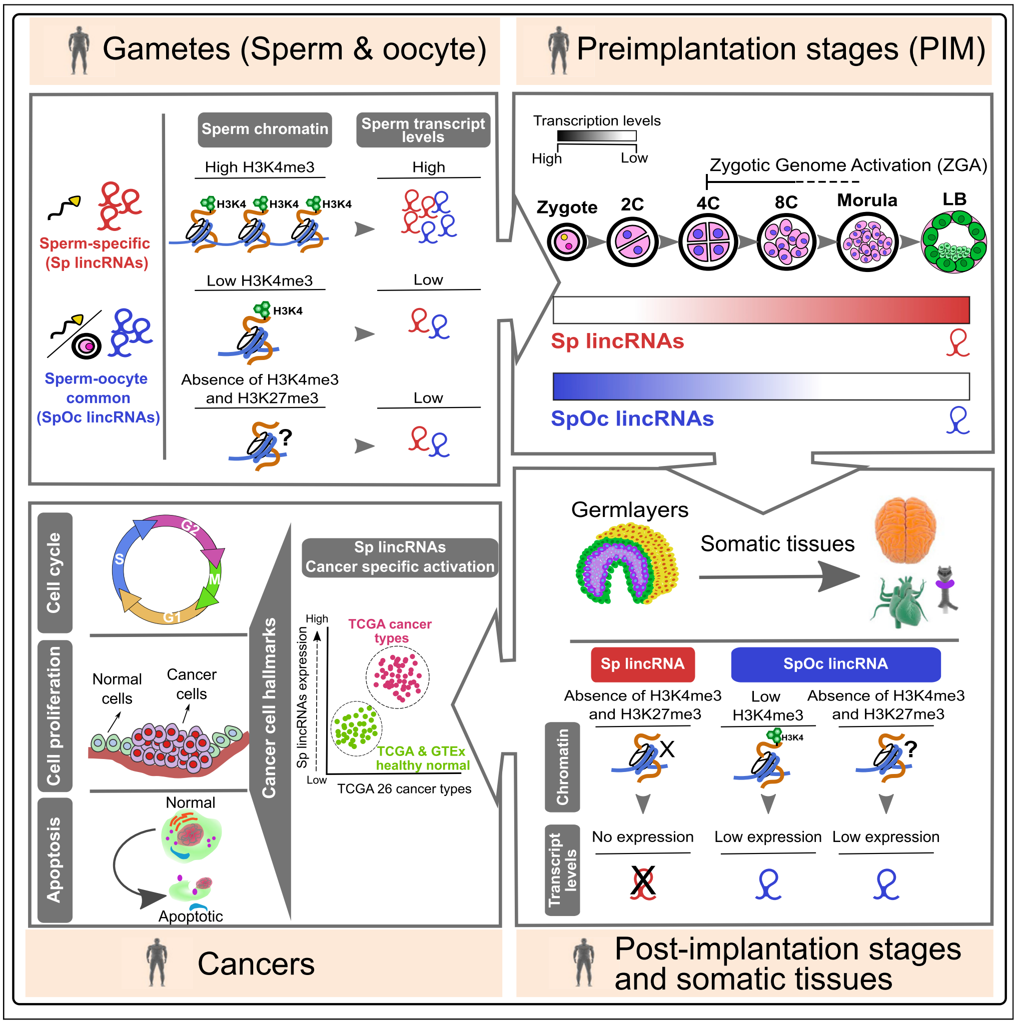
iScience

CellPress
OPEN ACCESS

Article

Sperm Originated Chromatin Imprints and LincRNAs in Organismal Development and Cancer



Santhilal Subhash, Meena Kanduri, Chandrasekhar Kanduri

kanduri.chandrasekhar@gu.se

HIGHLIGHTS

Sp-lincRNAs carry distinct chromatin structures correlating with transcript levels

Sp-lincRNAs are active during ZGA in preimplantation developmental stages

SpOc-lincRNAs are active in pre-ZGA and decline at the onset of ZGA

Sp-lincRNAs are silent post implantation but show aberrant cancer-specific activation

Subhash et al., iScience 23, 101165
June 26, 2020 © 2020 The Authors.
https://doi.org/10.1016/j.isci.2020.101165

iScience

CellPress OPEN ACCESS

Article

Sperm Originated Chromatin Imprints and LincRNAs in Organismal Development and Cancer

Santhilal Subhash,[1] Meena Kanduri,[2] and Chandrasekhar Kanduri[1,3,*]

SUMMARY

Importance of sperm-derived transcripts and chromatin imprints in organismal development is poorly investigated. Here using an integrative approach, we show that human sperm transcripts are equally important as oocyte. Sperm-specific and sperm-oocyte common transcripts carry distinct chromatin structures at their promoters correlating with corresponding transcript levels in sperm. Interestingly, sperm-specific H3K4me3 patterns at the lincRNA promoters are not maintained in the germ layers and somatic tissues. However, bivalent chromatin at the sperm-specific protein-coding gene promoters is maintained throughout the development. Sperm-specific transcripts reach their peak expression during zygotic genome activation, whereas sperm-oocyte common transcripts are present during early preimplantation development but decline at the onset of zygotic genome activation. Additionally, there is an inverse correlation between sperm-specific and sperm-oocyte lincRNAs throughout the development. Sperm-lincRNAs also show aberrant activation in tumors. Overall, our observations indicate that sperm transcripts carrying chromatin imprints may play an important role in human development and cancer.

INTRODUCTION

Maternal or oocyte-derived transcripts are known to regulate maternal to zygotic transition (MZT), which involves zygotic genome activation (ZGA) and maternal transcript degradation. These are the two important events necessary for subsequent preimplantation embryonic development (Xia et al., 2019). On the other hand, sperm with its compacted nucleus and minimal cytoplasm has long been considered transcriptionally incompetent and a passive vehicle that transmits merely paternal genome to the oocyte during fertilization. However, with the advent of high-throughput technologies, it is now clear not only that sperm contains RNA but also that its genome is highly structured and organized into distinct chromatin states dictated by positioned nucleosomes alongside protamine-enriched compact chromatin regions (Brykczynska et al., 2010; Hammoud et al., 2009; Paradowska et al., 2012). These observations suggest that not only oocyte but sperm also may take part in dictating genome organization and gene expression during MZT. Supporting this notion, overexpression of histone demethylation machinery (KDM1A/LSD1), which maintains H3K4 methylation levels, during spermatogenesis, had developmental defects that lasted for two subsequent generations (Siklenka et al., 2015). Thus, genetic and epigenetic information from both sperm and oocytes seem to have equal stakes in oocyte to embryo transition and also intergenerational transfer of acquired traits. Despite our increased understanding of sperm-dependent developmental regulation, how sperm-marked epigenetic states and sperm-derived transcripts contribute to gene expression at the embryonic and the adult stages of mammalian development remains unknown. Hence there is a need for an extensive investigation of transcriptionally competent chromatin regions across the sperm genome and this would immensely help in understanding their contribution to mammalian development. In this study, we attempt to integrate 17,705 human and mouse high-throughput sequencing samples comprising transcriptome (RNA sequencing [RNA-seq]) and histone chromatin profile datasets (H3K4me3 and H3K27me3 chromatin immunoprecipitation sequencing [ChIP-seq]) to explore the role of sperm-encoded information in development and disease regulation (Table S1). This study, in particular, emphasizes long noncoding RNAs (lncRNAs) as they have been shown as important lineage-specific developmental regulators, having cell-type- and stage-dependent functions (Cabili et al., 2011; Guttman et al., 2011; Zhang et al., 2014). Moreover, on lncRNAs, there exists only sporadic information regarding their significance in organism development. Therefore, there is a need for an integrative study to realize

[1] Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg 40530, Sweden

[2] Department of Clinical Chemistry and Transfusion Medicine, Institute of Biomedicine, Sahlgrenska University Hospital 413 45, Gothenburg, Sweden

[3] Lead Contact

*Correspondence: kanduri.chandrasekhar@gu.se

https://doi.org/10.1016/j.isci.2020.101165

the importance of these transcripts, alongside protein-coding transcripts, through tracing their expression throughout the mammalian development.

Cancer testis antigens (CTAs) are a group of testis-specific proteins that show predominant expression on cancer cells. This observation suggests a functional connection between testis-specific expression and cancer. Majority of the CTA genes are preferentially expressed in testicular cell types such as spermatogonia or spermatocytes. Hence, we were particularly interested in characterizing whether sperm-derived RNAs show any preferential expression in different tumor types, which would in turn pave the way for understanding the contribution of sperm-derived transcripts in tumor development and progression (Simpson et al., 2005).

RESULTS

Transcriptome-wide Identification of Sperm- and Oocyte-Associated Transcripts

At the outset, we analyzed the transcriptome profiles of sperm, oocyte, and preimplantation embryo stages (two cell to late blastocyst) in human and mouse (Fan et al., 2015; Hammoud et al., 2014; Yan et al., 2013; Zhang et al., 2017) (Figure 1A). We included analysis on, in addition to protein-coding genes, long intergenic noncoding RNAs (lincRNAs) and other classes of noncoding transcripts (Other-ncRNAs), excluding lincRNAs, from the Ensembl transcript annotation (Figures S1A and S1D). For further interpretation of the data, we considered only intergenic transcripts or lincRNAs to avoid unnecessary noise from sequencing reads due to overlapping transcripts from the majority of non-stranded transcriptome datasets used in this study. This approach will increase the reliability and reproducibility of the results.

Our analysis show that sperm and oocyte, from both mouse and human, harbor comparable number of transcripts, and in particular, mouse harbors nearly twice as many transcripts in sperm compared with oocyte (Figures 1B and 1C). There were 653 lincRNA and 1,906 protein-coding gene (PCG) transcripts that were specifically present in human sperm, and hence we termed them as Sp-lincRNAs and Sp-PCGs, respectively. We found 531 lincRNAs and 741 PCGs to be human oocyte-specific (Oc-lincRNAs and Oc-PCGs), whereas 275 lincRNAs and 8,468 PCG transcripts were found to be expressed commonly in human sperm and oocyte (SpOc-lincRNAs and SpOc-PCGs). Transcripts that are inconsistently expressed between donors, within a gamete, were considered as miscellaneous and not considered for further analysis (Table S2) (Figure 1B). Like in humans, we categorized lincRNAs and PCGs from mouse sperm and oocyte expression data in a similar manner. In mouse, 338 lincRNA and 3,251 PCG transcripts were sperm specific (Sp-lincRNAs and Sp-PCGs); 97 lincRNAs and 194 PCG transcripts were oocyte specific (Oc-lincRNAs and Oc-PCGs). There were 23 lincRNA and 6,206 PCG transcripts commonly expressed between sperm and oocyte in mouse (SpOc-lincRNAs and SpOc-PCGs) (Table S2) (Figure 1C). Expression levels of SpOc-lincRNAs and SpOc-PCGs were higher in oocyte compared with sperm both in human and mouse (Figures 1D and 1E). Overall, there are a greater number of Sp-lincRNAs and Sp-PCGs compared with Oc-lincRNAs and Oc-PCGs in both mouse and human, which predicts a greater significance for the sperm-derived transcripts in mammalian development.

Human Sperm Chromatin Is Well Structured and Correlates with Transcriptional Status

Previously, it has been shown that paternal traits acquired in response to external stimuli can be inherited by offspring which, in particular, emphasizes a functional role for sperm chromatin structures in transgenerational inheritance (Gapp et al., 2018; Zhang et al., 2018). Consistent with the latter notion, based on ChIP-seq data, it was predicted that nearly 4% of the sperm genome is occupied by nucleosomes and that a significant portion of the retained nucleosomes were modified, specifically at noncoding RNA loci, imprinted genes, and developmental regulators (Hammoud et al., 2009; Jung et al., 2017). Considering these observations, we performed chromatin structure analyses at the promoter regions of Sp and SpOc transcripts to check whether they carry any distinct chromatin signatures and how they are maintained during the rest of the organismal development. Chromatin profiles of active (H3K4me3 enriched) and inactive (H3K27me3 enriched) histone marks at the Sp and SpOc gene promoters revealed several clusters based on the patterns of histone modifications (Hammoud et al., 2009; Jung et al., 2017) (Table S2). Analyses of human Sp-lincRNAs and SpOc-lincRNA promoters revealed three clusters: (1) high H3K4me3 (high-K4), (2) low H3K4me3 (low-K4), and (3) promoters lacking both H3K4me3 and H3K27me3 ($K4^-K27^-$) (Figure 2A). In Sp-PCG and SpOc-PCG promoters, we found chromatin clusters with high-K4, low-K4, bivalent chromatin (enriched with both H3K27me3 and H3K4me3 marks), and $K4^-K27^-$ (Figure 2B). In mouse, however, Sp-lincRNAs contain only a single cluster with high H3K4me3 at their promoters (Figure 2C). In

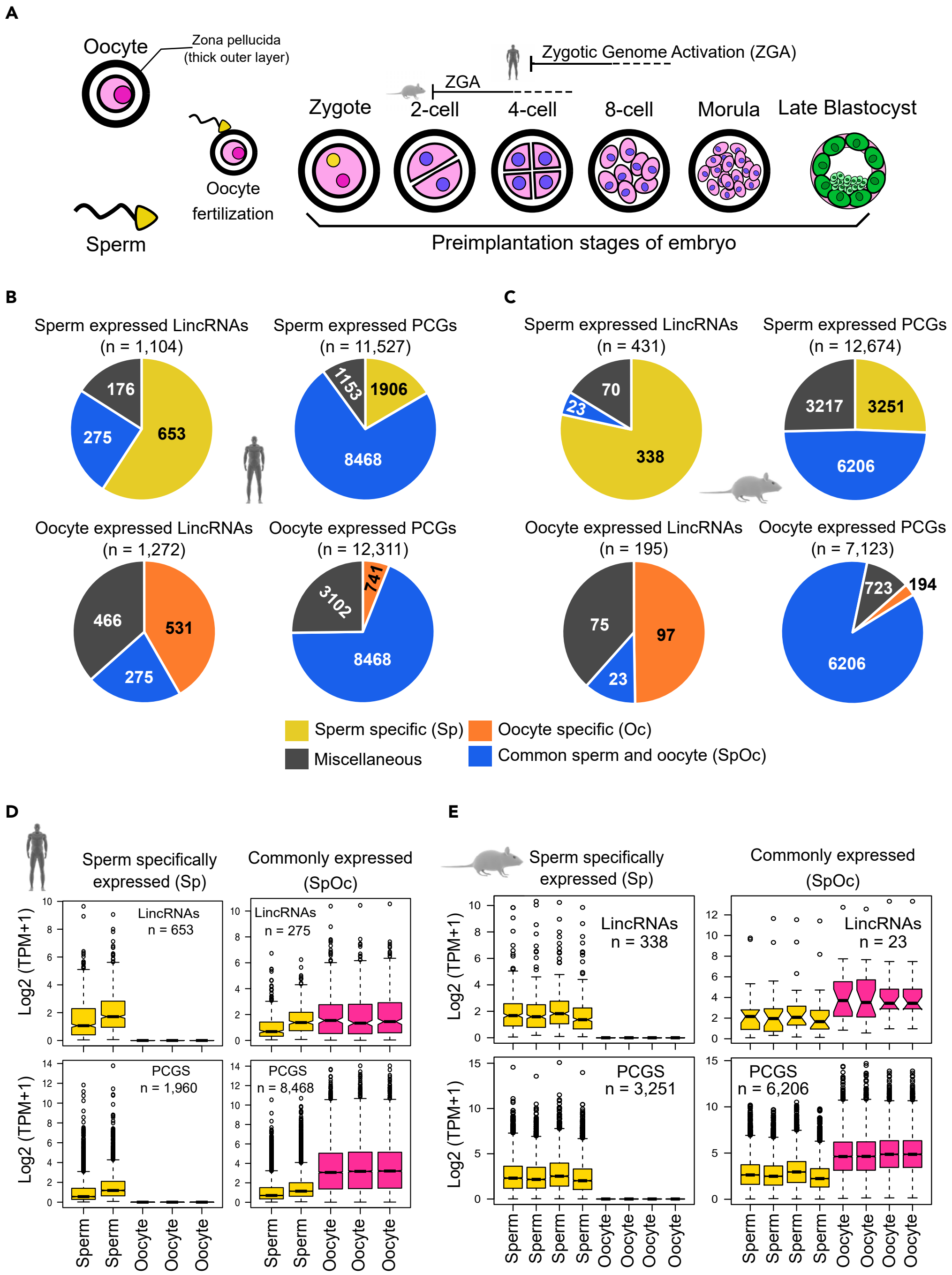


Figure 1. Comparative Transcriptome Analysis Reveals Similar Number of Transcripts Expressed in Sperm and Oocyte

(A) A schematic of male and female germ cells and preimplantation embryo stages used for comparative transcriptomic analyses from both human and mouse. See also Table S1.

