRNA Biogenesis and DNA Methylation. *Cell Rep* **12**, 1234-1243 (2015). https://doi.org:10.1016/j.celrep.2015.07.036

30   Pezic, D., Manakov, S. A., Sachidanandam, R. & Aravin, A. A. piRNA pathway targets active LINE1 elements to establish the repressive H3K9me3 mark in germ cells. *Genes Dev* **28**, 1410-1428 (2014). https://doi.org:10.1101/gad.240895.114

31   Ernst, C., Odom, D. T. & Kutter, C. The emergence of piRNAs against transposon invasion to preserve mammalian genome integrity. *Nat Commun* **8**, 1411 (2017). https://doi.org:10.1038/s41467-017-01049-7

32   Aravin, A. A. *et al.* Cytoplasmic compartmentalization of the fetal piRNA pathway in mice. *PLoS Genet* **5**, e1000764 (2009). https://doi.org:10.1371/journal.pgen.1000764

33   Stein, P. *et al.* Essential Role for endogenous siRNAs during meiosis in mouse oocytes. *PLoS Genet* **11**, e1005013 (2015). https://doi.org:10.1371/journal.pgen.1005013

34   Fuentes, D. R., Swigut, T. & Wysocka, J. Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. *Elife* **7** (2018). https://doi.org:10.7554/eLife.35989

35   Todd, C. D., Deniz, O., Taylor, D. & Branco, M. R. Functional evaluation of transposable elements as enhancers in mouse embryonic and trophoblast stem cells. *Elife* **8** (2019). https://doi.org:10.7554/eLife.44344

36   Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57-63 (2012). https://doi.org:10.1038/nature11244

37   Ai, Z. *et al.* Kruppel-like factor 5 rewires NANOG regulatory network to activate human naive pluripotency specific LTR7Ys and promote naive pluripotency. *Cell Rep* **40**, 111240 (2022). https://doi.org:10.1016/j.celrep.2022.111240

38   Grow, E. J. *et al.* Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**, 221-225 (2015). https://doi.org:10.1038/nature14308

39   Yu, H. *et al.* Dynamic reprogramming of H3K9me3 at hominoid-specific retrotransposons during human preimplantation development. *Cell Stem Cell* **29**, 1031-1050 e1012 (2022). https://doi.org:10.1016/j.stem.2022.06.006

40   Kurimoto, K. *et al.* Complex genome-wide transcription dynamics orchestrated by Blimp1 for the specification of the germ cell lineage in mice. *Genes Dev* **22**, 1617-1635 (2008). https://doi.org:10.1101/gad.1649908

41   Ohinata, Y. *et al.* Blimp1 is a critical determinant of the germ cell lineage in mice. *Nature* **436**, 207-213 (2005). https://doi.org:10.1038/nature03813

42   Hancock, G. V., Wamaitha, S. E., Peretz, L. & Clark, A. T. Mammalian primordial germ cell specification. *Development* **148** (2021). https://doi.org:10.1242/dev.189217

43   Chen, D. *et al.* Human Primordial Germ Cells Are Specified from Lineage-Primed Progenitors. *Cell Rep* **29**, 4568-4582 e4565 (2019). https://doi.org:10.1016/j.celrep.2019.11.083

44   Lawson, K. A. *et al.* Bmp4 is required for the generation of primordial germ cells in the mouse embryo. *Genes & Development* **13**, 424-436 (1999). https://doi.org:DOI 10.1101/gad.13.4.424

45   Ying, Y., Liu, X. M., Marble, A., Lawson, K. A. & Zhao, G. Q. Requirement of Bmp8b for the generation of primordial germ cells in the mouse. *Mol Endocrinol* **14**, 1053-1063 (2000). https://doi.org:10.1210/mend.14.7.0479

46   Ying, Y. & Zhao, G. Q. Cooperation of endoderm-derived BMP2 and extraembryonic ectoderm-derived BMP4 in primordial germ cell generation in the mouse. *Dev Biol* **232**, 484-492 (2001). https://doi.org:10.1006/dbio.2001.0173

47   de Sousa Lopes, S. M. *et al.* BMP signaling mediated by ALK2 in the visceral endoderm is necessary for the generation of primordial germ cells in the mouse embryo. *Genes Dev* **18**, 1838-1849 (2004). https://doi.org:10.1101/gad.294004

48   de Sousa Lopes, S. M., Hayashi, K. & Surani, M. A. Proximal visceral endoderm and extraembryonic ectoderm regulate the formation of primordial germ cell precursors. *BMC Dev Biol* **7**, 140 (2007). https://doi.org:10.1186/1471-213X-7-140

49   Sasaki, K. *et al.* Robust In Vitro Induction of Human Germ Cell Fate from Pluripotent Stem Cells. *Cell Stem Cell* **17**, 178-194 (2015). https://doi.org:10.1016/j.stem.2015.06.014

50   Irie, N. *et al.* SOX17 is a critical specifier of human primordial germ cell fate. *Cell* **160**, 253-268 (2015). https://doi.org:10.1016/j.cell.2014.12.013

51   Hayashi, K., Ohta, H., Kurimoto, K., Aramaki, S. & Saitou, M. Reconstitution of the mouse germ cell specification pathway in culture by pluripotent stem cells. *Cell* **146**, 519-532 (2011). https://doi.org:10.1016/j.cell.2011.06.052

52   Magnusdottir, E. *et al.* A tripartite transcription factor network regulates primordial germ cell specification in mice. *Nat Cell Biol* **15**, 905-915 (2013). https://doi.org:10.1038/ncb2798

53   Wang, X. *et al.* The chromatin accessibility landscape reveals distinct transcriptional regulation in the induction of human primordial germ cell-like cells from pluripotent stem

cells. *Stem Cell Reports* **16**, 1245-1261 (2021). https://doi.org:10.1016/j.stemcr.2021.03.032

54   Sybirna, A. *et al.* A critical role of PRDM14 in human primordial germ cell fate revealed by inducible degrons. *Nat Commun* **11**, 1282 (2020). https://doi.org:10.1038/s41467-020-15042-0

55   Kagiwada, S., Kurimoto, K., Hirota, T., Yamaji, M. & Saitou, M. Replication-coupled passive DNA demethylation for the erasure of genome imprints in mice. *EMBO J* **32**, 340-353 (2013). https://doi.org:10.1038/emboj.2012.331

56   Ohno, R. *et al.* A replication-dependent passive mechanism modulates DNA demethylation in mouse primordial germ cells. *Development* **140**, 2892-2903 (2013). https://doi.org:10.1242/dev.093229

57   Seisenberger, S. *et al.* The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Mol Cell* **48**, 849-862 (2012). https://doi.org:10.1016/j.molcel.2012.11.001

58   Hajkova, P. *et al.* Chromatin dynamics during epigenetic reprogramming in the mouse germ line. *Nature* **452**, 877-881 (2008). https://doi.org:10.1038/nature06714

59   Kim, S. *et al.* PRMT5 protects genomic integrity during global DNA demethylation in primordial germ cells and preimplantation embryos. *Mol Cell* **56**, 564-579 (2014). https://doi.org:10.1016/j.molcel.2014.10.003

60   Seki, Y. *et al.* Extensive and orderly reprogramming of genome-wide chromatin modifications associated with specification and early development of germ cells in mice. *Dev Biol* **278**, 440-458 (2005). https://doi.org:10.1016/j.ydbio.2004.11.025

61   Yu, Z., Liu, J., Deng, W. M. & Jiao, R. Histone chaperone CAF-1: essential roles in multi-cellular organism development. *Cell Mol Life Sci* **72**, 327-337 (2015). https://doi.org:10.1007/s00018-014-1748-3

62   Choi, J., Kim, T. & Cho, E. J. HIRA vs. DAXX: the two axes shaping the histone H3.3 landscape. *Exp Mol Med* **56**, 251-263 (2024). https://doi.org:10.1038/s12276-023-01145-3

63   Gill, M. E., Hu, Y. C., Lin, Y. & Page, D. C. Licensing of gametogenesis, dependent on RNA binding protein DAZL, as a gateway to sexual differentiation of fetal germ cells. *Proc Natl Acad Sci U S A* **108**, 7443-7448 (2011). https://doi.org:10.1073/pnas.1104501108

64   Chen, H. H. *et al.* DAZL limits pluripotency, differentiation, and apoptosis in developing primordial germ cells. *Stem Cell Reports* **3**, 892-904 (2014). https://doi.org:10.1016/j.stemcr.2014.09.003

65   Nicholls, P. K. *et al.* Mammalian germ cells are determined after PGC colonization of the nascent gonad. *Proc Natl Acad Sci U S A* **116**, 25677-25687 (2019). https://doi.org:10.1073/pnas.1910733116

66   Leitch, H. G. *et al.* Rebuilding pluripotency from primordial germ cells. *Stem Cell Reports* **1**, 66-78 (2013). https://doi.org:10.1016/j.stemcr.2013.03.004

67   Hill, P. W. S. *et al.* Epigenetic reprogramming enables the transition from primordial germ cell to gonocyte. *Nature* **555**, 392-396 (2018). https://doi.org:10.1038/nature25964

68   Huang, T. C. *et al.* Sex-specific chromatin remodelling safeguards transcription in germ cells. *Nature* **600**, 737-742 (2021). https://doi.org:10.1038/s41586-021-04208-5

69   Ishiguro, K. I. Mechanism of initiation of meiosis in mouse germ cells. *Curr Top Dev Biol* **151**, 1-26 (2023). https://doi.org:10.1016/bs.ctdb.2022.04.005

70   Bowles, J. *et al.* Retinoid signaling determines germ cell fate in mice. *Science* **312**, 596-600 (2006). https://doi.org:10.1126/science.1125691

71   Ohta, K. *et al.* Male differentiation of germ cells induced by embryonic age-specific Sertoli cells in mice. *Biol Reprod* **86**, 112 (2012). https://doi.org:10.1095/biolreprod.111.095943

72   Li, J. *et al.* Accurate annotation of accessible chromatin in mouse and human primordial germ cells. *Cell Res* **28**, 1077-1089 (2018). https://doi.org:10.1038/s41422-018-0096-5

Chapter 2

Embryo Models as a Platform for Investigation of Transposable Element Function in

Embryonic Development.

Transposable elements in early human embryo development and embryo models

Jonathan A. DiRusso [1,2,3,4] and Amander T. Clark [1,2,3,4,*]

Transposable elements (TEs), long discounted as 'selfish genomic elements,' are increasingly appreciated as the drivers of genomic evolution, genome organization, and gene regulation. TEs are particularly important in early embryo development, where advances in stem cell technologies, in tandem with improved computational and next-generation sequencing approaches, have provided an unprecedented opportunity to study the contribution of TEs to early mammalian development. Here, we summarize advances in our understanding of TEs in early human development and expand on how new stem cell-based embryo models can be leveraged to augment this understanding.

Addresses

[1] Department of Molecular, Cell and Developmental Biology, University of California, 90095 Los Angeles, CA, USA
[2] Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, 90095 Los Angeles, CA, USA.
[3] Molecular Biology Institute, University of California, 90095 Los Angeles, CA, USA
[4] Center for Reproductive Science, Health and Education, University of California, 90095 Los Angeles, CA, USA

Corresponding author: Clark, Amander T. (clarka@ucla.edu)
* Twitter account: @clarklabucla1

Introduction

A considerable portion of mammalian genomes are composed of Transposable Elements (TEs). In humans, the most recent assemblies identify ~53% of the genome as being composed of TEs [1,2]. There are two major classes of TEs: retrotransposons, which involve an RNA intermediate that is reverse transcribed during the process of transposition, and DNA-only TEs, which do not use reverse transcriptase for transposition. The most prolific colonization of mammalian genomes has been achieved by retrotransposons, which can broadly be classified into long terminal repeat (LTR) and non-LTR TEs (Figure 1). Non-LTR TEs include families such as long interspersed nuclear element (LINE) and short interspersed nuclear elements (SINE), while LTR transposons include the endogenous retrovirus (ERV) families. ERV integrants consist of pro-viral sequences, often flanked by regulatory LTRs, as well as solo-LTRs, which have lost their associated pro-virus through the recombination of LTRs or degradation [3]. In humans, most non-LTR and LTR-family TEs are no longer capable of transposition, while some families such as the LINE1 Human Specific (L1Hs), Alu SINE elements, and the composite retrotransposon family SINE-Variable Number of Tandem Repeat (VNTR)-Alu (SVA) element (SINE-VNTR-Alu or SVA), a SINE derivative, remain mobile in humans [4].

During early embryogenesis, dynamic changes in RNA expression from TEs are associated with key developmental progressions in totipotent and pluripotent cells including at the time of zygotic genome activation (ZGA), the conversion of totipotent cells in the morula to pluripotent pre-implantation epiblast cells of the blastocyst, as well as during early embryo development including formation of primordial germ cells (PGCs) [5–12]. Studying TE expression and epigenetic regulation in human pre-implantation embryos is possible but limited due to the small number of embryos donated to research following *in vitro* fertilization (IVF). As described below, most studies on TEs in early human embryos involve tracking RNA expression and chromatin changes. However, a major discovery revealed that pro-viruses originating from ancient human endogenous retrovirus K (HERVK) pro-viral genomic integrants are capable of generating viral particles during blastocyst formation when studied *in vitro* [13].

Evaluating the role of TEs in post-implantation human embryo development poses even more challenges, especially due to the highly limited availability of human embryo samples. To fill this gap in knowledge, human pluripotent stem cells (hPSCs), namely human embryonic stem cells (hESCs) and induced PSCs (hiPSCs), are used to model pre- and post-implantation states of

Figure 1
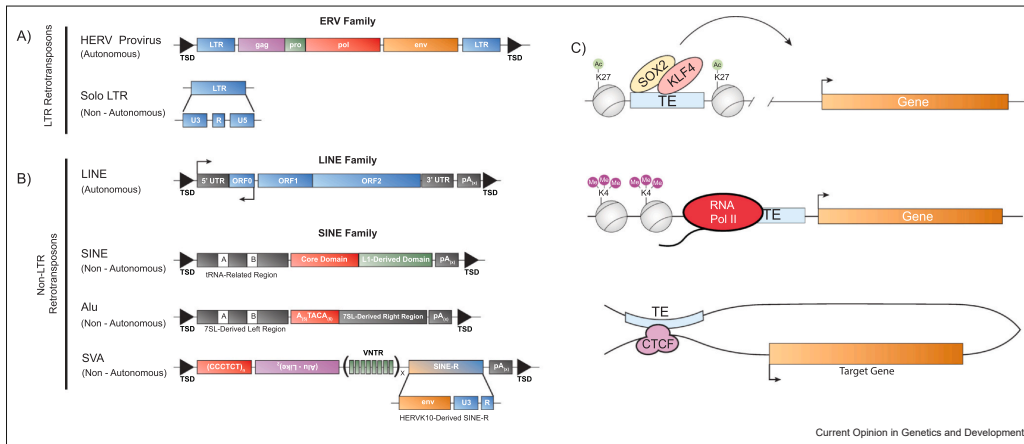
Schematics of retroviral class elements in humans**.** **(a)** (Upper) A typical human endogenous retrovirus (HERV) including (from left to right): Target Site Duplication (TSD), LTR, a proviral genome with gag, pro, pol and env genes (Lower) A schematic of the LTR and it's regulatory regions. Solo LTRs are highly related to those found in the provirus. **(b)** (Top) LINE containing two open reading frames in the sense orientation and one in the anti-sense orientation. (Middle) SINE, a nonautonomous element that is reliant on LINE activity for mobility. The Alu element is a subfamily of SINE and is nonautonomous but remains mobile in the human genome. (Bottom) The SINE-VTNR-Alu (SVA) element, which contains a SINE core repeat and Alu-like region (antisense to the remainder of the element), a VNTR repetitive region and a SINE-R region, which is derived from HERVK10 and contains homology to LTR5Hs. **(c)** Modes of cis-regulatory TE activity showing (Upper) enhancer activity where H3K27ac-marked TEs are bound by transcription factors (Middle) Promoter-acting TE, where an upstream integrant acts as a promoter for a downstream gene. (Lower) A TE acts as a 3D genome regulatory element.

pluripotency, emerging somatic lineages, and PGCs. As described below, the field has used relatively simple 2D models to generate foundational knowledge regarding expression, epigenetic state, and transcription factor binding to TEs in early human development. However stem cell-based embryo models (embryo models), which recapitulate various three-dimensional aspects of early human embryo development, represent the next frontier for evaluating the functional role of TEs in regulating human biology during this key stage of embryo development.

It is now appreciated that some TEs can serve as cis-regulatory elements, and therefore have the potential to drive cell-type-specific gene expression. In both mice and humans, extensive binding of transcription factors to ERV-family elements, especially the key regulators of PSC self-renewal (NANOG, OCT4, and KLF4), have driven the hypothesis that TEs contribute to gene regulation in the pluripotent state of mammals [14]. Further supporting this view, many ERV LTRs harbor multiple pluripotency factor binding sites, and chromatin immunoprecipitation followed by sequencing (ChIP-seq) has shown that ERV LTRs are, in many cases, co-bound by multiple pluripotent transcription factors [14,15]. In

addition to being transcription factor-bound, specific families of ERVs are also enriched with chromatin and histone modifications associated with active states of gene expression [11,16–18]. Evidence for a role of TEs in 3D genome organization exists as well, as CCCTC-binding factor (CTCF) binding sites are found in TEs of the LINE, SINE, and LTR superfamilies. These sites are thought to be functional, as alterations in 3D genome organization are observed when TEs are altered via Clustered Regularly Interspaced Short Palindromic Repeats/Cas9 (CRISPR/cas9) editing [19,20]. These observations support a role for TEs as potential cis-regulatory enhancer/promoter elements in regulating states of pluripotency in the early human embryo.

The epigenetic silencing of ERVs, and therefore the decommissioning of active ERV LTR-family enhancer/promoter elements during early embryo development likely involves Kruppel-associated box zinc-finger proteins (KRAB-ZFPs), a broad family of DNA binding proteins that bind TEs and work in tandem with TRIM28 (KAP1) to facilitate the deposition of repressive H3K9me3 heterochromatin at the targeted ERVs [21–24]. The enrichment of H3K9me3 at ERVs is highly dynamic during early human embryo development, beginning at the four-cell

stage before ZGA and continuing through blastocyst formation [7,8]. This is important as the human genome is largely demethylated during these embryonic stages [25], suggesting that H3K9me3 likely serves as the predominant repressive epigenetic modification responsible for dynamically silencing TEs in the early embryo. Supporting this hypothesis, ultra-low input ChIP-seq for H3K9me3 in pre-implantation human embryos has revealed that intergenic ERV LTRs are progressively enriched with H3K9me3 from the four-cell stage to formation of the blastocyst [7,8]. Notably, some primate and hominoid-specific TEs including LTR5Hs, LTR7B, and SINE-R-VNTR-Alu_D (SVA_D) are not marked with H3K9me3 by the eight-cell stage, and these are the TEs which are enriched for pluripotency factor motifs. In particular, SVA_D elements are highly enriched for DUX (double homeobox) motifs, suggesting a cis-regulatory role for these elements [7,8]. To evaluate the necessity of TE remodeling at the eight-cell stage, Yu et al. performed CRISPR-interference (CRISPRi)-mediated repression of SVA_D elements in human embryos at the 4-cell stage, which resulted in a developmental block at the 8-cell stage, [8] indicating SVA_D expression or the chromatin state at SVA_D is involved in human ZGA.

Similarly, there is evidence that TEs function in the morula as totipotency is extinguished in order to give rise to trophectoderm and inner cell mass (ICM) cells of blastocysts. Specifically, the chromatin accessibility of morula cells reveals two distinct chromatin states; 1) cells where regions of high chromatin accessibility are enriched for transcription factor motifs involved in trophectoderm formation (GATA and TEAD families), and 2) cells where regions of high chromatin accessibility are enriched for transcription factor motifs associated with pluripotency (NANOG, SOX2, and KLF4). Interestingly, the cells of the putative outer morula (the ones fated to become trophectoderm), acquire H3K9me3 at hominoid-specific ERVs, including those of the endogenous retrovirus K (ERVK) family (MER11B, MER11C LTRs and HERK9-int) and ERV1 family (LTR12). The footprinting analysis of the putative ICM reveals that these ERV LTRs are bound by pluripotency factors OCT4 and SOX2 [8]. Taken together, these observations support the hypothesis that targeted silencing of the aforementioned ERVs safeguards the specification of extraembryonic lineages by precluding the ectopic activation of the pluripotency program [8]. Collectively, these results suggest that active epigenetic remodeling of TEs is important for early human embryonic development.

Although RNA expression and epigenetic modifications can be mapped to TE families and subfamilies, a caveat exists when studying the role of individual TE integrants and the regulation of neighboring gene expression. The repetitive nature of TEs, combined with the short-read lengths acquired using traditional Illumina sequencing precludes accurate mapping to unique genomic sites [26]. Furthermore, in the case of mobile elements, such as L1Hs and Alu, a precise mapping of new integrants across individuals using a reference genome is impossible as the reference would likely not include the position of person-specific integrations. Still, computational advances and long-read sequencing (discussed below) are beginning to address these shortfalls, promising better platforms for interrogating the gene regulatory role of TEs in human biology.

In addition to TE regulation at the chromatin level, it is now appreciated the N6-methyladenosine modification to endogenous viral mRNA is also a potent modality of ERV (and LINE1) regulation. Evidence in mouse PSCs has shown that ERV RNAs transcribed from the retrotransposition-active intracisternal A-particle (IAP) family are heavily marked by m6A due to the activity of the METTL3/MEETL14 heterodimer. m6A-marked ERV RNAs are then bound by YTH N6-methyladenosine RNA-binding protein F1/2/3 (YTHDF1/2/3), m6A readers, which in turn results in the clearance of viral mRNAs [27]. While little is known about the role of m6A in regulating ERVs, SVAs or LINE elements in the early human embryo, it is apparent that the analysis of m6A at TEs during early human embryo development and human PSCs is needed.

Evaluating transposable elements using human pluripotent states in culture

Evaluating TE expression, epigenetic regulation, and function in pre-implantation and early post-implantation human embryos is challenging. Therefore, TE analysis in hESCs and hiPSCs has emerged as an important *in vitro* model to understand the role of TEs in pluripotency. Pluripotency is a spectrum of states from naïve to primed, with three major states of pluripotency (naïve, formative, and primed) successfully captured as self-renewing stem cells *in vitro*. A fourth state of pluripotency, called latent pluripotency, occurs in PGCs and is captured *in vitro* with the differentiation of PGC-like cells (PGCLCs) [28–36] (Figure 2). The culture of human PSCs in naïve conditions recapitulates the transcriptome and key epigenomic features of cells from the morula at Carnegie stage (CS) 2 and pre-implantation epiblasts at CS3-CS4 (Figure 2). The primed and formative culture conditions recapitulate the post-implantation epiblast at CS4-CS5 [37,38]. The differentiation of human PGCLCs (hPGCLCs) *in vitro* results in the formation of germ cells equivalent to those found in post-implantation embryos between CS5-CS7 (Figure 2) [10,39,40]. In the following section, we will highlight key studies evaluating TEs in the four states of human pluripotency.

Figure 2
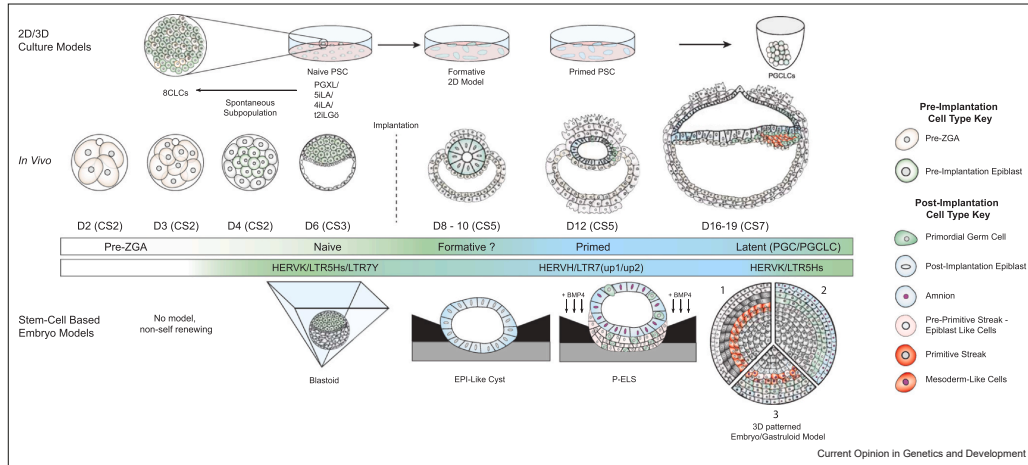
In vitro and stem cell-based embryo models. Top) 2D culture *in vitro* models representing various states in embryonic development paired with their embryonic time-matched counterparts below (*In vivo*). From left to right: 8CLCs, non-renewing transient subpopulations in naïve human PSCs cultures. Naïve PSCs which represent the inner cell mass of the pre-implantation epiblast. Naïve cells are marked by the high expression of HERVK and functional reliance on LTRT5Hs. Formative cells can be stably cultured but have not yet been identified *in vivo* human embryos, thus the post-implantation time point they may correspond to is unknown (?). Primed PSCs represent cells of the post-implantation epiblast, and are marked by the high expression of HERVH and the functional activity of LTR7(up1/2) integrants. Far-right is the PGCLC aggregate model, in which PGCLCs occupy a state of latent pluripotency. CS3-CS7 embryos are shown as a lateral cross-section. Bottom) Stem cell-based embryo models (Left to right): Blastoids represent the pre-implantation embryo. The P-ELS model involves gradients of Bone Morphogenetic Protein 4 (BMP4) exposure driving the formation of the amnion, hPGCLCs and the expression of T in the pre-streak epiblast. Two variations of a 3D-printed embryo model and a gastruloid are shown. 1) lineages derived by Warmflash 2014, from outer to inner are trophectoderm-like, endoderm-like, primitive-streak-like, and ectoderm-like. 2) The model of Jo et al., 2022, where they observe (from outer to inner) amnion-like cells, PGCLCs, and ectoderm-like cells. 3) A gastruloid model representative of BMP4-treated gastruloids in Minn et al., 2020 (from outer to inner): amnion-like and trophectoderm-like cells, endoderm and PGCLCs, mesoderm and primitive streak like, epiblast-like and ectoderm-like cells. Pluripotent pre-implantation lineages are noted in the key as well as primitive steak. Other cells represented in the models are shown in gray.

At the RNA level, it is now appreciated that the different states of pluripotency *in vitro* express differing repertoires of TEs, with corresponding ChIP-seq and chromatin accessibility data revealing the epigenetic status of TEs in each state [17,31,32]. For example, naïve PSCs exhibit high expression of the HERVK provirus (HERVK-int) and LTR5Hs, a *Hominoidea*-specific LTR of the HERVK(HML2) family [31]. The RNA expression of LTR5Hs in naïve human PSCs is coupled with the hypomethylation and enrichment of H3K27ac at these TEs. Interestingly, LTR5A and 5B, two older members of the HERVK(HML2) family first identified in *Hominoidea* (Apes) and *Catarrhini* (New World Monkeys), respectively, are not highly expressed or marked with H3K27ac in the naïve state. The members of the SVA family, specifically SVA_D, are also expressed by naïve human PSCs, and these integrants are correspondingly enriched in H3K27ac [16,31]. This is of particular interest, as the HERVK10-derived SINE-R region of the composite SVA_D element shares

significant homology with LTR5Hs (Figure 1), meaning this region of SVA_D likely recruits the same transcription factors and KRAB-ZFPs as LTR5Hs [16,41]. Functional studies in naïve human PSCs using CRISPRi to target LTR5Hs and SVA_D results in widespread gene deregulation, especially of genes involved in 3D genome organization, cell polarity, and lineage restriction [16,42]. Thus, naïve human PSCs do not simply express LTR5Hs and SVA_D, but are also reliant on chromatin remodeling and/or RNA expression of these TEs for the proper maintenance of the naïve state.

Naïve hESCs also have high accessibility and expression of LTR7Y, a regulatory element of the HERVH provirus. While some LTR7B and some LTR7 elements exhibit H3K27ac enrichment in the naïve state, it is the LTR7Y elements that are highly enriched in H3K27ac, and are bound by KLF4, KLF5, and NANOG [16–18]. Notably, different naïve culture conditions lead to differences in LTR7 family expression, with 5i/L/A

conditions resulting in higher LTR7Y expression and H3K27ac, while 3i/L naïve conditions lead to higher expression of LTR7. This is likely a result of 5i/L/A conditions and 3i/L naïve conditions capturing different states on the pluripotency spectrum. Indeed, the responsiveness of LTR7, LTR7Y, and LTR7B to different naïve culture conditions likely suggests they are highly dynamic and sensitive to developmental progression through the different states of pluripotency in the early embryo.

As the human blastocyst implants during CS4 to CS5, the naïve pluripotent cells of the pre-implantation epiblast convert to a state of pluripotency referred to as formative. This is a state competent to differentiate into somatic cells and PGCs following exposure to appropriate differentiation cues [32]. In the mouse embryo, formative pluripotency corresponds to cells of the post-implantation epiblast at the early egg-cylinder stage. In humans, the location of formative cells in the post-implantation embryo is not yet known but likely corresponds to post-implantation pluripotent cells somewhere between day 8–12 post-fertilization. Despite this unknown, formative cells have been reported in culture [32]. Formative cells share some TE overlap with naïve human PSCs, including the expression of HERVK, however, formative cells uniquely express LTR6A [32]. Curiously, formative cells share the expression of a number of KRAB-ZFPs with naïve cells, including *ZNF676*, *ZNF560*, and *ZNF528*. However, much of the formative gene transcriptional profile is shared with the primed state [32]. This similarity in gene expression profiles elevates LTR6A as a key marker of the formative state, and underscores the utility of TEs as markers of developmental populations which are hard to capture *in vivo*.

Primed hPSCs, which recapitulate post-implantation epiblast cells, express pro-viral integrants of the HERVH and HERVK families [17,43]. Furthermore, the epigenome of HERVH-associated LTR regulatory sequence LTR7 are hypomethylated and enriched with H3K27ac in the primed state [17,31,44–47]. In a recent phylogenic analysis of HERVH and LTR7, new subfamilies were identified, and their evolutionary trajectories were defined. Of particular interest are *Homininae*-specific LTR7up1 and LTR7up2, which were previously clustered with LTR7. The re-analysis of LTR7Y, LTR7up1, and LTR7up2 using these updated phylogenies revealed that LTR7Y elements are expressed by naïve human PSCs, whereas it is the LTR7up families that are highly expressed in primed cells, with LTR7up1 and LTR7up2 accounting for most of the observed LTR7 transcription in this state. Consistent with these observations, LTR7up1 and LTR7up2 are also enriched with H3K27ac and are bound by pluripotency factors NANOG, SOX2, and FOXP1 in primed cells [48]. Altogether, this supports the idea that

young TEs of the LTR7up families likely acting as cis-regulatory regions in primed state pluripotency.

Latent pluripotency is a feature of PGCs. During embryo implantation and the progression of pluripotent cells from the naïve to primed state, hPGCs are specified [49,50]. Based on primate models, this is speculated to occur at CS5, coincident with amnion and extra-embryonic mesoderm formation but initiated before the formation of the primitive streak [39,51]. Studying hPGC specification *in vivo* is extremely challenging due to the rarity of early human post-conceptus tissues donated to research. A recent paper profiling single cells of a CS7 post-implantation human embryo demonstrated that hPGCs exhibit a suite of TEs with similarities to naïve cells [9,10,40]. Specifically, this data set revealed that CS7 hPGCs *in vivo* express HERVK, LTR5Hs, and SVA_D [40].

In order to model the expression and function of TEs during hPGC specification, the differentiation of hPGCLCs from primed human PSCs is used (Figure 2) [33–35,52]. The induction of hPGCLCs *in vitro* leads to a dynamic change in TE expression relative to the primed state, with one of the most significant changes involving the up-regulation of LTR5Hs. At the epigenetic level in hPGCLCs, LTR5Hs sequences become hypomethylated, enriched in H3K27ac, and bound by NANOG, SOX15, SOX17, and TFAP2C transcription factors all of which are necessary for hPGC specification and maintenance [10,49,53,54]. Using CRISPRi to dampen LTR5Hs accessibility before hPGCLC induction results in a significant reduction in the competency of human PSCs to differentiate into hPGCLCs [10]. This suggests that LTR5Hs is necessary either for the acquisition or maintenance of latent pluripotency upon hPGCLC formation.