(B and C) Venn diagrams showing the number of sperm and oocyte expressed long intergenic noncoding RNAs (lincRNAs) and protein-coding genes (PCGs) from human (B) and mouse (C) represented in three categories: sperm or oocyte (Oc) specific, commonly expressed in sperm and oocyte (SpOc), and Miscellaneous (inconsistently expressed between replicates samples from gametes). See also Table S2.

Figure 1. *Continued*

(D and E) Boxplots from human (D) and mouse (E) showing the expression of sperm-specific and sperm-oocyte expressed (SpOc) lincRNAs and PCGs in sperm and oocyte. Box plots represent low expression range (lower whiskers), higher expression range, (upper whisker), median, inter quartile range (IQR), and the extreme expression values.
See also Figure S1 and Table S3.

the case of SpOc-lincRNAs, there were only a few candidates to be used for clustering (Figure 1C and Table S2). Mouse Sp-PCGs and SpOc-PCGs showed chromatin clusters having high-K4, low-K4, bivalent with high-K27, and bivalent with low-K27 marks (Figure 2D). We next compared the transcript levels of human and mouse Sp and SpOc transcripts with the chromatin modification patterns at their promoters. Interestingly, the transcript levels of human Sp (Sp-lincRNAs and Sp-PCGs) and SpOc (Sp-lincRNAs and SpOc-PCGs) transcripts correlated with the levels of H3K4me3 over their promoters (Figures 2A, 2B, 2E, and 2F). A higher transcript level was seen for the promoters with high-K4 compared with low-K4, bivalent, and K4^{-}K27^{-} promoters. Interestingly, however, human SpOc-lincRNAs did not show cluster-specific transcript abundance in oocyte as seen in human sperm confirming that these chromatin profiles and transcript abundance patterns are sperm specific (Figures 2A and 2E). In the case of mouse, there was no correlation between chromatin patterns and the levels of sperm-specific (Sp-lincRNAs and Sp-PCGs) and commonly expressed transcripts (SpOc-lincRNAs and SpOc-PCGs) (Figures 2C, 2D, 2G, and 2H) (Tables S2 and S3). These observations indicate that there is a correlation between H3K4me3 enrichment at the promoters and the levels of their encoded transcripts in the human sperm, but such correlation was not evident in the mouse sperm. Similar analysis was performed on other classes of noncoding RNAs (Other-ncRNAs) by excluding lincRNAs and PCGs, and we did not find any correlation between chromatin patterns at their promoters and the encoded transcripts (Figures S1B–S1D). Thus, for further analysis, we have considered only lincRNAs to avoid unnecessary noise in the data due to ambiguously assigned reads to the overlapping transcripts from the majority of non-stranded transcriptome datasets used in this study.

Transcripts from Sperm-Derived Chromatin Clusters Have Distinct Biological Functions in Human

Intriguingly, PCGs from individual sperm-derived chromatin clusters fall into distinct biological functions in human. PCG cluster with high-K4 levels showed functions related to spermatogenesis such as flagellated sperm motility, sperm chromatin condensation, DNA packaging, and spermatid nucleus differentiation. Considering the enrichment of spermatogenesis-related functions in high-K4 group PCGs, we expect that lincRNAs from high-K4 clusters may play an important role in spermatogenesis regulation (Figure 3A). The cluster of Sp-PCGs having bivalent marks was predicted to be involved in biological functions such as organismal development, extracellular matrix organization, skeletal muscle development, VEGFR signaling, animal organ morphogenesis, and dorsal spinal cord development (Figure 3A). This observation corroborates the previous suggestion that bivalent domains are functionally linked to lineage commitment (Maezawa et al., 2018; Voigt et al., 2013). PCGs cluster devoid of histone marks (K4^{-}K27^{-}) showed immune defense-related functions such as immune response, neutrophil chemotaxis, and defense response (Figure 3A). SpOc-PCGs with high-K4 and low-K4 did not show distinct functions (Figure 3B).

Mouse PCG promoters having bivalent chromatin were also enriched in developmental processes as seen in humans, whereas the other chromatin clusters did not reveal any distinct functional profiles (Figures 3C and 3D). Additionally, this bivalent chromatin cluster, unlike in humans, were also enriched with spermatogenesis-related functions. These observations collectively suggest that the bivalent PCG groups from both mouse and human sperm may play an important role in the post-implantation stages of embryos and during the lineage commitments.

Sperm Transcripts Show Temporal Expression during Preimplantation Development

Since temporal and stage-specific gene expression patterns are strongly linked to function, we investigated whether Sp and SpOc transcripts possess temporal expression during preimplantation stages of development. We found that Sp-lincRNA and Sp-PCG transcripts were present at a very low level during early preimplantation stages but get activated between four- and eight-cell stages coinciding with ZGA (Figure 4A). SpOc transcripts showed similar expression patterns throughout the preimplantation development except SpOc lincRNAs, which showed gradual decrease in their levels from four cell stage coinciding with the commencing of ZGA (Figure 4B). Interestingly, Sp transcripts show exclusive stage-specific expression during preimplantation development between the four-cell stage and late

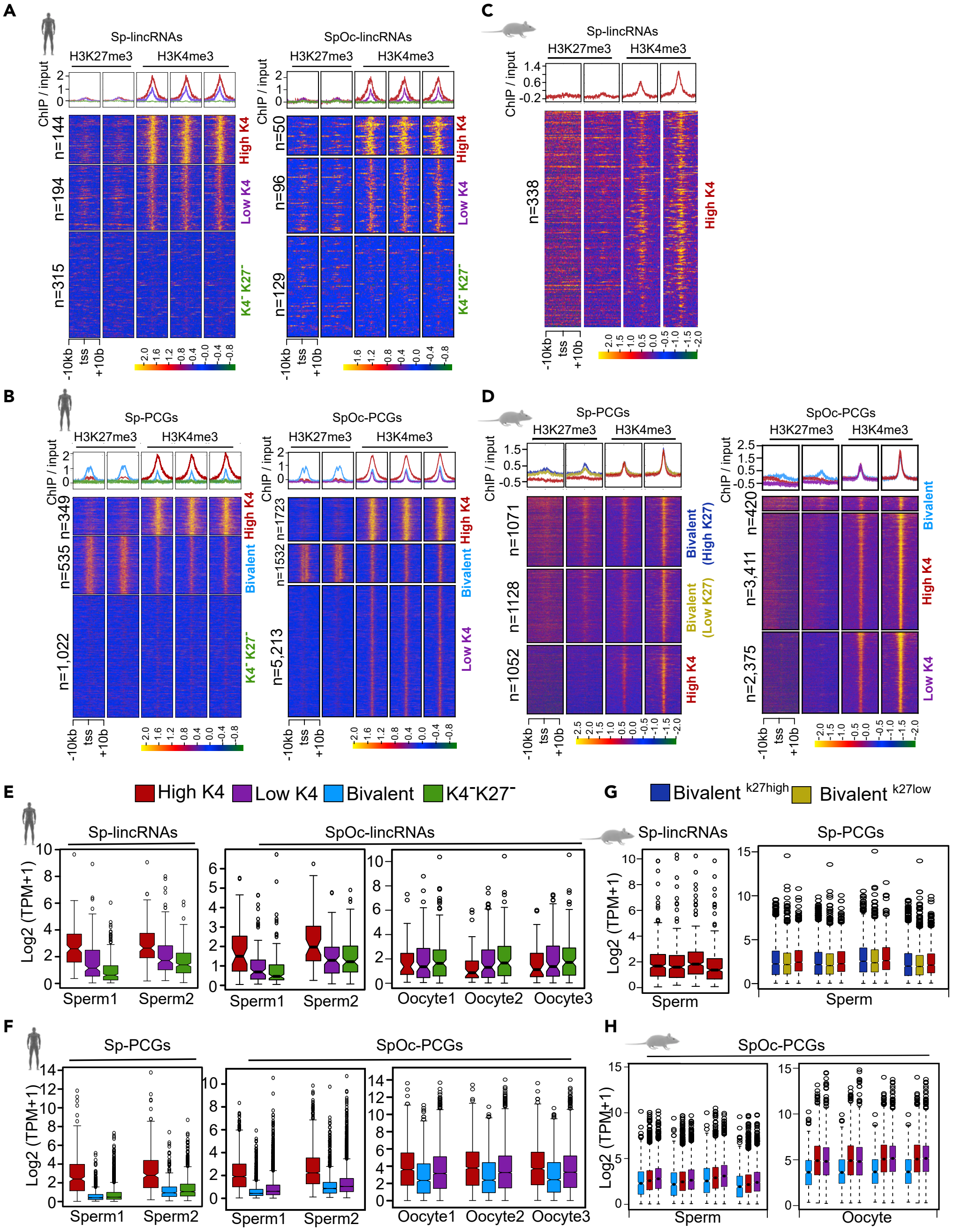


Figure 2. Sperm-Derived Chromatin Clusters Show Variable Expression Levels in Sperm

(A and B) Based on H3K27me3 and H3K4me3 enrichment, human Sp and SpOc lincRNA (A) and PCG (B) promoters (extended ±10 kb from transcription start site, TSS) were categorized into three optimal sperm-derived chromatin clusters. See also Table S2.

(C) Sp-lincRNA promoters (extended ±10 kb from transcription start site, TSS) from mouse show only high-K4 based on H3K27me3 and H3K4me3 enrichment. See also Table S2.

(D) Sp and SpOc-PCG promoters (extended ±10 kb from transcription start site, TSS) from mouse show three sperm-derived chromatin clusters based on H3K27me3 and H3K4me3 enrichment. See also Table S2.

(E and F) Expression status of Sp and SpOc-lincRNAs (E) and PCGs (F) from sperm-derived chromatin clusters in sperm and oocyte RNA-seq samples.

Figure 2. *Continued*

(G and H) Expression status of Sp lincRNAs and PCGs (G) and SpOc PCGs (H) from sperm-derived chromatin clusters in sperm and oocyte samples. Heatmaps showing sperm-derived chromatin clusters were generated using k-means clustering. Box plots represent low expression range (lower whiskers), higher expression range (upper whisker), median, inter quartile range (IQR), and the extreme expression values.
See also Figure S1 and Table S3.

blastocyst embryos, whereas SpOc transcripts show predominant expression between zygote and four-cell-stage embryos (Figures 4C and 4D). In particular, the expression of SpOc-lincRNAs, compared with SpOc-PCGs, is highly restricted to early preimplantation stages (Figures 4C and 4D). Exclusive stage-specific expression of Sp-lincRNAs suggests that they may play a critical role in preimplantation development.

We then investigated the temporal expression patterns of Sp and SpOc transcripts from the sperm-derived chromatin clusters during preimplantation development. Of note, Sp transcripts (Sp-lincRNAs and Sp-PCGs) from sperm-derived chromatin clusters show highly temporal expression patterns with no expression in early preimplantation-stage embryos and specific activation between four- and eight-cell stages of embryos (Figure S2A). Like in humans, in mouse, Sp-PCGs from sperm-derived chromatin clusters, which showed no expression during one- and two-cell early preimplantation embryos, become activated between two- and four-cell stages, coinciding with ZGA in mouse (Figure S2B).

Human SpOc transcripts from sperm-derived chromatin clusters showed a marginal decrease in expression between two- and four-cell stages. However, a marked decrease in expression was seen for transcripts from K4^{-}K27^{-} and low-K4 sperm-derived chromatin clusters from four-cell stage onward. More importantly, SpOc PCGs from low-K4 group become super activated during preimplantation development (Figure S2A), whereas in mouse, Sp and SpOc transcripts from most of the sperm-derived chromatin clusters show opposing expression patterns across preimplantation stages (Figure S2B). Collectively, in mouse and human, sperm-specific transcripts from sperm-histone-derived clusters were present at very low level during early preimplantation embryos but get activated during ZGA, whereas SpOc transcripts, from sperm-histone-derived clusters (except the low-k4 cluster from mouse), remain active during early preimplantation development before getting inactivated during ZGA (Figures S2A and S2B). It is well known that maternal transcript and proteins (maternal detritus) are removed prior to the onset of ZGA (Hamm and Harrison, 2018). However, our data show that not just maternal (Oc) transcripts but also paternal transcripts were removed prior to ZGA, indicating involvement of both maternal and paternal or sperm transcript degradation (SpOc-TD) during early preimplantation development (Figures 4A–4D). Thus, our data open up investigations to explore further the sperm-derived transcripts in preimplantation development.

Specific Loss of High-K4 and Low-K4 Chromatin Structures from the Sp-lincRNA Promoters in Post-implantation Development

During the process of gastrulation, the blastula begins to differentiate into specialized cell types giving rise to three distinct germ layers, namely, ectoderm (outer layer), mesoderm (middle layer), and endoderm (innermost layer) (Figure 5A). Each individual germ layer consists of multi-potent lineage-specific stem cells that can differentiate into different tissues types (Figure 5A). Therefore, we extended our analysis by investigating chromatin structure and expression patterns of Sp and SpOc transcripts in three germ layers and their derived somatic tissues (brain from ectoderm, thyroid from endoderm, and heart from mesoderm) (Chu et al., 2016; Davis et al., 2018; Friedman et al., 2018; Locke et al., 2015; Loh et al., 2014; Rada-Iglesias et al., 2011; Yan et al., 2013). We looked into the chromatin and expression patterns of Sp and SpOc transcripts from individual sperm-derived chromatin clusters during the post-implantation stages of the embryo and mature tissues. Strikingly, sperm-derived high-K4 and low-K4 chromatin structures from Sp-lincRNA promoters were lost in all the three germ layers and also the germ-layer-derived somatic tissues. Interestingly, however, the bivalent domains from Sp-PCG promoters maintained their bivalency (H3K4me3+H3K27me3) at their promoters in all the three germ layers and somatic tissues (Figures 5B, 5C, and S3A). Like high-K4 Sp-lincRNA promoters, high-K4 Sp-PCG promoters also lost their sperm-derived H3K4me3 levels in the germ layers and somatic tissues. Sperm-derived K4^{-}K27^{-} clusters of Sp-lincRNAs and Sp-PCG transcripts maintained their K4^{-}K27^{-} chromatin structures in the germ layers, but in somatic tissues there was a slight increase in H3K4me3 levels in Sp-PCG promoters (Figures 5B, 5C, and S3A).