Transposable element expression in totipotent-like cells in culture

Following fertilization (CS1), the maternal and paternal genomes undergo epigenetic reprogramming to create a totipotent state competent for ZGA, a process where embryonic transcription is activated in the diploid embryo [5,55,56]. In humans, ZGA occurs in the 8-cell embryo, while in the mouse, it occurs in the 2-cell embryo. In both mouse and human embryos, ERVL (MERVL/HERVL) family elements are broadly derepressed with ZGA, coincident with a high expression of *ZSCAN4* and the *TRIM* family of genes, as well as expression of *Dux/DUXA* [5,6,57,58]. Using these characteristics, human 8-cell-like cells (8CLCs) have been identified sporadically in hPSCs cultured under a variety of naïve culture conditions, including 4 inhibitor (4i), 5i, naïve human stem cell medium, PXGL (PD0325901, Gö6983, XAV939, LIF), t2iLGö (titrated 2 inhibitors of GSK3 and MAPK/Erk + LIF + Gö6983) [6] or from

naïve hPSCs reverted to a defined media called e4CL (enhanced 4 chemicals + LIF: PD0325901, IWR1, DZNep and TSA), which increases the subpopulation of 8CLCs [5]. Notably, the transcriptomes of 5i and PXGL 8CLCs are closest to those of e4CL, although e4CL was a much more potent generator of 8CLC cells [5]. 8CLCs have also been identified in populations of pre-epiblast-like PSCs (prEpiSCs), which are cultured in media containing inhibitors of MEK (GSK1120212), WNT (XAV939), PKC (Go6983), Src (A419259) and EED/H3K27me3 (DZNep) [12]. In all conditions, 8CLCs are transient and metastable, interconverting between 8CLCs and naïve PSCs *in vitro*. Therefore, like totipotent cells of the embryo, 8CLCs lack the capacity for self-renewal [56]. Despite these challenges, the reliable isolation of 8CLCs and the ability to identify them in single-cell transcriptomic experiments have enabled the characterization of their unique TE repertoire. TE expression in 8CLCs compared to the naïve cells reveals a general enrichment of ERVs including HERVK but with a marked increase in the expression of HERVL and its associated LTRs, MLT2A1, and MLT2A2 [5,6,12]. This is in line with TE expression in the 8-cell human embryo [5,6,12]. Importantly, neither expression nor chromatin accessibility of HERVL, MLT2A1, and MLT2A2 was observed in any other naïve hPSCs in culture, suggesting the expression of these LTRs is tightly associated with the 8CLC state and not naïve pluripotency [5,6,12,45,59,60].

Emerging technologies for understanding transposable elements in early human embryo development

Although human pre-implantation embryo development can be studied using IVF embryos, there are limitations in the number of embryos that can be used and, in some jurisdictions, research with human embryos is not allowed [61]. This has led to the emergence of embryo models as the critical avatars of human embryo development [62]. New embryo models recapitulate some key aspects of early human development, but not all. The closest representation of an intact embryo are blastoids, resembling pre-implantation blastocysts consisting of spatially organized pre-implantation epiblast, primitive endoderm, and trophoblast [63–66]. These emerging models represent a powerful tool in which the function of TEs can be evaluated using approaches such as CRISRi of SVA_D elements as in [7]. Although powerful, the blastoid model cannot recapitulate the embryonic events in an eight-cell embryo prior to blastocyst formation.

For postimplantation stages, embryo models including gastruloids [67,68], post-implantation amniotic sac embryoids (PASE), posteriorized-embryonic-like sacs (P-ELS) [69,70], and micropatterned self-organizing discs [36,65,71] have emerged. These tools enable a more thorough interrogation of TEs in lineage specification after embryo implantation, which is of particular importance as the divergence and origin of some post-implantation lineages, such as hPGCs, remains unclear. These models, coupled with advances in sequencing technologies, hold promise in understanding the role of TEs throughout development, including a more complex understanding of TE epigenetic reprogramming and its role in properly regulating early embryonic lineage restriction.

In tandem with advancing *in vitro* models, there has also been a rapid expansion of low-input and single-cell sequencing technologies. Single-cell RNA-seq has become important in examining both rare populations, such as 8CLCs in [5,6], as well as in-depth profiling of the transcriptome of rare clinical samples, such as SMART-seq of the human CS 7 embryo [40]. The third-generation long-read sequencing technologies also hold promise for a better characterization of the RNA expression and mapping of TEs, helping overcome long-standing challenges in mapping to specific genomic sites. For instance, single-cell long-read RNA-seq (CELLO-seq) [72] can be used to map TE transcripts to their original loci, helping to resolve questions about the cis- and trans-acting roles of TEs. While the higher error rate of third-generation sequencing still imposes some technical limitations on its use, it also harbors distinct advantages, such as the ability to read DNA 5-methylcytosine and RNA N6-methyladenosine natively without enzymatic modification to nucleic acids during library preparation [73,74].

Another significant advance in sequencing is the advent of single-cell multi-omics, which are powerful tools to understand how transcriptomes, chromatin accessibility, and epigenetic landscapes are remodeled at the single-cell level. Multi-omic approaches, such as low-input chromatin accessibility and transcriptome sequencing (Li-CAT-seq) as in [59], have helped link *in vitro* observations with *in vivo* data. These data are of particular importance when studying the expression and epigenetic state of TEs in early development. Here, multi-omic approaches which can map transcriptome, chromatin accessibility, and DNA methylome at the single-cell level such as scNMT-seq [75] and snmCAT-seq [76], hold promise in better understanding the potential roles of DNA methylation in early human development. These advances are particularly powerful when coupled with low-input chromatin profiling technologies such as CUTnTag and ultra-low input ChIP which have enabled high-quality profiling of chromatin with as few as 100 cells [8]. These advances help in understanding how the epigenetic regulation of TEs, including their activity as enhancers and the dynamics of their silencing during differentiation, contribute to mammalian development.

Conclusions

Here, we summarized major changes in the RNA expression and epigenetic state of TEs across early human embryo development and in PSCs *in vitro*. Importantly, we argue here that the distinct states of pluripotency can be defined by a unique repertoire of TEs, their expression profile and epigenetic state forming a molecular fingerprint that can be used to discriminate major transitions in embryo development including ZGA, different states of pluripotency and PGCs. While similar epigenetic profiling of other embryonic stages has not been examined in the same way, such an approach holds promise for improving our understanding of how expression and epigenetic reprogramming at TEs could shape other early embryonic lineage decisions in humans. Disrupting TE expression and chromatin accessibility with CRISPRi provides a glimpse into the functional role of TEs in early human development, and this functional work is likely to expand when coupled with embryo models. It is important to note, however, that the confirmation of key observations with embryo models will benefit from access to the equivalent stages of human embryo development where possible, and with appropriate regulatory oversight and public engagement as to the use of embryos in research.

CRediT authorship contribution statement

Amander T. Clark: Conceptualization, Supervision, Writing – reviewing & editing. **Jonathan A. DiRusso:** Writing – original draft presentation, reviewing & editing, Visualization.

Data Availability

No data were used for the research described in the article.

Declaration of Competing Interest

None.

Acknowledgements

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- •• of outstanding interest.

1. Hoyt SJ, Storer JM, Hartley GA, Grady PGS, Gershman A, de Lima LG, Limouse C, Halabian R, Wojenski L, Rodriguez M, *et al*.: **From telomere to telomere: the transcriptional and epigenetic state of human repeat elements**. *Science* 2022, **376**:eabk3112.

2. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, *et al*.: **The complete sequence of a human genome**. *Science* 2022, **376**:44-53.

3. Chalopin D, Naville M, Plard F, Galiana D, Volff JN: **Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates**. *Genome Biol Evol* 2015, **7**:567-580.

4. Autio MI, Bin Amin T, Perrin A, Wong JY, Foo RS, Prabhakar S: **Transposable elements that have recently been mobile in the human genome**. *BMC Genom* 2021, **22**:789.

5. •• Mazid MA, Ward C, Luo Z, Liu C, Li Y, Lai Y, Wu L, Li J, Jia W, Jiang Y, *et al*.: **Rolling back human pluripotent stem cells to an eight-cell embryo-like stage**. *Nature* 2022, **605**:315-324.
Mazid et al. model the 8C state in culture using enhanced naïve culture conditions. Similar to (6) and (12) they note that 8CLCs are not self-renewing, thought to be a feature of totipotency, and examine distinct controllers of the 8C-state, identifying TPRX1. A similar repetoire of LTR expression to those identified in (6) are also observed in this system as well, in line with observations from 8C *in vivo*

6. •• Taubenschmid-Stowers J, Rostovskaya M, Santos F, Ljung S, Argelaguet R, Krueger F, Nichols J, Reik W: **8C-like cells capture the human zygotic genome activation program in vitro**. *Cell Stem Cell* 2022, **29**:449-459.
Taubenschmid and colleagues demonstrate that naïve cells in culture give rise to a spontaneous population of 8C-like cells, marked by expression of DUX, ZSCAN and other markers of the 8C-state

7. • Xu R, Li S, Wu Q, Li C, Jiang M, Guo L, Chen M, Yang L, Dong X, Wang H, *et al*.: **Stage-specific H3K9me3 occupancy ensures retrotransposon silencing in human pre-implantation embryos**. *Cell Stem Cell* 2022, **29**:1051-1066.
Released back-to-back with Yu et al. (8), Xu. And colleagues map the dynamics of H3K9me3 acquisition in early human development. They observe that many TEs silenced by H3K9me3 specifically at the 8-cell stage were more likely to be marked with H3K27ac or H3K4me1 in subsequent stages of development. They also note bivalent domains in TEs which are thought to be regulatory are often silenced in trophectoderm cells, an observation shared by Yu. et al in the morula

8. •• Yu H, Chen M, Hu Y, Ou S, Yu X, Liang S, Li N, Yang M, Kong X, Sun C, *et al*.: **Dynamic reprogramming of H3K9me3 at hominoid-specific retrotransposons during human preimplantation development**. *Cell Stem Cell* 2022, **29**:1031-1050.
Yu. et al. demonstrate that H3K9me3 is aquired at hominid-specific TEs, often harboring pluripotency factor motifs in the outer cells of the morula, suggesting an early role for TEs in lineage bifurcation. Additionally, they observe that SVA elements are necessary during ZGA, and that loss of SVA_D accessibility during ZGA leads to defects in development

9. Ito J, Seita Y, Kojima S, Parrish NF, Sasaki K, Sato K: **A hominoid-specific endogenous retrovirus may have rewired the gene regulatory network shared between primordial germ cells and naive pluripotent cells**. *PLoS Genet* 2022, **18**:e1009846.

10. • Xiang X, Tao Y, DiRusso J, Hsu FM, Zhang J, Xue Z, Pontis J, Trono D, Liu W, Clark AT: **Human reproduction is regulated by retrotransposons derived from ancient Hominidae-specific viral infections**. *Nat Commun* 2022, **13**:463.
Xiang et al. demonstrate that LTR5Hs, a hominidae-specific LTR, is necessaery for PGCLC specification is epigenetically remodeled and bound by SOX15, SOX17, and NANOG, which are necessary for PGC fate. Silencing of LTR5Hs elements reduces PCGLC specification, supporting a functional role of LTR5Hs in germline specification

11. • Pontis J, Pulver C, Playfoot CJ, Planet E, Grun D, Offner S, Duc J, Manfrin A, Lutolf MP, Trono D: **Primate-specific transposable elements shape transcriptional networks during human development**. *Nat Commun* 2022, **13**:7178.
Pontis and colleagues demonstrate that early embryonic lineages, including during gastrulation, are marked by compliments of transposable elements and unique KRAB-ZFP expression

12. Yu X, Liang S, Chen M, Yu H, Li R, Qu Y, Kong X, Guo R, Zheng R, Izsvak Z, *et al*.: **Recapitulating early human development with 8C-like cells**. *Cell Rep* 2022, **39**:110994.

13. Grow EJ, Flynn RA, Chavez SL, Bayless NL, Wossidlo M, Wesche DJ, Martin L, Ware CB, Blish CA, Chang HY, *et al*.: **Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells**. *Nature* 2015, **522**:221-225.

14. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G: **Transposable elements have rewired the core regulatory network of human embryonic stem cells**. *Nat Genet* 2010, **42**:631-634.

15. Sundaram V, Choudhary MN, Pehrsson E, Xing X, Fiore C, Pandey M, Maricque B, Udawatta M, Ngo D, Chen Y, *et al*.: **Functional cis-regulatory modules encoded by mouse-specific endogenous retrovirus**. *Nat Commun* 2017, **8**:14550.

16. Pontis J, Planet E, Offner S, Turelli P, Duc J, Coudray A, Theunissen TW, Jaenisch R, Trono D: **Hominoid-specific transposable elements and KZFPs facilitate human embryonic genome activation and control transcription in naive human ESCs**. *Cell Stem Cell* 2019, **24**:724-735.

17. Goke J, Lu X, Chan YS, Ng HH, Ly LH, Sachs F, Szczerbinska I: **Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells**. *Cell Stem Cell* 2015, **16**:135-141.

18. Ai Z, Xiang X, Xiang Y, Szczerbinska I, Qian Y, Xu X, Ma C, Su Y, Gao B, Shen H, *et al*.: **Kruppel-like factor 5 rewires NANOG regulatory network to activate human naive pluripotency specific LTR7Ys and promote naive pluripotency**. *Cell Rep* 2022, **40**:111240.

19. Zhang Y, Li T, Preissl S, Amaral ML, Grinstein JD, Farah EN, Destici E, Qiu Y, Hu R, Lee AY, *et al*.: **Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells**. *Nat Genet* 2019, **51**:1380-1388.

20. Choudhary MNK, Quaid K, Xing X, Schmidt H, Wang T: **Widespread contribution of transposable elements to the rewiring of mammalian 3D genomes**. *Nat Commun* 2023, **14**:634.

21. Coluccio A, Ecco G, Duc J, Offner S, Turelli P, Trono D: **Individual retrotransposon integrants are differentially controlled by KZFP/KAP1-dependent histone methylation, DNA methylation and TET-mediated hydroxymethylation in naive embryonic stem cells**. *Epigenetics Chromatin* 2018, **11**:7.

22. Ecco G, Cassano M, Kauzlaric A, Duc J, Coluccio A, Offner S, Imbeault M, Rowe HM, Turelli P, Trono D: **Transposable elements and their KRAB-ZFP controllers regulate gene expression in adult tissues**. *Dev Cell* 2016, **36**:611-623.

23. Rowe HM, Jakobsson J, Mesnard D, Rougemont J, Reynard S, Aktas T, Maillard PV, Layard-Liesching H, Verp S, Marquis J, *et al*.: **KAP1 controls endogenous retroviruses in embryonic stem cells**. *Nature* 2010, **463**:237-240.

24. Imbeault M, Helleboid PY, Trono D: **KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks**. *Nature* 2017, **543**:550-554.

25. Smith ZD, Chan MM, Humm KC, Karnik R, Mekhoubad S, Regev A, Eggan K, Meissner A: **DNA methylation dynamics of the human preimplantation embryo**. *Nature* 2014, **511**:611-615.

26. Teissandier A, Servant N, Barillot E, Bourc'his D: **Tools and best practices for retrotransposon analysis using high-throughput sequencing data**. *Mob DNA* 2019, **10**:52.

27. Chelmicki T, Roger E, Teissandier A, Dura M, Bonneville L, Rucli S, Dossin F, Fouassier C, Lameiras S, Bourc'his D: **m(6)A RNA methylation regulates the fate of endogenous retroviruses**. *Nature* 2021, **591**:312-316.

28. Kalkan T, Bornelov S, Mulas C, Diamanti E, Lohoff T, Ralser M, Middelkamp S, Lombard P, Nichols J, Smith A: **Complementary activity of ETV5, RBPJ, and TCF3 drives formative transition from naive pluripotency**. *Cell Stem Cell* 2019, **24**:785-801.

29. Kalkan T, Olova N, Roode M, Mulas C, Lee HJ, Nett I, Marks H, Walker R, Stunnenberg HG, Lilley KS, *et al*.: **Tracking the embryonic stem cell transition from ground state pluripotency**. *Development* 2017, **144**:1221-1234.

30. Leitch HG, Nichols J, Humphreys P, Mulas C, Martello G, Lee C, Jones K, Surani MA, Smith A: **Rebuilding pluripotency from primordial germ cells**. *Stem Cell Rep* 2013, **1**:66-78.

31. Theunissen TW, Friedli M, He Y, Planet E, O'Neil RC, Markoulaki S,
•• Pontis J, Wang H, Iouranova A, Imbeault M, *et al*.:
Molecular criteria for defining the naive human pluripotent state. *Cell Stem Cell* 2016, **19**:502-515.

Theunissen et al demonstrate that the TE transcriptome is significantly different between primed and naïve states of pluripotency. These findings demonstrated that HERVK/LTR5Hs and SVA_D transcription are hallmarks of naïve hPSCs in culture, a observation that was later extended to the pre-implantation epiblast

32. Kinoshita M, Barber M, Mansfield W, Cui Y, Spindlow D, Stirparo
•• GG, Dietmann S, Nichols J, Smith A: **Capture of mouse and human stem cells with features of formative pluripotency**. *Cell Stem Cell* 2021, **28**:453-471.

Kinoshita et al. demonstrate that stable culture of formative mouse PSCs can be achieved, and that similar culture conditions can also induce hPSCs into a stable, self-renewing formative state. They also demonstrate that formative hPSCs are marked by their own compliment of hPSCs, as is common between differing states of pluripotency

33. Irie N, Weinberger L, Tang WW, Kobayashi T, Viukov S, Manor YS, Dietmann S, Hanna JH, Surani MA: **SOX17 is a critical specifier of human primordial germ cell fate**. *Cell* 2015, **160**:253-268.

34. Sasaki K, Yokobayashi S, Nakamura T, Okamoto I, Yabuta Y, Kurimoto K, Ohta H, Moritoki Y, Iwatani C, Tsuchiya H, *et al*.: **Robust In vitro induction of human germ cell fatefrom pluripotent stem cells**. *Cell Stem Cell* 2015, **17**:178-194.

35. Jo K, Teague S, Chen B, Khan HA, Freeburne E, Li H, Li B, Ran R,
•• Spence JR, Heemskerk I: **Efficient differentiation of human primordial germ cells through geometric control reveals a key role for Nodal signaling**. *Elife* 2022, **11**:e72811.

Jo et al. observe that hPGCLCs can be specified from hPSCs culture on a micropatterned surface after exposure to BMP4 and Nodal. Importantly, they demonstrate that duration and concentration of Nodal and BMP4 require careful titration, and demonstrate that changes in cytokine concentration or exposure time leads to specification of amnion or ectoderm, depending on conditions

36. Minn KT, Fu YC, He S, Dietmann S, George SC, Anastasio MA,
•• Morris SA, Solnica-Krezel L: **High-resolution transcriptional and morphogenetic profiling of cells from micropatterned human ESC gastruloid cultures**. *Elife* 2020, **9**:e59445.

Minn et al. demonstrate tha treatment of micropatterned hPSCs with BPM4 gives rise to a gastruloid with cells represent each of the germ layers, including EPI-like cells and PGCLCs

37. Liu X, Ouyang JF, Rossello FJ, Tan JP, Davidson KC, Valdes DS, Schroder J, Sun YBY, Chen J, Knaupp AS, *et al*.: **Reprogramming roadmap reveals route to human induced trophoblast stem cells**. *Nature* 2020, **586**:101-107.

38. Guo G, Stirparo GG, Strawbridge SE, Spindlow D, Yang J, Clarke J, Dattani A, Yanagida A, Li MA, Myers S, *et al*.: **Human naive epiblast cells possess unrestricted lineage potential**. *Cell Stem Cell* 2021, **28**:1040-1056.

39. Niu Y, Sun N, Li C, Lei Y, Huang Z, Wu J, Si C, Dai X, Liu C, Wei J, *et al*.: **Dissecting primate early post-implantation development using long-term in vitro embryo culture**. *Science* 2019, **366**:eaaw5754.

40. Tyser RCV, Mahammadov E, Nakanoh S, Vallier L, Scialdone A, Srinivas S: **Single-cell transcriptomic characterization of a gastrulating human embryo**. *Nature* 2021, **600**:285-289.

41. Raiz J, Damert A, Chira S, Held U, Klawitter S, Hamdorf M, Lower J, Stratling WH, Lower R, Schumann GG: **The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery**. *Nucleic Acids Res* 2012, **40**:1666-1683.

42. Fuentes DR, Swigut T, Wysocka J: **Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation**. *Elife* 2018, **7**:e35989.

43. Fuchs NV, Loewer S, Daley GQ, Izsvak Z, Lower J, Lower R: **Human endogenous retrovirus K (HML-2) RNA and protein expression is a marker for human embryonic and induced pluripotent stem cells**. *Retrovirology* 2013, **10**:115.

44. Barakat TS, Halbritter F, Zhang M, Rendeiro AF, Perenthaler E, Bock C, Chambers I: **Functionaldissection of the enhancer repertoire in human embryonic stem cells**. *Cell Stem Cell* 2018, **23**:276-288.

45. Gao L, Wu K, Liu Z, Yao X, Yuan S, Tao W, Yi L, Yu G, Hou Z, Fan D, *et al*.: **Chromatin accessibility landscape in human early embryos and its association with evolution**. *Cell* 2018, **173**:248-259.

46. Lu X, Sachs F, Ramsay L, Jacques PE, Goke J, Bourque G, Ng HH: **The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity**. *Nat Struct Mol Biol* 2014, **21**:423-425.

47. Wang J, Xie G, Singh M, Ghanbarian AT, Rasko T, Szvetnik A, Cai H, Besser D, Prigione A, Fuchs NV, *et al*.: **Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells**. *Nature* 2014, **516**:405-409.

48. Carter TA, 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**:e76257.
Carter et al. refine the phylogeny of the HERVH LTR, LTR7. In addition to new phylogeny, they demonstrate that LTR7 expression in the post-implantation epiblast is dominated by LTR7up1 and LTR7up2, which are hominidae-specific LTR7s, in line with observation that most regulatory TEs in the early embryo are evolutionarily young and not shared with lower primates

49. Chen D, Sun N, Hou L, Kim R, Faith J, Aslanyan M, Tao Y, Zheng Y, Fu J, Liu W, *et al*.: **Human primordial germ cells are specified from lineage-primed progenitors**. *Cell Rep* 2019, **29**:4568-4582.

50. Tang WW, Dietmann S, Irie N, Leitch HG, Floros VI, Bradshaw CR, Hackett JA, Chinnery PF, Surani MA, Unique A: **Gene regulatory network resets the human germline epigenome for development**. *Cell* 2015, **161**:1453-1467.

51. Sasaki K, Nakamura T, Okamoto I, Yabuta Y, Iwatani C, Tsuchiya H, Seita Y, Nakamura S, Shiraki N, Takakuwa T, *et al*.: **The germ cell fate of cynomolgus monkeys Is specified in the nascent amnion**. *Dev Cell* 2016, **39**:169-185.

52. Zheng Y, Yan RZ, Sun S, Kobayashi M, Xiang L, Yang R, Goedel A, Kang Y, Xue X, Esfahani SN, *et al*.: **Single-cell analysis of embryoids reveals lineage diversification roadmaps of early human development**. *Cell Stem Cell* 2022, **29**:1402-1419.

53. Wang X, Veerapandian V, Yang X, Song K, Xu X, Cui M, Yuan W, Huang Y, Xia X, Yao Z, *et al*.: **The chromatin accessibility landscape reveals distinct transcriptional regulation in the induction of human primordial germ cell-like cells from pluripotent stem cells**. *Stem Cell Rep* 2021, **16**:1245-1261.

54. Chen D, Liu W, Zimmerman J, Pastor WA, Kim R, Hosohama L, Ho J, Aslanyan M, Gell JJ, Jacobsen SE, *et al*.: **The TFAP2C-regulated OCT4 naive enhancer Is involved in human germline formation**. *Cell Rep* 2018, **25**:3591-3602.

55. Posfai E, Schell JP, Janiszewski A, Rovic I, Murray A, Bradshaw B, Yamakawa T, Pardon T, El Bakkali M, Talon I, *et al*.: **Evaluating totipotency using criteria of increasing stringency**. *Nat Cell Biol* 2021, **23**:49-60.

56. Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, Firth A, Singer O, Trono D, Pfaff SL: **Embryonic stem cell potency fluctuates with endogenous retrovirus activity**. *Nature* 2012, **487**:57-63.

57. Hendrickson PG, Dorais JA, Grow EJ, Whiddon JL, Lim JW, Wike CL, Weaver BD, Pflueger C, Emery BR, Wilcox AL, *et al*.: **Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons**. *Nat Genet* 2017, **49**:925-934.

58. De Iaco A, Planet E, Coluccio A, Verp S, Duc J, Trono D: **DUX-family transcription factors regulate zygotic genome activation in placental mammals**. *Nat Genet* 2017, **49**:941-945.

59. Liu L, Leng L, Liu C, Lu C, Yuan Y, Wu L, Gong F, Zhang S, Wei X, Wang M, *et al*.: **An integrated chromatin accessibility and transcriptome landscape of human pre-implantation embryos**. *Nat Commun* 2019, **10**:364.

60. Wu J, Xu J, Liu B, Yao G, Wang P, Lin Z, Huang B, Wang X, Li T, Shi S, *et al*.: **Chromatin analysis in human early development reveals epigenetic transition during ZGA**. *Nature* 2018, **557**:256-260.

61. Caulfield T, Zarzeczny A, McCormick J, Bubela T, Critchley C, Einsiedel E, Galipeau J, Harmon S, Huynh M, Hyun I, *et al*.: **The stem cell research environment: a patchwork of patchworks**. *Stem Cell Rev Rep* 2009, **5**:82-88.

62. Clark AT, Brivanlou A, Fu J, Kato K, Mathews D, Niakan KK, Rivron N, Saitou M, Surani A, Tang F, *et al*.: **Human embryo research, stem cell-derived embryo models and in vitro gametogenesis: Considerations leading to the revised ISSCR guidelines**. *Stem Cell Rep* 2021, **16**:1416-1424.

63. Yu L, Wei Y, Duan J, Schmitz DA, Sakurai M, Wang L, Wang K,
•• Zhao S, Hon GC, Wu J: **Blastocyst-like structures generated from human pluripotent stem cells**. *Nature* 2021, **591**:620-626.
Yu et. al introduce model system to derive human blastocyst embryoids, a so-called blastoid. The system devised by Yu. et. al relies on step-wise exposure of developing blastoids to medias which bias naïve hPSCs toward trophectoderm or hypoblast lineages. This step-wise process leads to formation of blastoids

64. Liu X, Tan JP, Schroder J, Aberkane A, Ouyang JF, Mohenska M,
•• Lim SM, Sun YBY, Chen J, Sun G, *et al*.: **Modelling human blastocysts by reprogramming fibroblasts into iBlastoids**. *Nature* 2021, **591**:627-632.
Liu. et al demonstrate that reprogramming of fibroblasts under specific minimal conditions followed by 3D culture in an Aggrewell and exposure to a media cocktail including HDAC inhibitors, WNT activation, ALK5/TGFb, and ALK4/NODAL inhibitors gives rise to induced blastoids (iBlastoids) which model the early pre-implantation embryo

65. Kagawa H, Javali A, Khoei HH, Sommer TM, Sestini G,
•• Novatchkova M, Scholte Op Reimer Y, Castel G, Bruneau A, Maenhoudt N, *et al*.: **Human blastoids model blastocyst development and implantation**. *Nature* 2022, **601**:600-605.
Kagawa et al. demonstrate that naïve hPSCs treated with a cocktail of Hippo, TGFb and ERK inhibitors in hydrogel wells self-assemble into blastoids. These blastoids, like other blastoid models, most closely model the pre-implantation embryo

66. Sozen B, Jorgensen V, Weatherbee BAT, Chen S, Zernicka-Goetz M: **Reconstructing aspects of human embryogenesis with pluripotent stem cells**. *Nat Commun* 2021, **12**:5550.

67. Moris N, Anlas K, van den Brink SC, Alemany A, Schroder J, Ghimire S, Balayo T, van Oudenaarden A, Martinez Arias A: **An in vitro model of early anteroposterior organization during human development**. *Nature* 2020, **582**:410-415.

68. Simunovic M, Metzger JJ, Etoc F, Yoney A, Ruzo A, Martyn I, Croft G, You DS, Brivanlou AH, Siggia ED: **A 3D model of a human epiblast reveals BMP4-driven symmetry breaking**. *Nat Cell Biol* 2019, **21**:900-910.

69. Shao Y, Taniguchi K, Townshend RF, Miki T, Gumucio DL, Fu J: **A pluripotent stem cell-based model for post-implantation human amniotic sac development**. *Nat Commun* 2017, **8**:208.

70. Zheng Y, Xue X, Shao Y, Wang S, Esfahani SN, Li Z, Muncie JM,
•• Lakins JN, Weaver VM, Gumucio DL, *et al*.: **Controlled modelling of human epiblast and amnion development using stem cells**. *Nature* 2019, **573**:421-425.
Zheng et al. updated the PASE model with the P-ELS model and find, upon induction using BMP4, that amnion-like cells and PGCLCs are specified in embyponic micro-patterned culture devices. Change to the compliment of cytokines and inhibitors added results in different lineage biases, enabling its use as a model to understand the interplay between signaling cues in the post-implantation epiblast

71. Warmflash A, Sorre B, Etoc F, Siggia ED, Brivanlou AH: **A method to recapitulate early embryonic spatial patterning in human embryonic stem cells**. *Nat Methods* 2014, **11**:847-854.

72. Berrens RV, Yang A, Laumer CE, Lun ATL, Bieberich F, Law CT, Lan G, Imaz M, Bowness JS, Brockdorff N, *et al*.: **Locus-specific expression of transposable elements in single cells with CELLO-seq**. *Nat Biotechnol* 2022, **40**:546-554.

73. Parker MT, Knop K, Sherwood AV, Schurch NJ, Mackinnon K, Gould PD, Hall AJ, Barton GJ, Simpson GG: **Nanopore direct RNA**

sequencing maps the complexity of Arabidopsis mRNA processing and m(6)A modification. *Elife* 2020, **9**:e49658.

74. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W: **Detecting DNA cytosine methylation using nanopore sequencing**. *Nat Methods* 2017, **14**:407-410.

75. Clark SJ, Argelaguet R, Kapourani CA, Stubbs TM, Lee HJ, Alda-Catalinas C, Krueger F, Sanguinetti G, Kelsey G, Marioni JC, *et al.*: **scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells**. *Nat Commun* 2018, **9**:781.

76. Luo C, Liu H, Xie F, Armand EJ, Siletti K, Bakken TE, Fang R, Doyle WI, Stuart T, Hodge RD, *et al.*: **Single nucleus multi-omics identifies human cortical cell regulatory genome diversity**. *Cell Genom* 2022, **2**:100107.

Chapter 3

Human reproduction is regulated by retrotransposons derived from ancient *Hominidae*-

specific viral infections

ARTICLE

Check for updates

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

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

Germ cells are essential to pass DNA from one generation to the next. In human reproduction, germ cell development begins with the specification of primordial germ cells (PGCs) and a failure to specify PGCs leads to human infertility. Recent studies have revealed that the transcription factor network required for PGC specification has diverged in mammals, and this has a significant impact on our understanding of human reproduction. Here, we reveal that the *Hominidae*-specific Transposable Elements (TEs) LTR5Hs, may serve as TEENhancers (TE Embedded eNhancers) to facilitate PGC specification. LTR5Hs TEENhancers become transcriptionally active during PGC specification both in vivo and in vitro with epigenetic reprogramming leading to increased chromatin accessibility, localized DNA demethylation, enrichment of H3K27ac, and occupation of key hPGC transcription factors. Inactivation of LTR5Hs TEENhancers with KRAB mediated CRISPRi has a significant impact on germ cell specification. In summary, our data reveals the essential role of *Hominidae*-specific LTR5Hs TEENhancers in human germ cell development.

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

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

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

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

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

The most recent expansion of human ERVs occurred over the last 5-20 million years in the HERVK (human mouse mammary tumor virus like-2, HML-2) group[27]. Even though HERVK(HML2) elements are also found in old world primates, distinct phylogenetic differences exist between those found in Hominidae relative to Hominoidea. For example, HERVK(HML2) elements which are found in both monkey and human genomes have LTR5A and LTR5B regulatory sequences, while the most recent Hominidae-specific HERVK(HML2) elements harbor the LTR5Hs regulatory sequence[27]. In addition, some HERVK(HML2) TEs contain intact open reading frames that can code for viral proteins[28], with LTR5Hs-regulated HERVK(HML2) provirus expression proposed to be a property of naïve human embryonic stem cells (hESCs)[29]. Grow and colleagues further hypothesize that expression of full-length LTR5Hs-regulated HERVK(HML2) proviruses may confer a critical immunoprotective effect in the human pre-implantation embryo by stimulating IFITM-1, a viral restriction factor, potentially protecting against HERVK(HML-2)-like retrovirus re-infection[29].

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

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

Results

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

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

33
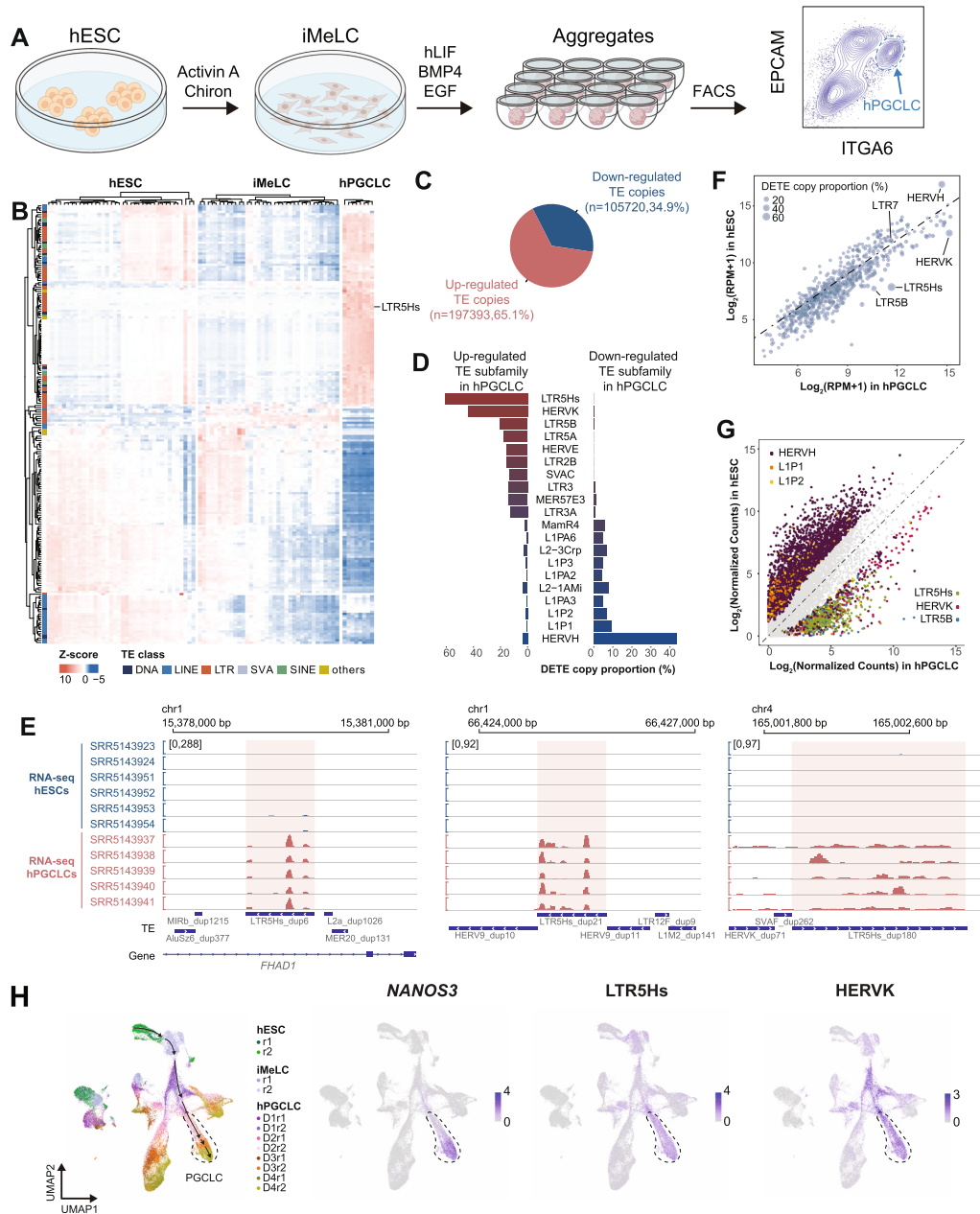
SQuIRE, Telescope, or TEtranscripts[32–35] followed by DETE calling with DESeq2[37] (Supplementary Fig. 1A and Supplementary Data 2–5). Using a four-fold difference and false discovery rate (FDR) <0.05 as a cut-off, we identified more up-regulated DETE copies in hPGCLCs compared to hESCs (65.1% for featureCounts; 66.1% for Telescope; 71.4% for TEtranscripts) except for SQuIRE

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

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

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

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

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

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

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

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

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

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

Increased chromatin accessibility of LTR5Hs in hPGCLCs. Given the potential enhancer role of TEs in regulating gene expression[30,42], we next evaluated changes in chromatin accessibilities during hPGCLC induction with our previously published Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) data[16]. Comparing ATAC-seq data of primed-state hESCs, iMeLCs, and hPGCLCs, we identified 31,276 and 90,201 ATAC-seq peaks that become specifically more accessible in hPGCLCs (referred as hPGCLC open regions, hPGCLC-ORs) and hESCs (referred as hESC open regions, hESC-ORs), respectively (Fig. 2A, Supplementary Fig. 5A, and Supplementary Data 6).

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

35
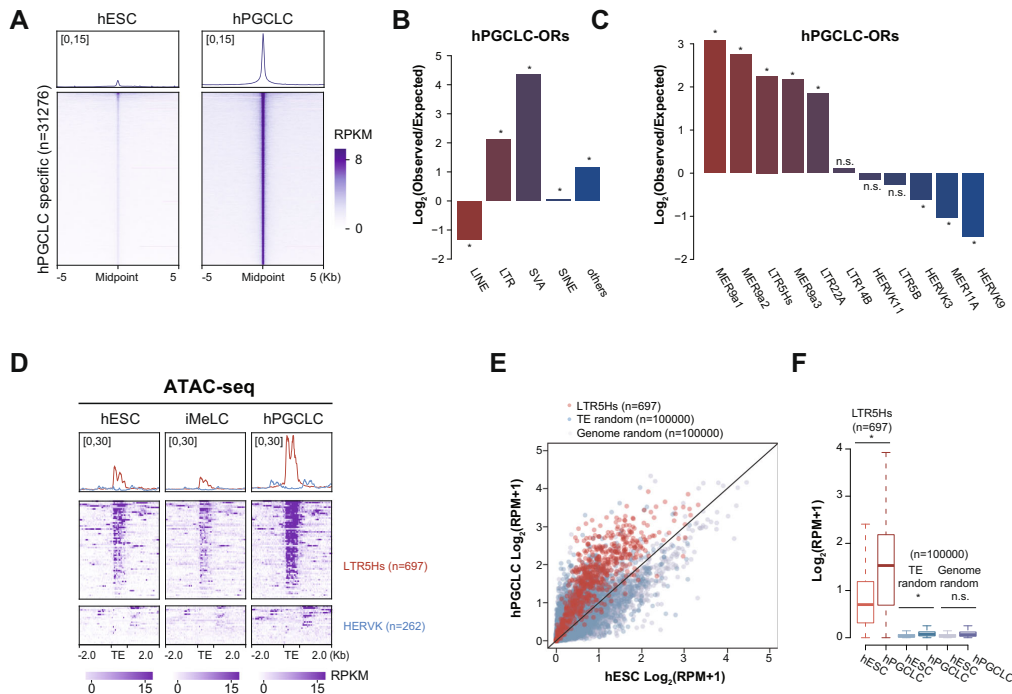
Fig. 2 Increased chromatin accessibility over LTR5Hs with hPGCLC induction. A Heatmap and metaplot for ATAC-seq signals over hPGCLC-ORs ($n = 31276$). **B**, **C** Enrichment of TE classes (**B**), and TE subfamilies within the ERVK family (**C**) for hPGCLC-ORs over random shuffled regions with comparable genomic distributions (*p-value < 0.05, binomial test; n.s. = not significant). **D** Heatmap and metaplot of ATAC-seq signals over all LTR5Hs ($n = 697$) and HERVK copies ($n = 262$) in hESCs, iMeLCs, and hPGCLCs. **E**, **F** Scatterplot (**E**) and boxplot (**F**) of ATAC-seq signals over LTR5Hs, randomly shuffled TE and genomic regions in hESCs and hPGCLCs (*p-value < 0.05, Welch Two Sample t-test; n.s. = not significant). In **F** the middle line represents the median; boxes represent the 25th (bottom) and 75th (top) percentiles; and whiskers represent the minimum and maximum points within 1.5× the interquartile range. Source data underlying **B**, **C**, and **F** are provided as a Source Data file.

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

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

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

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

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

36
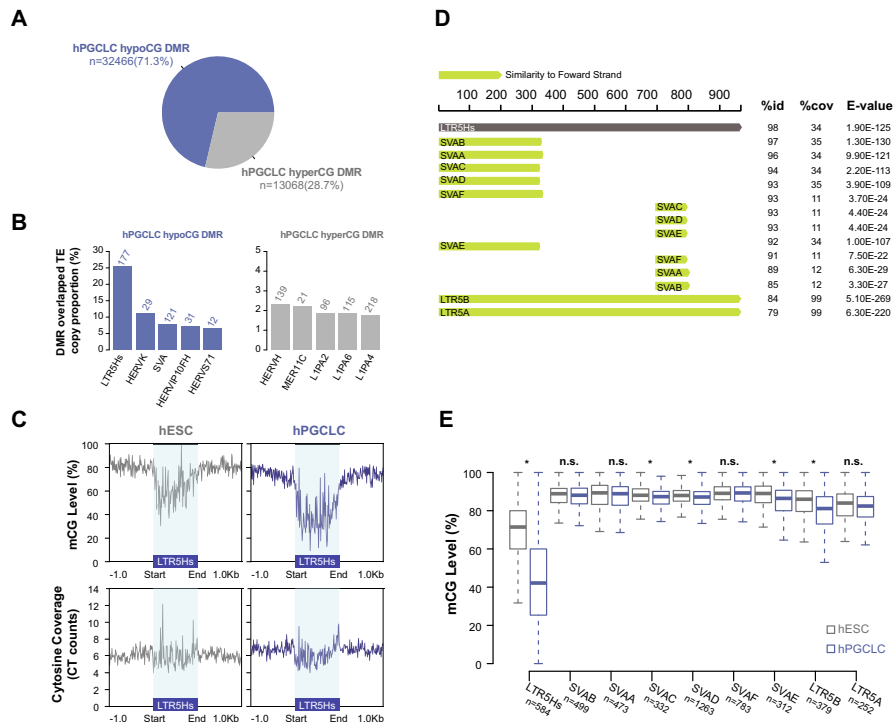
Fig. 3 Localized DNA demethylation over LTR5Hs in hPGCLCs. A Percentage of hypoCG and hyperCG DMRs in hPGCLCs compared to hESCs. **B** Bar plot showing enrichment of TE subfamilies in hyperCG or hypoCG DMRs as a proportion of total TE copy number. **C** Metaplot of aggregate CG methylation level (top) and cytosine coverage (bottom) over LTR5Hs in hESCs and hPGCLCs. Blue shaded rectangle region indicates annotated LTR5Hs regions. **D** The consensus sequence similarity of LTR5Hs and related TE clades from Dfam[45]. Percent identity between the entry consensus sequences (%id), percent shared coverage (%cov) and match e-value (E-value) are displayed on the right. **E** Boxplot of CG methylation level over LTR5Hs and related TE clades in hESCs and hPGCLCs. Only TE subfamilies with a copy number >100 are included in the plot. *p-value < 0.05, Welch Two Sample t-test; n.s. represents not significant. The middle line represents the median; boxes represent the 25th (bottom) and 75th (top) percentiles; and whisker bars represent the minimum and maximum points within the 1.5× interquartile range. Source data underlying **E** is provided as a Source Data file and originates from $n = 2$ biological replicates (separate experiments) of hESCs and hPGCLCs generated from the UCLA2 hESC line.

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

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

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

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

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

Motif analysis for NANOG and SOX17 ChIP-seq peaks in hPGCLCs validated the quality of our ChIP-seq experiments, with the most enriched motif as Nanog and the Oct4:Sox17 fused motif, respectively (Fig. 4A, B). Interestingly, we detected the
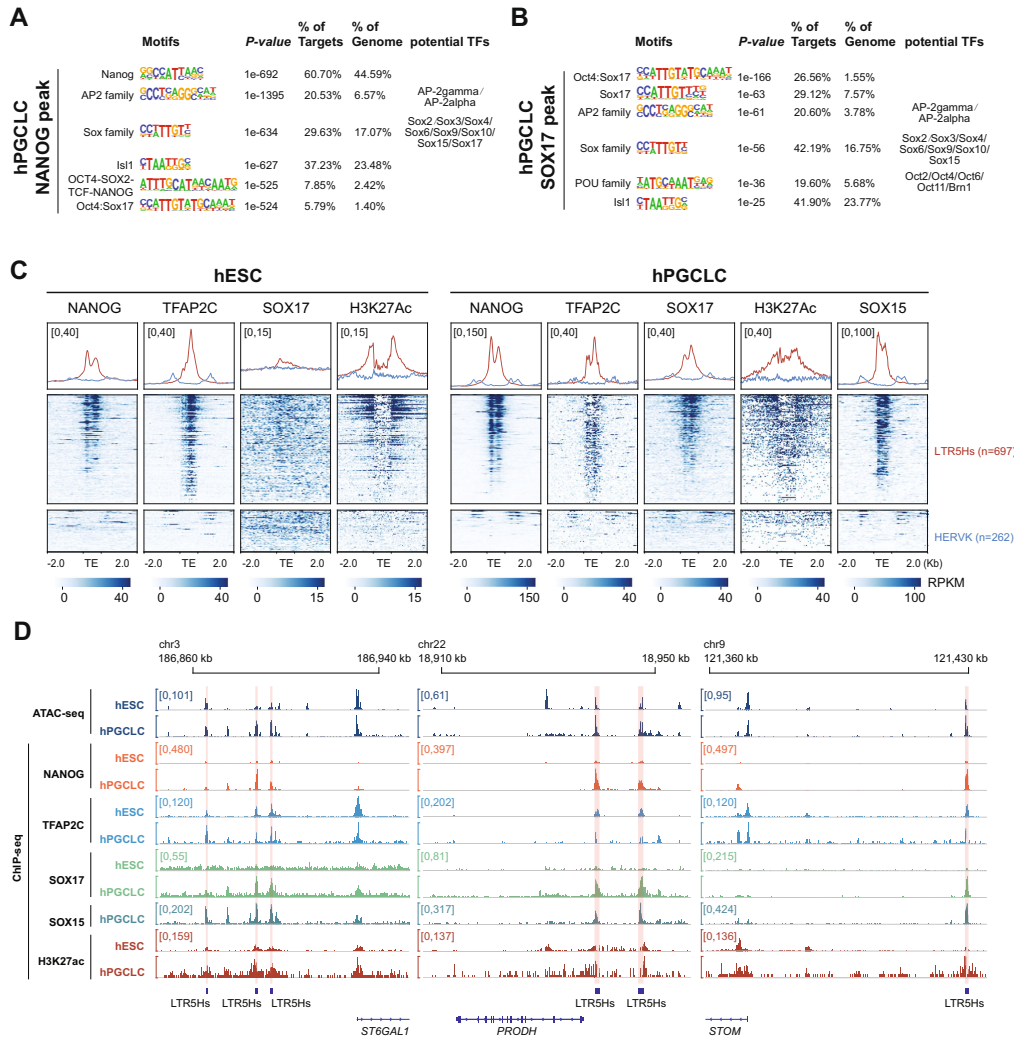
Fig. 4 LTR5Hs may serve as TEENhancers in hPGCLCs. A, B Top enriched motifs within NANOG (**A**) and SOX17 (**B**) ChIP-seq peaks in hPGCLCs. Hypergeometric test is performed using Homer[77]. **C** Heatmaps and metaplots of NANOG, TFAP2C, SOX17, and H3K27ac ChIP-seq signals in hESCs and hPGCLCs, and SOX15 CUT&Tag-seq signals in hPGCLCs over all LTR5Hs ($n = 697$) and HERVK ($n = 262$). **D** Screenshots for ATAC-seq signals, NANOG, TFAP2C, SOX17, and H3K27ac ChIP-seq signals in hESCs and hPGCLCs, and CUT&Tag-seq signals for SOX15 over LTR5Hs TEENhancers and their potential target genes (*ST6GAL1*, *PRODH*, and *STOM*).

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

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

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

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

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

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

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

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

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

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

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

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

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

Discussion

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

39
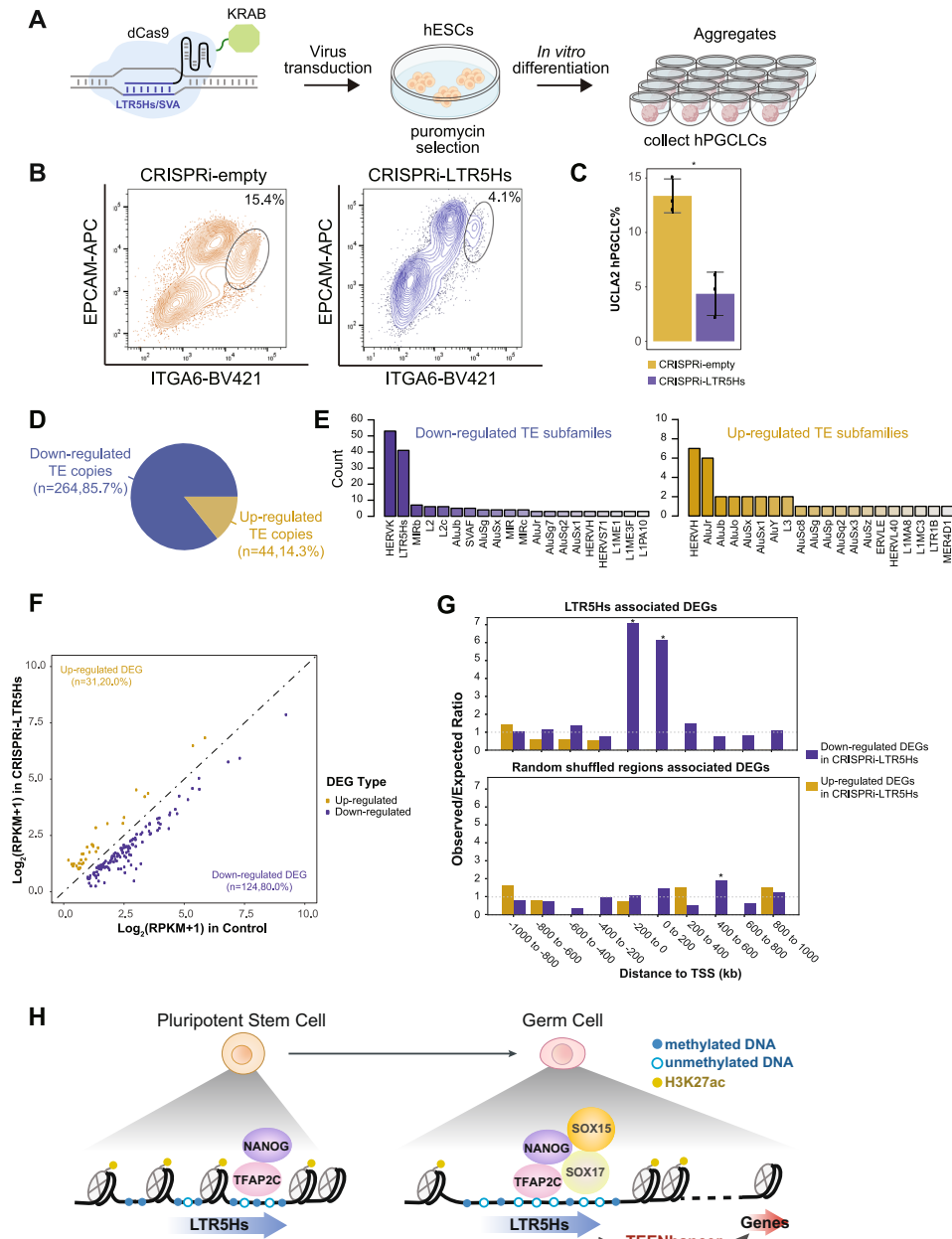
as the rhesus or cynomologous macaque. These relatively new *Hominidae* TEs have evolved species-specific functions which are unique to apes, and in some cases are unique to humans.

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

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

40

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

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

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

Despite lack of KLF4 activity and KLF17 expression in germ cell specification, hPGCLCs in vitro and hPGCs in vivo exhibit a naïve-like pluripotent molecular program that has similarities to the naïve state in pre-implantation human embryos. This includes two active X chromosomes in females, genome-wide DNA demethylation and expression of naïve pluripotent TFs including KLF4, TFCP2L1, and TFAP2C[9,16,52–54]. Similar to KLF4-KLF17, TFAP2C also regulates transcription and the identity of naïve pluripotent stem cells by opening naïve-specific enhancers to regulate neighboring gene expression[17,30]. Our results imply that the commissioning of LTR5Hs TEENhancers during hPGCLC induction is driven by the marking of these sites in primed hESCs by a basic network of TFAP2C and NANOG, which is then reinforced with the recruitment of SOX17 and SOX15 during hPGCLC induction. Further supporting this interpretation, time-resolved ATAC-seq during hPGCLC induction from Wang et al. shows Sox15 and Oct4:Sox17 motifs become preferentially open during the second day of the four-day hPGCLC differentiation protocol[7], roughly concomitant with enrichment of naïve-state gene profiles by hPGCLCs[2]. Thus, proper LTR5Hs TEENhancer activity may be necessary for acquisition of a naïve-like transcriptome during hPGCLC induction.

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

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

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

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

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

41

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

In addition to gene regulation at enhancers and promoters, human ERVs are also known to regulate 3-D chromatin architecture in pluripotent stem cells. Specifically, HERVH is highly expressed in primed pluripotent stem cells, and is involved in maintaining 3-D chromatin structure[59]. Given that HERVH is down-regulated and LTR5Hs sequences are up-regulated during hPGCLC induction, it is possible that LTR5Hs may also be required for the assembly of genome 3-D architecture in hPGCLCs, and therefore the failure to fully repress HERVH in the CRISPRi-LTR5Hs hPGCLCs.

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

Modern and archaic humans began to diverge ~500,000 years ago with modern humans becoming the dominant surviving human species ~50,000 years ago[63,64]. Extinction occurs when reproduction fails. Considering the contributions of TEs to the renewal of the genetic pool during evolution, one hypothesis could be that human-specific TEs, like LTR5Hs became beneficial to germ cell specification and consequently improved human reproductive fitness. As we have identified multiple TF networks that converge on LTR5Hs, it is also possible that other factors not profiled in this work contribute to the specification and reinforcement of PGC fate. Likewise, advances in recent techniques to model the early embryo could provide additional platforms to dissect the networks which delineate the naïve state networks in the pre-implantation embryo.

Methods

Ethics statement
The UCLA2 and UCLA1 hESC lines were derived at UCLA by the UCLA Pluripotent Stem Cell Core Facility following Institutional Review Board (IRB) and UCLA Embryonic Stem Cell Research Oversight (ESCRO) Committee Approvals. Informed consent was obtained after the embryo donors contacted the UCLA Broad Stem Cell Research Center to inquire about donating surplus embryos following in vitro fertilization. Embryo donors were not paid and were able to freely withdraw consent to use the embryos for stem cell research up to the point of hESC derivation when the embryo is destroyed. Informed consent was obtained from all embryo donors prior to sending frozen donated embryos to UCLA. Once derived, the hESC lines were authenticated using Affymetrix Genome-wide Human SNP Array 6.0 to detect Single Nucleotide Polymorphisms and Copy Number Variant (SNP/CNV) prior to distribution. The UCLA1 and UCLA2 hESC lines are provided to researchers de-identified, with all links and identifiers broken prior to distribution. All de-identified hESC lines used in this study are registered with the National Institute of Health Human Embryonic Stem Cell Registry and are available for research use with NIH funds. Mycoplasma test (Lonza, LT07-418) was performed every month. All experiments using the de-identified hESC lines were approved by the UCLA Embryonic Stem Cell Research Oversight Committee.

Cell culture
UCLA2 and UCLA1 hESC are cultured as previously described[16], briefly the hESCs are cultured in hESC media, which was composed of 20% knockout serum replacement (KSR) (Life Technologies, A3181502), 1x MEM Non-Essential Amino Acids (NEAA) (Fisher Scientific, 25-025-CI), 1x Penicillin/Streptomycin/Glutamine (Thermo Fisher, 10378016), 55 μM 2- Mercaptoethanol (Life Technologies, 21985-023), 10 ng/mL recombinant human FGF basic (Proteintech, HZ1285), and 50 ng/mL primocin (InvivoGen, ant-pm-2) in DMEM/F12 media (GIBCO, 11330-032). The primed hESCs were split by 1 mg/ml Collagenase type IV (GIBCO, 17104-019) and maintained routinely on mitomycin C (MMC)-inactivated mouse embryonic fibroblasts (MEFs). The hESCs were split every 7 days using Collagenase type IV (GIBCO, 17104-019). HEK293 cells were acquired from ATCC (Cat# CRL-3216). No lines used in this study belong to the International Cell Line Authentication Committee register of misidentified cell lines.

hPGCLC differentiation
Using the UCLA2 hESC line, the differentiation of hPGCLCs in vitro was performed as previously described[4,5]. Specifically, 0.05% trypsin-EDTA (Thermo Fisher Scientific, 25300120) was used to digest confluent hESCs cultured on mitomycin C inactivated mouse embryonic fibroblasts (MEFs) into single cells, followed by plating onto a 12-well-plate that had previously been coated with human plasma fibronectin (Life Technologies, 33016-015) for at least 1 hour (h). Cells were plated at cell density of 200,000 cells/well in 2 mL/well of incipient mesoderm-like cells (iMeLCs) medium, which is composed of 15% knockout serum replacement (KSR, Life Technologies, A3181502), 1x MEM Non-Essential Amino Acids (NEAA) (Fisher Scientific, 25-025-CI), 55 μM 2-Mercaptoethanol (Life Technologies, 21985-023), 1x Penicillin/Streptomycin/Glutamine (Thermo Fisher, 10378016), 1 mM sodium pyruvate (Life Technologies, 11360070), 50 ng/mL Activin A (PeproTech, AF-120-14E), 3 mM CHIR99021 (Reprocell, 04-0004-10), 10 mM ROCKi (Y27632, Stemgent, 04-0012-10), and 50 ng/mL primocin (InvivoGen, ant-pm-2) in Glasgow's minimal essential medium (GMEM) (Life Technologies, 11710035). After 24 h, iMeLCs were dissociated into single cells by 0.05% trypsin-EDTA (Thermo Fisher Scientific, 25300120), then plated into ultra-low cell attachment U-bottom 96-well plates (Corning, 7007) at a density of 3000 cells/well in 200 mL/well of hPGCLC medium, which is composed of 15% KSR (Life Technologies, A3181502), 1x MEM Non-Essential Amino Acids (NEAA) (Fisher Scientific, 25-025-CI), 55 μM 2- Mercaptoethanol (Life Technologies, 21985-023), 1x Penicillin/Streptomycin/Glutamine (Thermo Fisher, 10378016), 1 mM sodium pyruvate (Life Technologies, 11360070), 10 ng/mL recombinant human leukemia inhibitory factor (Sigma-Aldrich, LIF1010), 200 ng/mL human BMP4 (R&D systems, 314-BP), 50 ng/mL human epidermal growth factor (Fisher Scientific, 236EG200), 10 mM of ROCKi (Y27632, Stemgent, 04-0012-10), and 50 ng/mL primocin in GMEM (Life Technologies, 11710035). Day-4 hPGCLC aggregates were collected for further analysis.

Flow cytometry and fluorescence-activated cell sorting
hPGCLC aggregates were dissociated with 0.05% Trypsin-EDTA (Thermo Fisher Scientific, 25300120) for 10 minutes (min) at 37 °C. The dissociated cells were then stained with conjugated antibodies, washed with FACS buffer (1% BSA in PBS) and resuspended in FACS buffer with 7-AAD (BD PharMingen, 559925) as viability dye. The single-cell suspension was sorted for further experiments. For SOX17 ChIP-seq, all hPGCLCs were collected and sorted by BD FACSDiva v8.0.2. For NANOG ChIP-seq, 96 aggregates were sampled via FACS, while the remaining aggregates were dissociated in parallel before being fixed and flash frozen (see TF Chromatin Immunoprecipitation). The antibodies used in this study are: BV421 conjugated anti-human/mouse CD49f (ITGA6) (BioLegend; Cat#313624; RRID: AB_2562244; Lot#B274314) at 1/80; APC-conjugated anti-human CD326 (EPCAM) (BioLegend; Cat#324208; RRID: AB_756082; Lot#B284158) at 1/80.

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

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

42

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

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

CRISPR/dCas9-kRAB assay. Two gRNAs (gRNA55 and gRNA57) that targeted LTR5Hs were designed by Pontis and colleagues[30]. The two gRNAs were cloned into pLV-dCas9-KRAB-T2a-Puro (Addgene 71236), and the plasmid with no LTR5Hs gRNA was used as a control. Using a second-generation lentiviral system we generated dCas9-KRAB-gRNA55, dCas9-KRAB-gRNA57, and dCas9-KRAB-empty virus in HEK293T cells (ATCC, Manassas, Virginia, Cat# CRL-3216). Supernatants that contain lentivirus were then collected and ultracentrifuged. Confluent hESC were trypsinized with 0.05% trypsin at 37 °C for 5 min, then 200k cells were counted and collected to mix with the concentrated lentivirus. After mixture on nutator for two hours at room temperature, cells were transferred onto mitomycin C treated MEFs in hESC media with 10 mM ROCKi. After transduction into hESC, 1ug/mL puromycin was used to screen for positive cells for at least 5 days. Surviving cells were then used to perform downstream assays.

RNA-sequencing. RNA-seq was performed as previously described[65]. Briefly the hPGCLCs were directly sorted into 350 µL RLT lysis buffer (QIAGEN RNeasy micro kit, 220006-800). Total RNA was then extracted by RNeasy micro kit (Qiagen RNeasy micro kit, 220006-800). Total RNA was reverse transcribed and cDNA was amplified using Ovation RNA-Seq System V2 (Tecan, 7102-32) according to the manufacturer's instructions. Amplified cDNA was then sheared to ~200 bp length by Covaris S220 Focused ultrasonicator. RNA-seq libraries were constructed by using Ovation Rapid Library Systems (Tecan, 0319-32) and quantified by a KAPA library quantification kit (Kapa Biosystems, kk4824). Libraries were then subjected to pair-end sequencing on Illumina NovaSeq 6000 sequencer.

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

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

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

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

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

RNA-seq analysis. Other than methods used for the TE quantification, "STAR + featureCounts + DESeq2" method was applied for both TE and gene quantification and DETE/DEG calling in the article. Besides GSE93126 RNA-seq data[16] of hESC to hPGCLC differentiation (Supplementary Data 1), other RNA-seq datasets used in this article including RNA-seq of CRISPRi in hPGCLCs generated from this paper, RNA-seq data of hESC multilineage differentiation from previous publication[41] GSE16256 (Supplementary Data 1), and RNA-seq data of CRISPRi in naïve hESCs from previous publication[30] GSE117395 (Supplementary Data 1). For RNA-seq data quality control, alignment and track generation, FastQC[70] v0.11.8, STAR[36] v2.7.0e and deeptools[72] v3.4.3 were applied as described in "TE quantification methods comparison" section.

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

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

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

scRNA-seq analysis. Two biological replicates for the scRNA-seq data of hESC to hPGCLC differentiation (UCLA2 line) was downloaded from previous publication[2] GSE140021 (Supplementary Data 1). The reads were quantified by 10x Genomics Cell Ranger[74] v.3.1.0 to both gene and TE reference genome with default parameters. The generated cell-by-gene/TE unique molecular identifier (UMI) count matrix was analyzed in Seurat[75] R package v3.2.2. Due to limited coverage in scRNA-seq data, we aggregated reads from individual TE copies to TE subfamilies for downstream analysis.

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

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

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

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

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

TE enrichment analysis over hPGCLC- or hESC-ORs. TE annotation for hPGCLC- or hESC-ORs was conducted by Homer[77] v4.7 annotatePeaks.pl function using GRCh38 TE annotation file.

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

The expected TE occurrence was calculated by the average number of TE copies which were intersected with the 10,000 combined random shuffled hPGCLC/hESC-ORs. If the expected TE copy occurrence for certain TE subfamily was smaller than 1, it was rounded to 1. The observed TE occurrence was counted based on the TE annotation for hPGCLC/hESC-ORs. The value of $\log_2$ transformed "observed TE occurrence/expected TE occurrence" was used as enrichment score, with one-sided exact binomial test for statistical test.

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

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

WGBS analysis. Raw WGBS data were downloaded from previous publication[44] GSE139115 (Supplementary Data 1). Reads were aligned with BSMAP[68] v.2.74 by mapping reads to all four strands (-n 1), allowing maximum one equal best hits and less than two mismatches per read (-w 1, -v 2). Aligned reads in bam format for biological replicates of hESC and hPGCLCs were merged to increase the coverage using SAMtools[71] v1.9 merge function. Methratio.py script built in BSMAP were used to calculate cytosine counts only keeping unique mappings (-u), non-duplicated reads (-r), and reporting loci with zero methylation ratios (-z). Methylation level at CG sites was then calculated by $\#C/(\#C + \#T)$.

To visualize the CG methylation level and cytosine coverage over LTR5Hs and SVAD, only CG sites with ≥3 covered reads were retained. Wiggle tracks were generated with customized perl script and converted to bigwig with wigToBigWig[78] v4 followed by data visualization with deeptools[72] v3.4.3.

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

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

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

For all ChIP-seq data, quality control was performed by FastQC[70] v0.11.8. Then reads were aligned by STAR[36] v2.7.0e allowing maximal 1000 multiple mapped reads with no more than three mismatches and the best hit was kept (--outFilterMultimapNmax 1000 --outFilterMismatchNmax 3 --outSAMmultNmax 1). Splice junction is neglected by building STAR index without general feature format file and not allowing intron length (--alignIntronMax 1). PCR duplicates were removed using SAMtools[71] v1.9 rmdup function.

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

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

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

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

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

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

44

2. Chen, D. et al. Human primordial germ cells are specified from lineage-primed progenitors. *Cell Rep.* **29**, 4568–4582.e5 (2019).

3. Irie, N. et al. SOX17 is a critical specifier of human primordial germ cell fate. *Cell* **160**, 253–268 (2015).

4. Sasaki, K. et al. Robust in vitro induction of human germ cell fate from pluripotent stem cells. *Cell Stem Cell* **17**, 178–194 (2015).

5. Chen, D. et al. Germline competency of human embryonic stem cells depends on eomesodermin. *Biol. Reprod.* **97**, 850–861 (2017).

6. Kojima, Y. et al. Evolutionarily distinctive transcriptional and signaling programs drive human germ cell lineage specification from pluripotent stem cells. *Cell Stem Cell* **21**, 517–532.e5 (2017).