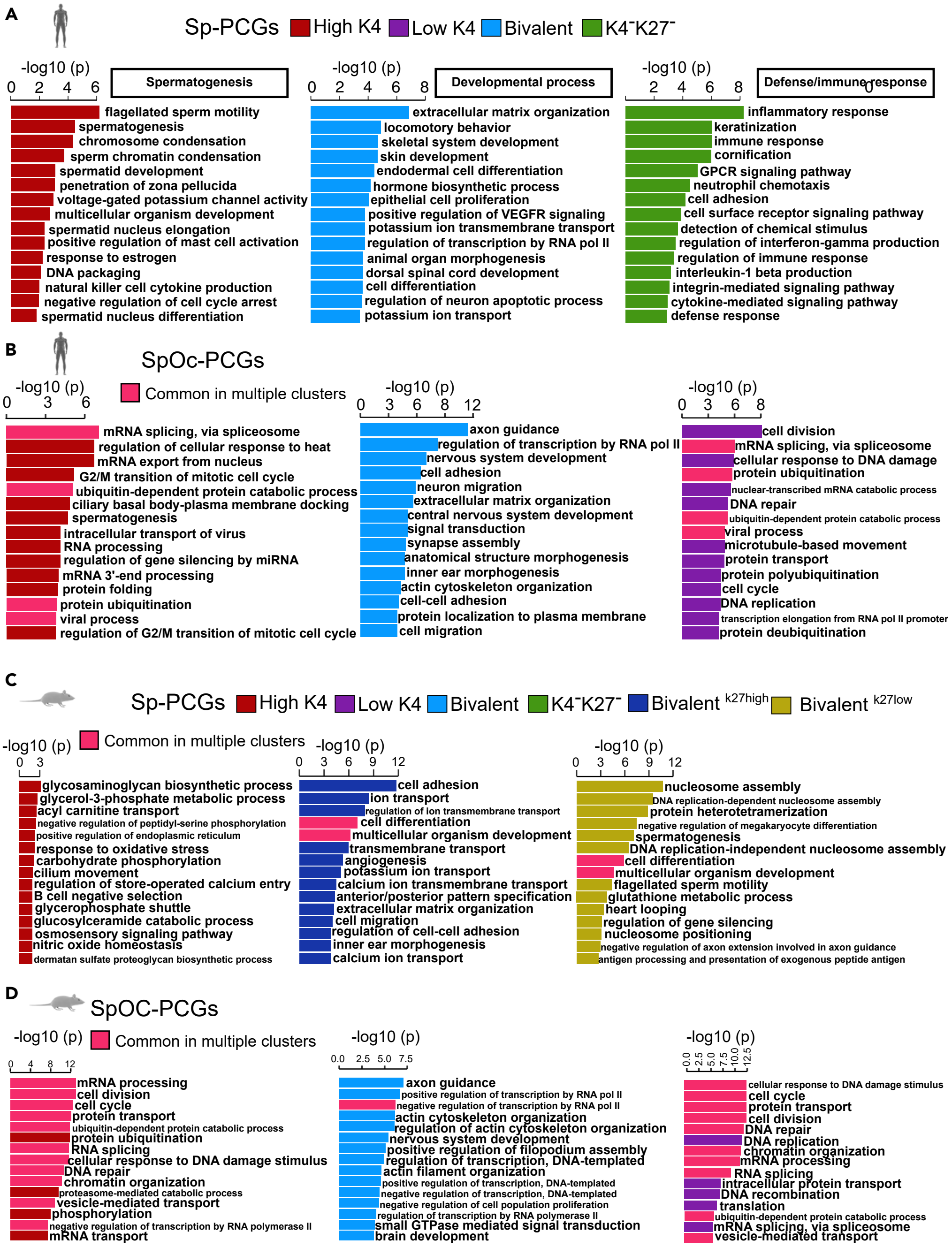


Figure 3. Sperm-Derived Chromatin Clusters Show Distinct Functional Profiles

(A and B) Sperm-derived chromatin clusters of Sp-PCGs (A) and SpOc-PCGs (B) from human with enriched biological functions derived from gene ontology terms ranked using GeneSCF.

(C and D) Sperm-derived chromatin clusters of Sp-PCGs (C) and SpOc-PCGs (D) from mouse with enriched biological functions derived from gene ontology terms ranked using GeneSCF. All gene ontology terms from the bar graphs were selected with enrichment p value < 0.05.

High-K4 and low-K4 sperm-derived chromatin clusters from SpOc-lincRNAs showed low levels of H3K4me3 at their promoters, whereas $K4^{-}K27^{-}$ clusters maintain the same chromatin structure at their promoters in all the three germ layers and somatic tissues. Interestingly, sperm-derived chromatin clusters from SpOc-PCG promoters showed higher enrichment of H3K4me3 in all the three germ layers and their derived somatic

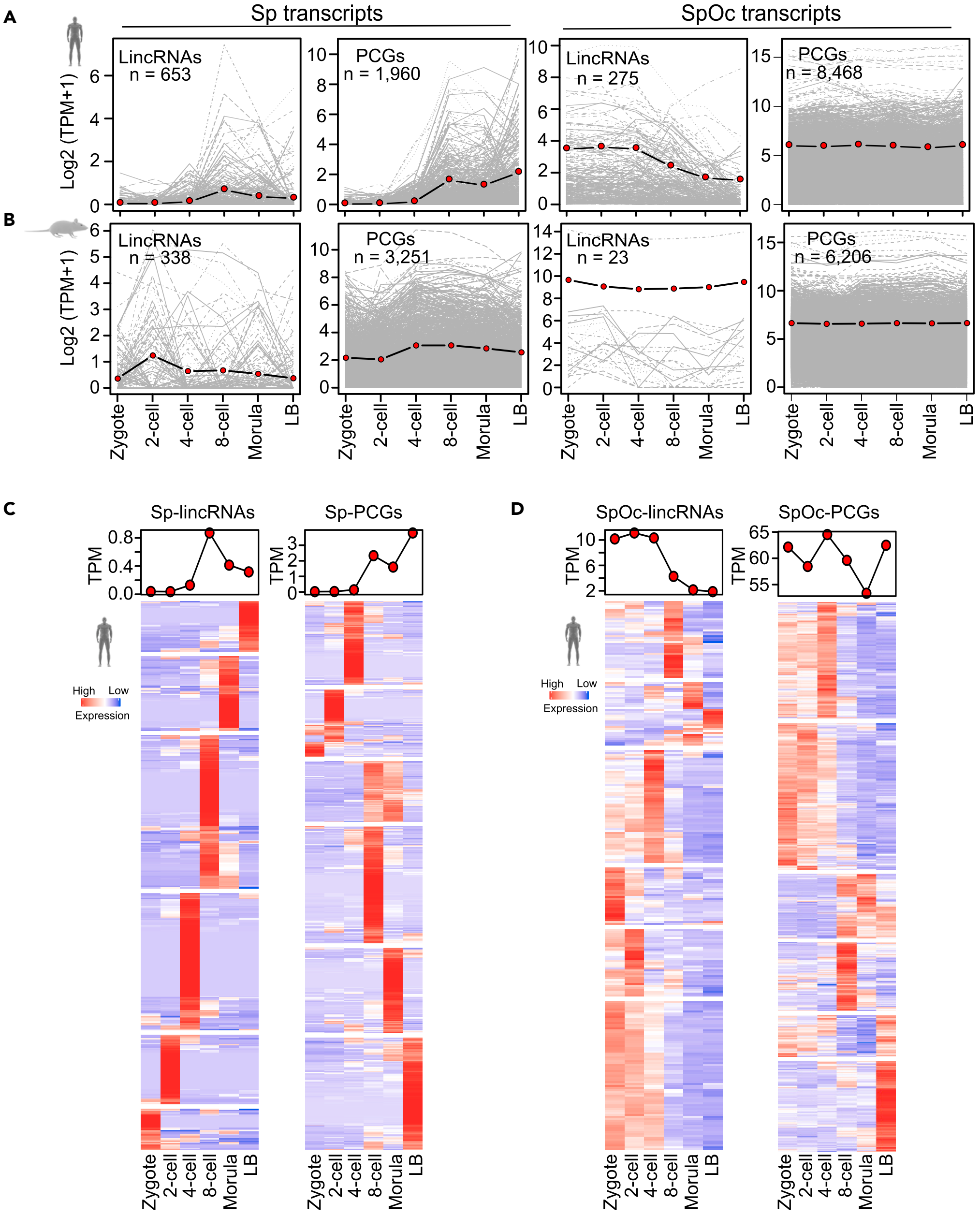


Figure 4. Sp and SpOc Show Temporal Expression during the Preimplantation Developmental Stages of Embryo

(A and B) Expression of Sp- and SpOc transcripts (lincRNAs and PCGs) from human (A) and mouse (B) during the preimplantation stages of developing embryo (zygote, two-cell, four-cell, and eight-cell, morula, and late blastocyst).

(C and D) Heatmaps showing stage-specific expression of Sp (C) and SpOc (D) transcripts (lincRNAs and PCGs) during preimplantation stage embryos. The expression values in plots (A and B) were log transformed to avoid extreme expression values. Expression profiles in heatmaps (C and D) were represented by *Z* score.

See also Figure S2 and Table S3.

tissues. However, sperm-derived bivalent domains from SpOc-PCG promoters maintained their bivalent chromatin structure in all the three germ layers but in the germ-layer-derived tissues only traces of bivalency were maintained (Figures 5D and 5E). Consistent with the lack of active histone marks, Sp-lincRNAs from all three sperm-derived chromatin clusters were not expressed in the three germ layers and somatic tissues (Figures 5F, 5G, and S3A). In contrast, Sp-PCGs having sperm-derived bivalent domain showed higher expression in germ layers as well as somatic tissues compared with the transcripts from high-K4

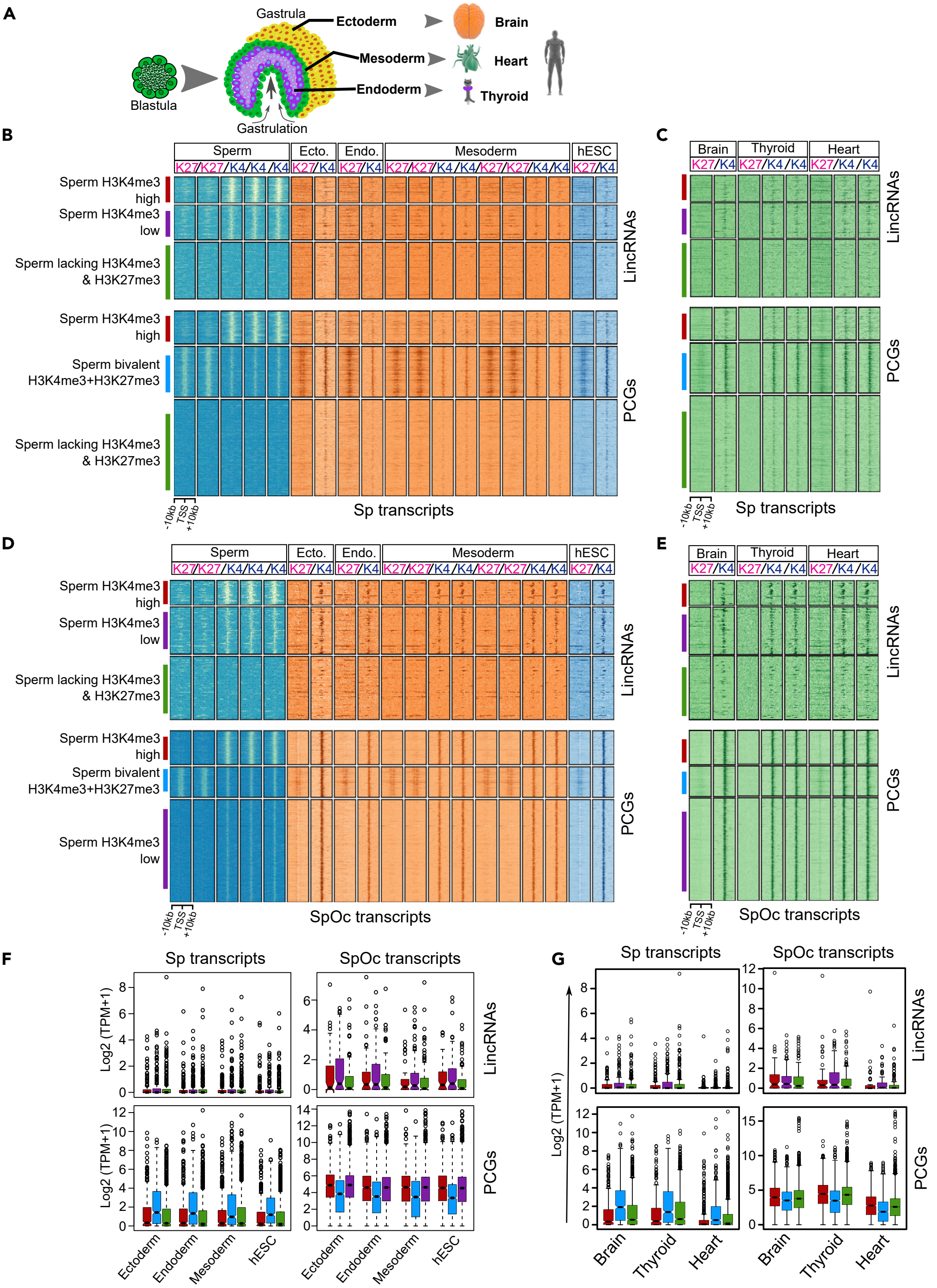


Figure 5. Sperm-Specific lincRNAs Are Not Critical for the Three Germ Layers Differentiation and Mature Tissue Formation

(A) Schematic of samples chosen from three germ layers (ectoderm, mesoderm, and endoderm) and matured tissues (brain, heart, and thyroid), derived from respective germ layers.

(B and C) Enrichment of H3K27me3 (K27) and H3K4me3 (K4) ChIP-seq signals from the three germ layers, human embryonic stem cells (hESC) (B), and the germ-layer-derived tissues (brain, thyroid, and heart) (C) over the promoters of Sp-lincRNA and PCGs (extended ±10 kb from TSS) from sperm-derived chromatin clusters.

Figure 5. *Continued*

(D and E) Enrichment of ChIP-seq H3K27me3 (K27) and H3K4me3 (K4) signals from the three germ layers, hESCs (D), and the germ-layer-derived tissues (brain, thyroid, and heart) (E) at the promoters of SpOc-lincRNAs and PCGs (extended ±10 kb from TSS).

(F) Expression of Sp- and SpOc-lincRNAs and PCGs in the three germ layers and hESCs.

(G) Expression of Sp- and SpOc-lincRNAs and PCGs in brain, thyroid, and heart tissues.

See also Figure S3 and Table S3.

and K4$^-$K27$^-$ sperm-derived chromatin clusters (Figures 5F, 5G, and S3A). This observation is consistent with the potential of bivalent domains in lineage commitment (Maezawa et al., 2018). The sperm-derived high-K4 and low-K4 chromatin clusters of SpOc-lincRNAs have low levels of H3K4me3 with the corresponding low expression levels in all the three germ layers (Figures 5F, 5G, and S3A). All three sperm-derived chromatin clusters of SpOc-PCGs with high H3K4me3 levels exhibit higher expression (Figures 5F, 5G, and S3A).

Overall, these observations suggest that Sp-lincRNAs appear to be crucial for sperm maturation and preimplantation stage embryogenesis because of their highly cell-type- and developmental-stage-specific expression, whereas Sp-PCGs, SpOc-lincRNAs, and SpOc-PCGs seem to play crucial role in the formation of the three germ layers and somatic tissues. In particular, Sp-PCG transcripts with a bivalent chromatin at their promoters were present at low level in sperm owing to the higher levels of H3K27me3 in relation to H3K4me3 at their promoters, and their increased expression during the formation of germ layers and somatic tissues is correlated with an increase in H3K4me3 levels.