7. Wang, X. et al. The chromatin accessibility landscape reveals distinct transcriptional regulation in the induction of human primordial germ cell-like cells from pluripotent stem cells. *Stem Cell Rep.* **16**, 1245–1261 (2021).

8. Yamaguchi, S. et al. Conditional knockdown of *Nanog* induces apoptotic cell death in mouse migrating primordial germ cells. *Development* **136**, 4011–4020 (2009).

9. Guo, F. et al. The transcriptome and DNA methylome landscapes of human primordial germ cells. *Cell* **161**, 1437–1452 (2015).

10. Ohinata, Y. et al. Blimp1 is a critical determinant of the germ cell lineage in mice. *Nature* **436**, 207–213 (2005).

11. Weber, S. et al. Critical function of AP-2gamma/TCFAP2C in mouse embryonic germ cell maintenance1. *Biol. Reprod.* **82**, 214–223 (2010).

12. Yamaji, M. et al. Critical function of Prdm14 for the establishment of the germ cell lineage in mice. *Nat .Genet.* **40**, 1016–1022 (2008).

13. Sybirna, A. et al. A critical role of PRDM14 in human primordial germ cell fate revealed by inducible degrons. *Nat. Commun.* **11**, 1282 (2020).

14. Hara, K. et al. Evidence for crucial role of hindgut expansion in directing proper migration of primordial germ cells in mouse early embryogenesis. *Dev. Biol.* **330**, 427–439 (2009).

15. Kanai-Azuma, M. et al. Depletion of definitive gut endoderm in Sox17-null mutant mice. *Development* **129**, 2367–2379 (2002).

16. Chen, D. et al. The TFAP2C-regulated OCT4 naive enhancer is involved in human germline formation. *Cell Rep.* **25**, 3591–3602.e5 (2018).

17. Pastor, W. A. et al. TFAP2C regulates transcription in human naive pluripotency by opening enhancers. *Nat. Cell Biol.* **20**, 553–564 (2018).

18. Friedli, M. & Trono, D. The Developmental Control of Transposable Elements and the Evolution of Higher Species. *Annu. Rev. Cell Dev. Biol.* **31**, 429–451 (2015).

19. Turner, G. et al. Insertional polymorphisms of full-length endogenous retroviruses in humans. *Curr. Biol.* **11**, 1531–1535 (2001).

20. Reus, K. et al. HERV-K(OLD): ancestor sequences of the human endogenous retrovirus family HERV-K(HML-2). *J. Virol.* **75**, 8917–8926 (2001).

21. Barbulescu, M. et al. Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. *Curr. Biol.* **9**, 861–S1 (1999).

22. Jha, A. R. et al. Human endogenous retrovirus K106 (HERV-K106) was infectious after the emergence of anatomically modern humans. *PLoS ONE* **6**, e20234–e20234 (2011).

23. Medstrand, P. & Mager, D. L. Human-specific integrations of the HERV-K endogenous retrovirus family. *J. Virol.* **72**, 9782–9787 (1998).

24. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).

25. Kunarso, G. et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42**, 631–634 (2010).

26. Wang, T. et al. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl Acad. Sci. USA* **104**, 18613–18618 (2007).

27. Subramanian, R. P., Wildschutte, J. H., Russo, C. & Coffin, J. M. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology* **8**, 90 (2011).

28. Fuchs, N. V. et al. Human endogenous retrovirus K (HML-2) RNA and protein expression is a marker for human embryonic and induced pluripotent stem cells. *Retrovirology* **10**, 115–115 (2013).

29. Grow, E. J. et al. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**, 221–225 (2015).

30. Pontis, J. et al. Hominoid-specific transposable elements and KZFPs facilitate human embryonic genome activation and control transcription in naive human ESCs. *Cell Stem Cell* **24**, 724–735.e5 (2019).

31. Teissandier, A., Servant, N., Barillot, E. & Bourc'his, D. Tools and best practices for retrotransposon analysis using high-throughput sequencing data. *Mobile DNA* **10**, 52 (2019).

32. Yang, W. R., Ardeljan, D., Pacyna, C. N., Payer, L. M. & Burns, K. H. SQuIRE reveals locus-specific regulation of interspersed repeat expression. *Nucleic Acids Res.* **47**, e27–e27 (2019).

33. Jin, Y., Tam, O. H., Paniagua, E. & Hammell, M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593–3599 (2015).

34. Bendall, M. L. et al. Telescope: characterization of the retrotranscriptome by accurate estimation of transposable element expression. *PLoS Comput. Biol.* **15**, e1006453 (2019).

35. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

36. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

37. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

38. Kelley, D. & Rinn, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.* **13**, R107 (2012).

39. Thompson, P. J., Macfarlan, T. S. & Lorincz, M. C. Long terminal repeats: from parasitic elements to building blocks of the transcriptional regulatory repertoire. *Mol. Cell* **62**, 766–776 (2016).

40. Tyser, R. C. V. et al. Single-cell transcriptomic characterization of a gastrulating human embryo. *Nature* **600**, 285–289 (2021).

41. Xie, W. et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).

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

43. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).

44. Gell, J. J. et al. An extended culture system that supports human primordial germ cell-like cell survival and initiation of DNA methylation erasure. *Stem Cell Rep.* **14**, 433–446 (2020).

45. Hubley, R. et al. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–D89 (2016).

46. Smela, M. P., Sybirna, A., Wong, F. C. K. & Azim Surani, M. Testing the role of sox15 in human primordial germ cell fate [version 2; peer review: 2 approved]. *Wellcome Open Res.* **4**, 122 (2019).

47. Chambers, I. et al. Nanog safeguards pluripotency and mediates germline development. *Nature* **450**, 1230–1234 (2007).

48. Ji, X. et al. 3D chromosome regulatory landscape of human pluripotent cells. *Cell Stem Cell* **18**, 262–275 (2016).

49. Collier, A. J. et al. Comprehensive cell surface protein profiling identifies specific markers of human naive and primed pluripotent states. *Cell Stem Cell* **20**, 874–890.e7 (2017).

50. Thakore, P. I. et al. Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nat. Methods* **12**, 1143–1149 (2015).

51. Guo, Y. et al. RAD: a web application to identify region associated differentially expressed genes. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btab075 (2021).

52. Hancock, G. V. et al. Divergent roles for KLF4 and TFCP2L1 in naive ground state pluripotency and human primordial germ cell development. *Stem Cell Res.* **55**, 102493 (2021).

53. Gkountela, S. et al. DNA demethylation dynamics in the human prenatal germline. *Cell* **161**, 1425–1436 (2015).

54. Tang, W. W. C. et al. A unique gene regulatory network resets the human germline epigenome for development. *Cell* **161**, 1453–1467 (2015).

55. Kim, T.-K. et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).

56. De Santa, F. et al. A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. *PLoS Biol* **8**, e1000384 (2010).

57. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).

58. Theunissen, T. W. et al. Molecular criteria for defining the naive human pluripotent state. *Cell Stem Cell* **19**, 502–515 (2016).

59. Zhang, Y. et al. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat. Genet.* **51**, 1380–1388 (2019).

60. Hedlund, M., Ng, E., Varki, A. & Varki, N. M. α2-6–linked sialic acids on N-glycans modulate carcinoma differentiation In vivo. *Cancer Res.* **68**, 388–394 (2008).

61. Wang, Y.-C. et al. Glycosyltransferase ST6GAL1 contributes to the regulation of pluripotency in human pluripotent stem cells. *Sci. Rep.* **5**, 13317 (2015).

62. Liu, X. et al. Reprogramming roadmap reveals route to human induced trophoblast stem cells. *Nature* **586**, 101–107 (2020).

63. Sankararaman, S., Mallick, S., Patterson, N. & Reich, D. The combined landscape of denisovan and neanderthal ancestry in present-day humans. *Curr. Biol.* **26**, 1241–1247 (2016).

64. Wolf, A. B. & Akey, J. M. Outstanding questions in the study of archaic hominin admixture. *PLoS Genet.* **14**, e1007349 (2018).

65. Tao, Y. et al. TRIM28-regulated transposon repression is required for human germline competency and not primed or naive human pluripotency. *Stem Cell Rep.* **10**, 243–256 (2018).

66. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2019).

67. Howe, K. L. et al. Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).

68. Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinform.* **10**, 232 (2009).

69. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

70. Andrews, S. A quality control tool for high throughput sequence data. BibSonomy (2010).

71. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

72. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. DeepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).

73. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

74. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017).

75. Stuart, T. et al. Comprehensive Integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).

76. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

77. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

78. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).

79. Catoni, M., Tsang, J. M., Greco, A. P. & Zabet, N. R. DMRcaller: a versatile R/Bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts. *Nucleic Acids Res.* **46**, e114 (2018).

Acknowledgements

Author contributions

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

Competing interests

The authors declare no competing interests.

Additional information

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

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

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

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

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

46

Chapter 4

TRIM28 Safeguards Primordial Germ Cell Differentiation by Suppressing an Aberrant 2C-like Transcriptome

TRIM28 safeguards Primordial Germ Cell Differentiation by Preventing Entry into a 2C State

Authors: Jonathan A. DiRusso[1-4], Lingyu Zhan[5-7], Yu Tao[1,3], Allison Lynn Wang[1], Xinyu Xiang[8], Alexander C. Robbins[1], Azra J. Cruz[1], Wanlu Liu[8-10], Amander T. Clark[1-4,11,*]

[1]- Dept. of Molecular, Cell and Developmental Biology, University of California Los Angeles; Los Angeles, CA 90095, USA

[2]- Molecular Biology Institute, University of California Los Angeles; Los Angeles CA 90095, USA

[3]-Eli and Edith Broad Center for Regenerative Medicine and Stem Cell Research, University of California Los Angeles; Los Angeles, CA 90095, USA

[4] – Center for Reproductive Science, Health and Education, University of California Los Angeles; Los Angeles, CA 90095, UCLA

[5] – University of California Los Angeles, Department of Psychiatry, University of California Los Angeles; Los Angeles CA 90095, USA

[6] – Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles; Los Angeles CA 90095, USA

[7]- The Collaboratory, Institute for Quantitative and Computational Biosciences, University of California Los Angeles, Los Angeles CA 90095

[8]- Zhejiang University-University of Edinburgh Institute (ZJU-UoE Institute), Zhejiang University School of Medicine, International Campus, Zhejiang University; Haining, 314400, China

[9]- Department of Orthopedic of the Second Affiliated Hospital of Zhejiang University, School of Medicine, Zhejiang University; Hangzhou, 310029, China

[10] - Dr. Li Dak Sum & Yip Yio Chin Center for Stem Cell and Regenerative Medicine, Zhejiang University; Hangzhou, Zhejiang 310058, China

[11]- Jonsson Comprehensive Cancer Center, University of California; Los Angeles, Los Angeles, CA 90095, USA

* - Lead author, clarka@ucla.edu

Abstract:

Mammalian Primordial Germ Cells (PGCs) are embryonic precursors to the adult germline and must facilitate high-fidelity transfer for genomic material from one generation to the next. Transposable elements (TEs) represent an ongoing threat to genomic fidelity and are therefore tightly controlled during embryonic germline development. Here we find that some TEs change in accessibility during normal PGC differentiation, while others are constitutively repressed by TRIM28, a master TE regulator. We find TRIM28 itself is regulated in a sex-specific manner and represses sex-specific TEs. In both testicular and ovarian PGCs TRIM28 protects against aberrant entry into a 2C-like transcriptome, which dysregulates PGC differentiation and results in a proliferative defect and incomplete activation of DAZL. This perturbs testicular and ovarian PGCs differently, as testicular PGCs fail to differentiate in embryonic life while ovarian PGCs inefficiently enter meiosis but lack healthy gametes by the onset of sexual maturity.

Introduction:

Sexual reproduction relies on the high-fidelity transfer of genetic information from one generation to the next, a task which is carried out by the germline. In sexually reproducing eukaryotes, genomic fidelity in the germline is antagonized by Transposable Elements (TEs), selfish genomic elements capable of transposition and expansion within the genome[1-3]. TEs are relatively ubiquitous within eukaryotic genomes, with only a few identified species lacking them altogether[4,5]. TEs are broadly classified as either Type 1, retrotransposons which move via an RNA intermediate, or Type 2, DNA transposons which move via a cut and paste mechanism[6,7]. Uncontrolled movement of ether Type 1 or Type 2 TEs represents a potent threat to genomic integrity and organismal fitness, placing TE control in the germline under heavy selective

pressure. Conversely, many TEs contribute positively to organismal fitness – a prominent example of this is in the mammalian placenta, which is reliant on syncytin genes derived from an endogenous retroviral *env* gene as well as a variety of retrotransposon-derived imprinted genes, which also have an essential role in placental formation[8,9]. TEs also act as species-specific enhancers in the placenta, implicating them in multiple aspects of the evolution of placental mammals[10-12]. In humans, a *Hominidae*-specific retrotransposon, LTR5Hs, acts as an enhancer during germline specification[13]. As a result of this dichotomy, the germline and early embryonic cells are endowed with a rich network of highly conserved mechanisms to control the deleterious effects of TEs while enabling their beneficial functions.

The most thoroughly characterized TE mobility-control mechanisms in the germline of sexually reproducing organisms are those that utilize small-RNA mediated silencing. In the *Drosophila* ovary, germline cells utilize piRNA targeting to silence TEs via an RNA-Induced Transcriptional Silencing (RITS)-like mechanism, leading both to loci- and transcript-level silencing[14-16]. A similar system is employed in pro-spermatogonial cells of the embryonic mouse testis, where piRNA-mediated silencing also drives targeted reacquisition of DNA methylation[17-22]. TE control without piRNAs have been described as well. For example, adult mouse oocytes utilize siRNA-mediated TE degradation to repress evolutionarily young retrotransposons[23]. While piRNA- and siRNA-mediated forms of repression act on the adult germline, how TE expression is regulated prior to sex determination in PGCs remains poorly characterized.

In sexually reproducing metazoans, germline development begins with specification of Primordial Germ Cells (PGCs), embryonic precursors to the adult germline. In the mouse, PGCs are specified from the post-implantation posterior epiblast and acquire expression of early PGC markers *Tfap2c*, *Prdm1* (*Blimp1*) and *Prdm14* and re-acquire core pluripotency factors including, *Nanog*, *Sox2* and *Oct4*[24-29]. These early PGCs migrate through dorsal mesentery towards the developing genital ridges[30,31]. During this migration, PGCs undergo genome-wide DNA demethylation and global re-arrangements of repressive epigenetic marks, including H3K9me3,

50

H3K9me2, H3K27me3 and H2A/H4R3me2s[32-36]. Following colonization of the developing gonad, PGCs upregulate DAZL, which is necessary to fully suppress the pre-sex determination PGC program and promote proper differentiation in both sexes[37,38]. Following expression of DAZL, PGCs are developmentally fate-restricted to the germline[39]. Sex-specific differentiation of PGCs is promoted by cues from the gonadal niche. Stimulation by retinoic acid in females promotes PGC entry into meiosis, resulting in meiotic germ cells.[40] In the testis, Sertoli cells within the developing testicular cords protect PGCs from retinoic acid while NOTCH signaling between Sertoli cells promotes entry of testicular PGCs into mitotic arrest before differentiating into pro-spermatogonia[41-43]. In both testicular and ovarian PGCs, timing of PGC differentiation relies on proper regulation of the PGC epigenome, with E12.5 representing a window in which perturbation to the epigenome can drive either precocious or failed differentiation.

In mouse pluripotent stem cells and in early mouse embryos, the Tripartite Motif Containing 28 (*Trim28* a.k.a *Kap1*)/KRAB-ZFP system is a major regulator of TE expression, particularly those of the Long term repeat (LTR) retrotransposon subclass[44-47]. When acting to repress LTRs, TRIM28 is targeted to sequence-specific sites in the genome by Krüppel-associated box domain zinc finger proteins (KRAB-ZFPs). Once bound, TRIM28 is then able to recruit the histone methyltransferase SETDB1, which catalyzes H3K9me3 at these sites to repress targeted loci[48]. In PGCs, TEs are repressed by a diversity of mechanisms, including SETDB1-mediated H3K9me3, PRC2-mediated H3K27me3 and Dnmt1-mediated DNA methylation, however the consequences of TE derepression are poorly defined.[49-52].

In the current study, we examine how control of TEs by *Trim28* safeguards the capacity of PGCs to differentiate into either meiotic germ cells or pro-spermatogonia. To do so, we used a PGC-specific conditional TRIM28 knockout (TCKO) mouse model. We show that TRIM28 represses unique TE targets in both testicular and ovarian germ cells both directly and indirectly, and that derepression of TEs is coupled with a misregulated transcriptome. Functionally, we show that TRIM28 loss results in improper differentiation, including defects in expression of DAZL and

sex-specific defects in silencing the PGC program. We also show that ovarian and testicular PGC differentiation is differently sensitive to TRIM28 loss, with testicular PGCs failing to differentiate entirely while ovarian PGCs heterogeneously enter meiosis show defects in meiotic progression and severe defects in the adult germline. Therefore, *Trim28* safeguards the germline by protecting the transcriptional program of PGCs as they undergo differentiation.

Results

TE Expression is Dynamic in PGCs

To understand how TEs are regulated as the germline matures, we first set out to characterize the changes in TE accessibility in the mouse as PGCs advance from migratory PGCs to differentiated germ cells. To do so, we used ATAC-seq from PGCs collected via FACS using an Oct4-eGFP at E10.5, prior to PGC entry into the developing genital ridge, and E14.5, at which point the chromatin accessibility landscape reflects differentiated germ cells[53]. From both XY and XX samples, we identified 73,136 TEs across the ERV, LINE and SINE superfamilies which are accessible in at least one condition. We then identified elements which were accessible at E10.5 but not E14.5 (Open to Close, OC, 12,469 in XY, 27,053 in XX), or those which were accessible at E14.5 but not E10.5 (Closed to Open, CO, 16,202 in XY, 10,368 in XX). (Fig 1A). As we found more accessible TEs in XX PGCs than XY PGC at E10.5, we wanted to rule out the possibility that X-chromosome dynamics drove this observation. To do so we examined the number of integrants detected per chromosome to the total number of ERV, LINE or SINE integrants on that chromosome (Supp. 1A). We did not find a significant increase in the ratio detected in the X chromosome compared to autosomes, therefore ruling out X-chromosome dynamics as a significant driver of this observation.

We next moved to characterize those integrants which transitioned from open to closed or closed to open to better understand how these TEs could be embedded into the cis-regulatory network of PGCs. To analyze this, we performed motif enrichment analysis on identified TEs,

across all superfamilies, from both XY and XX PGCs (Fig. 1B). Consistent with possible cis-regulatory roles for these TEs, we found motifs for AP2-gamma and AP2-alpha (TFAP2C and TFAP2A, respectively), as well as the OCT-SOX-NANOG and Zic2 motifs, to be significantly enriched in TEs which progress from open to closed. To further examine whether these TEs may have functional relevance, we examined the enrichment of H3K27ac, a marker of active enhancers and promoters. To do so, we used published data from PGC-Like Cells, an *in vitro* PGC model, which recapitulates ~E10.5 PGCs *in vivo*. We found H3K27ac peaks overlapped at 13.6% of OC integrants in XY PGCs and in 8.2% of OC integrants in XX PGCs across ERVs, LINEs and SINEs (Fig. 1C). In line with these observations, we found pluripotency-associated families RLTR13D6 and RLTR9E to be enriched, with ~10% of RLTR13D6 integrants transition from open to closed in XY and 18% in XX PGCs[54,55]. For RLTR9E, ~10% of all integrants transitioned from open to closed in both XY and XX PGCs (Fig. 1D). In both sexes the top enriched family was RLTR9D2, which has not been functionally evaluated but is enriched for H3K27ac and H3K4me1 in mESCs[55]. Thus, it is possible that proper silencing of these families mark progression from PGC identity to differentiated germ cell during which time the pluripotency program is extinguished.

Conversely, TEs which transition from closed at E10.5 to open at E14.5 were enriched for DMRT1, DMRT6 and CTCF motifs in both testicular and ovarian germ cells (Fig. 1B). DMRT1 has been implicated in germline differentiation and DMRT6 in maintenance of the pro-spermatogonial transcriptional program[56]. As expected, these sites had little enrichment for H3K27ac in d6 PGCLCs, which would agree with these TEs being associated with differentiated germ cells (Fig. 1C). Among the top enriched TE subfamilies are RLTR10C and RLTR10B2. Of interest is RLTR10B2, which has been functionally validated as an enhancer necessary for the mitosis-to-meiosis transition in spermatocytes[57]. We also found enrichment of RLTR10C, which is upregulated in spermatocytes as well, but any cis-regulatory function has not been mechanistically assessed (Fig 1D). Interestingly, when we analyzed these integrants for H3K27ac

enrichment using previously published ChIP-seq data from testicular and ovarian PGCs at E13.5, we found that only 4.3% of H3K27ac peaks overlapped with TEs which transitioned from closed to open in testicular PGCs and 1.9% in ovarian germ cells (Supp. 1D). Thus, while the TEs which become accessible are associated with later-stage function by motif enrichment, they do not appear to acquire enhancer-like properties at E13.5 as marked by acquisition of H3K27ac.

Next, we asked whether large changes in TE accessibility during PGC development was a phenomenon specific to the germline or if a similar change is found in the gonadal soma, which also undergoes rapid development and differentiation during this window. To do so, we used published ATAC-seq data of gonadal soma from E10.5 and E13.5[58]. Like PGCs, we found that testicular and ovarian soma had unique TE dynamics (Supp. 1E). A motif analysis found that TEs which are open at E10.5 and close by E13.5 are enriched for Hoxc9 and Hoxa9, while those that become accessible by E13.5 are enriched for Nr5a2, Dmrt1/6, and COUPTFII (Nr2f2) motifs, which aligns with the developmental progression of gonadal soma (Supp. 1F). Thus, large changes in TE repertoire may be a conserved property of gonadal development.

Given that ERVs are known to harbor transcription factor motifs, we hypothesized that a driver of some these changes in the chromatin accessibility of TEs may be TRIM28. TRIM28 is an epigenetic scaffolding protein which is recruited to loci by KRAB-ZFPs, and in turn coordinates with NuRD and SETDB1 to catalyze repressive heterochromatin at targeted loci. Thus, we hypothesized that our observed shift in TE accessibility as PGCs progress from E10.5 to E14.5 should also be marked with a concomitant change in KRAB-ZFP repertoire. To test this, we performed RNA-seq on both testicular and ovarian PGCs at E11.5, E12.5 and E13.5. We compared differential expression of KRAB-ZFPs between E11.5 and E13.5, identifying 40 DE KRAB-ZFPs in XY PGCs between E11.5 and E13.5 and 49 DE KRAB-ZFPs in XX PGCs (Fig. 1E, 1F). Of those, 19 were shared by both XY and XX PGCs (Supp. 1G). Of those only in XY PGCs, Zfp985 and Zfp59 have both been identified in later-stage male germline development, including Pachytene spermatocytes and round spermatid, respectfully. In females, Prdm9 is a known

regulator of double-strand breaks during meiosis, while Zfp987 has been characterized as a repressor of LTR family elements[59,60]. Finally, we found KRAB-ZFPs which targeted LINEs (Gm14406, Gm14412), and ERVK family (Gm15446) elements to be differentially expressed in both sexes as PGCs differentiate[60]. Therefore, the transition between early and differentiated PGCs is marked by large-scale changes to TE accessibility and, contemporaneously, changes to the repertoire of KRAB-ZFPs.

<u>TRIM28 Regulation of TEs is Sex-Specific</u>

We began by characterizing the expression of TRIM28 in PGCs as they differentiate. To do so, we assessed the abundance of TRIM28 at E12.5, E13.5 and E16.5. We noticed that the abundance of TRIM28 decreased in ovarian germ cells between E12.5 and E16.5 (Fig. 2A). To better quantify this, we compared the abundance of TRIM28 in germ cells to that of the surrounding gonadal soma. In testicular germ cells, we found the relative abundance of TRIM28 to be consistent at each time point. Ovarian germ cells had significantly lower relative abundance of TRIM28 at E12.5 compared to testicular germ cells. Moreover, the relative abundance of TRIM28 in ovarian germ cells decreased significantly between E12.5 and E16.5. (Fig. 2B). Thus, TRIM28 is regulated in a sex-specific manner in PGCs as they differentiate. Based on these results, we wondered whether the role of TRIM28 may be different in ovarian and testicular PGCs as PGCs undergo sex-specific differentiation.

To better understand which TEs TRIM28 was regulating and if there were sex-specific differences in TRIM28 activity, we performed ChIP-seq of TRIM28 at E12.5, as PGCs are undergoing determination as they differentiate. Consistent with our findings showing lower TRIM28 abundance in ovarian PGCs, ChIP-seq identified fewer TRIM28 binding sites in ovarian PGCs (1,406 peaks) compared to testicular PGC (5,689 peaks). Interestingly, we found little enrichment of TRIM28 or H3K9me3 over the dynamic TEs which progress from open at E10.5 to closed at E14.5 (Supp. 2A), suggesting that TRIM28 does not act directly at the dynamic loci in the germline at this stage and instead is likely acting at the sites which are constitutively closed.

To identify TEs that are regulated by TRIM28, we conditionally mutated *Trim28* in PGCs using *Blimp1-Cre*. We refer to this as a TRIM28 Knockout (TCKO) as TRIM28 is absent in the PGCs (Supp Fig. 2B,C). In addition, we bred an *Oct4-eGFP* reporter allele into the strain to enable FACS isolation of PGCs at E10.5 through P1, covering the window during which the repertoire of accessible TEs in the germline changes rapidly between E10.5 and E14.5. (Fig 1A)

As TRIM28 is a well-known regulator of ERVs, we focused our analysis only on ERV elements. In agreement with our initial analysis, we identified more accessible ERVs in XX PGCs than their XY counterparts at E10.5 (Fig. 2C). We also found both XY and XX PGCs to have ERVs which become ectopically accessible as early as E10.5. These ectopically accessible ERVs were not bound by TRIM28 and showed no H3K9me3 enrichment (Fig. 2D). Importantly, we found them to be closed successfully in the absence of TRIM28 and are therefore indirectly regulated by TRIM28. We found this to be true of ERVs which are accessible in control conditions as well, in agreement with our observation that changes to TE accessibility in PGCs are independent of TRIM28 function prior to E12.5.

In contrast to E10.5 PGCs, we discovered 3,638 ERV loci which were ectopically accessible in both testicular and ovarian PGCs (Fig. 2C). Both ectopically accessible ERVs identified only in testicular germ cells and those shared with both testicular and ovarian germ cells were enriched for TRIM28 at E12.5 and H3K9me3 at E10.5 and E13.5, indicating that this subset of ERVs are directly regulated by TRIM28 and are constitutively silenced during PGC differentiation (Fig. 2E). Critically, differentially accessible ERVs that are unique to ovarian germ cells at E14.5 were not bound by TRIM28 in control PGCs, and were not enriched in H3K9me3 (Fig. 2D, Supp. Fig. 2E). Taken together, these results demonstrate that TRIM28 plays a direct role in regulating ERVs in both ovarian and testicular PGCs as they gain competence to differentiate, but that some ERVs are silenced independent of TRIM28-direct activity in ovarian PGCs. This sexual dimorphism in the control of LTRs aligns with a reduced TRIM28 abundance

in ovarian PGCs as they differentiate, where other mechanisms of TE control likely act at these loci.

As ERVs directly repressed by TRIM28 were not derepressed until E14.5, we next asked at which point TRIM28-repressed ERVs were derepressed in our TCKO PGCs. We reasoned that that RNA from TRIM28-controlled ERVs likely becomes derepressed around E12.5 as PGCs differentiate. Comparing expression of ERVs in control and TCKO PGCs via RNA-seq we show that ERVs are rapidly derepressed between E11.5 and E12.5 (Fig. 2E). Starting at E12.5, both testicular and ovarian PCGs showed robust derepression of ERVs, with the largest proportional derepression in IAP subfamilies (Supp. 2F). These ERVs are evolutionarily young and expected to be under the most robust control by TRIM28 and KRAB-ZFPs. Between ovarian and testicular PGCs the dynamics of derepression did differ slightly, with ovarian PGCs having 35% of detected LTRs suppressed relative to controls, compared to 6% in testicular PGCs. While it is unclear how loss of TRIM28 would drive TE repression, this result does demonstrate response to loss of TRIM28 between ovarian and testicular PGCs at E12.5. In sum, our results here show that loss of TRIM28 drives TE derepression at E12.5, a critical timepoint during with DAZL is upregulated and PGCs become committed to germ cell differentiation.

TRIM28 Represses a 2C-Like Transcriptome

In mESCs, loss of TRIM28 is associated with a disrupted transcriptome, ultimately leading to loss of self-renewal[44]. Here, we evaluated whether a similar phenomenon might be occurring in TCKO PGCs. To examine this, we first performed GO-enriched Gene Set Enrichment Analysis (GSEA) at E12.5, when the major TE depression phenotype first emerges. In testicular PGCs, the top enriched terms included repression of genes related to rRNA production and biogenesis, while in ovarian PGCs DNA replication, repair and histone modification were likewise repressed terms (Fig. 3A). As disruption to ribosome biogenesis and cell-cycle progression have both been demonstrated to induce a 2-Cell (2C)-like, an *in vitro* recapitulation of the mouse embryo during ZGA, we wondered whether TCKO PGCs may be upregulated 2C-associated transcripts[61,62].

To assess this, we compared changes in gene expression in TCKO PGCs to 2C-Like Cells identified in (63). We found that TCKO PGCs had a robust correlation (XY r=0.7, XX r = 0.76) at E12.5 and E13.5 (0.67 in testicular PGCs, 0.76 in ovarian PGCs) (Fig. 3B, Supp. Fig. 3B). Of note, we found *Zscan4* and *Dux* (*Duxf3*) to be upregulated, as well as MERVL-int and its associated LTR, MT2a_Mm (Fig. 3C, Supp. 3A). During exit from the 2-cell state, TRIM28 is thought to pair with *Line1* RNAs to promote silencing of *Dux,* after which development of the embryo can progress [64,65]. To evaluate the chromatin state of *Duxf3*, we evaluated our ATAC-Seq data and identified *Duxf3* as being accessible in both control and TCKO PGCs (Fig. 3D). Conversely, *Zscan4c* gains accessibility in TCKO PGCs, in line with derepression of a 2C-like program (Fig. 3D). To confirm that *Dux* is normally accessible in E14.5 germ cells we also examined *Duxf3* accessibility in published DNAse-seq (53) and found the same result (Supp. 3D). As *Zscan4c* has been shown to promote a 2C-like transcript, we next asked if a putative Zscan4 enhancer identified in (66) as sensitive to TRIM28 loss was repressed by TRIM28 and H3K9me3[67]. Indeed, we found TRIM28 and H3K9me3 enriched at the putative enhancer, linking TRIM28 to repression of a 2C-like program in PGCs (Supp. Fig. 3C). Although it is unclear whether TCKO PGCs can become bona-fide 2CLCs, TRIM28 loss in PGCs does result in the emergence of a 2C-like transcriptome during a major fate change as PGCs acquire competency to differentiate.

A hallmark sign of the 2C- state is slowed cell cycle progression[68]. Given that PGCs enter a rapid mitosis between E11.5 and E13.5, we asked whether there is a difference in the number of TCKO PGCs during this window[69,70]. We found a significant difference in PGC number staring at E12.5 and this was sustained to at least E16.5 (Fig. 3F). To test whether TCKO PGCs at were in cycle, we performed immunofluorescence for Ki67, a marker of proliferating cells. At E13.5 both testicular and ovarian TCKO PGCs had a significantly higher abundances of Ki67 compared to controls (Fig. 3G), in line with normal mitotic arrest in testicular germ cells and meiotic entry in ovarian germ cells. To ensure that PGCs were not undergoing apoptosis or acquiring double-

stranded DNA breaks (DSBs), we examined the abundance of cPARP, a marker of apoptosis, and yH2Ax, a marker of DSBs. We found no evidence of wide-spread DSBs or evidence of apoptosis (Supp. 3E-G).

Taken together, our results show that PGCs are dependent on TRIM28 to suppress a 2C-like transcriptome at E12.5, a timepoint at which PGCs pass through a critical fate determination step – the acquisition of DAZL and the capacity to differentiate. Additionally, perturbation to the rapid PGC expansion which occurs between E11.5 and E13.5 threatens the developmental competency of these PGCs, as the mitotic programs of PGCs is intimately connected to their differentiation programs.

TRIM28 Is Required for PGC Competency

Given that TE de-repression, failed PGC expansion, and emergence of a 2C-transcriptome occurs as TCKO PGCs are about to differentiate, we next evaluated the germline competency marker *Dazl*. At E12.5 nearly all control PGCs are DAZL+ whereas only 10.71% of ovarian and 51.83% of testicular TCKO PGCs are positive for DAZL (Fig. 4A,B). At E13.5, DAZL is still negative in a large fraction of PGCs, with only 69.2% of testicular TCKO germ cells expressing DAZL and 20.74% in ovarian TCKO germ cells (Fig. 4A,B).

To evaluate whether defects in DAZL expression were the result of transcriptional or post-transcriptional mis-regulation, we examined the RNA-seq and discovered significantly reduced levels of *Dazl* RNA at E12.5 (Fig 4C). An alternate gonadal PGC gene *Ddx4* (VASA) was not differently expressed. At E13.5, *Dazl* was significantly reduced in ovarian but not testicular TCKO germ cells (Fig. 4C). As our transcriptomic results closely matched the dynamics of DAZL observed via IF, with testicular germ cell having a less severe defect in DAZL expression, we reasoned that defects in DAZL expression are likely due to differences in transcription, rather than by post-transcriptional or translational control.