Transcripts from Sperm-Derived High-K4 Chromatin Clusters Show Higher Expression in Round Spermatids

The prevailing view has been that sperm, despite harboring active chromatin structures, lacks transcriptional activity. Therefore, we wanted to investigate whether these transcripts are derived directly from mature sperm or were already present during pre-meiotic, meiotic, and post-meiotic spermatogenic cell types and transmitted to sperm. For that, we used RNA-seq samples from pre-meiosis (A-dark and A-pale spermatogonia), meiosis (leptotene/zygotene, early pachytene, and late pachytene), and post-meiosis (round spermatid) stages of spermatogenesis (Jan et al., 2017). Sp-lincRNAs, Sp-PCGs, and SpOc-lincRNAs from high- and low-K4 sperm-derived chromatin clusters were expressed more in round spermatid during spermatogenesis and show more testis-specific expression (Figures 6A, 6B, S3B, and S3C), whereas Sp-lincRNA, Sp-PCG, SpOc-lincRNA, and SpOc-PCG transcripts from sperm-derived K4$^-$K27$^-$ and bivalent chromatin clusters were not expressed in round spermatids; rather, they showed cell-type-specific expression in other pre-meiotic and meiotic spermatogenic cell types with less specificity toward testis (Figures 6A, 6B, S3B, and S3C), indicating that these transcripts may be generated in mature sperm.

In summary, tracing the expression dynamics from different spermatogenic cell types, gametes (matured sperm and oocyte), and preimplantation (zygote to late blastocyst) and post-implantation stages (germ layers and matured tissues) reveal that Sp and SpOc transcripts have opposing expression patterns throughout the human development (Figure 6C).

Transcription Factors Dictate the Expression of Sp- and SpOc-Transcripts in Preimplantation Development

It is known that histone modifications along with *cis*- or *trans*-acting regulatory elements can control the transcription of genes (Luo et al., 2016). Transcription factors are known to play an important role in modulating gene expression either by activating or repressing the transcription. During early embryonic development there are many transcription factors involved in the transition of a zygote into fully developed embryo. Since Sp- and SpOc-transcripts are temporally expressed during preimplantation embryonic stages, we wanted to know what kind of transcription-factor-binding sites are enriched over these promoters. We used sequences from promoters (±250 bp from transcription start site, TSS) of these transcripts to find enriched motifs by matching with the known transcription factor motif sequences generated by HOMER using published ChIP-seq datasets. Among these factors, ATF1 (Activating Transcription Factor 1) and EHF (ETS Homologous Factor) were enriched significantly in the promoters of a greater number of Sp-lincRNA promoters (Figure S4A). Expression of genes encoding ATF1 and EHF transcription factors correlated with the expression of their Sp-lincRNA targets during ZGA (between four-cell and blastocyst) (Figures 7A and 7B). Similarly, Sp-PCG promoters were enriched with Maz and Sp5-binding sites, and also

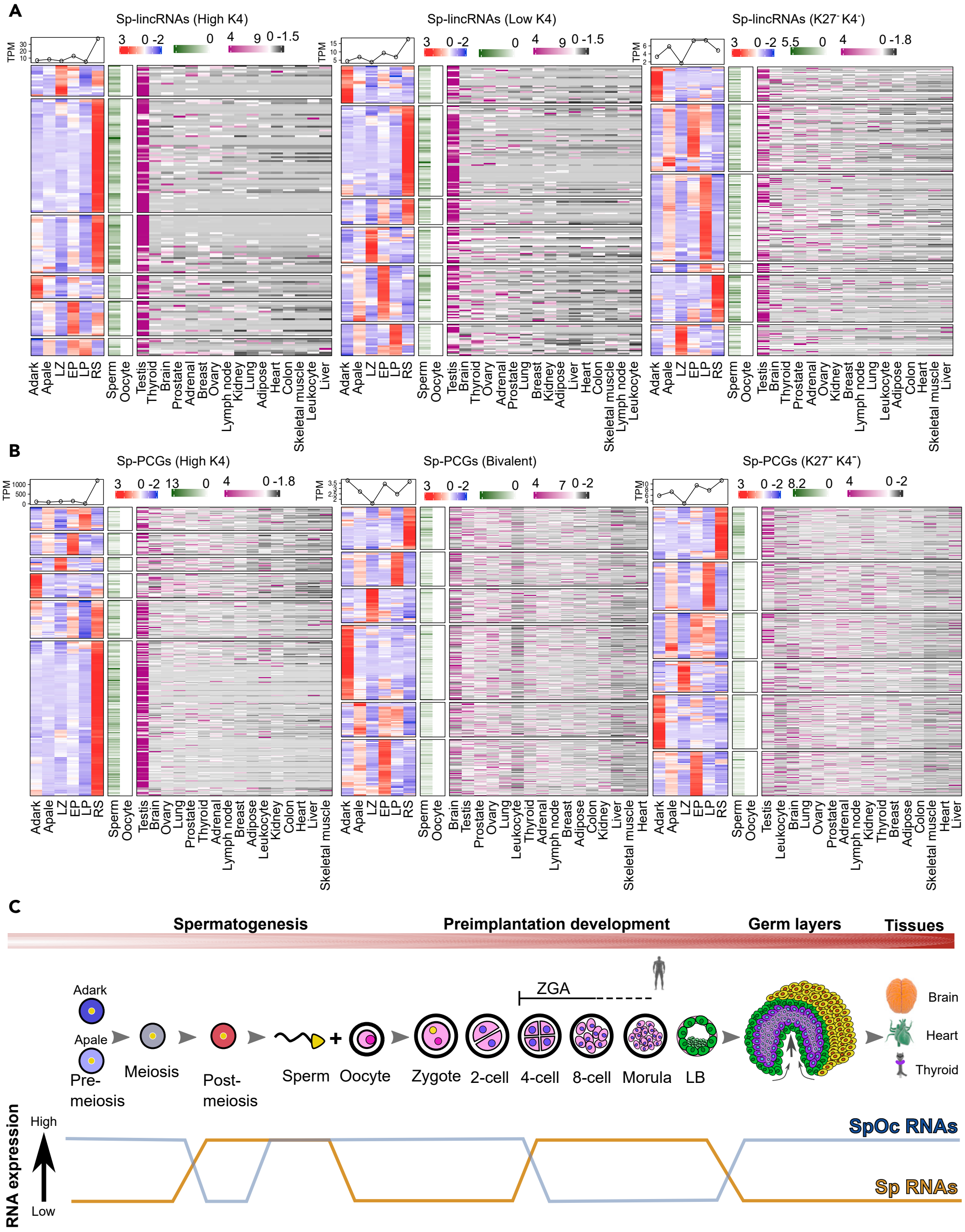


Figure 6. High-K4 Sp-lincRNAs and Sp-PCGS Are Abundantly Expressed in the Spermatid Stage of Spermatogenesis

(A and B) Heatmaps showing the expression of Sp-lincRNAs (A) and PCGs (B) from individual sperm-derived chromatin clusters during different stages of spermatogenesis starting from pre-miosis (A-dark and A-pale spermatogonia), meiosis (leptotene/zygotene, early pachytene, and late pachytene), and post-meiosis (round spermatid). This heatmap is followed by status in sperm, oocyte, and 16 different tissues from human body map 2.0.

(C) Model depicting observed expression patterns of sperm-specific and sperm-oocyte (SpOc) transcripts in gametes, preimplantation embryos, germ layers, and mature tissues.

See also Figure S3.

the expression of genes encoding these TFs corresponds to their target Sp-PCGs expression during ZGA (Figures S4B, 7A, and 7B). Sp-PCG transcripts from the sperm-derived $K4^{-}K27^{-}$ chromatin cluster of Sp-PCG transcripts have an Sp5 transcription factor motif, which is previously shown to be an important element in

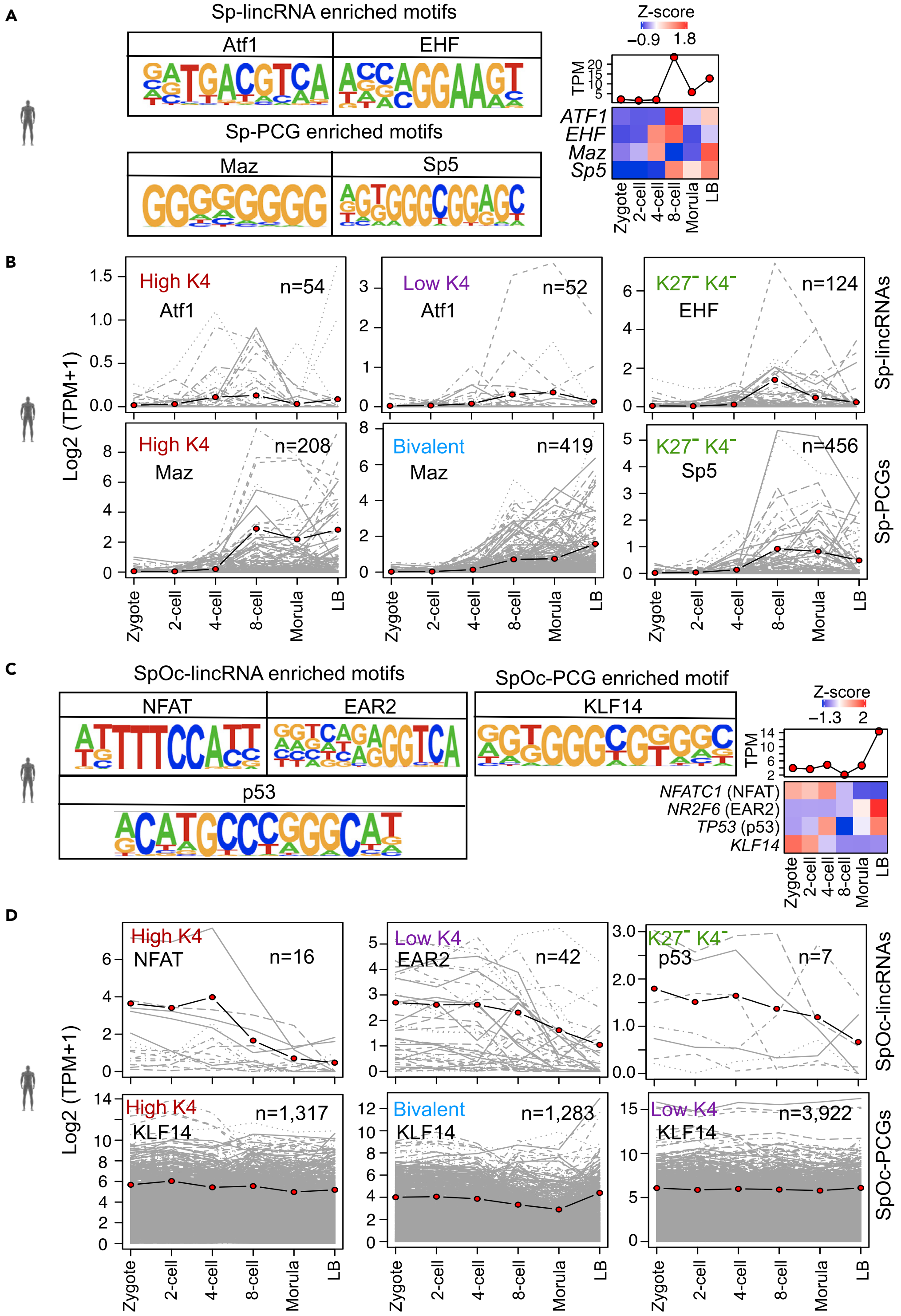


Figure 7. Preimplantation-Specific Transcription Factors Dictated the Expression of Sp-lincRNAs and Sp-PCGs during Maternal to Zygotic Transition

(A) Transcription factor motifs enriched at the promoters of Sp-lincRNAs (Atf1 and EHF) and Sp-PCGs (Maz and Sp5). Heatmap showing the expression status of genes encoding these transcription factors in germ cells and preimplantation stages of embryo.

Figure 7. *Continued*

(B) Expression of Sp transcripts, enriched with the indicated transcription factors, during preimplantation stages of embryo (two-, four-, and eight-cell, morula, and late blastocyst, LB).

(C) Transcription factor motifs enriched at the promoters of SpOc-lincRNAs (NAFT, EAR2, and p53) and SpOc-PCGs (KLF14). Heatmap with the expression status of genes encoding these transcription factors in germ cells and preimplantation stages of embryo.

(D) Expression of SpOc transcripts enriched with the indicated transcription factors during preimplantation stages of embryonic development (two-, four-, and eight-cell, morula, and late blastocyst, LB).

See also Figures S4 and S5 and Table S3.

maintaining gene expression patterns during embryonic development (Treichel et al., 2001). We found NFAT (Nuclear Factor of Activated T Cells), EAR2 (ErbA-related protein 2), p53 (Phosphoprotein 53), and KLF14 (Kruppel Like Factor 14) transcription factors to be enriched in the promoters of SpOc-lincRNA and SpOc-PCG transcripts from sperm-derived chromatin clusters (Figures S5A, S5B, and 7C). Among these, only the expression of NFAT and KLF14 transcription factors was correlated with their target gene expression during early stages of preimplantation development (between zygote to four-cell) (Figures 7C and 7D). We did not see any correlation between the other transcription factors, EAR2 and p53, and their target gene expression during the preimplantation embryonic stages (Figures 7C and 7D). *KLF14* is known to have maternally derived expression, and it is necessary for embryonic and extra-embryonic tissue development (Parker-Katiraee et al., 2007). Also, NFAT deficiency is known to cause embryonic lethality in mice (Chuvpilo et al., 2002; Mak et al., 2011). These observations collectively suggest that the expression of most of the Sp and SpOc transcripts is driven by important transcription factors known to have a key role during embryonic development.

Sp-lincRNAs Show Aberrant Expression in Cancer and Controls Cancer Cell Hallmarks

Previous studies have shown that RNAs involved in organism development and having testis-specific expression are known to take part in cancer development and progression (Aiello and Stanger, 2016; Hosono et al., 2017). To demonstrate the importance of Sp and SpOc transcripts in cancer, we used patient-derived RNA-seq samples from The Cancer Genome Atlas (TCGA). A comparison was made between TCGA tumors (N = 4,809) and corresponding TCGA healthy samples. Additionally, for a set of TCGA tumors (N = 4,035) for which there were no available TCGA normal samples, we used normal tissue samples from the GTEx consortium. Among the Sp and SpOc lincRNAs and PCGs, Sp-lincRNAs from sperm-derived chromatin cluster high-K4 showed more aberrant expression compared with the Sp-lincRNAs from low-K4 and K4$^-$K27$^-$ clusters (Figure 8A); thus, the level of their deregulation in cancers correlated with the extent of H3K4me3 enrichment at their promoters. However, Sp-PCGs from sperm-derived chromatin clusters did not reveal such correlation as seen with the Sp lincRNAs. SpOc lincRNAs and PCGs from sperm-derived chromatin clusters were less deregulated in cancers compared with the Sp-lincRNAs (Figure 8B). One common feature among the Sp and SpOc lincRNAs and PCGs is that transcripts from sperm-derived high-K4 clusters showed more deregulation than transcripts from lowK4, bivalent, and k4$^-$K27$^-$ chromatin clusters (Figures 8A–8D and S6–S8). A detailed look into the expression patterns of individual genes revealed some important previously known candidates from each chromatin cluster of Sp and SpOc transcripts (Figures 9A and S9A). Of note, we found an intergenic lncRNA *LINC01518* from Sp-lincRNAs group to be deregulated in 21 of the 26 cancers, showing high levels of deregulation in lung (LUAD and LUSC), prostate (PRAD), testis (TCGT), and uterine or ovary (UCS, UCEC, and OV)-related cancers.