We next asked whether this could be direct effect of TRIM28. Analysis of TRIM28 ChIP-Seq revealed that the *Dazl* gene body is enriched in TRIM28 at E12.5 (Fig. 4D). In contrast, other

expressed germline genes, *Ddx4* (Vasa) and *Prdm1*, were not enriched in TRIM28. (Fig. 4D, Supp. 4A). Therefore, it is possible TRIM28 has some direct action on the DAZL locus itself.

Given that a null mutation in *Dazl* leads to persistence of pluripotent gene expression in the germline, we examined the abundance of pluripotency proteins NANOG, SOX2 and TFAP2C (AP2-gamma) using immunofluorescence[39]. SOX2 and NANOG were appropriately downregulated between E12.5 and E13.5 (Supp. 4B-E) in line with an orderly exit from pluripotency (Supp. 4D-E), however at E14.5 we found some ovarian TCKO germ cells remain NANOG or SOX2 positive. In testicular germ cells, we found control testicular germ cells to be partially positive for NANOG and SOX2 at E14.5 and therefore examined E16.5 pro-spermatogonia for NANOG or SOX2 and found it to be absent both in controls and TCKO germ cells (Supp. Fig. 4G). In contrast, TFAP2C protein but not RNA exhibited higher abundance in TCKO testicular PGCs at E13.5 (Fig. 4J,K). No difference was observed at E12.5 (Supp. 4H-J). Therefore, a deletion of TRIM28 is associated with inefficient silencing of the pluripotency program in testicular and ovarian PGCs. Together, these results demonstrate that TRIM28 is associated with proper DAZL induction and pluripotent gene repression, a critical fate-restriction event as PGCs differentiate. Given these results, we next asked whether TRIM28-knockout PGCs were able to differentiate.

TRIM28 has Sex-specific Effects on Germ Cell Differentiation

We next evaluated germ cell differentiation at E16.5. To evaluate testicular germ cell differentiation, we analyzed two pro-spermatogonia genes, MILI (*Piwil2*) and MIWI2 (*Piwil4*). For ovarian germ cell differentiation, we evaluated γH2Ax (H2AXpS139) and synaptonemal complex protein 3 (SYCP3).

Consistent with abnormalities in germ cell differentiation, TCKO germ cells at E16.5, failed to express MILI and MIWI2, whereas nearly all control pro-spermatogonia expressed both proteins with the expected subcellular localization (Fig. 5A)[71]. In TCKO pro-spermatogonia, VASA also failed to show the distinct punctate appearance indicative of co-localization with MILI, as

60

observed in controls. RNA-seq data from E13.5 testicular PGCs shows other members of the piRNA pathway, including *Tdrd1* and *Mael*, had significantly reduced expression well (Fig. 5E). Additionally, a GO analysis of significantly downregulated genes in TCKO PGCs at E13.5 found GO terms associated with Developmental Processes (GO 0032502, 0009888, 0050793) to be highly enriched (Supp. Fig. 5A,B). Importantly, testicular germ-cell specific genes such as *Nanos2* were not significantly different. Thus, while testicular PGCs fail to differentiate by E16.5, they do still initiate aspects of the male germ line program. We also examined a P1 TCKO knockout and found VASA+ germ cells to be rare. While these germ cells did express MIWI2, expression was weak and lacked the distinct localization typical of MIWI2 (Supp. 5G). Given these results, we conclude that TRIM28 knockout results in testicular infertility and disordered differentiation.

In embryonic ovaries at E16.5 we found that roughly half of the TCKO ovarian germ cells expressed SCP3 at E16.5 (Fig. 5C, D). Although ovarian TCKO germ cells were positive for SCP3, synaptonemal complex formation in TCKO germ cells was not obvious at this time point, in contrast, control meiotic germ cells exhibited SCP3 loaded onto condensed chromosomes as is typical during the pachytene stage of meiotic prophase I. (Supp. Fig. 5D). RNA-seq from E13.5 ovarian PGCs showed TCKO PGCs also showed attenuation of the meiotic program, including reduced expression of Stra8 and Tex15, indicative of inefficient differentiation of ovarian PGCs to meiotic germ cells (Fig. 5E).

Finally, we asked if there was any evidence of oocyte formation in the adult TCKO ovaries at sexual maturity. While a TCKO animals failure to thrive post-birth, we were able to rear one TCKO female to 6.8 weeks of age. In comparison to a control littermate, we found the TCKO ovary to be much smaller (Fig. 5H, Supp 5F). Control ovaries had STELLA-expressing oocytes in primordial follicles as well as secondary and antral follicles. Conversely, one of the two TCKO ovaries lacked oocytes altogether but did contain normal soma and granulosa cells, marked by Nr2f2 (COUP-TFII) and Foxl2, respectively. The second TCKO ovary contained two activated

follicles with atretic appearing STELLA+ oocytes but otherwise lacked a reserve of primordial follicles. (Fig. 5G, Supp. Fig. 5F).

Discussion:

Historically it was thought that TEs in PGCs are under durable epigenetic repression. Instead, we show that the chromatin state of TEs in PGCs is highly dynamic with some TE's stably closed as expected, while others change in accessibility as PGCs differentiate. We speculate that TE regulation in the germline has important functional consequences given that TEs harbor motifs for pluripotent transcription factors (for example TFAP2C and OCT4-SOX-NANOG motifs) which predictably switch from open to closed as germ cells initiate sex-specific differentiation, a stage associated with repression of pluripotent transcription factors[37,39]. In particular, our TRIM28 knockout data in PGCs highlights the importance of TE regulation, particularly at E12.5, the stage at which PGCs progress through determination as they differentiate, defined by expression and function of DAZL. Critically, E12.5 is also a major convergence point for PGC epigenetic control. Indeed, PGC-specific deletions of *Dnmt1*, *Prmt5*, *Eed*, *Ezh2*, and *Setdb1, and Ring1b* lead to either accelerated or stalled PGC differentiation at or around E12.5[35,49-52,72,73]. In this study, we contribute new knowledge on TRIM28, a master regulator of young TE repression at the point when PGCs acquire competency. Using ChIP-seq at E12.5 we show TRIM28 is bound to ERVs that are closed and enriched in H3K9me3 at E10.5 and E13.5, in agreement with them being targeted for durable repression. Although both male and female germ cells are affected by TRIM28 loss, subsequent developmental defects differ between both sexes.

TRIM28 represses over 3,600 ERVs in the germline of both sexes, however there are nearly 4,500 additional ERVs which are significantly more accessible in the absence of TRIM28 in testicular germ cells only. This sex-specific difference in TRIM28 aligns with our observation that TRIM28 becomes less abundant in the ovarian germ line after E12.5, in line with TRIM28 repression being a bona fide element of ovarian PGC differentiation into meiotic germ cells. A PGC-specific knockout of *Setdb1*, the downstream effector of TRIM28, also results in weaker TE

de repression in ovarian germ cells, indicating that TRIM28/SETDB1 is likely compensated by other mechanisms of TE repression in the female germline, whereas TRIM28/SETDB1 remains a major regulator of ERVs following differentiation in the testicular germline[49]. One possible compensatory mechanism of ERV repression in ovarian germ cells is *Ezh2*/H3K27me3, as loss of *Ezh2* (but not *Eed*) results in derepression of young ERVs, similar some repressed by TRIM28[50,72]. In contrast, as *Ezh2* knockout in testicular germ cells does not result in ERV derepression, it is likely TRIM28 is a major regulator of young ERVs in the male germline until repression of these ERVs is reinforced by the piRNA pathway, although it is clear that both piRNA and TRIM28-mediated TE repression are required independently, therefore ruling out the ability of one of these two systems to fully compensate for the other .

Proper PGC differentiation entails expression of the competency factor DAZL. When *Dazl* is knocked out, normal repression of pluripotency genes is disrupted, and germ cells fail to differentiate correctly[37-39]. In TCKO PGCs, we discovered that a major phenotype involved the absence of DAZL from the majority of PGCs at E12.5 with modest gains in DAZL expression by E13.5, with the ovarian germline showing a more exacerbated failure to induce DAZL. Although a reduction in proliferation-coupled loss of DNA methylation at the *Dazl* locus could explain this result in part, we do find ovarian and testicular PGCs have similarly reduced PGC expansion but differing severities of *Dazl* misregulation, suggesting other regulators contribute to our observed DAZL phenotype[70,74,75]. A failure to induce *Dazl* did not seem to slow the initial decrease in the pluripotency program, however we did observe that some ovarian PGCs fail to repress SOX2 and NANOG by E14.5. A similar defect was not seen in TCKO testicular germ cells at E16.5. Using TRIM28-ChIP, our data suggests that TRIM28 may have some undefined function at the *Dazl* locus, being enriched across the *Dazl* gene body at a time when *Dazl* is highly expressed. Given this result, we speculate that TRIM28 may act on a regulator of *Dazl*. TRIM28 enrichment along the gene body seems unique to *Dazl*, as another germline gene expressed at a similar time, *Ddx4* (*Vasa*), is not bound by TRIM28. While it is unlikely TRIM28 is acting as a repressor, it is possible

TRIM28 acts on a regulator of *Dazl* through its E3-SUMO ligase activity, which has been shown to positively regulate the stability of other proteins[76-79]. Among these is FOXL2, which TRIM28 stabilizes in granulosa cells via E3-SUMO ligation, loss of which results in non-canonical sex-reversal of TRIM28 mutant granulosa cells into Sertoli cells[80]. In PGCs, we note that sex-reversal does not occur as ovarian PGCs continue to enter meiosis and testicular germ cells fail to show upregulation of genes related to meiosis. It would be informative in future studies to identify targets of TRIM28 SUMOylation in PGCs, and this may lead to the identification of new regulators of PGC differentiation.

While transcriptional misregulation and cell death are often observed following genome-wide derepression of TEs, we found neither stress or apoptotic pathways were upregulated in TCKO PGCs. Instead, the response to loss of TRIM28 was stalled proliferation and entrance into a 2C-like transcriptional state. Both *in vitro* and *in vivo* evidence has previously linked TRIM28 to regulation of the 2C state via *Dux*, a highly conserved transcription factor which activates during Zygotic Genome Activation in mammals[81]. In ESCs, TRIM28 is hypothesized to scaffold on LINE1 RNA with *Nucleolin* to direct repression of *Dux*[64,65], and TRIM28 KO induces expression of 2C-genes[82]. Interestingly, the *Dux* locus is accessible in PGCs under wild type conditions but does become significantly upregulated in TCKO testicular PGCs at E12.5 and in ovarian PGCs at E13.5. In contrast, *Zscan4c*, which is likewise a marker of the 2C-state and which drives 2C-state MERVL expression, was significantly upregulated in TCKO PGCs[67]. *Zscan4c* is a member of the collective cluster referred to as ZSCAN4 on chromosome 7. We find that PGCs likely repress a TRIM28-sensitive putative enhancer of Zscan4, as we found this putative enhancer to be enriched for TRIM28 and H3K9me3 in PGCs[66]. In TRIM28 knockout PGCs, germ cell number fails to increase from E12.5, and this is associated with the timing of major TEs de-repression and emergence of the 2C phenotype. Therefore, with an active *Dux* locus, PGCs appear poised to enter a 2C state at E12.5 without TRIM28, which in turn results in a disordered PGC transcriptome and the emergence of a transient 2C-like transcriptome.

In sum, our work demonstrates that TRIM28 is a sex-specific regulator that becomes exquisitely important for PGC biology at E12.5, safeguarding PGC determination and therefore PGC differentiation in both sexes. The loss of TRIM28 in PGCs has only minor consequences before E12.5, with normal numbers of PGCs entering the genital ridge and limited effects on TE or gene expression or TE accessibility. At E12.5, the PGC epigenetic convergence point and stage in which *Dazl* should be expressed in all PGCs, the TRIM28 phenotype reveals itself. Facilitated by the normal expression of *Dux* in PGCs at E12.5, a TRIM28 loss results in massive upregulation of TEs and derepression of a 2C-like transcriptome that should normally be silenced during germ cell differentiation of both sexes. This leads to simultaneous effects on the ability of PGCs to rapidly proliferate and express *Dazl*, thus disrupting the normal progression of PGC differentiation and therefore preventing PGCs from differentiating correctly. We use this observation to put forth the hypothesis that TRIM28 serves not only as a direct defense against expression of young TEs in PGCs, but also as a mechanism protecting PGCs from entering into a 2C-like transcriptome.

Author Contributions:

ATC conceived the project. JAD and YT maintained mouse lines. JAD and YT designed all experiments. JAD, YT, ALW, ACR, and AJC conducted experiments. YT bred and developed the TRIM28 mouse line. JAD, XY and LZ analyzed data. XY and WL provided feedback on bioinformatic code and manuscripts. JAD and ATC drafted the manuscript. We thank ALW, XY, WL for their detailed feedback.

<u>Declaration of Interests:</u>

Authors declare no competing interests.
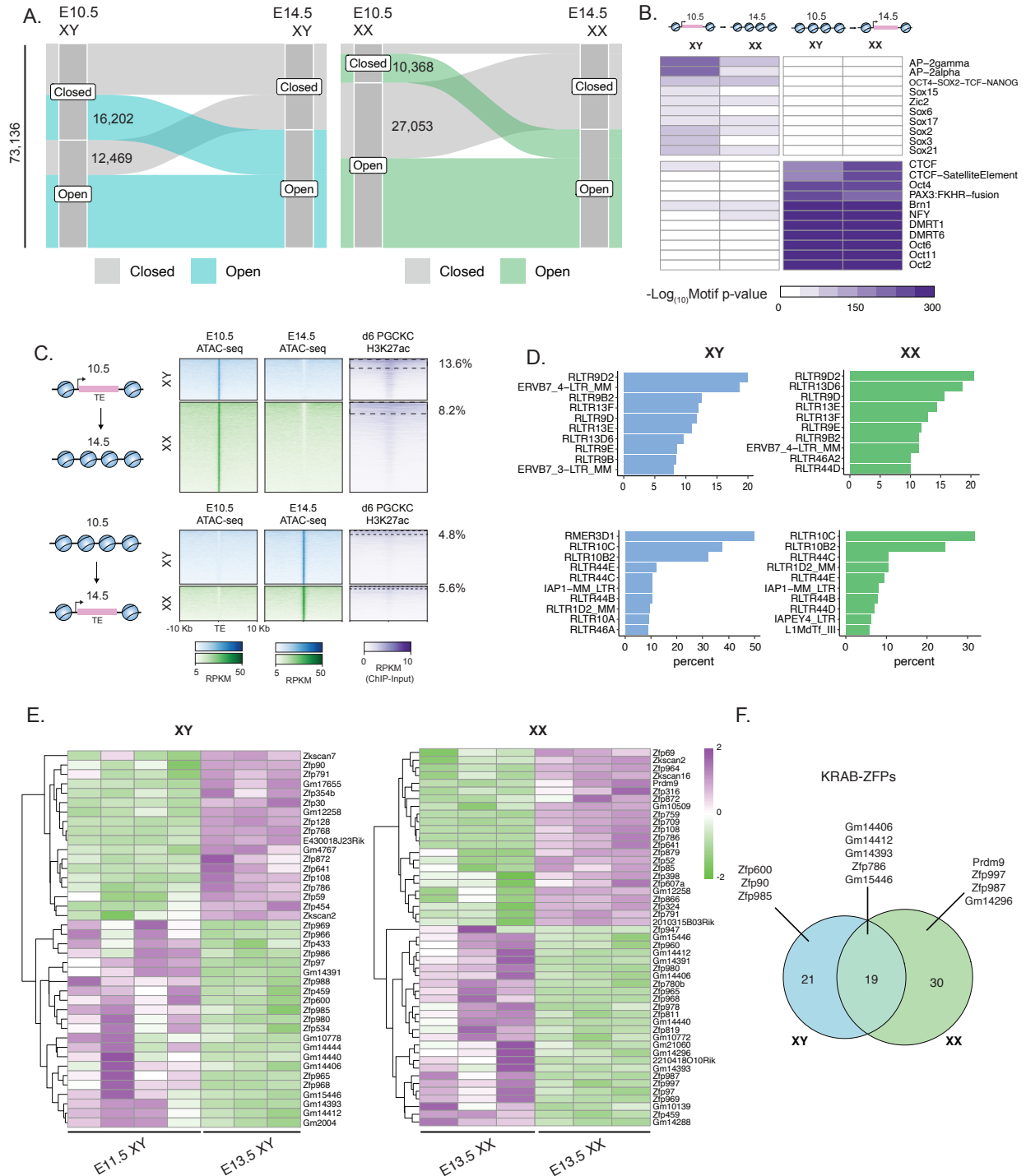
<u>Figure 1: Transposable Element Accessibility Changes During PGC Differentiation</u> A. Alluvial plots showing the chromatin accessibility state of TEs (ERV, LINE, SINE) overlapping with ATAC peaks at E10.5 and E14.5 in XY (left) and XX (right) samples. B. Motif enrichment of TEs identified as processing from open to closed (left) or closed to open (right) in either sex. P-values from HOMER. C. Chromatin accessibility of TEs which either progress from open to closed (top) or closed to open (bottom) and E10.5 and E14.5, enrichment of H3K27ac in *in vitro* PGCs. Percentages represent overlap with identified peaks. D. Percentage of TEs within families for which the largest proportion of elements are represented. Top, TEs open to close, bottom, TEs close to open. E. Heatmap showing differential gene expression between E11.5 and E13.5 testicular PGCs (left) and ovarian PGCs (right). F. Venn diagram showing sex-specific and shared KRAB-ZFPs.

Figure 2:
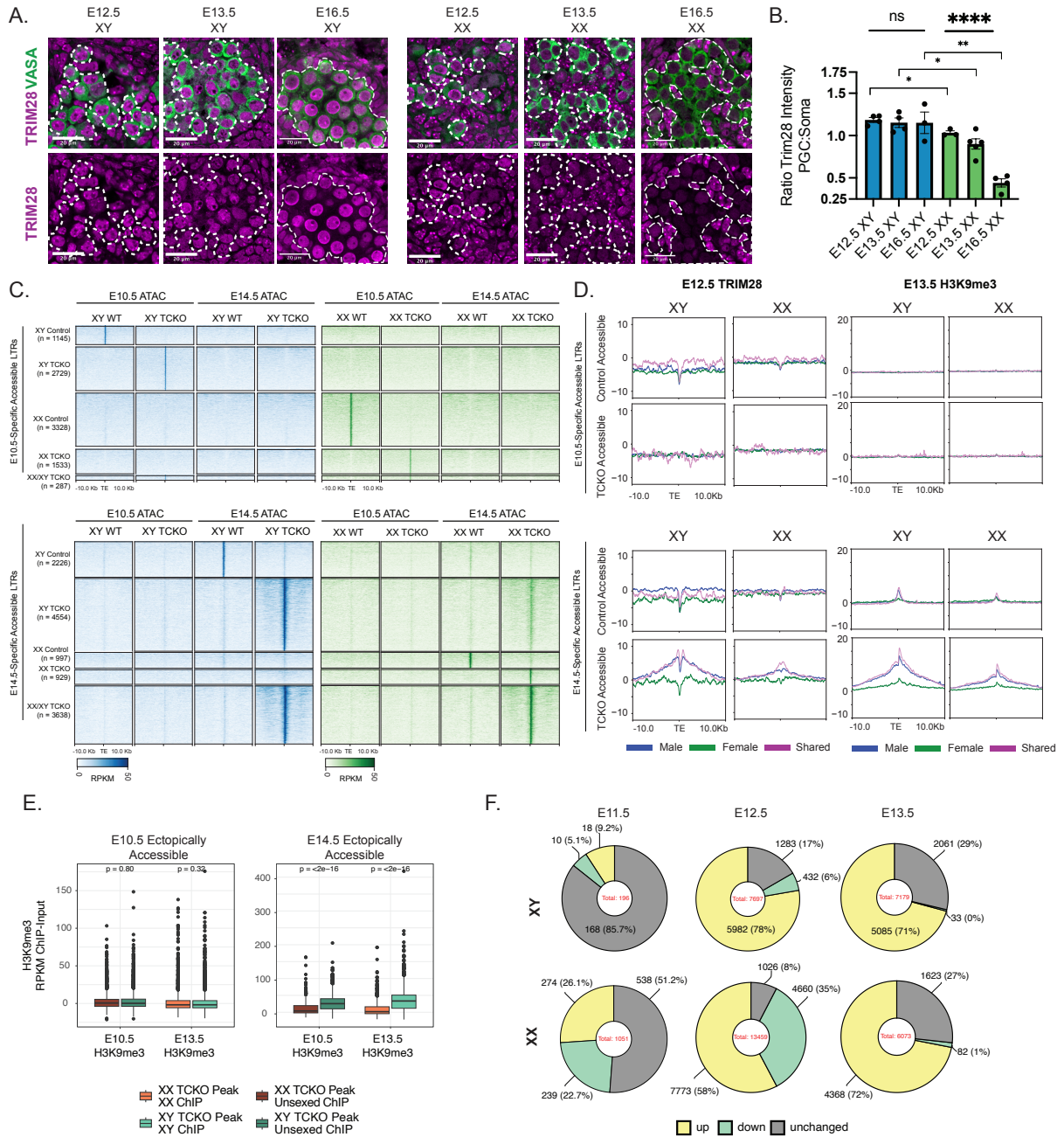
Figure 2: TRIM28 Repression of Transposable Elements is Sex-Specific A. Immunofluorescence showing TRIM28 at E12.5, E13.5 and E16.5 in XY (left) and XX (right) samples. Germ cells are marked in green (VASA) and TRIM28 in magenta. Scale 20 μm. N ≥ 3 for all conditions. B. Quantification of TRIM28 abundance in PGCs relative to gonadal soma. Each dot represents a biological replicate. Pairwise comparisons are t-tests and three-way comparisons are ANOVA tests. Error bars SEM. C. Heatmaps showing ERV accessibility in control and TRIM28 knockout PGCs at E10.5 (top) and E14.5 (bottom). D. Profile plots of ChIP-seq data of TRIM28 at E12.5 (left) or H3K9me3 at E13.5 (right) at ERVs accessible in control PGCs (Top) or TCKO PGCs (bottom). Trends for peaks identified in XY, XX and Shared (XY and XX) are shown in blue, green and purple respectively. Until are Input-subtracted normalized to read count (DeepToops). E. H3K9me3 enrichment at LTRs identified at E10.5 (left) and E14.5 (right) at E13.5. Units are input-subtracted RPKM. Significance tested using unpaired 2-sided t-test. Box-plot shows median, with hinges equal to 25th and 75th percentiles. Whiskers extend to 1.5 IQR with outliers shown. F) Proportional representation of TEs which are upregulated (FDR < 0.005, Log2FC > 2 using DESeq2), downregulated (FDR < 0.005, Log2FC < 0.25) and unchanged (Log2Fold Change 0.99 to 1.005, FDR which can be calculated by DESeq2). TEs not matching any criteria are not represented. Total number of TE integrants represented in the middle of each pie chart. Number of DETEs and percentage of TEs (from a total of TEs matching any criteria) shown in each pie section. For statistical tests, ns = p>0.05, * = p<0.05, ** = p<0.01, *** = p<0.001, ****=p<0.0001.

Figure 3:
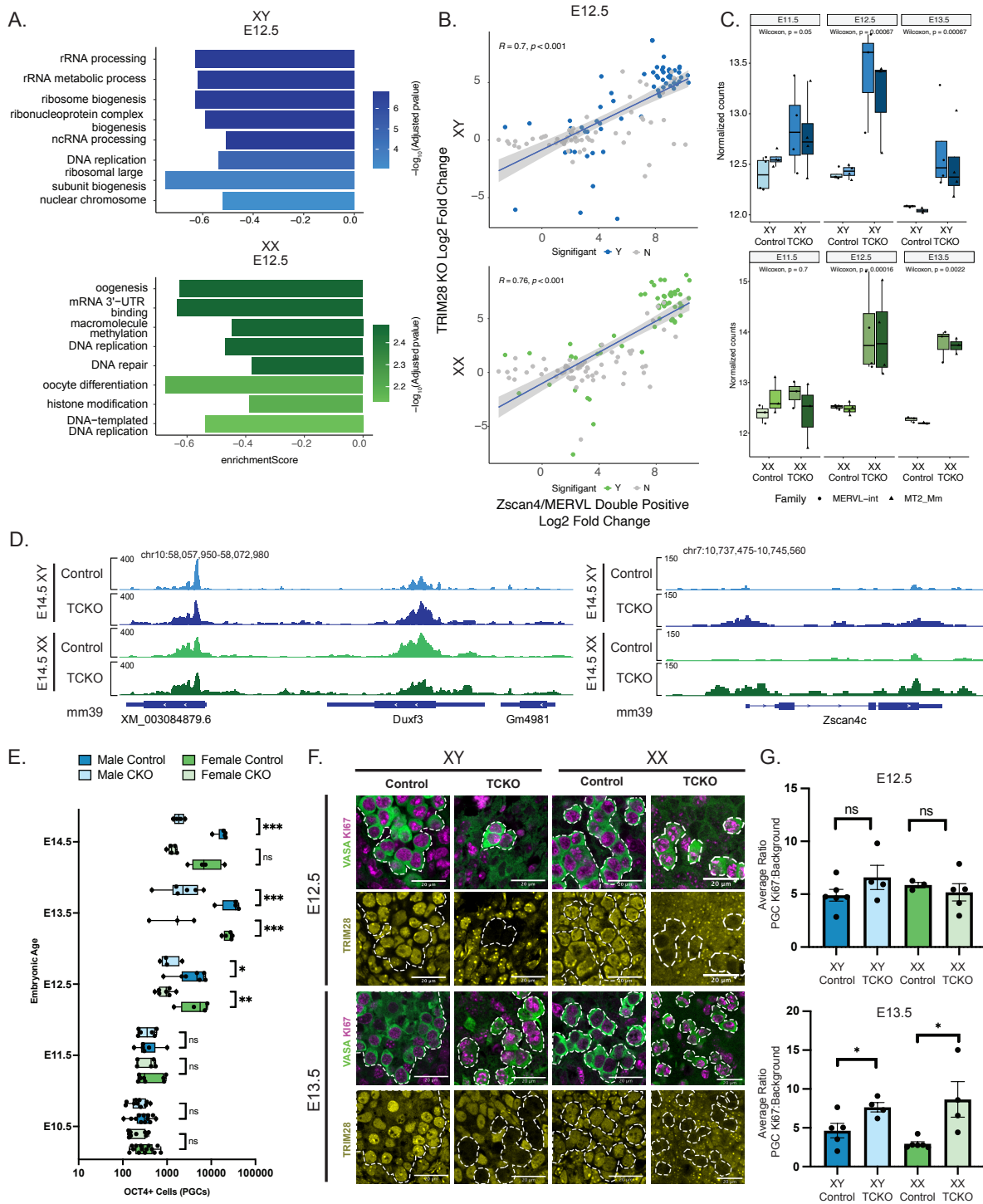
Figure 3: TRIM28 Suppresses Emergence of a 2C-Like Transcriptome A. Top enriched GO-enriched GSEA analysis results in testicular (top) and ovarian (bottom) PGCs. B. Correlation of Log2 Fold Change 2C-transcriptome genes identified in (CITE) between mESCs /2CLCs (x-axis) and Control/TCKO PGCs (y-axis) at E12.5. Colored dots are significant in TRIM28 DEG analysis ($|Log2FC| > 2$, pad < 0.05). R correlation calculated by smplot2 using pearson correlation. C. Normalized RNA-seq (rlog, DESeq2) of MERVL-int and MT2_Mm at E11.5, E12.5 and E13.5 in control and TCKO PGCs. P-value is Wilcoxon t-test. D. Browser track plots of ATAC-seq at E14.5 in control and TCKO XY (top) and XX (bottom) PGCs at the *Dux* locus (left) and *Zscan4c* (right). Tracks are RPKM normalized. E. Box plot of Oct4-eGFP+ PGCs isolated at each indicated timepoint from XY or XX embryos (E10.5 or E11.5) or embryonic testes and ovaries (E12.5, E13.5 and E14.5). Y-axis is log(10) scaled. Significance testing by t-test. F. Representative images of Ki67 and TRIM28 staining in testicular (XY) and ovarian (XX) PGCs at E12.5 and E13.5. VASA marks germ cell population. VASA, Ki67 and TRIM28 brightness enhanced for clarity. G. Quantification of background-corrected Ki67 intensity in E12.5 (top) and E13.5 (bottom) testicular and ovarian PGCs, with the average ratio of each biological replicate plotted. Significance testing by Welch's t-test. Error bars in SEM. At least n=3 biological samples. For statistical tests, ns = p>0.05, * = p<0.05, ** = p<0.01, *** = p<0.001, ****=p<0.0001.
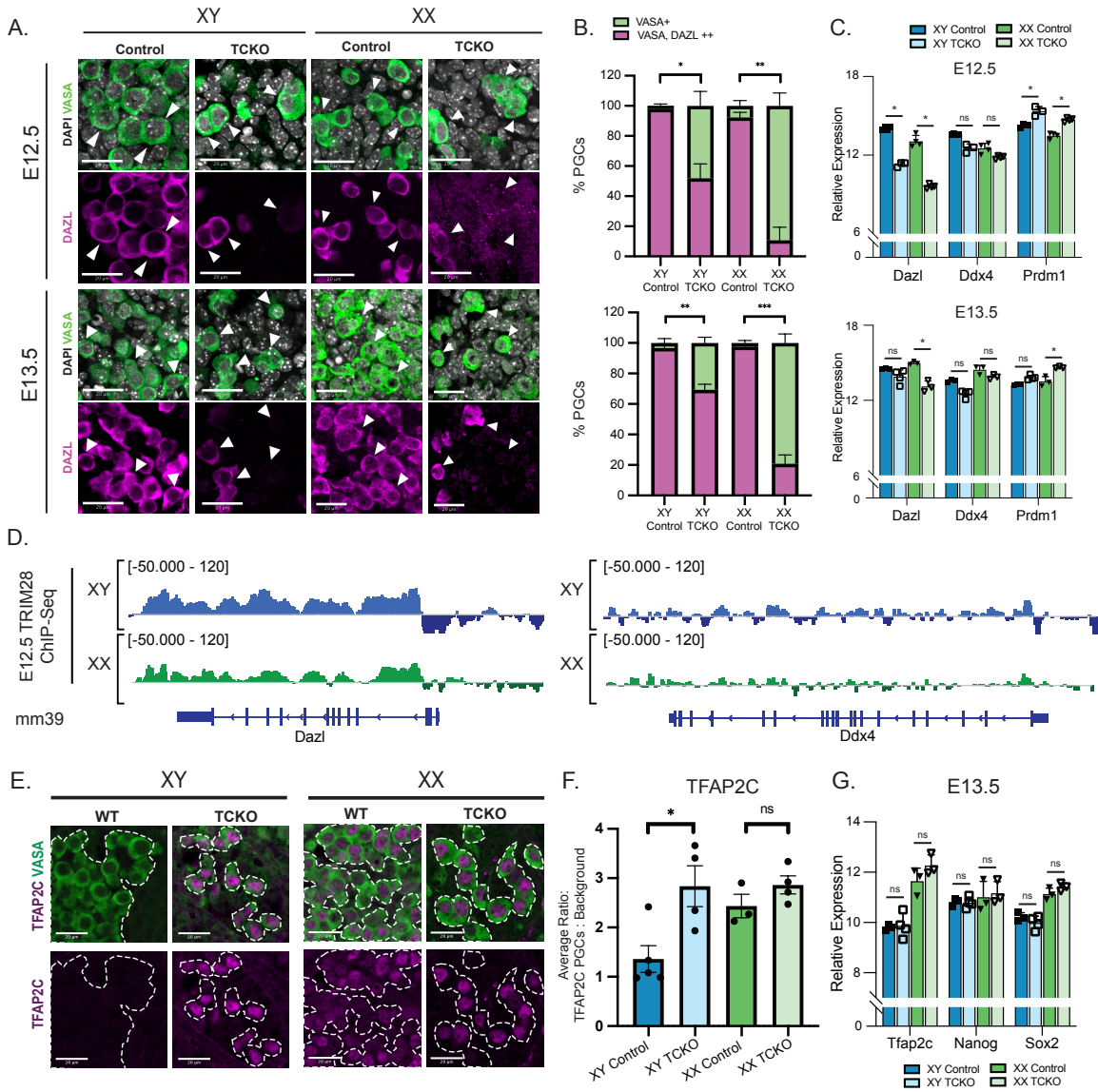
Figure 4:

Figure 4: TRIM28 is Required for Timely Acquisition of Competency A. Representative images of embryonic testes and ovaries at E12.5 and E13.5 showing PGCs (VASA, green) and DAZL (magenta). VASA, DAZL and DAPI brightness enhanced for clarity. B) Quantification of percentage of PGCs which are positive for DAZL and VASA (magenta) or only VASA (green) in testicular (XY) and ovarian (XX) PGCs at E12.5 (top) (XX Control n = 4, XX TCKO n = 3, XY Control n = 3, XY TCKO n=3) and E13.5 (bottom) (XX Control n=4, XX TCKO n=4, XY Control n=3, XY TCKO n = 3). Significance testing by Welch's t-test. Error bars in SEM. C) Normalized expression of *Dazl*, *Ddx4*, and *Prdm1* at E12.5 (top) and E13.5 (bottom). DESeq2 rlog normalization. Significance testing by DESeq2. Error bars in SD. D) Gene tracks showing Input-subtracted read-count normalized abundance of TRIM28 over the *Dazl* locus (left) and *Ddx4* locus (right) in E12.5 testicular (top) and ovarian (bottom) PGCs. E) Representative images of embryonic testes and ovaries with TFAP2C (magenta) and VASA (green). VASA and TFAP2C brightness has been enhanced for clarity. F) Quantification of TFAP2C intensity over background in PGCs, each dot represents the mean of one biological replicate. Significance by 2-sided unpaired Welch's t-test. (XY Control n = 5, XY TCKO n = 4, XX Control n = 3, XX TCKO n = 4) G) Normalized expression of *Tfap2c*, *Nanog*, *Sox2* in E13.5 PGCs. Significance by DESeq2. Error bars in SD. For t-tests, ns = p>0.05, * = p<0.05, ** = p<0.01, *** = p<0.001, ****=p<0.0001. For C and G, * = significant by DESeq2, abs(Log2FC) > 1.5, FDR < 0.05.