Since Sp lincRNAs show higher expression in tumors compared with normal, we wanted to investigate whether they possess oncogenic properties. To this end, we investigated the expression of three Sp-lincRNAs (*P4HA3-AS1*, *LINC01518*, and *AP001476.1*), representing each of the three sperm-derived chromatin clusters, in human embryonic kidney cell line HEK-293 and cervical carcinoma cell line HeLa. We found that the Sp lincRNAs show higher expression in HeLa cells compared with the HEK293 cells (Figures 9B and S9B). We therefore next investigated the effect of their downregulation in HeLa cells, using siRNAs, on important cancer cell hallmarks such as cell proliferation, cell cycle progression, and apoptosis. Downregulation of the Sp lincRNAs significantly affected all the three cancer cell hallmarks, indicating that the three chosen Sp lincRNAs possess oncogenic properties (Figures 9C–9F and S9C).

These results suggest that Sp- and SpOc-lincRNAs, which are temporally expressed during preimplantation embryo development coinciding with ZGA and SpOc-TD, appear to play a crucial role in cancer development and progression.

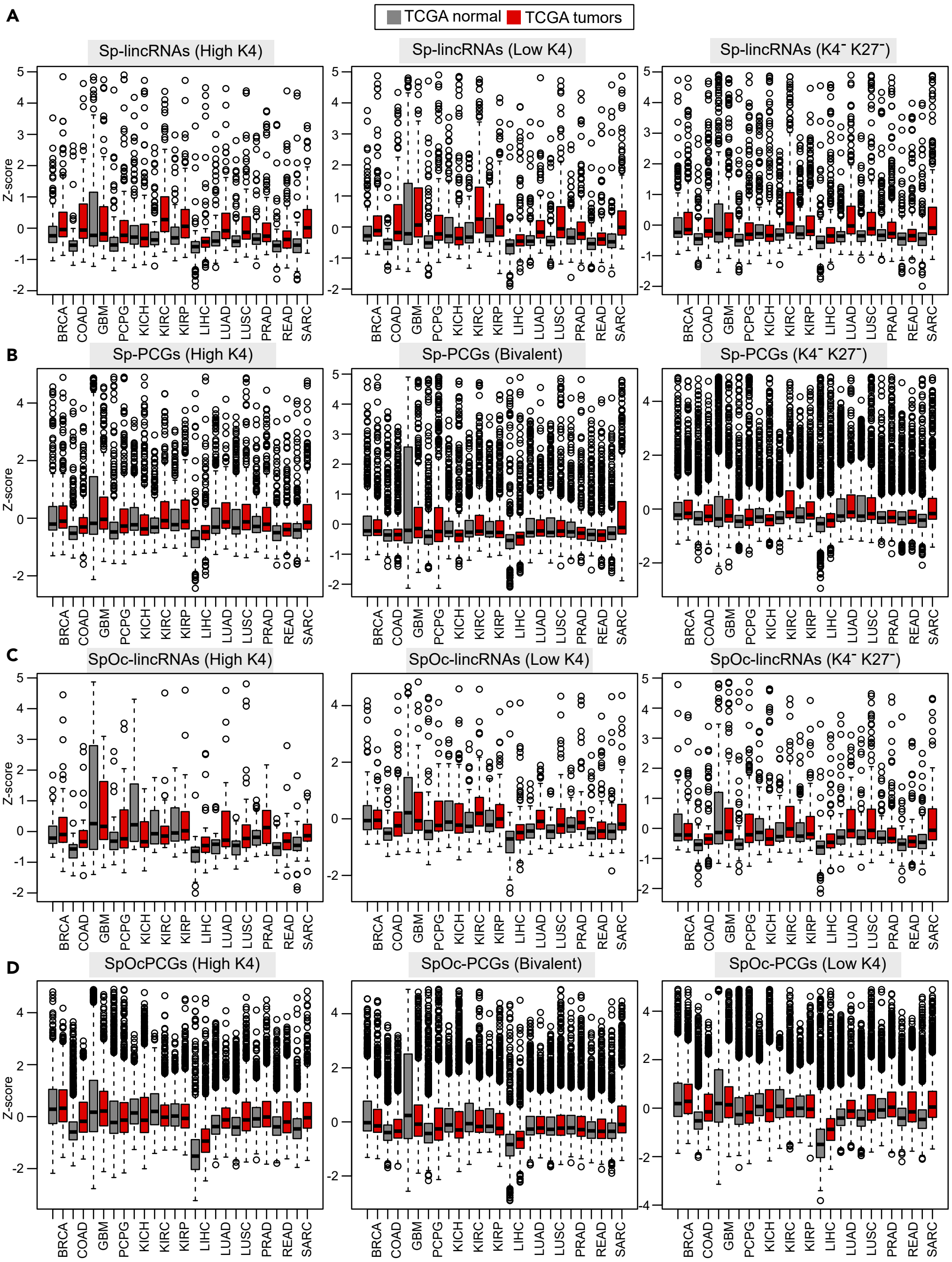


Figure 8. Sp-lincRNAs and Sp-PCGs Show Aberrant Expression in Different Cancers

(A–D) Expression status of Sp-lincRNAs (A), Sp-PCGs (B), SpOc-lincRNAs (C), and SpOc-PCGs (D) in different tumors and corresponding healthy tissues from TCGA patient cohort. The *Z* score in the plot is derived from the normalized TPM expression values. See also Figures S6–S8 and Table S3.

DISCUSSION

Our study explores the potential role of genes that encode sperm-specific transcripts and possess sperm-inherited chromatin structures in mammalian development and cancer. Intriguingly, in humans, we found a comparable number of transcripts present in both sperm and oocyte, whereas in mouse, sperm

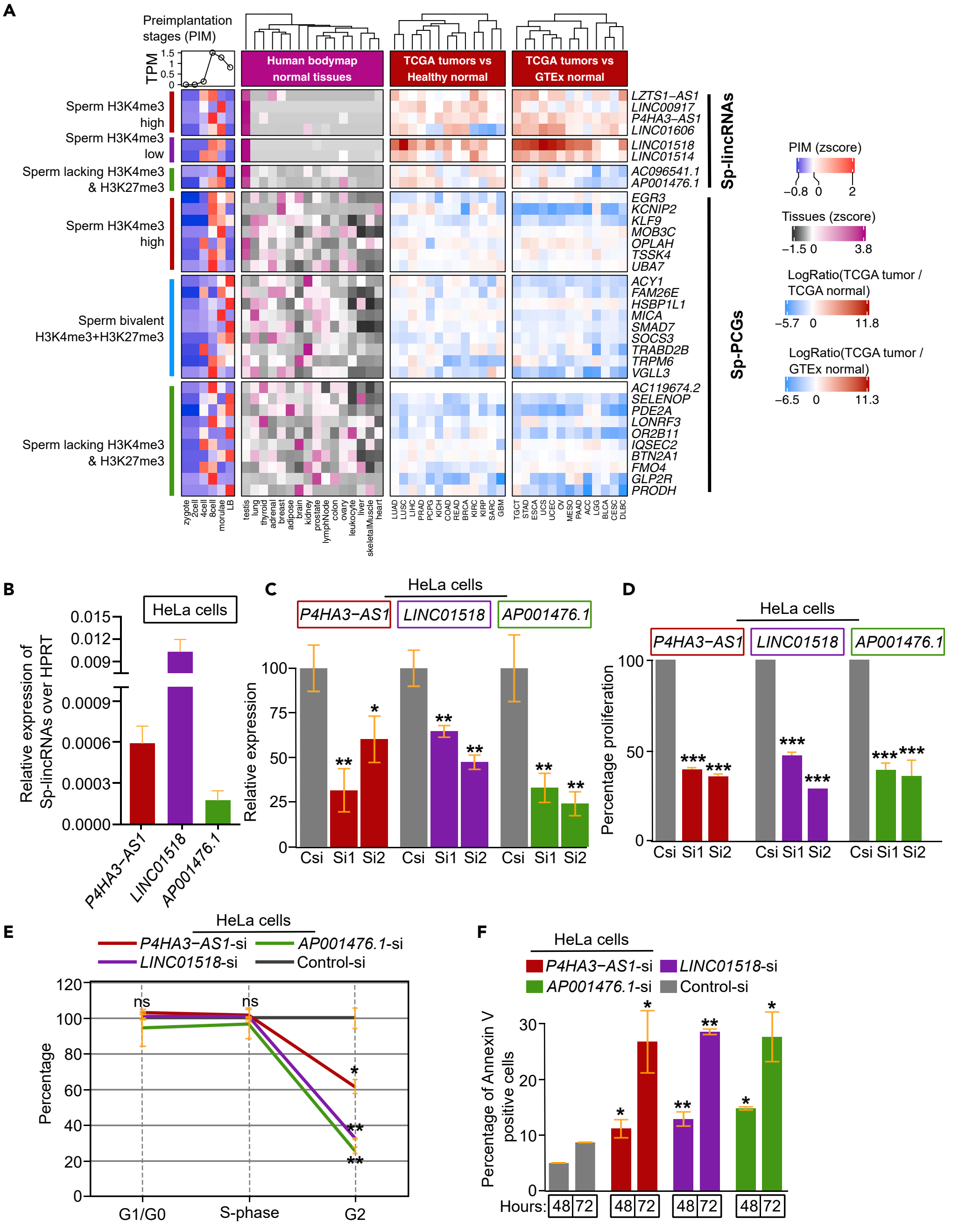


Figure 9. Sp and SpOc lincRNAs Show Tumor-Specific Activation

(A) Status of Sp transcripts in germ cells, preimplantation stage embryos, the human body map tissues, TCGA tumors compared with the corresponding healthy samples, or TCGA tumors compared with GTEx health samples. *Z* score in the plots is derived from the normalized TPM expression values. The log fold change is calculated by comparing the expression of tumors with the health samples expression.

(B) Relative expression of Sp transcripts in HeLa cell lines.

(C) Percentage of relative expression levels of three Sp-lincRNAs that are downregulated using two different siRNAs and control siRNAs in HeLa cells.

(D) MTT assay showing the percentage of proliferative cells for HeLa cells transfected with siRNAs for Sp-lincRNAs compared with the respective control siRNA samples.

(E) Line graph showing the cell-cycle profiles of HeLa cells transfected with Sp-lincRNA siRNA and control siRNA.

Figure 9. *Continued*

(F) Bar graphs representing the percentage of Annexin V-positive cells after 48 and 72 h transfection of HeLa cells with siRNA for three different Sp-lincRNAs and control siRNA samples. For each gene, the p value is calculated using two different siRNA-transfected samples. * indicates p value < 0.05, ** indicates p value <0.01, and *** indicates p value <0.001. Data from plots are represented as mean ± SD.
See also Figure S9 and Tables S3 and S4.

has a greater number of transcripts compared with the oocyte. Our observations are consistent with previous data where a pool of 3,281 poly(A) transcripts were identified based on the microarray analysis of RNA from the ejaculated sperm of nine individuals (Ostermeier et al., 2002). The presence of significant number of sperm transcripts is interesting because sperm has been considered transcriptionally incompetent. This observation in particular gains importance considering that nearly 4% of the sperm genome retains histone-enriched chromatin, and more importantly, there are large islands of H3K4me3 at the promoter regions of sperm transcripts. The presence of RNA and active chromatin structures obviously points to potential of sperm chromatin supporting the transcriptional events. Consistent with the latter tenet, we observed correlation between the sperm transcript abundance and the extent of H3K4me3 enrichment at their promoters. Furthermore, previous investigations have shown sperm transcript alterations in response to motility, capacitation, and cryopreservation (Ren et al., 2017). However, this is a far-fetched conclusion considering that there is a lack of data supporting the active transcription machinery at the H3K4me3 enriched promoters in sperm. This raises an important question, from where do sperm transcripts originally originate? Previously, it was suggested that the transcript accumulation in sperm is not a stochastic event rather a preservation of RNA through an organized process throughout spermiogenesis (Krawetz, 2005). Consistent with the latter notion, our data show that promoters of transcripts from sperm-derived chromatin cluster with high H3K4me3 show higher expression in spermatid, compared with the other spermatogenic cell types. This observation may indicate that sperm could inherit these transcripts from spermatid. However, it is not clearly evident how the accumulation of transcripts in sperm from the other sperm-derived chromatin clusters (low-K4 and K4$^-$K27$^-$) occurs, even though their expression is restricted to meiotic or pre-meiotic spermatogenic cell types. These observations indicate that low levels of transcription may persist during sperm maturation, and this notion is consistent with the data that sperm transcripts differs in caput and cauda and moreover, as discussed earlier, sperm RNA alterations in response to motility and capacitation (Conine et al., 2018; Ren et al., 2017). Thus, future investigations along similar lines will be fruitful.

The next important question is whether genes encoding sperm-specific transcripts and/or the transcripts per se have any role in preimplantation development. Multiple lines of evidence suggest that sperm transcripts play an important role in early preimplantation development. For example, defective blastocyst embryos generated from dicer knockout germ cells can be rescued with sperm RNA (Yuan et al., 2016). Similarly, embryos developed from caput sperm had post-implantation developmental defects compared with cauda sperm, which give rise to embryos that develop to term. Interestingly, microinjection of cauda-specific small RNAs into caput-derived embryos rescued post-implantation embryonic lethality (Conine et al., 2018). Additionally, sperm RNA has been shown to play an important role in modulating the RNA levels during fertilization, first cleavage, and blastocyst of early stages of embryo development (Alves et al., 2019; Bohacek and Rassoulzadegan, 2020). In line with this, our gene ontology analysis of sperm transcripts from high-K4 chromatin cluster reveals biological functions related to spermatogenesis, penetration to zona pellucida, ion channel activity, etc. Sperm penetration to the zona pellucida and ion channel activity are the most important functions during fertilization, and it seems that sperm transcripts play an important role in these biological processes. These observations, collectively, not only emphasize the importance of sperm transcripts in early preimplantation development but also highlight the global transcriptome dynamics during sperm maturation. Thus, the identification of a significant number of sperm transcripts in our investigation signifies the importance of sperm transcripts in early preimplantation development. Interestingly, sperm-derived bivalent chromatin clusters were enriched with biological functions related to developmental processes (Tomizawa et al., 2018), which is consistent with previous studies on the association of genes harboring bivalent chromatin promoters with developmental functions. These kind of distinct clusters of biological processes dictated by sperm transcripts were not seen in sperm-oocyte-expressed transcripts. In mouse, on the other hand, there was no correlation between chromatin structure at the promoters and the levels of their encoded transcripts in sperm, and also, we did not find any distinct biological functions enriched for individual clusters as seen in the human sperm. These observations clearly indicate that the human sperm genome is highly structured to execute important developmental functions and that sperm-specific lincRNAs appear to have a greater biological role in sperm maturation and preimplantation development.

Another interesting aspect of the current investigation is the temporal expression of genes that encode sperm-specific and sperm-oocyte transcripts during preimplantation development. Genes that encode sperm-specific transcripts carry distinct chromatin structures and show stage-specific expression across preimplantation stages. This exclusive stage-specific expression is not seen with genes that encode sperm-oocyte transcripts, rather the majority of these transcripts show expression during early preimplantation stages, primarily between two- and four-cell stages and start to decline or degrade during the onset of ZGA. Highly temporal expression of genes that encode sperm-specific transcripts in the preimplantation stages signifies their importance during early preimplantation development. Moreover, genes that encode sperm-specific and sperm-oocyte transcripts have crucial non-overlapping roles in preimplantation development. It is reasonable to assume that the sperm-inherited chromatin imprints together with the help of preimplantation development associated transcription factors may bring in the observed stage-specific expression (Bui et al., 2011). Indeed, this seems to be the case as sperm transcripts harbor motifs for transcription factors that show specific expression during early embryonic development. Thus, it is more likely that an interplay between sperm-inherited chromatin imprints and early embryonic-specific transcription factors may in part contribute to the stage-specific expression of the genes that encode sperm and sperm-oocyte transcripts.

Strikingly, the expression of Sp-lincRNAs as well as their promoter epigenetic profiles were lost in the three germ layers and their derived somatic tissues. This again emphasizes the importance of the genes that encode sperm lincRNAs in fertilization and preimplantation development. On the other hand, genes that encode Sp-PCGs and SpOc transcripts (SpOc-lincRNAs and SpOc-PCGs) continue to express in all the three germ layers and the germ-layer-derived multiple tissues. Comparison of the histone profiles of sperm, germ layers, and mature tissues revealed that Sp-PCGs and SpOc-PCGs retain their sperm-specific bivalent chromatin in the germ layers and somatic tissues; however, the ratio of H3K4me3 and H3K27me3 is different from that in the sperm. These observations indicate that the sperm-derived bivalent chromatin imprints at the Sp-PCGs and SpOc-PCGs promoters are maintained during the germ layers formation and these sperm-inherited chromatin imprints may play an important role in crucial developmental decisions by regulating spatiotemporal expression of Sp-PCGs and SpOc-PCGs. Thus, our study found two sets of genes: one set, sperm-specific genes with high-K4 promoters, whose transcripts are enriched in sperm, activated during ZGA and transcriptionally silenced in the three germ layers and somatic tissues. In the other set of genes, comprising Sp-PCGs, SpOc-lincRNAs, and SpOc-PCGs with high-K4, low-K4, K4$^-$K27$^-$, and bivalent (H3K4me3/H3K27me3) promoters, transcripts are enriched in sperm, and they express throughout preimplantation development, the germ layers, and in multiple somatic tissues. Thus, sperm-encoded information seems to take part in important developmental decisions.

Previous studies have proposed a functional link between early embryonic gene expression program and cancer development and progression (Bouckenheimer et al., 2016; Shah et al., 2018). In particular, genes active during preimplantation development are deregulated in several cancers. Moreover, cancer/testis expression has long been serving as basis for identifying diagnostic and prognostic markers for several cancers. Our data on Sp-lincRNAs having high-K4 at the promoters in sperm are particularly striking as they show higher deregulation in multiple cancers compared with lincRNAs from the other sperm-derived chromatin clusters (bivalent or K4$^-$K27$^-$). Interestingly, the level of their deregulation in cancer matches with the sperm H3K4me3 levels at the promoters. For example, lincRNAs with high-K4 levels at their promoters showed more deregulation in cancer compared with the lincRNAs with low-K4, indicating that mechanisms that establish sperm-specific chromatin imprints are recapitulated in cancer. Their higher expression in tumors compared with the corresponding normal tissues indicate that they may behave as oncogenes. We have also tested the oncogenic properties of selected lncRNAs from high-K4, low-K4, and K4$^-$K27$^-$ clusters in HeLa cell line by measuring their effect on crucial cancer cell hallmarks that define the oncogenic drivers such as cell proliferation, cell cycle progression, and apoptosis. Loss of function of these lncRNAs in HeLa cells using siRNAs resulted in decrease in cell proliferation, cell cycle progression, and apoptosis, indicating that these sperm lincRNAs harbor oncogenic properties. Thus, Sp-lincRNAs may serve as comprehensive resource for diagnostic/prognostic markers and potential therapeutic targets. SpOc-lincRNAs also showed deregulation in multiple cancers but to a lesser extent compared with the Sp-lincRNAs. Surprisingly, Sp-PCGs and SpOc-PCGs were mostly downregulated in cancers, and this in particular is consistent with their role in immune response regulation as immune response genes are mostly suppressed in many cancers.

In sum, the abundance of RNA, in particular lincRNAs, in sperm necessitates further investigation on their importance in sperm maturation, fertilization, and preimplantation development and cancer. More importantly, our study laid a strong basis for further investigation on the functional role of sperm-inherited chromatin imprints in organismal development.

Limitations of the Study

Lack of chromatin data (ChIP-seq) for different stages of human preimplantation embryos limited our possibility to explore whether sperm-derived chromatin imprints are preserved during preimplantation development. Exploring this would have added an in-depth resolution to our analysis.

Resource Availability

Lead Contact

Correspondence to lead contact Chandrasekhar Kanduri, E-mail: kanduri.chandrasekhar@gu.se.

Materials Availability

This study did not generate any new sequencing data. Analysis was performed using public datasets.

Data and Code Availability

The processed datasets generated during this study are available publicly at Mendeley Data repository with the following DOI https://dx.doi.org/10.17632/695c4zvr6d.1.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101165.

ACKNOWLEDGMENTS

The results shown here are in whole or part based upon datasets collected from GEO, ENCODE, GTEx, The Cancer Genome Atlas (TCGA), and dbGAP repositories. The brain chromatin profiles are obtained from controlled dataset from dbGAP project phs000791.v1.p1. We thank Susann Li for helping in flow cytometry analysis. All computations were performed on resources provided by Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) high-performance computing (HPC), which is part of Swedish National Infrastructure for Computing (SNIC). This work was supported by the grants from Knut and Alice Wallenberg Foundation (KAW2014.0057); Swedish Foundation for Strategic Research (RB13-0204); Swedish Cancer Research Foundation (Cancerfonden: Kontraktno. CAN2018/591); Swedish Research Council (2017-02834); Barncancerfonden (PR2018-0090); IngaBritt och Arne Lundbergs Forskningsstiftelse and LUA/ALF (to C.K.).

AUTHOR CONTRIBUTIONS

S.S. and C.K. conceptualized and designed the study; S.S. performed data collection and computational analysis and implemented pipelines; M.K. performed functional assays; S.S. and C.K. contributed to the data interpretation and wrote the manuscript.

DECLARATION OF INTERESTS

A patent application has been filed based on the findings presented in this study.

Received: January 10, 2020
Revised: April 4, 2020
Accepted: May 8, 2020
Published: June 26, 2020

REFERENCES

Aiello, N.M., and Stanger, B.Z. (2016). Echoes of the embryo: using the developmental biology toolkit to study cancer. Dis. Model. Mech. *9*, 105–114.

Alves, M.B.R., de Arruda, R.P., De Bem, T.H.C., Florez-Rodriguez, S.A., Sa Filho, M.F., Belleannee, C., Meirelles, F.V., da Silveira, J.C., Perecin, F., and Celeghini, E.C.C. (2019). Sperm-borne miR-216b modulates cell proliferation during early embryo development via K-RAS. Sci. Rep. *9*, 10358.

Bohacek, J., and Rassoulzadegan, M. (2020). Sperm RNA: quo vadis? Semin. Cell Dev Biol *97*, 123–130.

Bouckenheimer, J., Assou, S., Riquier, S., Hou, C., Philippe, N., Sansac, C., Lavabre-Bertrand, T., Commes, T., Lemaitre, J.M., Boureux, A., et al. (2016). Long non-coding RNAs in human early embryonic development and their potential in ART. Hum. Reprod. Update *23*, 19–40.

Brykczynska, U., Hisano, M., Erkek, S., Ramos, L., Oakeley, E.J., Roloff, T.C., Beisel, C., Schubeler, D., Stadler, M.B., and Peters, A.H. (2010). Repressive and active histone methylation mark distinct promoters in human and mouse spermatozoa. Nat. Struct. Mol. Biol. *17*, 679–687.

Bui, H.T., Wakayama, S., Mizutani, E., Park, K.K., Kim, J.H., Van Thuan, N., and Wakayama, T. (2011). Essential role of paternal chromatin in the regulation of transcriptional activity during mouse preimplantation development. Reproduction *141*, 67–77.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927.

Chu, L.F., Leng, N., Zhang, J., Hou, Z., Mamott, D., Vereide, D.T., Choi, J., Kendziorski, C., Stewart, R., and Thomson, J.A. (2016). Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. Genome Biol. *17*, 173.

Chuvpilo, S., Jankevics, E., Tyrsin, D., Akimzhanov, A., Moroz, D., Jha, M.K., Schulze-Luehrmann, J., Santner-Nanan, B., Feoktistova, E., Konig, T., et al. (2002). Autoregulation of NFATc1/A expression facilitates effector T cells to escape from rapid apoptosis. Immunity *16*, 881–895.

Conine, C.C., Sun, F., Song, L., Rivera-Perez, J.A., and Rando, O.J. (2018). Small RNAs gained during epididymal transit of sperm are essential for embryonic development in mice. Dev. Cell *46*, 470–480.e3.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46*, D794–D801.

Fan, X., Zhang, X., Wu, X., Guo, H., Hu, Y., Tang, F., and Huang, Y. (2015). Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. Genome Biol. *16*, 148.

Friedman, C.E., Nguyen, Q., Lukowski, S.W., Helfer, A., Chiu, H.S., Miklas, J., Levy, S., Suo, S., Han, J.J., Osteil, P., et al. (2018). Single-cell transcriptomic analysis of cardiac differentiation from human PSCs reveals HOPX-dependent cardiomyocyte maturation. Cell Stem Cell *23*, 586–598.e8.

Gapp, K., van Steenwyk, G., Germain, P.L., Matsushima, W., Rudolph, K.L.M., Manuella, F., Roszkowski, M., Vernaz, G., Ghosh, T., Pelczar, P., et al. (2018). Alterations in sperm long RNA contribute to the epigenetic inheritance of the effects of postnatal trauma. Mol. Psychiatry, 1–13.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295–300.

Hamm, D.C., and Harrison, M.M. (2018). Regulatory principles governing the maternal-to-zygotic transition: insights from *Drosophila melanogaster*. Open Biol. *8*, 180183.

Hammoud, S.S., Low, D.H., Yi, C., Carrell, D.T., Guccione, E., and Cairns, B.R. (2014). Chromatin and transcription transitions of mammalian adult germline stem cells and spermatogenesis. Cell Stem Cell *15*, 239–253.

Hammoud, S.S., Nix, D.A., Zhang, H., Purwar, J., Carrell, D.T., and Cairns, B.R. (2009). Distinctive chromatin in human sperm packages genes for embryo development. Nature *460*, 473–478.

Hosono, Y., Niknafs, Y.S., Prensner, J.R., Iyer, M.K., Dhanasekaran, S.M., Mehra, R., Pitchiaya, S., Tien, J., Escara-Wilke, J., Poliakov, A., et al. (2017). Oncogenic role of THOR, a conserved cancer/testis long non-coding RNA. Cell *171*, 1559–1572.e20.

Jan, S.Z., Vormer, T.L., Jongejan, A., Roling, M.D., Silber, S.J., de Rooij, D.G., Hamer, G., Repping, S., and van Pelt, A.M.M. (2017). Unraveling transcriptome dynamics in human spermatogenesis. Development *144*, 3659–3673.

Jung, Y.H., Sauria, M.E.G., Lyu, X., Cheema, M.S., Ausio, J., Taylor, J., and Corces, V.G. (2017). Chromatin states in mouse sperm correlate with embryonic and adult regulatory landscapes. Cell Rep. *18*, 1366–1382.

Krawetz, S.A. (2005). Paternal contribution: new insights and future challenges. Nat. Rev. Genet. *6*, 633–642.

Locke, W.J., Zotenko, E., Stirzaker, C., Robinson, M.D., Hinshelwood, R.A., Stone, A., Reddel, R.R., Huschtscha, L.I., and Clark, S.J. (2015). Coordinated epigenetic remodelling of transcriptional networks occurs during early breast carcinogenesis. Clin. Epigenetics *7*, 52.

Loh, K.M., Ang, L.T., Zhang, J., Kumar, V., Ang, J., Auyeong, J.Q., Lee, K.L., Choo, S.H., Lim, C.Y., Nichane, M., et al. (2014). Efficient endoderm induction from human pluripotent stem cells by logically directing signals controlling lineage bifurcations. Cell Stem Cell *14*, 237–252.

Luo, S., Lu, J.Y., Liu, L., Yin, Y., Chen, C., Han, X., Wu, B., Xu, R., Liu, W., Yan, P., et al. (2016). Divergent lncRNAs regulate gene expression and lineage differentiation in pluripotent cells. Cell Stem Cell *18*, 637–652.

Maezawa, S., Hasegawa, K., Yukawa, M., Kubo, N., Sakashita, A., Alavattam, K.G., Sin, H.S., Kartashov, A.V., Sasaki, H., Barski, A., et al. (2018). Polycomb protein SCML2 facilitates H3K27me3 to establish bivalent domains in the male germline. Proc. Natl. Acad. Sci. U S A *115*, 4957–4962.

Mak, M.C., Lam, K.M., Chan, P.K., Lau, Y.B., Tang, W.H., Yeung, P.K., Ko, B.C., Chung, S.M., and Chung, S.K. (2011). Embryonic lethality in mice lacking the nuclear factor of activated T cells 5 protein due to impaired cardiac development and function. PLoS One *6*, e19186.

Ostermeier, G.C., Dix, D.J., Miller, D., Khatri, P., and Krawetz, S.A. (2002). Spermatozoal RNA profiles of normal fertile men. Lancet *360*, 772–777.

Paradowska, A.S., Miller, D., Spiess, A.N., Vieweg, M., Cerna, M., Dvorakova-Hortova, K., Bartkuhn, M., Schuppe, H.C., Weidner, W., and Steger, K. (2012). Genome wide identification of promoter binding sites for H4K12ac in human sperm and its relevance for early embryonic development. Epigenetics *7*, 1057–1070.

Parker-Katiraee, L., Carson, A.R., Yamada, T., Arnaud, P., Feil, R., Abu-Amero, S.N., Moore, G.E., Kaneda, M., Perry, G.H., Stone, A.C., et al. (2007). Identification of the imprinted KLF14 transcription factor undergoing human-specific accelerated evolution. PLoS Genet. *3*, e65.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.

Ren, X., Chen, X., Wang, Z., and Wang, D. (2017). Is transcription in sperm stationary or dynamic? J. Reprod. Dev. *63*, 439–443.

Shah, K., Patel, S., Mirza, S., and Rawal, R.M. (2018). Unravelling the link between embryogenesis and cancer metastasis. Gene *642*, 447–452.

Siklenka, K., Erkek, S., Godmann, M., Lambrot, R., McGraw, S., Lafleur, C., Cohen, T., Xia, J., Suderman, M., Hallett, M., et al. (2015). Disruption of histone methylation in developing sperm impairs offspring health transgenerationally. Science *350*, aab2006.

Simpson, A.J., Caballero, O.L., Jungbluth, A., Chen, Y.T., and Old, L.J. (2005). Cancer/testis antigens, gametogenesis and cancer. Nat. Rev. Cancer *5*, 615–625.

Tomizawa, S.I., Kobayashi, Y., Shirakawa, T., Watanabe, K., Mizoguchi, K., Hoshi, I., Nakajima, K., Nakabayashi, J., Singh, S., Dahl, A., et al. (2018). Kmt2b conveys monovalent and bivalent H3K4me3 in mouse spermatogonial stem cells at germline and embryonic promoters. Development *145*, dev169102, https://doi.org/10.1242/dev.169102.

Treichel, D., Becker, M.B., and Gruss, P. (2001). The novel transcription factor gene Sp5 exhibits a dynamic and highly restricted expression pattern during mouse embryogenesis. Mech. Dev. *101*, 175–179.