Figure 5:

A.

	VASA MILI	VASA MIWI2
XY WT		
XY TCKO		

B.

% MILI + ***
XY Control / XY TCKO

% MIWI2+ **
XY Control / XY TCKO

C. XY, E13.5

$-\text{Log}_{10} P$ vs Log_2 fold change
Trim28, Tdrd1, Mael, Shoc1, Ddx4, Prdm1, Foxa1, Nanos2

D.

	VASA SCP3	VASA γH2Ax
XX WT		
XX TCKO		

E.

% SCP3 + *
XX Control / XX TCKO

% γH2Ax + **
XX Control / XX TCKO

F. XX, E13.5

$-\text{Log}_{10} P$ vs Log_2 fold change
Trim28, Dazl, Tdx15, Stra8, Prdm1, Sycp1, Meioc

G.

	Stella (Dppa3)	Nr2f2 (COUP TF II)	FoxL2	Merge
Week 7 XX Control				
Week 7 XX TCKO				

H.

Area (cm^2)
XX Control / XX TCKO

Figure 5: TRIM28 is Necessary for Successful Gametogenesis A) Representative image of embryonic testes at E16.5 showing expression of MILI (left) and MIWI2 (right) in magenta. Prospermatogonia are labeled with VASA (green). Scalebar represents 20 μm. Fluorescence intensity adjusted for clarity. B) Quantification of percentage of MILI+ (left) and MIWI2+ (right) in E16.5 testes. Each point is one biological replicate. (Control n = 4, TCKO n = 3 for MILI, Control n = 5, TCKO n = 3 for MIWI2.) Significance by Welch's 2-sided t-test. Error bars show SEM. C) Representative image of embryonic ovaries at E16.5 showing expression of SCP3 (left) and yH2Ax (right) in magenta. Meiotic germ cells are labeled with VASA (green). Scalebar represents 20 μm. Fluorescence intensity adjusted for clarity. D) Quantification of percentage of SCP3+ (left) and yH2Ax+ (right) in E16.5 ovaries. Each point is one biological replicate. Control N = 3, TCKO N = 3. Significance by Welch's 2-sided t-test. Error bars show SEM. E and F) volcano plots showing derepression of genes in e) testicular PGCs at E13.5 and f) ovarian PGCs at E13.5. Select genes are labeled. P-adjusted values and Log2 Fold Change from DESeq2. G) IF images of week 7 (day 48) ovaries from 1 control female and 1 TCKO female. Ovarian soma (Nr2f2) and granulosa cells (Foxl2) and in yellow and magenta respectively. Oocyte are marked with Dppa3 (Stella, green). Each TCKO ovary is shown. Scale bar represents 100 μm. H) Measurement of ovary area from 2D-image, in cm^2. Each dot is a single ovary from one biological replicate. (n=1 both conditions). For t-tests, ns = p>0.05, * = p<0.05, ** = p<0.01, *** = p<0.001, ****=p<0.0001.

Figure 4-6 (S1):
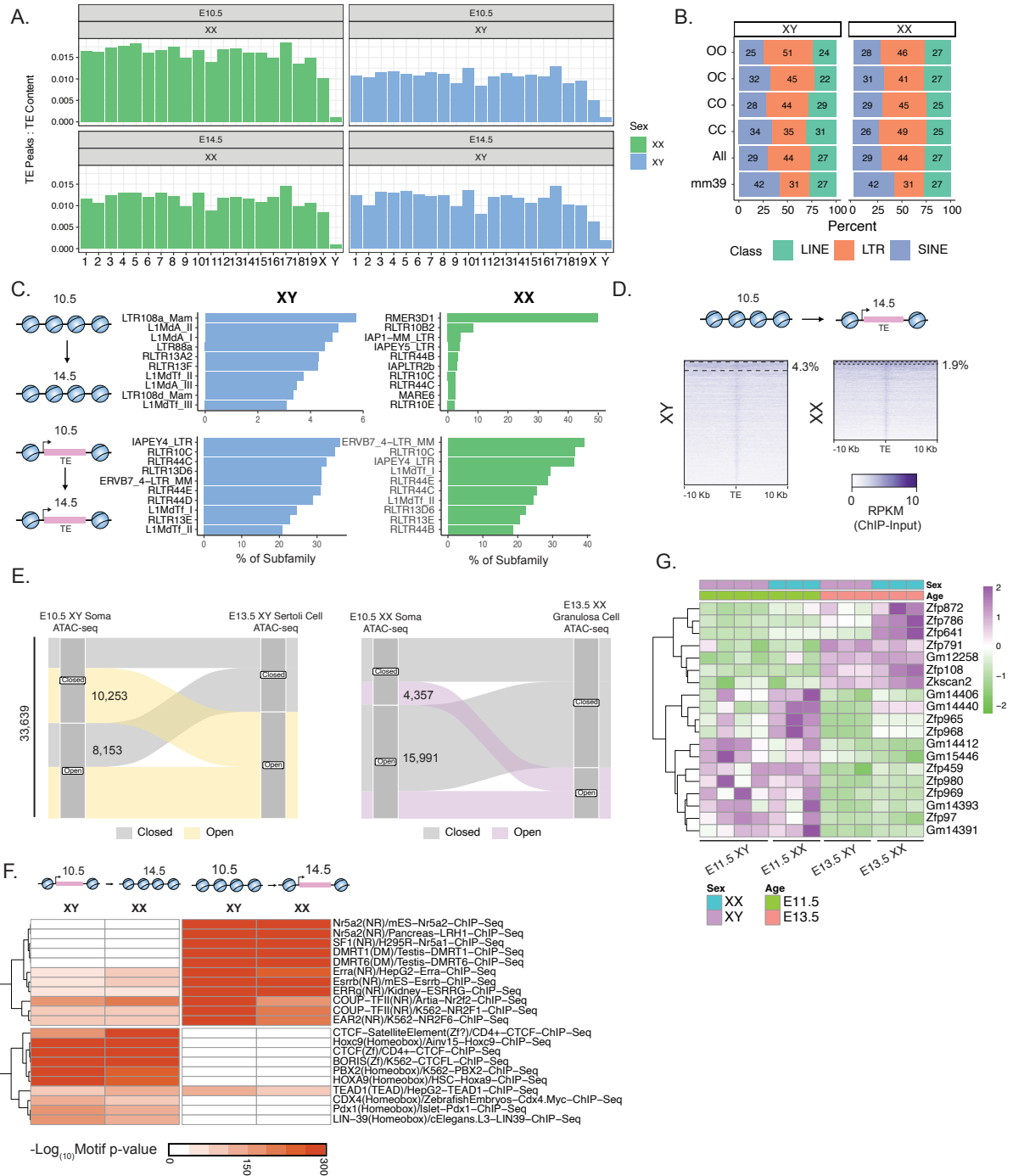
Supplemental Figure 1: Transposable Elements are Dynamin in both PGCs and Gonadal Tissue

A) Proportion of peaks called overlapping with TEs relative to the TE load (ERV, LINE and SINE) of each chromosome at E10.5 and E14.5 Color denotes sex. B) Percentage of peaks overlapping with TEs by category (Open to Closed, Closed to Open, Closed to Close and Open to Open) as well as all peaks overlapping with TEs (aggregate) and all TEs in the mm39 repeatmasker database. C. Plots showing the most represented TEs which are constitutively closed (top) or open (bottom) in XY and XX PGCs. Percentage refers to percentage of all elements in a subfamily represented. D. H3K27ac enrichment in E13.5 gonads (XY or XX) in peaks that transition from closed to open. Dotted line and percentage refer to the overlap between H3K27ac peaks and identified TEs. E) Alluvial plots showing the change in TE accessibility in E10.5 and E13.5 soma from XY and XX gonads. F) Motif analysis via HOMER showing enrichment of TE-embedded motifs in TEs which transition from open to closed (left) and close to open (right) in gonadal soma between E10.5 and E13.5. G. Heatmap showing expression of KRAB-ZFPs which are significantly different between E11.5 and E13.5 in both XY and XX germ cells.

Figure 4-7 (S2)
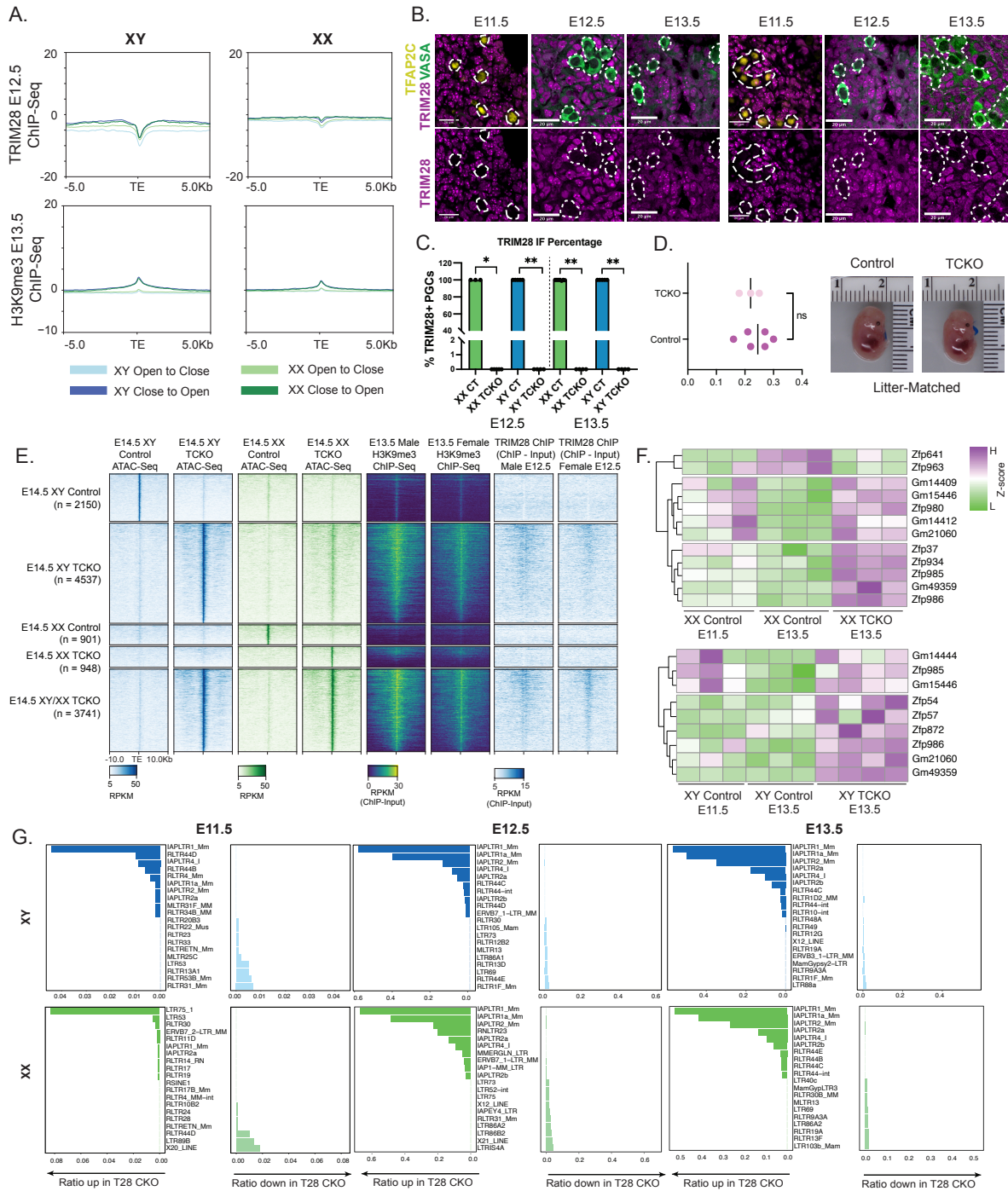
Supplemental Figure 2: Dynamic TEs in PGCs are not Orchestrated en masse by TRIM28 A)

TRIM28 ChIP-seq enrichment at E12.5 (Top) and H3K9me3 enrichment at E13.5 (bottom) TEs

which transition from open to closed or closed to open. In testicular and ovarian PGCs. Y axis is

input-subtracted normalized read count. B) Representative IF showing TRIM28 abundance at

E11.5, E12.5 and E13.5 via IF. C) Quantification of TRIM28 knockout. D) Mass (left) and

morphology (right) of E14.5 embryos from Control and TCKO-PGC harboring embryos.

Embryos shown are litter-matched. E) Heatmap showing read density over ATAC-peaks and

read density of H3K9me3 and TRIM28 ChIP-seq. ChIP units are input-subtracted RPKM. F)

DEG analysis of KRAB-ZFPs significantly misregulated in XY (top) and XX (bottom) PGCs at

E13.5. Scale is z-score of normalized (rlog, DESeq2) expression. G) Quantification showing

subfamily-level derepression of TE RNAs at E11.5, E12.5 and E13.5. Derepressed TEs are

represented on the left, repressed TEs on the right. X-axis represents ratio of elements detected

to all elements in subfamily.

Figure 4-8 (S3):
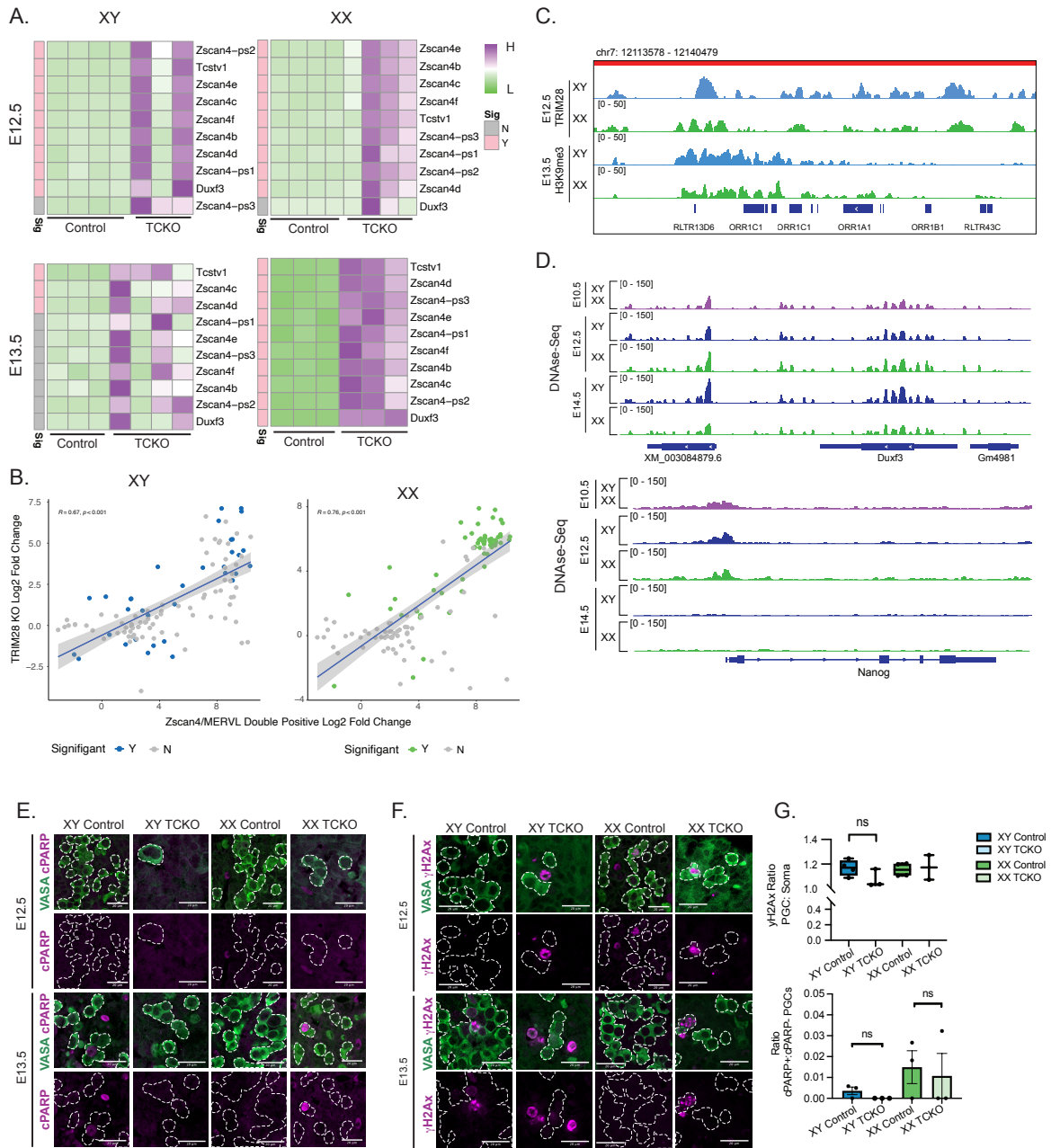
Supplemental Figure 3: mPGCs are primed to enter a 2C-like transcriptome A) Heatmaps

showing per-replicate z-scores of key marker 2C-genes in E12.5 and E13.5. Genes which are

significant in each condition are indicated. B) Correlation plots with TCKO PGCs and 2CLCs at

E13.5 as in Fig. 3. Correlation is Pearson correlation calculated using smplot2. C. Gene tracks

showing E12.5 TRIM28 ChIP (top) and E13.5 H3K9me3-ChIP (bottom) at a putative Zscan4

enhancer identified in Le, R. et al., (2021) Cell Stem Cell. Tracks are readcount-normalized

input-subtracted (deepTools). D) Gene tracks of DNAse-seq from Li, J.et al. (2018) Cell Res. at

E10.5, E12.5 and E13.5 from XY and XX PGCs at the *Dux* locus. D) Gene tracks as above for

the *Nanog* locus. E) Representative images of cPARP (magenta) in PGCs (green) at E12.5 and

E13.5. Scale bar represents 20 μm. E) Representative images of yH2Ax (magenta) in E12.5

and E13.5 PGCs. Scale bars are 20 μm. F) Quantification of yH2Ax (top) and cPARP (bottom)

ratios at E12.5. Significance testing by Welch's t-test for samples with n ≥ 3. Error bars are

SEM. Each dot represents a biological replicate. Boxplot line represents the mean, box bounds

are the first and third quartiles. Whiskers represent maximum and minimum values within 1.5x

the first (minimum) or third (maximum) quartile.

Figure 4-9 (S4):
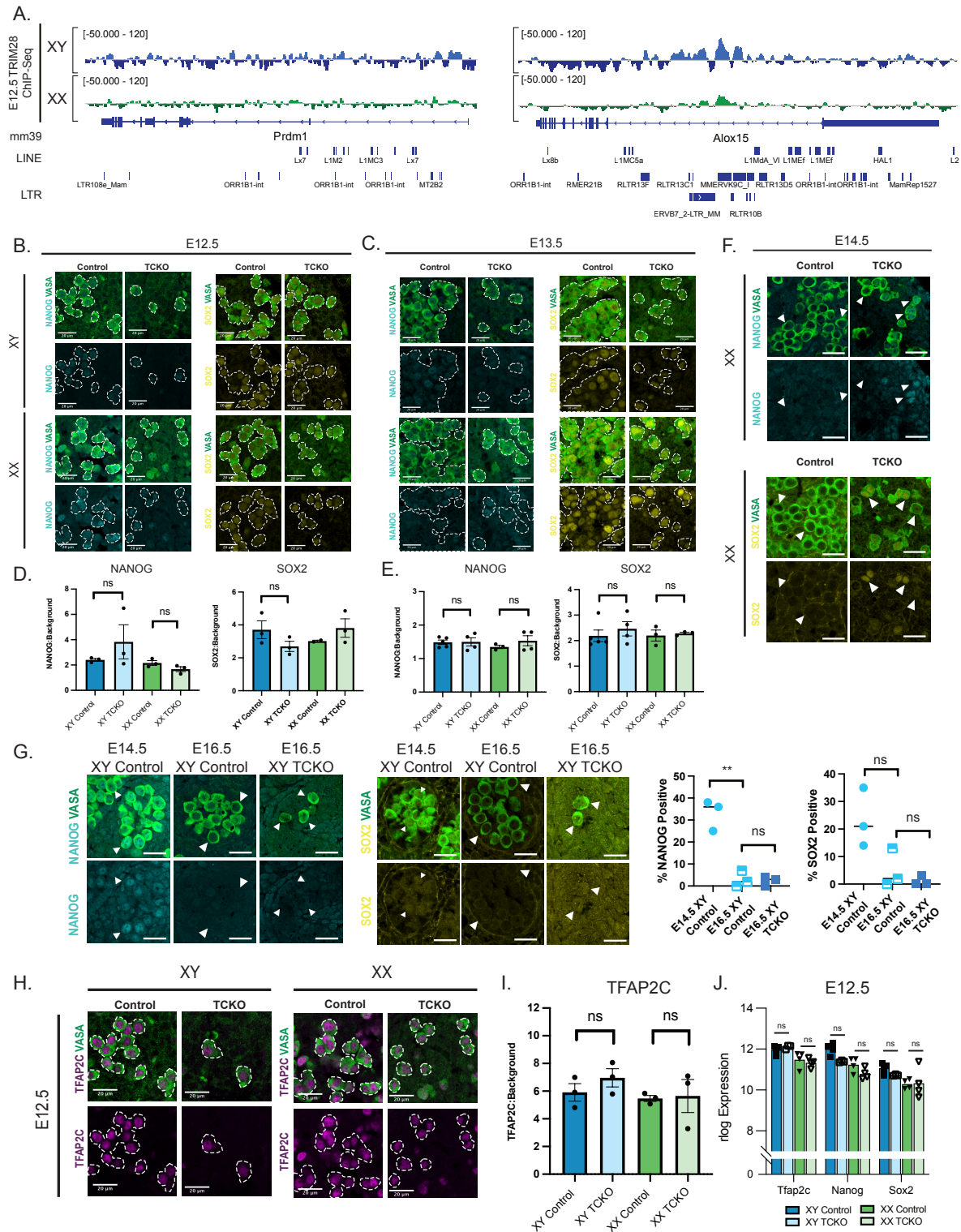
Supplemental Figure 4: Downregulation of PGC factors is altered following loss of TRIM28 A)

Gene tracks showing enrichment of TRIM28 at E12.5 (Input-subtracted, read-count normalized)

at *Prdm1* (left) and *Alox15* (right). Repeatmasker tracks for ERV, LINE and SINE elements

shown below. *Alox15* has intronic ERVs which are repressed by TRIM28. B) Representative

images showing abundance of NANOG (left, Cyan) and SOX2 (right, yellow) in PGCs at E12.5.

Scale bar represents 20 μm. C) As in B, but at E13.5. D) Quantification of NANOG and SOX2

relative abundance at E12.5. Y-axis represents ratio of NANOG or SOX2 in PGCs relative to

soma. Significance testing by 2-sided unpaired Welch's t-test. Individual points are one

biological replicate. Error bars show SEM. E) Same as in D) but for E13.5. F) Representative

images of E14.5 ovaries from control (left) and TCKO (right) gonads. Arrows indicated PGCs for

orientation. Scalebar represents 20 μm. n=2 each condition. G) (Right) Representative images

of testes at E14.5 and E16.5 in control (E14.5 and E16.5) and TCKO (E16.5) conditions. N=3 for

each condition. Scale bar represents 20 μm. (Left) Quantification of percentage of NANOG or

SOX2+ PGCs from the listed condition. N =3 replicates each. Line represents mean. Each point

is 1 biological replicate. Stats are student's t-test. H) Representative images of TFAP2C

(magenta) in testicular and ovarian PGCs at E12.5. Scale bar represents 20 μm. n =3 for all

conditions. I) Quantification of TFAP2C abundance in PGCs at E12.5. Significance testing by

unpaired 2-sided Welch's t-test. Individual points are average of all PGCs in biological replicate.

Error bars show SEM. J) Normalized expression of *Tfapc2c*, *Nanog* and *Sox2.* Statistical testing

by DESeq2. Normalized expression in rlog. For t-tests, ns = $p > 0.05$, * = $p<0.05$, ** = $p<0.01$,

*** = $p<0.001$, ****=$p<0.0001$

Figure 4-10 (S5):
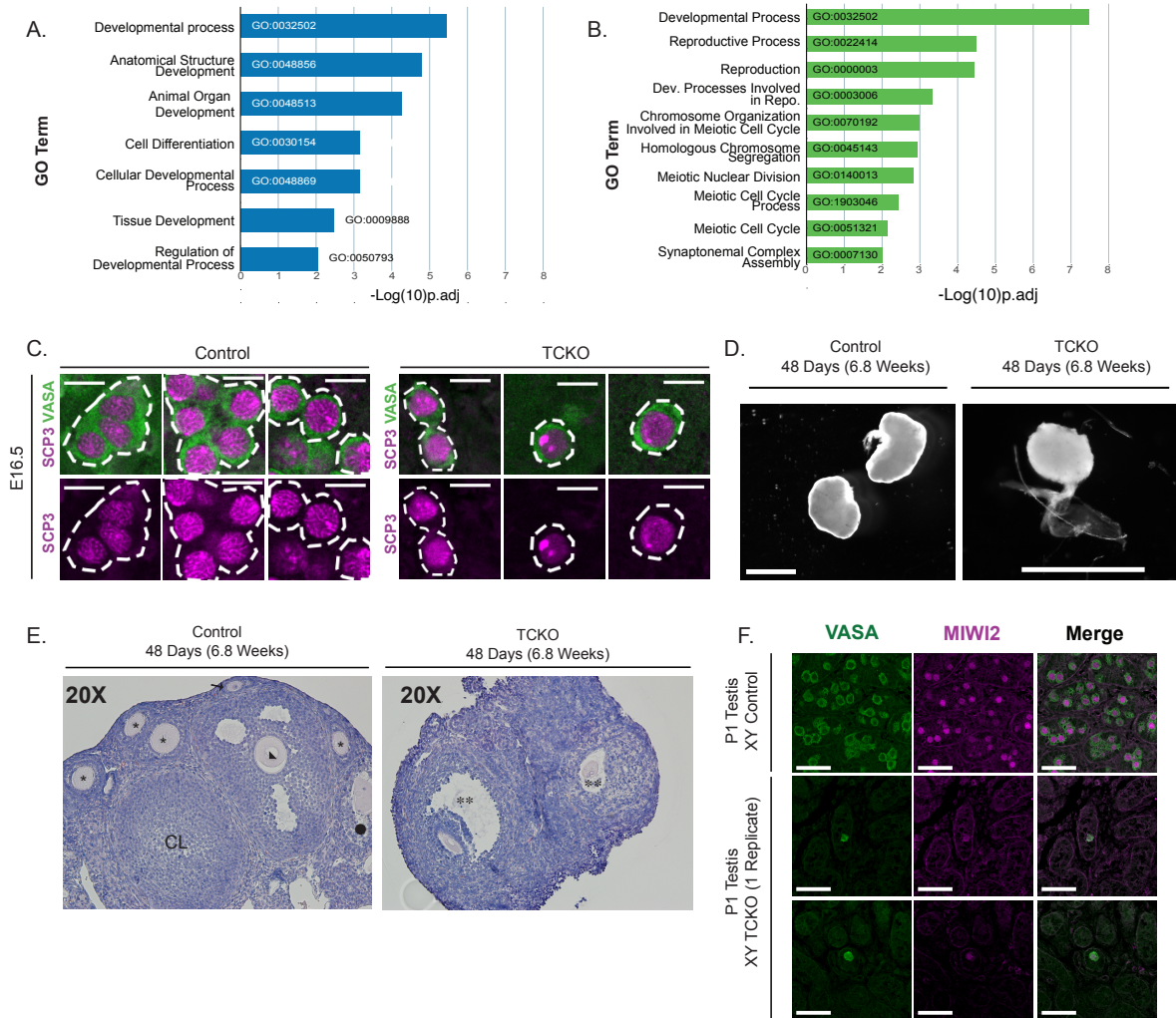
Supplemental Figure 5: A and B) GO analysis of E13.5 testicular PGC differential gene

expression using downregulated genes in A) testicular germ cells and B) ovarian germ cells. GO

number inset. C) Close-up view of E16.5 meiotic germ cells showing Sycp3 localization at E16.5

in control (left) and TCKO (right) PGCs. Each panel represents 1 biological replicate (n=3 each).

D) Images of ovaries isolated from n=1 48 day-old (~7 weeks) showing whole ovary

morphology. Scalebar represents 1500 μm (TCKO ovary is taken at a higher magnification). E)

H & E stain of control and TCKO ovary at 48 days. Control ovary shows primary follicle (arrow),

secondary follicles (asterisks), antral follicle (arrowhead) and an activated follicle (circle). TCKO

ovary shows two malformed follicles, one which appears antral-like (left, two asterisks) and one

which appears activated (right, one asterisk). F) IF images from P1 male testes from control (n =

3, top) and TCKO (n=1, bottom), showing reduced PGC number and improperly localized MIWI2

expression. Scale bars represent 50 μm.