Voigt, P., Tee, W.W., and Reinberg, D. (2013). A double take on bivalent promoters. Genes Dev. *27*, 1318–1338.

Xia, W., Xu, J., Yu, G., Yao, G., Xu, K., Ma, X., Zhang, N., Liu, B., Li, T., Lin, Z., et al. (2019). Resetting histone modifications during human parental-to-zygotic transition. Science *365*, 353–360.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat. Struct. Mol. Biol. *20*, 1131–1139.

Yuan, S., Schuster, A., Tang, C., Yu, T., Ortogero, N., Bao, J., Zheng, H., and Yan, W. (2016). Sperm-borne miRNAs and endo-siRNAs are important for fertilization and preimplantation embryonic development. Development *143*, 635–647.

Zhang, K., Huang, K., Luo, Y., and Li, S. (2014). Identification and functional analysis of long non-coding RNAs in mouse cleavage stage embryonic development based on single cell transcriptome data. BMC Genomics *15*, 845.

Zhang, X., Gao, F., Fu, J., Zhang, P., Wang, Y., and Zeng, X. (2017). Systematic identification and characterization of long non-coding RNAs in mouse mature sperm. PLoS One *12*, e0173402.

Zhang, Y., Zhang, X., Shi, J., Tuorto, F., Li, X., Liu, Y., Liebers, R., Zhang, L., Qu, Y., Qian, J., et al. (2018). Dnmt2 mediates intergenerational transmission of paternally acquired metabolic disorders through sperm small non-coding RNAs. Nat. Cell Biol. *20*, 535–540.

iScience, Volume *23*

Supplemental Information

Sperm Originated Chromatin Imprints and LincRNAs in Organismal Development and Cancer

Santhilal Subhash, Meena Kanduri, and Chandrasekhar Kanduri

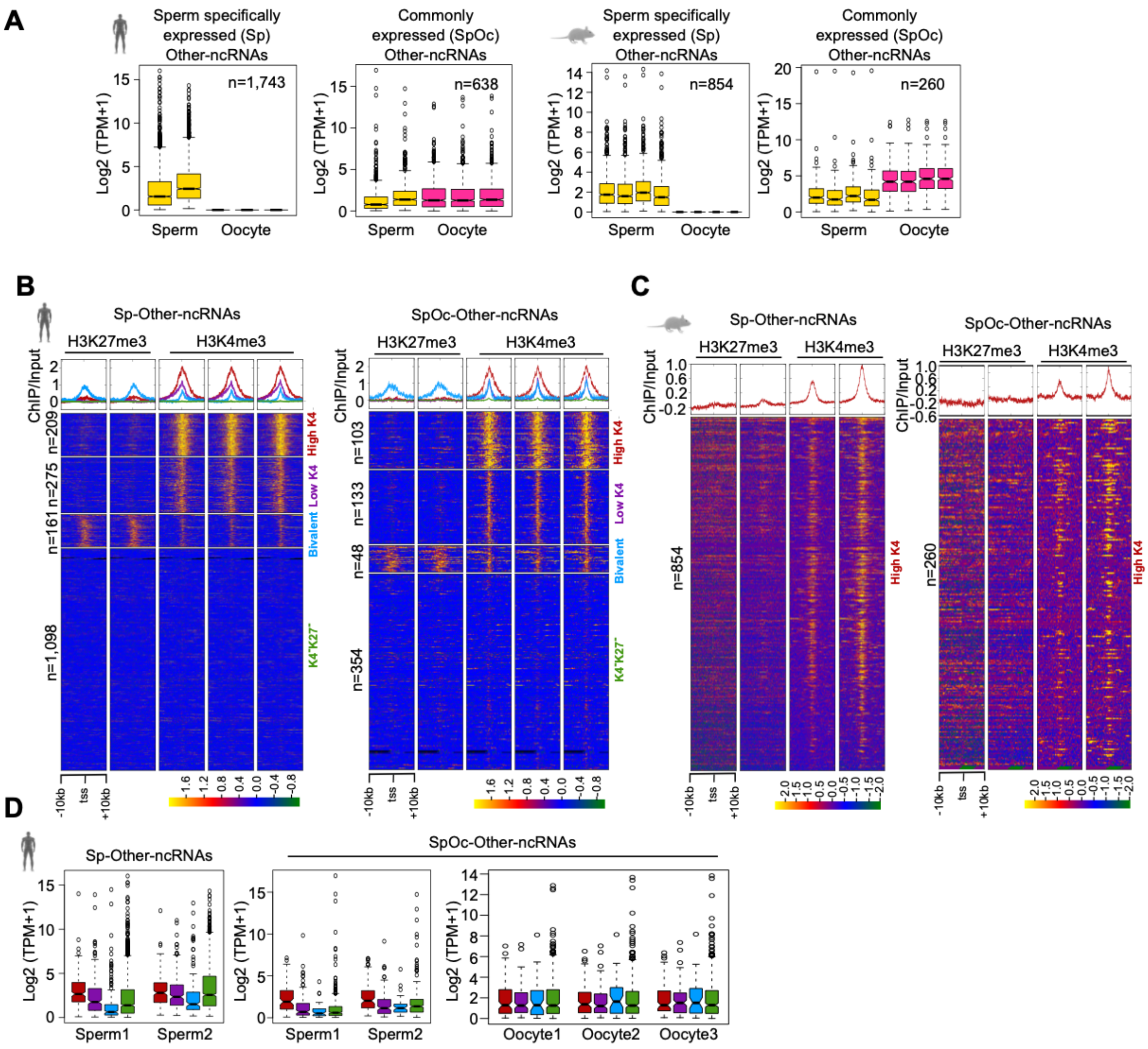


Figure S1. Transcriptome and chromatin profiles of Other-ncRNAs, Related to Figure 1. A) Boxplots from human (**left**) and mouse (**right**) showing the expression of sperm-specific and sperm-oocyte expressed (SpOc) lincRNAs and PCGs in sperm and oocyte. Box plots represent low expression range (lower whiskers), higher expression range (upper whisker), median, inter quartile range (IQR) and the extreme expression values. **B-C**) Based on H3K27me3 and H3K4me3 enrichment, Sp- and SpOc-Other-ncRNAs promoters (extended ± 10 kb from transcription start site, TSS) from human (**B**) and mouse (**C**) were categorized into three optimal sperm derived chromatin clusters. **D**) Expression status of human Sp and SpOc-Other-ncRNAs from sperm derived chromatin clusters in sperm and oocyte RNA-seq samples.

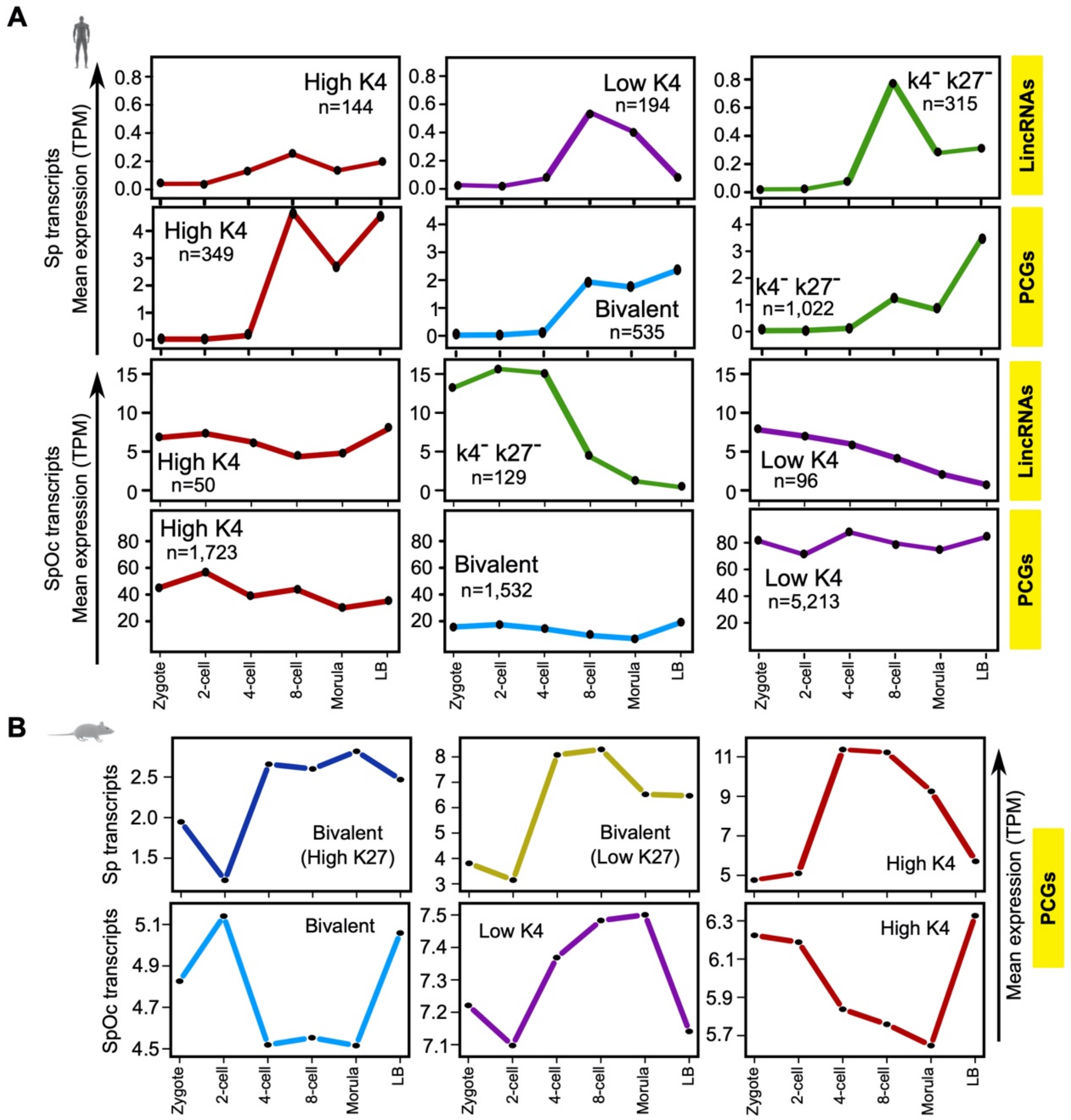


Figure S2. Sperm derived chromatin clusters show distinct stage-specific and non-overlapping gene expression during preimplantation development, Related to Figure 4. A-B) Sp- and SpOc transcripts (lincRNAs and PCGs) from sperm derived chromatin clusters showing cluster-wise expression during the preimplantation stages of developing embryo in human (**A**) and mouse (**B**). Mean expression is calculated from normalized TPM values.

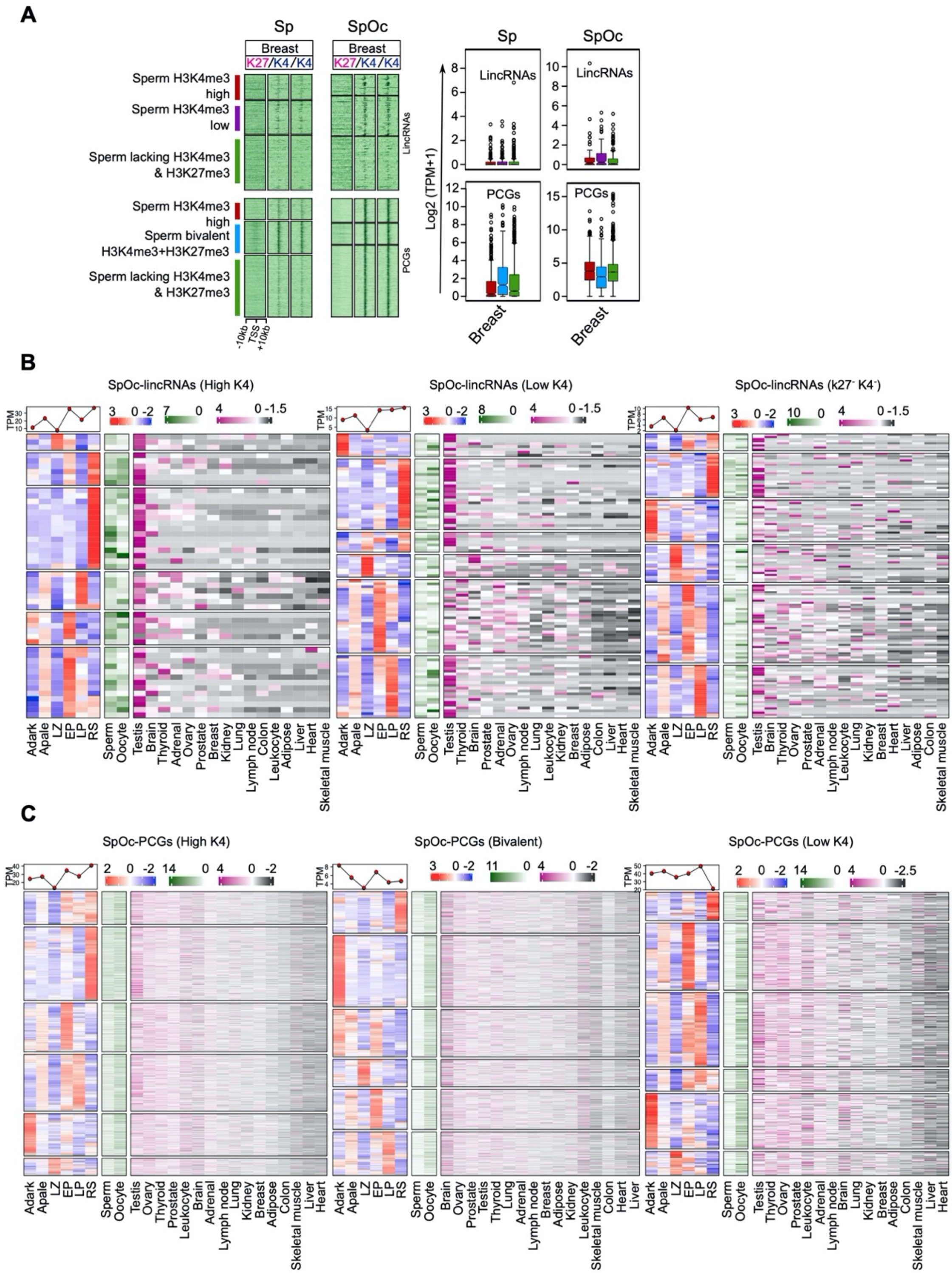


Figure S3. Sp and SpOc transcripts status in matured tissue and during spermatogenesis, Related to Figure 5 and Figure 6. A) Enrichment of H3K27me3 (K27) and H3K4me3 (K4) ChIP-seq signals over the promoters (± 10 kb) of Sp and SpOc lincRNAs and PCGs from sperm derived chromatin clusters in the ectoderm derived mature tissue (breast). Boxplots showing the expression of human Sp and SpOc transcripts (lincRNAs and PCGs) in normal breast tissue. **B-C**) Heatmaps showing the expression of SpOc-lincRNAs (**B**) and PCGs (**C**) from sperm derived chromatin clusters in spermatogenic cell types, sperm, oocyte and in 16 different tissues from the human body map 2.0.