Tables
Table 4-1: Antibodies used in this study with their dilutions

Antibody	Manufacturer	Catalog #	Concentration	Block	Primary	Secon.
MVH (VASA)	R and D	AF2030	1:100	12.5% NDS	2.5% NDS	0% NDS
TRIM28	Abcam	ab10484	1:200	12.5% NDS	2.5% NDS	0% NDS
DAZL	Abcam	ab215718	1:200	12.5% NDS	2.5% NDS	0% NDS
SCP3	Abcam	ab97672	1:200	12.5% NDS	2.5% NDS	0% NDS
NANOG	Abcam	ab214549	1:200	12.5% NDS	2.5% NDS	0% NDS
SOX2	Abcam	ab97959	1:200	12.5% NDS	2.5% NDS	0% NDS
TFAP2C	Santa Cruz	Sc-12762	1:200	12.5% NDS	2.5% NDS	0% NDS
PIWIL2 (MILI)	Abcam	ab36764	1:100	SuperBlock	Solution 1	Solution 2
PIWIL4 (MIWI2)	Thermo Scientific	PA5-31448	1:100	SuperBlock	Solution 1	Solution 2
cPARP	Cell Signaling Technology	9544	1:200	12.5% NDS	2.5% NDS	0% NDS
yH2Ax	Cell Signaling Technology	9718	1:200	12.5% NDS	2.5% NDS	0% NDS
Ki67	BD Pharmagen	556003	1:100	12.5% NDS	2.5% NDS	0% NDS
FOXL2	Novus	NB100-1277	1:200	12.5% NDS	2.5% NDS	0% NDS
NR2F2	Perseus	PP-H7147-00	1:200	12.5% NDS	2.5% NDS	0% NDS
DPPA3	Abcam	ab19878	1:500	12.5% NDS	2.5% NDS	0% NDS

Table 4-2: Genotyping Oligos Used in this Study

Target	Name	5' – 3'
Prdm1-cre	Blimp1-cre F	GCCGAGGTGCGCGTCAGTAC
	Blimp1-cre R	CTGAACATGTCCATCAGGTTCTTG
Trim28$^{f/f}$	Trim28 F	GGAATGGTTGTTCATTGGTG
	Trim28 R1 (No Excision)	ACCTTGGCCCATTTATTGATAAAG
	Trim28 R2 (Excision)	GCGAGCACGAATCAAGGTCAG
Sex	SMCX-1 (X)	CCGCTGCCAAATTCTTTGG
	SMCY-1 (Y)	TGAAGCTTTTGGCTTTGAG
Oct4-eGFP	Oct4-eGFP1	GATCACCTGGGGGTTTGAGAA
	Oct4-eGFP2	CAAGGCAAGGGAGGTAGACA
	Oct4-eGFP3	AGGAACTGCTTCCTTCACGA

Table 4-3: Online Datasets Used in this Study

Repository	Accession	SRA Record	Simple Name	Library Type
GEO	GSE60204	SRR1539456	D6_PGCLC_H3K27ac_Rep1	ChIP-seq
		SRR1539457	D6_PGCLC_H3K27ac_Rep2	ChIP-seq
		SRR1539503	D6_PGCLC_Input_Rep1	ChIP-seq
		SRR1539504	D6_PGCLC_Input_Rep2	ChIP-seq
GEO	GSE109770	SRR6519362	E10.5_PGC_Rep1	DNAse-seq
		SRR6519363	E10.5_PGC_Rep1	DNAse-seq
		SRR6519364	E12.5_PGC_XX_Rep1	DNAse-seq
		SRR6519365	E12.5_PGC_XX_Rep2	DNAse-seq
		SRR6519367	E12.5_PGC_XY_Rep1	DNAse-seq
		SRR6519368	E12.5_PGC_XY_Rep2	DNAse-seq
		SRR6519374	E14.5_PGC_XX_Rep1	DNAse-seq
		SRR6519375	E14.5_PGC_XX_Rep2	DNAse-seq
		SRR6519376	E14.5_PGC_XY_Rep1	DNAse-seq
		SRR6519377	E14.5_PGC_XY_Rep2	DNAse-seq
GEO	GSE141182	SRR10560100	E10.5_PGC_H3K9me3_Rep1	ChIP-seq
		SRR10560104	E10.5_PGC_H3K9me3_Rep2	ChIP-seq
		SRR13296472	E10.5_PGC_H3K9me3_Rep3	ChIP-seq
		SRR10560101	E10.5_PGC_Input_Rep1	ChIP-seq
		SRR10560105	E10.5_PGC_Input_Rep2	ChIP-seq
		SRR13296477	E10.5_PGC_Input_Rep3	ChIP-seq
		SRR10560108	E13.5_XX_PGC_H3K9me3_Rep1	ChIP-seq
		SRR10560114	E13.5_XX_PGC_H3K9me3_Rep2	ChIP-seq
		SRR13296474	E13.5_XX_PGC_H3K9me3_Rep3	ChIP-seq
		SRR10560106	E13.5_XX_PGC_Input_Rep1	ChIP-seq

		SRR10560112	E13.5_XX_PGC_Input_Rep2	ChIP-seq
		SRR13296479	E13.5_XX_PGC_Input_Rep3	ChIP-seq
		SRR10560111	E13.5_XY_PGC_H3K9me3_Rep1	ChIP-seq
		SRR10560117	E13.5_XY_PGC_H3K9me3_Rep2	ChIP-seq
		SRR13296476	E13.5_XY_PGC_H3K9me3_Rep3	ChIP-seq
		SRR10560109	E13.5_XY_PGC_Input_Rep1	ChIP-seq
		SRR10560115	E13.5_XY_PGC_Input_Rep2	ChIP-seq
		SRR13296481	E13.5_XY_PGC_Input_Rep3	ChIP-seq
DDBJ	DRA006633	DRR126197	E13.5_XY_PGC_H3K27ac_Rep1	ChIP-Seq
		DRR126198	E13.5_XY_PGC_H3K27ac_Rep2	ChIP-Seq
		DRR126199	E13.5_XY_PGC_Input	ChIP-Seq
		DRR126183	E13.5_XX_PGC_H3K27ac_Rep1	ChIP-Seq
		DRR126184	E13.5_XX_PGC_H3K27ac_Rep2	ChIP-Seq
		DRR126248	E13.5_XX_PGC_Input	ChIP-Seq
GEO	GSE118755	SRR7719557	E10.5_Gonadal_Soma_XY_Rep1	ATAC-Seq
		SRR7719558	E10.5_Gonadal_Soma_XY_Rep2	ATAC-Seq
		SRR7719555	E13.5_Gonadal_Soma_XY_Rep1	ATAC-Seq
		SRR7719556	E13.5_Gonadal_Soma_XY_Rep2	ATAC-Seq
		SRR7719559	E10.5_Gonadal_Soma_XX_Rep1	ATAC-Seq
		SRR7719560	E10.5_Gonadal_Soma_XX_Rep2	ATAC-Seq
		SRR7719553	E13.5_Gonadal_Soma_XX_Rep1	ATAC-Seq
		SRR7719554	E13.5_Gonadal_Soma_XX_Rep2	ATAC-Seq
GEO	GSE75751	SRR2980403	MERVL/Zscan4+ Rep 1	RNA-seq
		SRR2980404	MERVL/Zscan4+ Rep 2	RNA-seq
		SRR2980405	MERVL/Zscan4+ Rep 3	RNA-seq
		SRR2980406	E14 mESC NegativeControl_Rep1	RNA-seq

		SRR2980407	E14 mESC NegativeControl_Rep2	RNA-seq
		SRR2980408	E14 mESC NegativeControl_Rep3	RNA-seq

Methods:

RESOURCE AVAILIBILITY:

Lead contact: Reagents may be requested from the lead author, Amander Clark (clarka@ucla.edu)

Materials Availability: No novel reagents were made for this work. Relevant model systems are available from commercial suppliers.

Data and Code Availability: Raw data are available from the corresponding author upon request. Sequencing data generated in this work is available from NCBI GEO under Accession: GSE266178. No custom code was used in this project. FIJI (ImageJ) macros and code for data visualization is accessible at https://github.com/dirussoja/TRIM28_PGCs. All code available by request from the corresponding author.

EXPERIMENTAL MODEL DETAILS:

Animal Care

All animal experiments were approved by the UCLA Institutional Animal Care and Use Committee (IACUC), also known as the Chancellor's Animal Research Committee (ARC). All mice were housed in a standard research facility with a 12-hour light/dark cycle and at-will access to standard chow and water. All animals for this study were either bred in UCLA's on-site facilities or in a Charles River off-site colony facility, both of which adhere to national laboratory animal care guidelines.

Animal Models

Trim28[flox/flox] mice were established in ([55]). Trim28[flox/flox] (Jackson Laboratory, Cat # 018552) mice on a mixed C57BL/6 129/Sv background. Trim28[flox/flox] was crossed with Oct4-IRES-eGFP mice ([48]) on a C57BL/6 x 129S4/SvJae background (Jackson Laboratory) and backcrossed to get

Trim28$^{flox/flox}$;Oct4-eGFP$^{+/+}$ mice. Trim28$^{flox/flox}$;Oct4-eGFP$^{+/+}$ mice were outcrossed to CD1 (Charles River, Cat # 022) with one backcross. Prdm1-cre ([24]) mice on a B6CBAF1 background (Jackson Laboratory, 008827) maintained by breeding heterozygous Prdm1-cre males with pre-breeding C57/BL6 females (Jackson Laboratory, Cat# 000664). Prdm1-cre, Trim28$^{\Delta/+}$, Oct4-IRES-eGFP$^{+/-}$ males were generated by crossing Prdm1-cre males to Trim28$^{flox/flox}$, Oct4-IRES-eGFP$^{+/-}$ females. Male Prdm1-cre, Trim28$^{\Delta/+}$, Oct4-IRES-eGFP$^{+/-}$ were crossed with Trim28$^{flox/flox}$;Oct4-eGFP$^{+/+}$ females to generate Prdm1-cre, Trim28$^{flox/\Delta}$ (TCKO) and Trim28$^{flox/+}$ (control) progeny. For all experiments, impregnation was assessed by presence of a vaginal plug, which was considered E0.5. Embryos were harvested for dissection at E10.5, E11.5, E12.5, E13.5, E14.5 and E16.5 in accordance with primary and secondary euthanasia requirements as set out by UCLA IACUC. Embryo sample genotypes were confirmed by PCR for Prdm1-cre, Trim28^{flox} and sex (Table S2).

METHOD DETAILS:

Immunofluorescence

Aorta-Gonad-mesonephros of E11.5 and gonads of E11.5, E12.5, E13.5, E16.5, P1 embryos or 48-day adult ovaries were isolated via dissection and fixed in 4% formaldehyde (Thermo Scientific, PI28908 diluted in 1X PBS) overnight at 4C. Samples were then rinsed with 1x PBS and treated for 5 minutes with hematoxylin (Sigma-Aldrich, GHS116) to aid in sectioning. Samples were rinsed once more with 1x PBS and embedded in HistoGel (Epredia, HG-4000) and stored in 70% ethanol. Paraffin embedding and sectioning was either performed by UCLA Tissue Pathology Core Laboratory (TPCL) or by ethanol dehydration followed by xylene and paraffin (Leica Paraplast X-tra, 39603002) perfusion. Samples were prepared as 5 μm sections. Sections were deparaffinized by immersion in xylenes (Fisher Scientific) followed by immersion in 100% (v/v), 95% (v/v), 70% (v/v) and 50% (v/v) ethanol solutions before final immersion in ddH2O and 1x PBS. For antigen retrieval, samples were immersed in a pH 6 solution of 10 mM Sodium Citrate (Sigma Aldrich, S1804) and 0.5% Tween-20 (Fisher Scientific, BP337) at 95C for 40 minutes.

Samples were then allowed to cool to room temperature for 20 minutes. Sections were washed with 1x PBS followed by 1x PBS + 0.2% Tween-20 (PBS-T) and permeabilized by washing with 1x PBS + 0.5% Triton-X (MP Biomedicals, 194854) for 20 minutes at room temperature. Samples were washed three times with 1x PBS-T and blocked for 45 minutes at room temperature in 12.5% Normal Donkey Serum in PBS-T (Jackson Immunological Laboratories, 017-000-001). After blocking samples were incubated with primary antibodies in a 2.5% NDS PBS-T solution overnight at 4C in a humid chamber. The following day samples were washed with PBS-T three times for five minutes each. Secondary antibodies conjugated to AlexFluor 488, 594 or 647 (Jackson ImmunoResearch) diluted 1:200 in a 1x PBS-T solution were added to samples which were then allowed to sit overnight at 4C in a humid chamber. For E16.5 gonadal sections the same protocol was followed except samples were blocked for 45 minutes at room temperature using SuperBlock (Thermo Scientific, Ref# 37580), primary antibodies were diluted in Solution 1 (EMD Millipore, KP31812) and secondary antibodies were diluted in Solution 2 (EMD Millipore, KP31855). Samples were then washed 3x with 1x PBS-T and then treated with 1x DAPI (Life Technologies, D1306) in PBS-T for 10 minutes. Samples were mounted in Prolong Gold mounting media (Invitrogen, P36934) and stored to 4-10C prior to imaging.

Microscopy

All samples were images on Zeiss LSM880 Axio Observer Z.1 microscopes using Zen Black 2.3 SP1 or on a Zeiss LSM800 Axio Observer using Zen Blue. All sample images were taken as Z-stacks at 40x or 20x magnification. Image processing for publication was performed using FIJI (FIJI is Just Image J, NIH) to perform maximum-intensity projections, adjust brightness for publication, and add scale bars. Adobe Photoshop was used to crop regions of interest and overlay arrows, labels and scalebars prior to publication.

Image Analysis

Abundance of TRIM28 in PGCs versus soma and yH2Ax abundance was assessed using Imaris 8.3.1 (Bitplane). In brief, spots were assigned and manually verified for soma and PGCs. Soma

and PGCs for each technical replicate were averaged and the resulting ratio for each biological replicate reported. cPARP, DAZL, MIWI2, MILI, E16.5 NANOG and SOX2 and E16.5 oocyte γH2AX and SCP3 quantification were done by removing unnecessary z-slices using FIJI (ImageJ, NIH) and counting positive and negative PGCs. Ratios reported are taken from totals across technical replicates for each biological replicate. Ki67, Nanog, Tfap2c, and Sox2 relatiev intensity quantification were performed using FIJI (ImageJ, NIH) by selecting PGCs and a selection of somatic cells in a supervised manner. The mean gray intensity of individual PGCs was taken, normalized to somatic cells within technical replicates and normalized values plotted. Each PGC quantification is normalized to the somatic average within that technical replicate. Scripts for ImageJ analyses are available upon request. Statistical analysis and graphs generated using Prism 10.2.2.

FACS Isolation of PGCs

Embryonic trunks (E10.5, E11.5) or embryonic gonads (E12.5, E13.5 or E14.5) were dissected from embryos in 1x PBS and placed in 0.05% Trypsin (Gibco, 25300054) at 37˚C/5% CO2 for 5-minute intervals until they were visibly dissociated. Trypsin was quenched using 1x High Glucose DMEM (Gibco, 11965092) + 10% FBS (Gibco, 26140079) and centrifuged at 5,000 rpm for 5 minutes. Samples were washed 1x with 1x DPBS (Gibco, 14190144) + 1% BSA (Sigma-Aldrich, A3311) on ice and centrifuged again at 1200 rpm for 5 minutes. Samples were resuspended in 1x PBS + 1% BSA on ice and passed through a 100 um filter and treated 7AAD (BD Pharmingen, 559925) viability dye. FACS sorting was performed by the UCLA Broad Stem Cell Research Center FACS core using a BD ARIA SORP sorting for GFP+, 7AAD negative cells. For RNA-seq, cells were sorted into Qiagen Buffer RLT (Qiagen, 79216). For ATAC-seq, PGCs were sorted into 1x PBS + 1% BSA.

Isolation of PGCs for ChIP-seq

PGCs were isolated from 10 E12.5 time-mated breeding of pure-breeding CD1 mice (Charles River, 022). Entire ovaries or testes from E12.5 embryos, including the mesonephrous, were

94

dissected and placed in high glucose DMEM (Gibco, 11965092) + 10% FBS (Gibco, 26140079) on ice. The mesonephros was removed via surgical blades and collected in high glucose DMEM (Gibco, 11965092) + 10% FBS (Gibco, 26140079) with 1x DNase I (Invitrogen, 18047-019) and then washed twice with 1X DPBS (Gibco, 14190144). Resulting ovaries were dissociated in 0.05% Trypsin (Gibco, 25300054) for 10 minutes at 37C. Trypsin was quenched using high glucose DMEM + 10% FBS as above. The resulting suspension was passed through at 70 um cell strainer and then washed using the same media. The pellet was then resuspended in 1X DPBS, 0.1% BSA (Sigma-Aldrich, A3311), 0.5% EDTA (Thermo Fisher Scientific, AM9260G). Suspension was incubated with MACS antibodies against SSEA1 (Miltenyi, 130-094-530) (testicular and ovarian samples) and CD31 (Miltenyi, 130-097-418) (ovarian samples only) at 4°C for 30 min (ovarian samples) or 20 minutes (testicular samples). The resulting cells were centrifuged at 1600 rpm for 5 minutes and resuspended in the same buffer as above and loaded onto an MS column on a MiniMACS magnet. 1 mL of the buffer described above was added to the column and the cells released using the supplied plunger. The resulting suspension was counted and cells fixed as described in the ChIP methods.

Library Preparation for RNA-seq

RNA from PGCs sorted into Qiagen Buffer RLT was extracted using Qiagen RNeasy (Qiagen, 74004) according to the manufacturer's instructions. cDNA libraries were prepared using Ovation RNA-seq System V2 (Tecan Genomics, 7102) according to manufacturer's instructions. cDNA libraries were purified using Beckman Coulter Agencourt beads (Beckman Coulter, A63881). Purified cDNA was amplified by SPIA and purified using Qiagen MinElute (Qiagen, 28206) according to manufacturer's instructions and eluted in 50 uL of low EDTA TE buffer. Libraries were then sonicated to 200 bp fragments using a Covaris S2 (Covaris, Woburn MA) sonicator. Sonicated cDNA was purified using Qiagen MinElute (Qiagen, 28206) according to the manufacturer's instructions and the cDNA eluted in 10 uL of low EDTA TE buffer. 8uL of eluate was then repaired and indexed for sequencing using Encore Rapid Library System (Tecan

Genomics, 0319 and 0320) according to manufacturer's instructions. Final indexed libraries were purified using Agencourt XP beads (Beckman Coulter, A63881) and the final libraries suspended in low-EDTA TE buffer. Library concentration was assessed using KAPA Library Quantification Kit for Illumina platforms (Roche Sequencing, KK4824) according to the manufacturer's instructions. Libraries were sequenced on an Ilumina NovaSeq 6000 at 2 x 50 bp (E13.5) or 2x100 bp (E11.5 and E12.5).

Library Preparation for ATAC-seq

We followed a low-input OMNI-ATAC-seq protocol as in (CITE). For E10.5 samples, all PGCs were collected. For E14.5 samples, we collected only 1500 PGCs per sample, to avoid differences in input quantity between TCKO and control gonads. Immediately following isolation via FACS, sorted PGCs in 1x PBS + 1% BSA were pelleted by centrifugation at 750 rcf for 10 minutes at 4C. Supernatant was aspirated an resuspended in 400 uL of pre-chilled RSB buffer on ice: 10 mM Tris-HCl pH 7.4 (Teknova, 15074), 10 mM NaCl (Invitrogen, AM9760G), 3 mM $MgCl_2$ (Invitrogen, AM9530G). Suspension was re-pelleted by centrifugation at 750 RCF for 10 minutes at 4C. Following centrifugation, supernatant was removed and pellet was resuspended in 10 uL of Tn5 Enzyme Mix on ice: 0.33X PBS, 10 mM Tris-HCl ph 7.6, 5 mM $MgCl_2$, 5% Dimethyl Formamide (Acros Organics, 327171000), 3mM Tn5 Enzyme (Illumina, 20034197), 0.1% Digitonin (Promega, G9441), 0.1% Tween-20 (Sigma Aldrich, 11332465001), 0.1% NP-40(Sigma Aldrich, 11332473001). Reaction was processed on a thermoshaker at 37C, 1000 rpm for 30 minutes. After reaction, samples were placed on ice and diluted with 10 uL of pre-chilled ice cold water. Fragmented DNA was isolated using NEB Monarch DNA purification system (NEB, T1030S). 100 uL of Monarch DNA Cleanup Binding Buffer was added to each sample and passed through the column by centrifugation for 1 minutes at 13500 rpm. Flow-through was discarded and column washed two times with 200 uL of DNA Wash Buffer followed by centrifugation for 1 minute at 13500 rpm. Samples were eluted in 20 uL of water. Eluate was mixed on ice with 2x NEBNext High-Fidelity 2x PCR mix (NEB, M0541L) and 25 uM of Indexing forward and reverse primers

96

based on Illumina Nextera Indexing Kit (IDT). Resulting reaction was amplified for 5 cycles, then split into two reactions prior to 4 and 8 more cycles of amplification. Amplified DNA was purified using NEB Monarch DNA Purification kit as above using 200 uL of DNA binding buffer and eluting in 21 uL of Low-EDTA EB buffer. Sample quality was assessed using Tapestation 4200 using a D1000 ScreenTap (Agilent Technologies, 5067-5582) and quantified by Quibit HS (Thermo Fisher Scientific, Q32854). Samples were sequenced on an Illumina NovaSeq 6000 (E14.5, SP 2x50 bp) or NovaSeq X (E10.5, 2x100 bp).

Chromatin Immunoprecipitation (ChIP)

After isolation via MACS as described above, samples were rinsed with 1x PBS (Gibco, 14190144). Cells were pelleted by centrifugation at 1500 rpm for 5 minutes. Supernatant removed. Cells were resuspended in 1% paraformaldehyde (Thermo Scientific Pierce, PI28906) at room temperature rotating for 10 minutes. Fixation was quenched using 0.14 M Glycine (Fisher Scientific, G45212) rotating at room temperature for 10 minutes. Cells were pelleted at 3000 rpm for 5 minutes, supernatant removed and flash frozen in liquid nitrogen and stored at -80C. Fixed PGCs were thawed on ice and resuspended in 10 mM Tris-HCl pH 8 (Invitrogen 15568025), 0.25% Triton X-100 (MP Biomedicals, ICN19485480), 10 mM EDTA (Sigma-Aldrich, T9285), 0.5 mM EGTA (bioWorld, 40520008), Roche cOmplete, 4693116001), 1mM PMSF (Thermo Scientific, 36978) and incubated rotating at room temperature for 15 minutes. Cells were pelleted by centrifugation at 4000 rpm for 5 minutes at 4C. Supernatant was removed and cells resuspended in 10 mM Tris-HCl pH 8.0 (Invitrogen 15568025), 200 mM NaCl (Invitrogen, AM9760G), 10 mM EDTA (Sigma-Aldrich, T9285), 0.5 mM EGTA (bioWorld, 40520008), Roche cOmplete, 4693116001), 1mM PMSF (Thermo Scientific, 36978) for 10 minutes rotating at 4C. Cells were pelleted by centrifugation at 4000 rpm, for 5 minutes at 4C. Supernatant was discarded and pellet resuspended in 650 uL of 10 mM Tris-HCl pH 8.0 (Invitrogen 15568025), 10 mM EDTA (Sigma-Aldrich, T9285), 0.5 mM EGTA (bioWorld, 40520008), Roche cOmplete, 4693116001), 1mM PMSF (Thermo Scientific, 36978). Supernatant was transferred to a 12mm sonication tube

(Covaris, 520130) and sonicated using a Covaris S2 with the following program: Intensity = 5, Cycles/burst = 200, duty cycle = 5% with a 4x 30" on, 30" off, 30" on, 30" off periodicity for an effective sonication time of 4 minutes. Sonicated lysate was centrifuged at 14000 rpm for 10 minutes at 4C. 10% of clarified sonicate was kept as an input sample at -80C. Samples to be immunoprecipitated were pre-cleared with Protein A Dynabeads (Invitrogen, 10001D) in a 16.7 mM Tris-HCl pH 8.0 (Invitrogen 15568025), 0.01% SDS (Ambion, AM9820), 1.1% Triton-X 100 (MP Biomedicals, ICN19485480), 1.2 mM EDTA (Sigma-Aldrich, T9285), 167 mM NaCl (Invitrogen, AM9760G) solution for 2 hours rotating at 4C. Supernatant was collected and 4 ug of Abcam ab10484 added. Antibody and supernatant were left to incubate overnight rotating at 4C. Protein A dynabeads were added to mixture as before and incubated rotating at 4C for 2 hours. Supernatant was the removed and discarded and dynabeads washed 2 times with 50 mM HEPES pH 7.9 (Fisher Bioreagents, BP310), 1% Triton-X100 (MP Biomedicals, ICN19485480), 0.1% Deoxycholate, 1mM EDTA (Sigma-Aldrich, T9285), 140 mM NaCl (Invitrogen, AM9760G) for 4 minutes rotating at 4C. Beads were then washed two times with 2x with Tris-EDTA for 4 minutes rotating at 4C. Supernatant was discarded and beads resuspended in 50 mM Tris-HCl pH 8 (Invitrogen 15568025), 1mM EDTA (Sigma-Aldrich, T9285), 1% SDS (Ambion, AM9820) for elution. Elution was performed on a thermoshaker at 65C shaking at 1400 rpm for 10 minutes. Supernatant was collected and the elution procedure repeated on beads. First and second elution were pooled. Input fractions and eluted chromatin were incubated overnight at 65C. Samples were then treated with 15 ug of RNAseA (Invitrogen, 12091021) for 30 minutes at 37C. Samples were then treated with 100 ug of Proteinase K for 2 hours at 56C. DNA was then purified using Quiagen MinElute (Qiagen, 28206) according to instructions with a final elution of 12 uL.

ChIP-seq Library Preparation

ChIP-seq libraries were prepared using the Ovation UltraLow System V2 (Tecan Genomics, 0344) according to manufacturer's instructions. For each library, the number of amplification cycles was based off of quantification by Quibit dsHS (Thermo Fisher Scientific, Q32854). Final amplified and

indexed libraries were purified using Agencourt XP beads (Beckman Coulter, A63881) and library quantity determined using Quibit HS assuming an average fragment length of 250 bp. Libraries were sequenced on an Illumina Novaseq 6000 at 2x100 bp.

Bioinformatic Analysis

Data Retrieval, Reference Genome and Repeatmasker

Mouse reference genome GRCm39 and the corresponding ENSEMBL gene annotation file GRCm39.109 were downloaded from Ensembl[1] and utilized for all genomics analyses. TE annotation file for GRCm39 from repeatmasker (http://repeatmasker.org/) were used for TE related analyses. To retrieve data from US NCBI SRA, we used SRA Tools v2.10.9.

RNA-seq

Quality control and read quantification were adopted and optimized from a previous study.[2] Specifically, we used FastQC v0.11.9[3] to perform quality control on raw RNA-seq data. The alignment was performed with STAR v2.7.10a[4] with the optimized setting (--outFilterMultimapNmax 1000, --outSAMmultNmax 1, --outFilterMismatchNmax 3, and --alignIntronMax 1). We sorted and indexed the aligned bam output using SAMtools v1.15[5] with the default setting. The read quantification was performed separately for genes and transposable elements. We used the featureCounts function from Subread v2.0.2[6] to process the BAM outputs from the previous step. For TE quantification, we included multimapped reads and quantified the individual duplicates (transcript_id), as well as individual subfamilies (gene_id).

The raw read counts were input into and processed by R 4.2.2[7]. Genes and TEs with no reads across all samples were removed, and RPKM values were calculated. To identify DETEs and DEGs, we filtered out TEs or genes with RPKM mean less than 1 in both treatment and control groups. Significant TEs and genes were defined with two log2-transformed fold-change thresholds (1.5 and 2) and FDR < 0.05. Both TE and gene expression counts were normalized using the rlogtranformation function and were compared both within each sex and between sexes from all time-points.

ChIP-seq

ChIP-seq of TRIM28 were preprocessed using TrimGalore v0.6.10 with the arguments --stringency 3 --length 20 --paired --nextseq 20. Alignment was performed using STAR/2.7.9a with options --outFilterMultimapNmax 1000 --outFilterMismatchNmax 3 --alignIntronMax 1 --outSAMmultNmax 1. For TRIM28 ChIP-seq we invoked --alignEndsType EndToEnd. Deduplication was done using Picard Tools MarkDuplicates v2.25.0. The aligned BAM files were then processed by SAMtools v1.15 to sort by indexes. For H3K27ac ChIP-seq in d6 PGCLCs, reads were converted from color space to base space using fastq-dump -B.

Bigwig files were generated using Deeptools v3.5.1 bamCoverage with options –bs10 and -e and were normalized by RPKM or normalized read counts. Input-subtracted RPKM bigwig files were generated using bamCompare with --scaleFactorsMethod None --normalizeUsing RPKM -e --operation subtract. Tracks were visualized as heatmaps using the computeMatrix function using the center of the region as the reference point and a window size of 3,000, 5,000 or 10,000 bp up and down-stream invoking --missingDataAsZero. Heatmaps were then generated using plotHeatmap from DeepTools v3.5.1. Profiles were generated using plotProfiles from deepTools v3.5.1. Gene tracks produced using IGV 2.16.1.

ChIP-seq peaks were defined using the callpeak function from MACS v3.0.0[11] with an FDR threshold of 0.05 and an effective genome size compiled for GRCm39. Peaks were called individually for each biological replicate, combined, coordinate sorted using bedtools sort and overlapping peaks removed by bedtools merge -I -d 20 to produce a final list.

To calculate RPKM of H3K9me3 over TRIM28-derepressed ERVs we calculated RPKM from input-subtracted RPKM bigwig files using DeepTools multiBigwigSummary. For each sex, sex-specific and shared TCKO regions were combined. Graphs of H3K9me3 enrichment were plotted using ggplot2 (R 4.2.1) and ggpubr (R 4.3.3).

ATAC-seq Analysis

Adapter trimming on ATAC-seq libraries was performed using TrimGalore v0.6.10 with the parameters --clip_R1 18 --stringency 3 --length 20 --paired --nextseq 20. Trimmed reads were aligned with STAR 2.7.9a with options --outFilterMultimapNmax 1000 --outFilterMismatchNmax 3 --alignIntronMax 1 --outSAMmultNmax 1 --outFilterMatchNminOverLread 0.2. PCR duplicates were removed using Picard Tools MarkDuplicates v2.25.0. Biological replicates passing QC were merged using Samtools merge v1.15. Bigwig files were generated using Deeptools v3.5.1. Gene tracks produced using IGV 2.16.1.

For analysis of open to closed or closed to open chromatin, peaks of both sexes in control-only conditions were called using Genrich as above and merged above. We then calculated the RPKM of these peaks across all samples using multiBigWigSummary. We set an RPKM threshold of < 15 RPKM for "closed" and > 15 RPKM for "open" chromatin. We then binned, for each sex, peaks which were open at both time points, closed, or transitioned from closed to open or open to closed. Resulting peak files were stored as BED for downstream analysis and visualization of open to closed and closed to open peak classes was performed using computeMatrix –missingDataAsZero followed by plotHeatmap from deepTools v3.5.1. For gonadal soma ATAC-seq, we used the same approached but used RPKM < 10 as a cutoff for closed and RPKM > 10 as a cutoff for open chromatin. Subfamily tally was performed by dividing the number of unique integrants of a subfamily by the total number of integrants in the subfamily.

For TRIM28-specific ERV derepression analysis (Fig. 2), peaks were called using Genrich v0.6.1 with individual BAMs for biological replicates, with options -j -q 0.05. Peak files for all conditions were merged together and duplicate or bookended peaks removed with bedtools merge -I -d 10. RPKM of merged peak files were called using multiBigwigSummary. Peaks were classified based on higher RPKM value than all other conditions (across genotype, time and sex) and > 2 fold higher RPKM values than all other conditions in R 4.2.1. Resulting tables were exported and visualization performed using computeMatrix with –missingDataAsZero followed by plotHeatmap.

HOMER Motif Analysis

Motif analysis was performed using BED files of TEs which transitioned from open to closed or closed to open in our PGC or gonadal soma ATAC-seq analysis. Motif analysis was performed using HOMER v4.11.1. Motifs were called using findMotifsGenome.pl with --mset vertebrates against the mm39 genome. Motifs from knownMotifs were plotted using R 4.3.3 using pheatmap with p-values from HOMER transformed to -log10(pvalue) in R.

DNAse-Analysis

FASTQ files were aligned using STAR v2.7.9a with the parameters --outFilterMultimapNmax 1000 --outFilterMismatchNmax 3 --alignIntronMax 1 --outSAMmultNmax 1. PCR duplicates were removed using Picard Tools MarkDuplicates v2.25.0. Biological replicates passing QC were merged using Samtools merge v1.15. Bigwig files were generated using Deeptools v3.5.1. Peaks were called using MACS3 (v3.0.0a6) with parameters -f BAM -g 2654621783 -q 0.01. bigWig files were generated using deepTools v3.5.1 bamcoverage. Gene tracks produced using IGV 2.16.1.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical tests were performed using R (v4.2.1, v4.3.3) and Graphpad Prism (v 10.2.2-3). $p < 0.05$ used for all t-test and Welch's t-test as the cutoff for significance. For RNA-seq, significance was determined by DESeq2 with $|Log2FC| > 2$ and FDR (p.adj) < 0.05. Significance testing for ATAC-seq and ChIP-seq peak-calling determined by peak-calling program and are detailed in their respective methods. All graphs show individual Ns, and graph details are provided in each figure legend. Source data are provided as Supplemental Tables.

Table 4-4: KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Goat anti-Mouse Vasa Homolog	R and D	AF2030; RRID: AB_2277369
Rabbit anti-KAP1 (TRIM28)	Abcam	ab10484; RRID: AB_297223
Rabbit anti-DAZL	Abcam	ab215718; RRID: AB_2893177
Mouse anti-TFAP2C	Santa Cruz Biotechnology	sc-12762, RRID: AB_667770
Mouse anti-Ki67	BD	556003, RRID: AB_396287
Rabbit anti-Sox2	Abcam	ab97959, RRID: AB_2341193
Rabbit anti-Nanog	Abcam	ab214549, RRID: AB_3668944
Rabbit anti-Piwil2 (MILI)	Abcam	ab36764, RRID: AB_777284
Rabbit anti-Piwil4 (MIWI2)	Thermo Fisher	PA5-31448, RRID: AB_2548922
Rabbit anti-γH2Ax	Cell Signaling Technology	9718, RRID: AB_2118009
Mouse anti-Sycp3	Abcam	ab97672, RRID: AB_10678841
Goat anti-FoxL2	Novus	NB100-1277, RRID: AB_2106188
Mouse anti-Nr2f2 (COUP TF II)	Perseus	PP-H7147-00, RRID: AB_2155627
Rabit anti-Dppa3 (STELLA)	Abcam	ab19878, RRID: AB_2246120
Critical commercial assays		
Ovation RNA-Seq System V2	Tecan Genomics	7102
Ovation UltraLow System V2	Tecan Genomics	0344NB-32
Illumina Tagment DNA Enzyme and Buffer Small Kit	Illumina	20034197
Monarch PCR and DNA Clean-UP	New England Biolabs	T1030S
RNeasy Micro	Qiagen	74004
Qiagen MinElute Reaction Cleanup Kit	Qiagen	28204
Agencourt AMPure XP	Beckman Coulter	A63881
NEBNext High-Fidelity 2X PCR Master Mix	New England Biolabs	M0541L
Invitrogen SYBR Green I 10,000X	Invitrogen	S7563
High Sensitivity D1000 Screen Tape	Agilent	5067-5584
KAPA Library Quantification Kits for Next-Generation Sequencing	Kapa Biosystems (Roche)	kk4824
Quibit dsDNA High Sensitivity	Life Technologies	Q32854
Dynabeads Protein A Beads	Invitrogen	10001D
Dynabeads Protein G Beads	Invitrogen	10003D
Deposited data		
RNA-seq of PGCs at E11.5, E12.5 and E13.5	This Paper	GEO GSE266176

ATAC-seq of E10.5 and E14.5 PGCs	This Paper	GEO: GSE266172
ChIP-seq of PGCs at E12.5	This Paper	GEO: GSE266173
ATAC-Seq of gonadal soma	Garcia-Moreno et al., *Dev. Biol.*, 2019	GEO: GSE118755
DNAse-Seq of PGCs	Li et al., *Cell Research*., 2018	GEO: GSE109770
RNA-seq of mESCs and Zscan4/MERVL+ mESCs	Eckersley-Maslin et al., *Cell Reports*, 2016	GEO: GSE75751
H3K9me3 ChIP-seq of PGCs	Huang et al., *Nature*, 2021	GEO: GSE141180
H3K27ac ChIP-seq of d6 PGCLCs	Kurimoto et al., *Cell Stem Cell*, 2015	GEO: GSE60204
H3K27ac ChIP-seq of E13.5 PGCs	Kawabata et al., *Epigenomics*, 2019	DDBJ: DRA006633
Experimental models: Organisms/strains		
Mouse: Strain B6;129S4-*Pouf5f1*^{tm2Jae}/J	The Jackson Laboratory	RRID: IMSR_JAX:008214
Mouse: Strain B6.Cg-Tg(Prdm1-Cre)^{1Masu}/J	The Jackson Laboratory	RRID: IMSR_JAX:008827
Mouse: Strain B6.129S2(SJL)-Trim28$^{tm1.1lpc}$/J	The Jackson Laboratory	I RRID: MSR_JAX:018552
Mouse Strain: CD1	Charles River	RRID: IMSR_CRL:022
Oligonucleotides		
Primers are listed in Table S1.	NA	NA
Software and algorithms		
STAR 2.7.9a	Dubin et al., *Bioinformatics* 2013	RRID: SCR_004463
Featurecounts	Liao et al., *Bioinformatics* 2014	RRID: SCR_012919
Bedtools 2.30.0	Quinlan et al., *Bioinformatics*, 2010	RRID: SCR_006646
Samtools 1.15.0	Li et al., *Bioinformatics*, 2009	RRID: SCR_002105
R 4.2.1	R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, 2023	RRID: SCR_001905
DESeq2	Love et al., *Genome Biology* 2014	RRID: SCR_015687
ggplot2	Wickham, *ggplot2: Elegant Graphics for Data Analysis*. 2016	RRID: SCR_014601
DeepTools v3.5.1	Ramírez et al., *Nucleic Acids Research*, 2016	RRID: SCR_016366
MACS 3.0.0	Zhang et al., *Genome Biology*, 2008	RRID: SCR_013291
Genrich	John M. Gaspar, https://github.com/jsh58/Genrich	RRID: SCR_025320

SRA Toolkit	US NCBI, https://github.com/ncbi/sra-tools/	RRID: SCR_024350
HOMER	Heinz et al., *Mol Cell*, 2010	RRID: SCR_010881
FIJI	Schindelin et al., *Nat. Methods*, 2012	RRID: SCR_002285
Picard tools v2.25.0	Broad Institute, http://broadinstitute.github.io/picard	RRID: SCR_006525
pheatmap	Kolde R (2018). pheatmap: Pretty Heatmaps.	RRID: SCR_016418
Smplot2	Seung Hyun Min, https://github.com/smin95/smplot2	NA
TrimGalore	Martin, Marcel, *EMBnet journal*, 2011 *and* https://github.com/FelixKrueger/TrimGalore	RRID: SCR_011841
Zen Black 2.3 SP1	Zeiss	RRID: SCR_013672
Zen Blue	Zeiss	RRID: SCR_013672
Graphpad Prism	Dotmatics/Graphpad	RRID: SCR_002798
Other		
Zeiss LSM 880	Zeiss	RRID: SCR_020925
Agilent 4200 Tapstation	Agilent	RRID: SCR_018435
Biorad CFX Connect Thermocycler	Biorad	RRID: SCR_017251
Biorad T100 Thermocycler	Biorad	RRID: SCR_021921
Illumina NovaSeq 6000	Illumina	RRID: SCR_016387
Illumina NovaSeq X Plus	Illumina	RRID: SCR_024569
Covaris S2 Sonicator	Covaris	N/A

References:

1   Orgel, L. E. & Crick, F. H. Selfish DNA: the ultimate parasite. *Nature* **284**, 604-607 (1980). https://doi.org:10.1038/284604a0

2   Doolittle, W. F. & Sapienza, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**, 601-603 (1980). https://doi.org:10.1038/284601a0

3   Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* **8**, 272-285 (2007). https://doi.org:10.1038/nrg2072

4   Wells, J. N. & Feschotte, C. A Field Guide to Eukaryotic Transposable Elements. *Annu Rev Genet* **54**, 539-561 (2020). https://doi.org:10.1146/annurev-genet-040620-022145

5   Kissinger, J. C. & DeBarry, J. Genome cartography: charting the apicomplexan genome. *Trends Parasitol* **27**, 345-354 (2011). https://doi.org:10.1016/j.pt.2011.03.006

6   Finnegan, D. J. Eukaryotic transposable elements and genome evolution. *Trends Genet* **5**, 103-107 (1989). https://doi.org:10.1016/0168-9525(89)90039-5

7   Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**, 973-982 (2007). https://doi.org:10.1038/nrg2165

8   Cornelis, G. *et al.* Retroviral envelope syncytin capture in an ancestrally diverged mammalian clade for placentation in the primitive Afrotherian tenrecs. *Proc Natl Acad Sci U S A* **111**, E4332-4341 (2014). https://doi.org:10.1073/pnas.1412268111

9   Dupressoir, A. *et al.* A pair of co-opted retroviral envelope syncytin genes is required for formation of the two-layered murine placental syncytiotrophoblast. *Proc Natl Acad Sci U S A* **108**, E1164-1173 (2011). https://doi.org:10.1073/pnas.1112304108

10   Lynch, V. J., Leclerc, R. D., May, G. & Wagner, G. P. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat Genet* **43**, 1154-1159 (2011). https://doi.org:10.1038/ng.917

11   Chuong, E. B., Rumi, M. A., Soares, M. J. & Baker, J. C. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat Genet* **45**, 325-329 (2013). https://doi.org:10.1038/ng.2553

12   Senft, A. D. & Macfarlan, T. S. Transposable elements shape the evolution of mammalian development. *Nat Rev Genet* **22**, 691-711 (2021). https://doi.org:10.1038/s41576-021-00385-1

13   Xiang, X. *et al.* Human reproduction is regulated by retrotransposons derived from ancient Hominidae-specific viral infections. *Nat Commun* **13**, 463 (2022). https://doi.org:10.1038/s41467-022-28105-1

14   Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* **128**, 1089-1103 (2007). https://doi.org:10.1016/j.cell.2007.01.043

15   Malone, C. D. *et al.* Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. *Cell* **137**, 522-535 (2009). https://doi.org/10.1016/j.cell.2009.03.040

16   Wang, S. H. & Elgin, S. C. Drosophila Piwi functions downstream of piRNA production mediating a chromatin-based transposon silencing mechanism in female germ line. *Proc Natl Acad Sci U S A* **108**, 21164-21169 (2011). https://doi.org/10.1073/pnas.1107892109

17   Kuramochi-Miyagawa, S. *et al.* Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. *Development* **131**, 839-849 (2004). https://doi.org/10.1242/dev.00973

18   Girard, A., Sachidanandam, R., Hannon, G. J. & Carmell, M. A. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442**, 199-202 (2006). https://doi.org/10.1038/nature04917

19   Carmell, M. A. *et al.* MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev Cell* **12**, 503-514 (2007). https://doi.org/10.1016/j.devcel.2007.03.001

20   Aravin, A. A. *et al.* A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol Cell* **31**, 785-799 (2008). https://doi.org/10.1016/j.molcel.2008.09.003

21   De Fazio, S. *et al.* The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements. *Nature* **480**, 259-263 (2011). https://doi.org/10.1038/nature10547

22   Pezic, D., Manakov, S. A., Sachidanandam, R. & Aravin, A. A. piRNA pathway targets active LINE1 elements to establish the repressive H3K9me3 mark in germ cells. *Genes Dev* **28**, 1410-1428 (2014). https://doi.org/10.1101/gad.240895.114

23   Stein, P. *et al.* Essential Role for endogenous siRNAs during meiosis in mouse oocytes. *PLoS Genet* **11**, e1005013 (2015). https://doi.org/10.1371/journal.pgen.1005013

24   Ohinata, Y. *et al.* Blimp1 is a critical determinant of the germ cell lineage in mice. *Nature* **436**, 207-213 (2005). https://doi.org/10.1038/nature03813

25   Magnusdottir, E. *et al.* A tripartite transcription factor network regulates primordial germ cell specification in mice. *Nat Cell Biol* **15**, 905-915 (2013). https://doi.org/10.1038/ncb2798

26   Leitch, H. G. & Smith, A. The mammalian germline as a pluripotency cycle. *Development* **140**, 2495-2501 (2013). https://doi.org/10.1242/dev.091603

27   Weber, S. *et al.* Critical function of AP-2 gamma/TCFAP2C in mouse embryonic germ cell maintenance. *Biol Reprod* **82**, 214-223 (2010). https://doi.org/10.1095/biolreprod.109.078717

28   Yamaji, M. *et al.* PRDM14 ensures naive pluripotency through dual regulation of signaling and epigenetic pathways in mouse embryonic stem cells. *Cell Stem Cell* **12**, 368-382 (2013). https://doi.org:10.1016/j.stem.2012.12.012

29   Nakaki, F. *et al.* Induction of mouse germ-cell fate by transcription factors in vitro. *Nature* **501**, 222-226 (2013). https://doi.org:10.1038/nature12417

30   Molyneaux, K. A., Stallock, J., Schaible, K. & Wylie, C. Time-lapse analysis of living mouse germ cell migration. *Dev Biol* **240**, 488-498 (2001). https://doi.org:10.1006/dbio.2001.0436

31   Molyneaux, K. A., Wang, Y., Schaible, K. & Wylie, C. Transcriptional profiling identifies genes differentially expressed during and after migration in murine primordial germ cells. *Gene Expr Patterns* **4**, 167-181 (2004). https://doi.org:10.1016/j.modgep.2003.09.002

32   Hajkova, P. *et al.* Chromatin dynamics during epigenetic reprogramming in the mouse germ line. *Nature* **452**, 877-881 (2008). https://doi.org:10.1038/nature06714

33   Seisenberger, S. *et al.* The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Mol Cell* **48**, 849-862 (2012). https://doi.org:10.1016/j.molcel.2012.11.001

34   Seki, Y. *et al.* Extensive and orderly reprogramming of genome-wide chromatin modifications associated with specification and early development of germ cells in mice. *Dev Biol* **278**, 440-458 (2005). https://doi.org:10.1016/j.ydbio.2004.11.025

35   Li, Z. *et al.* The Sm protein methyltransferase PRMT5 is not required for primordial germ cell specification in mice. *EMBO J* **34**, 748-758 (2015). https://doi.org:10.15252/embj.201489319

36   Kim, S. *et al.* PRMT5 protects genomic integrity during global DNA demethylation in primordial germ cells and preimplantation embryos. *Mol Cell* **56**, 564-579 (2014). https://doi.org:10.1016/j.molcel.2014.10.003

37   Gill, M. E., Hu, Y. C., Lin, Y. & Page, D. C. Licensing of gametogenesis, dependent on RNA binding protein DAZL, as a gateway to sexual differentiation of fetal germ cells. *Proc Natl Acad Sci U S A* **108**, 7443-7448 (2011). https://doi.org:10.1073/pnas.1104501108

38   Chen, H. H. *et al.* DAZL limits pluripotency, differentiation, and apoptosis in developing primordial germ cells. *Stem Cell Reports* **3**, 892-904 (2014). https://doi.org:10.1016/j.stemcr.2014.09.003

39   Nicholls, P. K. *et al.* Mammalian germ cells are determined after PGC colonization of the nascent gonad. *Proc Natl Acad Sci U S A* **116**, 25677-25687 (2019). https://doi.org:10.1073/pnas.1910733116

40   Bowles, J. *et al.* Retinoid signaling determines germ cell fate in mice. *Science* **312**, 596-600 (2006). https://doi.org:10.1126/science.1125691

41   Zhao, J. *et al.* Cell-fate transition and determination analysis of mouse male germ cells throughout development. *Nat Commun* **12**, 6839 (2021). https://doi.org:10.1038/s41467-021-27172-0

42   Garcia, T. X. & Hofmann, M. C. NOTCH signaling in Sertoli cells regulates gonocyte fate. *Cell Cycle* **12**, 2538-2545 (2013). https://doi.org:10.4161/cc.25627

43   Ohta, K. *et al.* Male differentiation of germ cells induced by embryonic age-specific Sertoli cells in mice. *Biol Reprod* **86**, 112 (2012). https://doi.org:10.1095/biolreprod.111.095943

44   Rowe, H. M. *et al.* TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. *Genome Res* **23**, 452-461 (2013). https://doi.org:10.1101/gr.147678.112

45   Rowe, H. M. *et al.* KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature* **463**, 237-240 (2010). https://doi.org:10.1038/nature08674

46   Ecco, G. *et al.* Transposable Elements and Their KRAB-ZFP Controllers Regulate Gene Expression in Adult Tissues. *Dev Cell* **36**, 611-623 (2016). https://doi.org:10.1016/j.devcel.2016.02.024

47   Imbeault, M., Helleboid, P. Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550-554 (2017). https://doi.org:10.1038/nature21683

48   Schultz, D. C., Ayyanathan, K., Negorev, D., Maul, G. G. & Rauscher, F. J., 3rd. SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev* **16**, 919-932 (2002). https://doi.org:10.1101/gad.973302

49   Liu, S. *et al.* Setdb1 is required for germline development and silencing of H3K9me3-marked endogenous retroviruses in primordial germ cells. *Genes Dev* **28**, 2041-2055 (2014). https://doi.org:10.1101/gad.244848.114

50   Huang, T. C. *et al.* Sex-specific chromatin remodelling safeguards transcription in germ cells. *Nature* **600**, 737-742 (2021). https://doi.org:10.1038/s41586-021-04208-5

51   Hill, P. W. S. *et al.* Epigenetic reprogramming enables the transition from primordial germ cell to gonocyte. *Nature* **555**, 392-396 (2018). https://doi.org:10.1038/nature25964

52   Hargan-Calvopina, J. *et al.* Stage-Specific Demethylation in Primordial Germ Cells Safeguards against Precocious Differentiation. *Dev Cell* **39**, 75-86 (2016). https://doi.org:10.1016/j.devcel.2016.07.019

53   Li, J. *et al.* Accurate annotation of accessible chromatin in mouse and human primordial germ cells. *Cell Res* **28**, 1077-1089 (2018). https://doi.org:10.1038/s41422-018-0096-5

54   Sundaram, V. *et al.* Functional cis-regulatory modules encoded by mouse-specific endogenous retrovirus. *Nat Commun* **8**, 14550 (2017). https://doi.org:10.1038/ncomms14550

55   Todd, C. D., Deniz, O., Taylor, D. & Branco, M. R. Functional evaluation of transposable elements as enhancers in mouse embryonic and trophoblast stem cells. *Elife* **8** (2019). https://doi.org:10.7554/eLife.44344

56   Zhang, T. & Zarkower, D. DMRT proteins and coordination of mammalian spermatogenesis. *Stem Cell Res* **24**, 195-202 (2017). https://doi.org:10.1016/j.scr.2017.07.026

57   Sakashita, A. *et al.* Endogenous retroviruses drive species-specific germline transcriptomes in mammals. *Nat Struct Mol Biol* **27**, 967-977 (2020). https://doi.org:10.1038/s41594-020-0487-4

58   Garcia-Moreno, S. A. *et al.* Gonadal supporting cells acquire sex-specific chromatin landscapes during mammalian sex determination. *Dev Biol* **446**, 168-179 (2019). https://doi.org:10.1016/j.ydbio.2018.12.023

59   Baudat, F. *et al.* PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* **327**, 836-840 (2010). https://doi.org:10.1126/science.1183439

60   Otsuka, K., Sakashita, A., Maezawa, S., Schultz, R. M. & Namekawa, S. H. KRAB-zinc-finger proteins regulate endogenous retroviruses to sculpt germline transcriptomes and genome evolution. *bioRxiv* (2023). https://doi.org:10.1101/2023.06.24.546405

61   Yu, H. *et al.* rRNA biogenesis regulates mouse 2C-like state by 3D structure reorganization of peri-nucleolar heterochromatin. *Nat Commun* **12**, 6365 (2021). https://doi.org:10.1038/s41467-021-26576-2

62   Zhu, Y. *et al.* Cell cycle heterogeneity directs spontaneous 2C state entry and exit in mouse embryonic stem cells. *Stem Cell Reports* **16**, 2659-2673 (2021). https://doi.org:10.1016/j.stemcr.2021.09.003

63   Eckersley-Maslin, M. A. *et al.* MERVL/Zscan4 Network Activation Results in Transient Genome-wide DNA Demethylation of mESCs. *Cell Rep* **17**, 179-192 (2016). https://doi.org:10.1016/j.celrep.2016.08.087

64   Xie, S. Q. *et al.* Nucleolar-based Dux repression is essential for embryonic two-cell stage exit. *Genes Dev* **36**, 331-347 (2022). https://doi.org:10.1101/gad.349172.121

65   Percharde, M. *et al.* A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. *Cell* **174**, 391-405 e319 (2018). https://doi.org:10.1016/j.cell.2018.05.043

66   Le, R. *et al.* Dcaf11 activates Zscan4-mediated alternative telomere lengthening in early embryos and embryonic stem cells. *Cell Stem Cell* **28**, 732-747 e739 (2021). https://doi.org:10.1016/j.stem.2020.11.018

67   Zhang, W. *et al.* Zscan4c activates endogenous retrovirus MERVL and cleavage embryo genes. *Nucleic Acids Res* **47**, 8485-8501 (2019). https://doi.org:10.1093/nar/gkz594

68   Eckersley-Maslin, M. A., Alda-Catalinas, C. & Reik, W. Dynamics of the epigenetic landscape during the maternal-to-zygotic transition. *Nat Rev Mol Cell Biol* **19**, 436-450 (2018). https://doi.org:10.1038/s41580-018-0008-z

69   Atchison, F. W., Capel, B. & Means, A. R. Pin1 regulates the timing of mammalian primordial germ cell proliferation. *Development* **130**, 3579-3586 (2003). https://doi.org:10.1242/dev.00584

70   Kagiwada, S., Kurimoto, K., Hirota, T., Yamaji, M. & Saitou, M. Replication-coupled passive DNA demethylation for the erasure of genome imprints in mice. *EMBO J* **32**, 340-353 (2013). https://doi.org:10.1038/emboj.2012.331

71   Aravin, A. A. *et al.* Cytoplasmic compartmentalization of the fetal piRNA pathway in mice. *PLoS Genet* **5**, e1000764 (2009). https://doi.org:10.1371/journal.pgen.1000764

72   Lowe, M. G. *et al.* EED is required for mouse primordial germ cell differentiation in the embryonic gonad. *Dev Cell* **57**, 1482-1495 e1485 (2022). https://doi.org:10.1016/j.devcel.2022.05.012

73   Yokobayashi, S. *et al.* PRC1 coordinates timing of sexual differentiation of female primordial germ cells. *Nature* **495**, 236-240 (2013). https://doi.org:10.1038/nature11918

74   Linher, K. *et al.* An epigenetic mechanism regulates germ cell-specific expression of the porcine Deleted in Azoospermia-Like (DAZL) gene. *Differentiation* **77**, 335-349 (2009). https://doi.org:10.1016/j.diff.2008.08.001

75   Ohno, R. *et al.* A replication-dependent passive mechanism modulates DNA demethylation in mouse primordial germ cells. *Development* **140**, 2892-2903 (2013). https://doi.org:10.1242/dev.093229

76   Ivanov, A. V. *et al.* PHD domain-mediated E3 ligase activity directs intramolecular sumoylation of an adjacent bromodomain required for gene silencing. *Mol Cell* **28**, 823-837 (2007). https://doi.org:10.1016/j.molcel.2007.11.012

77   Li, M., Xu, X., Chang, C. W. & Liu, Y. TRIM28 functions as the SUMO E3 ligase for PCNA in prevention of transcription induced DNA breaks. *Proc Natl Acad Sci U S A* **117**, 23588-23596 (2020). https://doi.org:10.1073/pnas.2004122117

78   Li, M. *et al.* SUMO2 conjugation of PCNA facilitates chromatin remodeling to resolve transcription-replication conflicts. *Nat Commun* **9**, 2706 (2018). https://doi.org:10.1038/s41467-018-05236-y

79   Georges, A. *et al.* SUMOylation of the Forkhead transcription factor FOXL2 promotes its stabilization/activation through transient recruitment to PML bodies. *PLoS One* **6**, e25463 (2011). https://doi.org:10.1371/journal.pone.0025463

80   Rossitto, M. *et al.* TRIM28-dependent SUMOylation protects the adult ovary from activation of the testicular pathway. *Nat Commun* **13**, 4412 (2022). https://doi.org:10.1038/s41467-022-32061-1

81   De Iaco, A. *et al.* DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat Genet* **49**, 941-945 (2017). https://doi.org:10.1038/ng.3858

82   Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57-63 (2012). https://doi.org:10.1038/nature11244

Chapter 5

Conclusions and Future Perspectives

Transposable elements occupy an unusual duality in the germline. On one hand, evolutionarily young TEs which are capable of transposition must be repressed to safeguard against damage to the genomic fidelity of gametes. At the same time, TEs can also act as important genomic regulatory elements, acting as enhancers, boundary elements and 3D genome elements during development and in adult tissues[1-4]. In the germline, TEs are under exceptional selective pressure. In this thesis, I show that epigenetic regulation of transposable elements is an important aspect of germline biology, but not in the rather binary terms in which this balance is often written.

In Chapter 3, we demonstrate that LTR5Hs, a Hominidae-specific ERV, acts as an enhancer during human germline specification using an *in vitro* PGCLC model of hPGC specification. Indeed, we found that LTR5Hs is remodeled to be more accessible in PGCs and that it is likewise bound by PGC factors SOX17, SOX15, NANOG and TFAP2C. In line with these observations, ectopic silencing of LTR5Hs also reduced the efficiency of PGCLC induction. Thus, LTR5Hs is comprised of functional regulatory integrants required for hPGCLC specification. Interestingly, LTR5Hs also acts as an enhancer in naïve PSCs but not in primed hPSCs [3,5]. PGCLCs can be induced from hPSCs cultured in 4i (CHIR99021, PD0325901, SB203580, SP600125) which inhibits GSK-3, MEK, p38-MAPK and JNK and transitions primed hPSCs towards a naïve hPSC state[6]. By using 4i-cultured hPSCs, hPGCLCs could be induced without inducing incipient Mesoderm Like Cell (iMeLCs), which is required in a similar PGCLC induction method that starts with primed hPSCs[7]. Thus, it is possible that culture conditions which push hPSCs towards naïve pluripotency result in the accessibility of loci needed for PGC specification, including LTR5Hs. Additionally, this provides further evidence that, as in the mouse, hPGCs exist in a state of latent pluripotency which has distinct aspects of the naïve program.

In chapter 4 we investigated how control of TEs contributes to proper PGC differentiation, for which we use a mouse *in vivo* model due to the inability of PGC-Like Cells of

either species to be reliably matured and differentiated. First we identified TEs which change in accessibility between E10.5 and E14.5 in control PGCs. Interestingly, we found that these TEs were enriched for stage (either pre-sex determination or differentiated) specific transcription factors motifs, but only minimally enriched in H3K27ac, suggesting that these integrants may not function as enhancers. We likewise found that ERVs which were specifically accessible in either sex under control conditions at E14.5 were not associated with higher expression of neighboring genes, suggesting they were not acting either as alternative promoters or enhancers. This aligns with other studies in the mouse, which have found that many TEs are decorated with enhancer-associated epigenetic marks, such as H3K27ac or H3K4me1/3, but fail to have enhancer activity. For instance, despite enrichment of enhancer-associated marks at RLTR9E and RLTR13D6 elements in mPSCs, silencing of RLTR13D6 elements does not disrupt mPSC gene expression profiles[8-10]. From our findings, this may be true as well in mouse PGCs.

Using a PGC-specific knockout of *Trim28* by *Prdm1*-cre mediated recombination of a floxed *Trim28* allele, we induce TRIM28 knockout prior to PGC differentiation and assessed the impact of TRIM28 loss on PGC differentiation. Using RNA-seq, we found the TRIM28-targeted TEs, especially IAP subfamilies, became derepressed at E12.5 in TRIM28 knockout PGCs. At the same time, we observed a 2C-like transcriptome, including upregulation of *Zscan4*, *Dux* and MERVL. These are all markers of the mouse 2-cell stage, a point in embryonic development during which mice undergo zygotic genome activation[11]. At E12.5, the PGC epigenome is highly demethylated and, unlike in mPSCs, *Dux* is accessible[12,13]. Thus, we speculate that mPGCs at E12.5 are poised to enter a 2C-like state, and loss of TRIM28 is sufficient to induce a 2C-like transcriptome. A possible mechanism explaining this is TRIM28-mediated repression of a putative distal enhancer of the *Zscan4* cluster[14]. Thus, loss of TRIM28 may have activated this enhancer and, in turn, activators of the 2C-state. Given that ectopic accessibility of *Dux* in mPSCs is sufficient to drive a 2C-like state, and 2C release *in vivo* has been shown in part to be driven by TRIM28-mediated repression of the *Dux* locus, it is unclear why *Dux* is normally

accessible in PGCs[12,13]. This is an interesting area of investigation, as it could act as a sensor

for misregulation of the PGC program, such as altered cell cycle progression or TE

derepression. It could also be the result of changes to the nucleolus or nucleolar membrane, as

Dux is tethered to the nucleolus and perturbation of nucleolar function by RNA Pol I inhibition in

mPSCs derepresses *Dux*[15]. In summary we show that mPGCs have the capacity to enter into a

2C-like transcriptome, which has been reported in mPSC and hPSCs (where it is 8C-like), but

not PGCs. Thus, latent pluripotency is compatible with entry into a 2C-transcriptome.

We find that like knockout of *Dnmt1*, *Eed*, *Ezh2* and *Tet1*, no phenotype manifests in

TRIM28 knockout PGCs until E12.5, the time at which PGCs begin to express *Dazl* and become

committed to the germline. These studies have collectively identified E12.5 as an inflection point

at which PCGs are no longer able to tolerate misregulation of their epigenome. Interestingly,

how PGCs respond to each epigenetic insult is different. While loss of PRC1/PRC2 and *Dnmt1*

result in precocious meiosis, loss *Tet1* reduces adult fertility and delays entry into meiosis[16-20].

Thus, PGC determination and subsequently PGC differentiation is reliant on and interconnected

network of epigenetic controls. I also show in this thesis that TRIM28 may contribute to the

regulation of *Dazl* directly, as it is highly enriched in the *Dazl* gene body. How *Trim28* regulates

Dazl, and if it is direct or through an intermediate, remains unknown. Interestingly, *Trim28*

knockout results in a defect with varying severities across sexes, with ovarian PGCs having a

more severe phenotype than testicular PGCs. Therefore, it is possible that there are also sex-

specific factors which link *Trim28* to regulation of *Dazl*.

Finally, we found that TRIM28 knockout PGCs fail to differentiate. While some ovarian

PGCs do enter meiosis, they fail to progress through prophase I of meiosis I regularly. In a rare

adult sample at 6.8 weeks, we found that TRIM28 knockout oocytes had mislocalized *Dppa3*

(*Stella*) and lacked primordial follicles. Thus, *Trim28* is required to form an ovarian reserve. The

mislocalization of *Dppa3* is of particular interest. *Trim28* is necessary for imprint acquisition in

oocytes and protects imprints in the early zygote[21,22]. *Dppa3* then protects these imprints from

Tet3-mediated demethylation[23,24]. It is possible that *Trim28* and *Dppa3* have coordinated activities in the oocyte, either directly or indirectly.

Unlike ovarian PGCs where some enter meiosis, testicular PGCs fail to differentiate altogether. In line with an increased dependence on *Trim28*, testicular germ cells at E14.5 had many more ERVs which became accessible following TRIM28 knockout compared to ovarian germ cells. It is likely that *Trim28* is important to establish the proper transcriptome in testicular germ cells, given genes near derepressed TEs were upregulated. This requirement for TRIM28 persists throughout adulthood, as loss of *Trim28* from adult spermatogonial stem cells results in symmetrical division of SSCs, depleting the pool of SSCs and thus causing premature infertility[25].

In sum, my thesis work demonstrates that transposable elements and their regulators are critical components of cell differentiation and can contribute to transcriptional fidelity and fate restriction. Here, I demonstrate this capacity in the germline. Given the capacity of transposable elements to work both as cis-regulatory elements which positively contribute to tissue development and health as well as their capacity to derail transcriptional programs, TEs and their regulators deserve careful consideration in both developmental and disease contexts.

References:

1   Bourc'his, D. & Bestor, T. H. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431**, 96-99 (2004). https://doi.org:10.1038/nature02886

2   Zhang, Y. *et al.* Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat Genet* **51**, 1380-1388 (2019). https://doi.org:10.1038/s41588-019-0479-7

3   Pontis, J. *et al.* Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. *Cell Stem Cell* **24**, 724-735 e725 (2019). https://doi.org:10.1016/j.stem.2019.03.012

4   Pontis, J. *et al.* Primate-specific transposable elements shape transcriptional networks during human development. *Nat Commun* **13**, 7178 (2022). https://doi.org:10.1038/s41467-022-34800-w

5   Theunissen, T. W. *et al.* Molecular Criteria for Defining the Naive Human Pluripotent State. *Cell Stem Cell* **19**, 502-515 (2016). https://doi.org:10.1016/j.stem.2016.06.011

6   Irie, N. *et al.* SOX17 is a critical specifier of human primordial germ cell fate. *Cell* **160**, 253-268 (2015). https://doi.org:10.1016/j.cell.2014.12.013

7   Sasaki, K. *et al.* Robust In Vitro Induction of Human Germ Cell Fate from Pluripotent Stem Cells. *Cell Stem Cell* **17**, 178-194 (2015). https://doi.org:10.1016/j.stem.2015.06.014

8   He, J. *et al.* Transposable elements are regulated by context-specific patterns of chromatin marks in mouse embryonic stem cells. *Nat Commun* **10**, 34 (2019). https://doi.org:10.1038/s41467-018-08006-y

9   Wolf, G. *et al.* KRAB-zinc finger protein gene expansion in response to active retrotransposons in the murine lineage. *Elife* **9** (2020). https://doi.org:10.7554/eLife.56337

10   Todd, C. D., Deniz, O., Taylor, D. & Branco, M. R. Functional evaluation of transposable elements as enhancers in mouse embryonic and trophoblast stem cells. *Elife* **8** (2019). https://doi.org:10.7554/eLife.44344

11   De Iaco, A. *et al.* DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat Genet* **49**, 941-945 (2017). https://doi.org:10.1038/ng.3858

12   Sun, X. *et al.* POGZ suppresses 2C transcriptional program and retrotransposable elements. *Cell Rep* **42**, 112867 (2023). https://doi.org:10.1016/j.celrep.2023.112867

13   Xie, S. Q. *et al.* Nucleolar-based Dux repression is essential for embryonic two-cell stage exit. *Genes Dev* **36**, 331-347 (2022). https://doi.org:10.1101/gad.349172.121

14   Le, R. *et al.* Dcaf11 activates Zscan4-mediated alternative telomere lengthening in early embryos and embryonic stem cells. *Cell Stem Cell* **28**, 732-747 e739 (2021). https://doi.org:10.1016/j.stem.2020.11.018

15   Yu, H. *et al.* rRNA biogenesis regulates mouse 2C-like state by 3D structure reorganization of peri-nucleolar heterochromatin. *Nat Commun* **12**, 6365 (2021). https://doi.org:10.1038/s41467-021-26576-2

16   Hargan-Calvopina, J. *et al.* Stage-Specific Demethylation in Primordial Germ Cells Safeguards against Precocious Differentiation. *Dev Cell* **39**, 75-86 (2016). https://doi.org:10.1016/j.devcel.2016.07.019

17   Lowe, M. G. *et al.* EED is required for mouse primordial germ cell differentiation in the embryonic gonad. *Dev Cell* **57**, 1482-1495 e1485 (2022). https://doi.org:10.1016/j.devcel.2022.05.012

18   Hill, P. W. S. *et al.* Epigenetic reprogramming enables the transition from primordial germ cell to gonocyte. *Nature* **555**, 392-396 (2018). https://doi.org:10.1038/nature25964

19   Huang, T. C. *et al.* Sex-specific chromatin remodelling safeguards transcription in germ cells. *Nature* **600**, 737-742 (2021). https://doi.org:10.1038/s41586-021-04208-5

20   Nicholls, P. K. *et al.* Mammalian germ cells are determined after PGC colonization of the nascent gonad. *Proc Natl Acad Sci U S A* **116**, 25677-25687 (2019). https://doi.org:10.1073/pnas.1910733116

21   Quenneville, S. *et al.* In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Mol Cell* **44**, 361-372 (2011). https://doi.org:10.1016/j.molcel.2011.08.032

22   Alexander, K. A., Wang, X., Shibata, M., Clark, A. G. & Garcia-Garcia, M. J. TRIM28 Controls Genomic Imprinting through Distinct Mechanisms during and after Early Genome-wide Reprogramming. *Cell Rep* **13**, 1194-1205 (2015). https://doi.org:10.1016/j.celrep.2015.09.078

23   Bian, C. & Yu, X. PGC7 suppresses TET3 for protecting DNA methylation. *Nucleic Acids Res* **42**, 2893-2905 (2014). https://doi.org:10.1093/nar/gkt1261

24   Nakamura, T. *et al.* PGC7/Stella protects against DNA demethylation in early embryogenesis. *Nat Cell Biol* **9**, 64-71 (2007). https://doi.org:10.1038/ncb1519

25   Tan, J. H. L., Wollmann, H., van Pelt, A. M. M., Kaldis, P. & Messerschmidt, D. M. Infertility-Causing Haploinsufficiency Reveals TRIM28/KAP1 Requirement in Spermatogonia. *Stem Cell Reports* **14**, 818-827 (2020). https://doi.org:10.1016/j.stemcr.2020.03.013