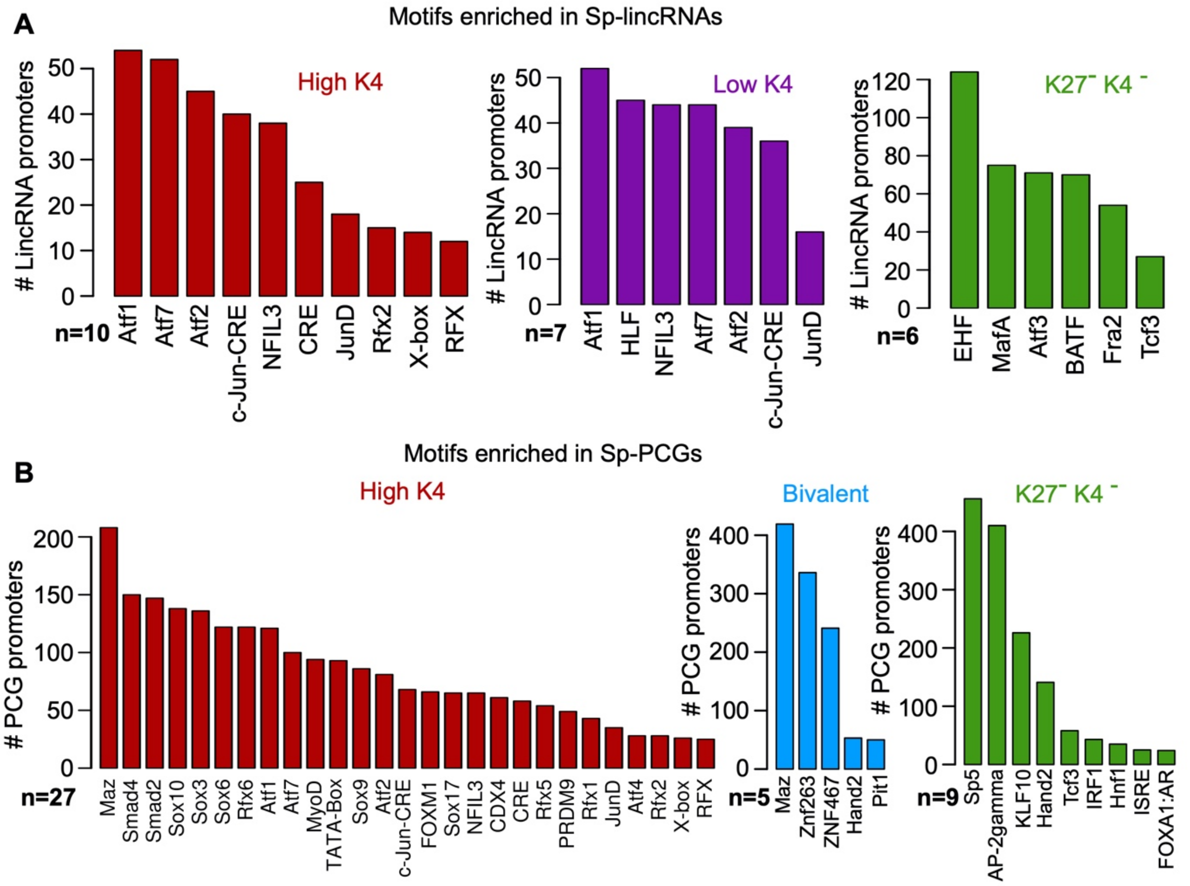


Figure S4. Frequency of transcription factor motifs in human Sp transcripts, Related to Figure 7. A-B) Bar graphs with the number of associated human Sp-lincRNA (**A**) and Sp-PCG (**B**) promoters from three sperm derived chromatin clusters of Sp transcripts enriched with the indicated transcription factor motifs.

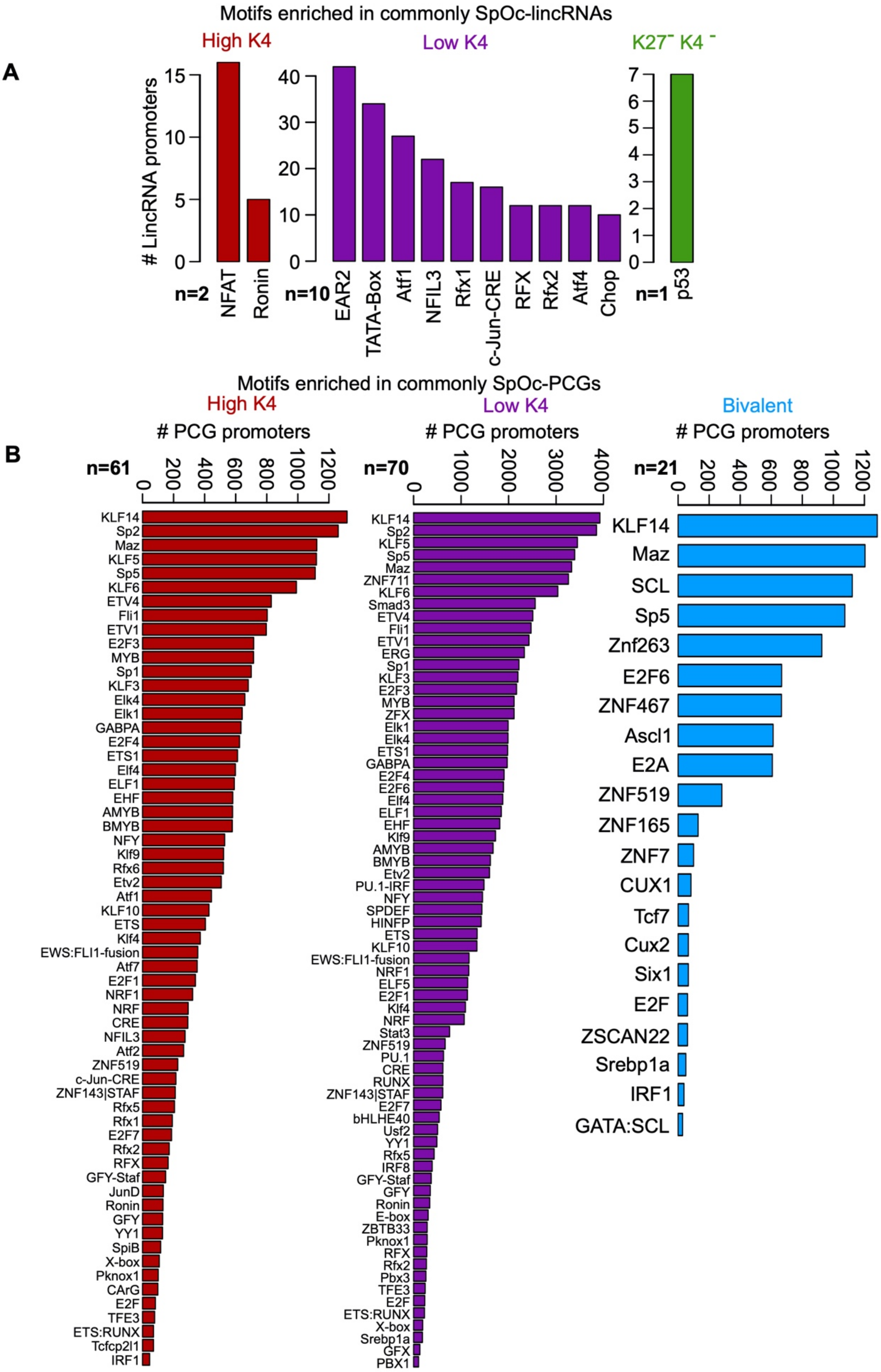


Figure S5. Frequency of transcription factor motifs in human SpOc expressed transcripts, Related to Figure 7. A-B) Bar graphs showing the number of associated human SpOc-lincRNA (**A**) and SpOc-PCG (**B**) promoters from three sperm derived chromatin clusters of SpOc transcripts enriched with the indicated transcription factor motifs.

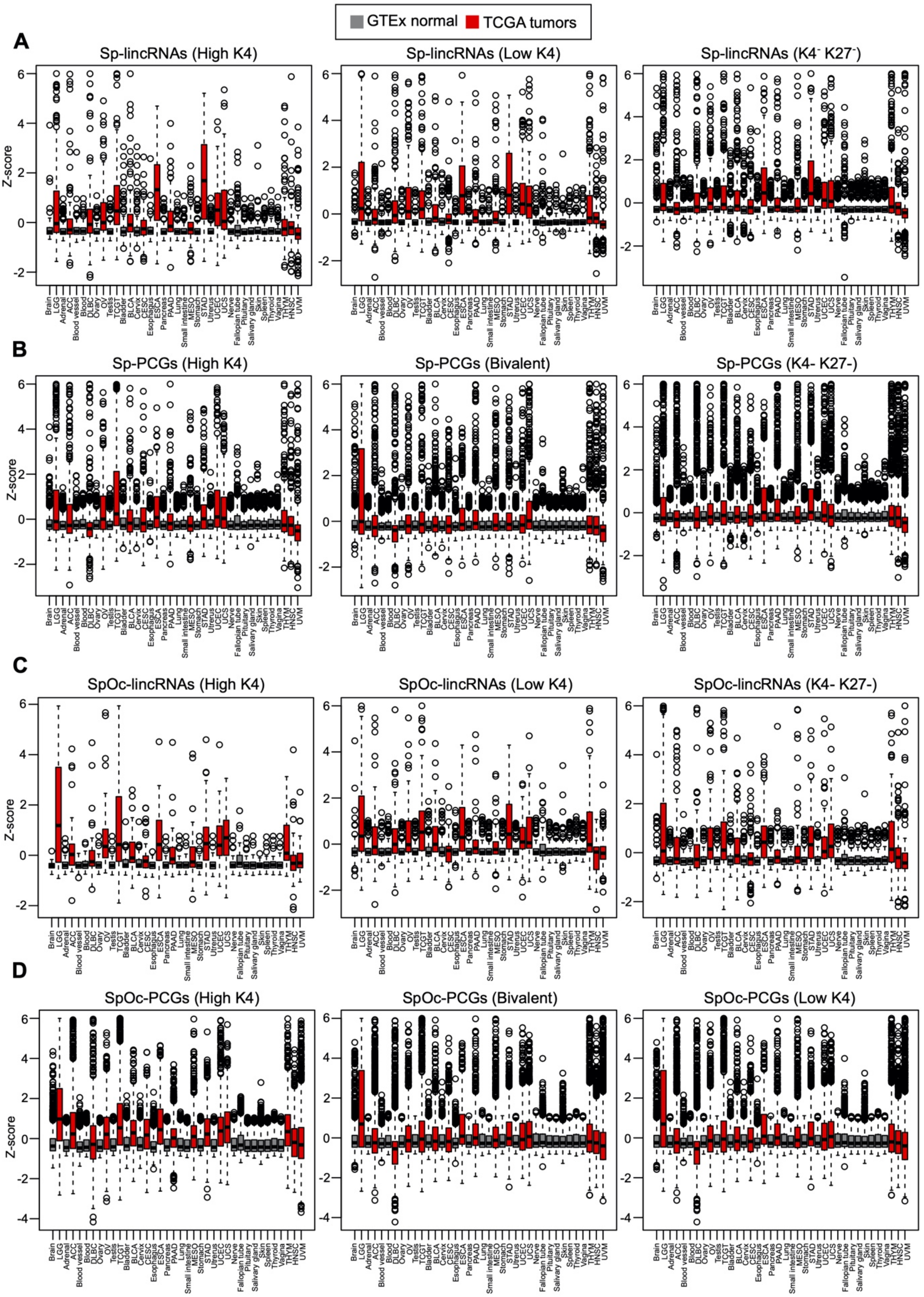


Figure S6. Aberrant expression of Sp and SpOc transcripts in different cancers, Related to Figure 8. A-D) Expression status of sperm derived chromatin clusters of Sp-transcripts (Friedman et al.) and SpOc-transcripts (**C-D**) in TCGA tumor patient cohorts and healthy samples from GTEx. The z-score in the plots were derived from the normalized TPM values.

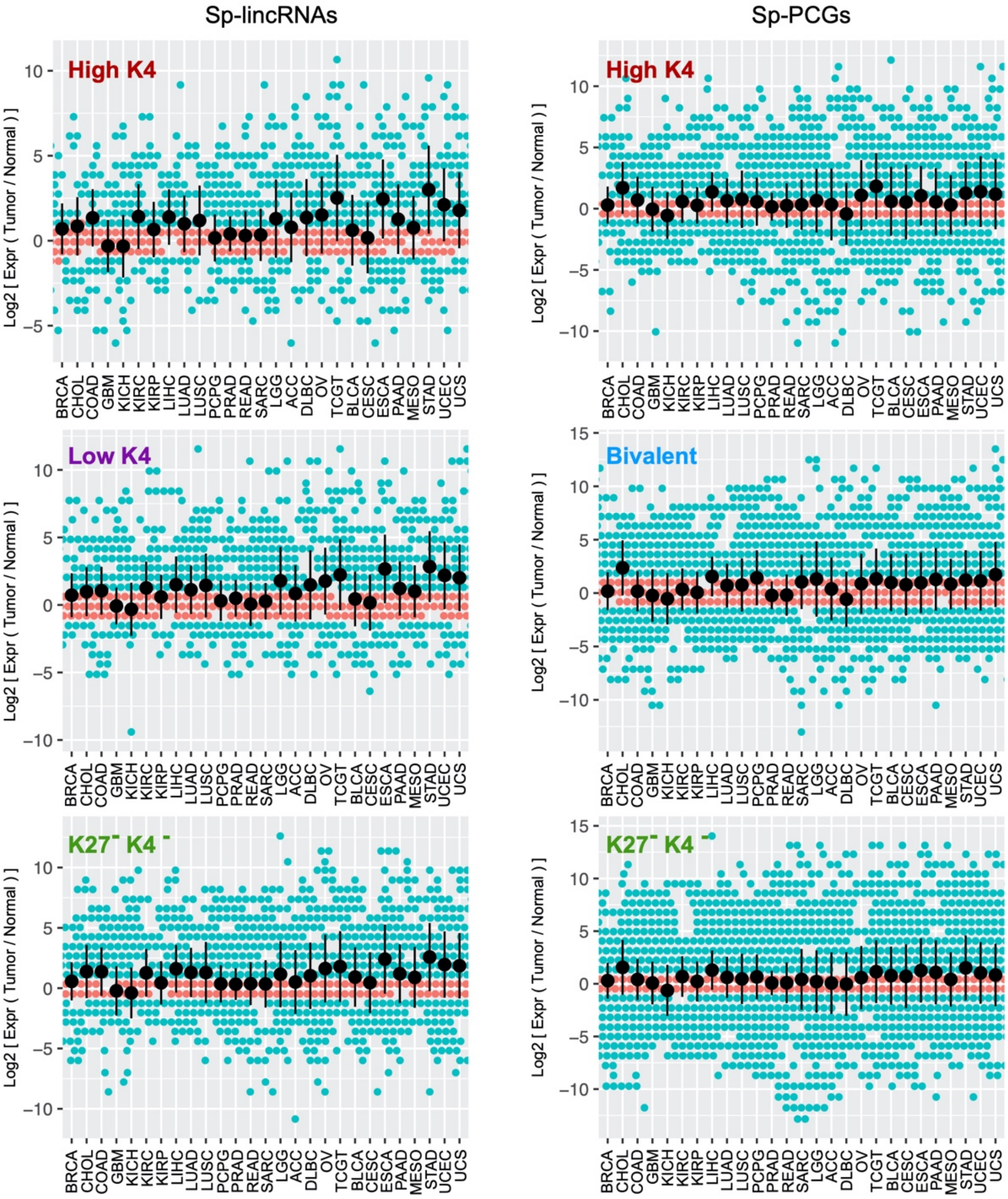


Figure S7. Directional regulation status of Sp transcripts in different cancers, Related to Figure 8. A-B) Fold differences of Sp-lincRNAs (**A**) and Sp-PCGs (**B**) from sperm derived chromatin clusters by comparing TCGA tumors with normal healthy tissues. The fold changes are calculated from normalized TPM values.

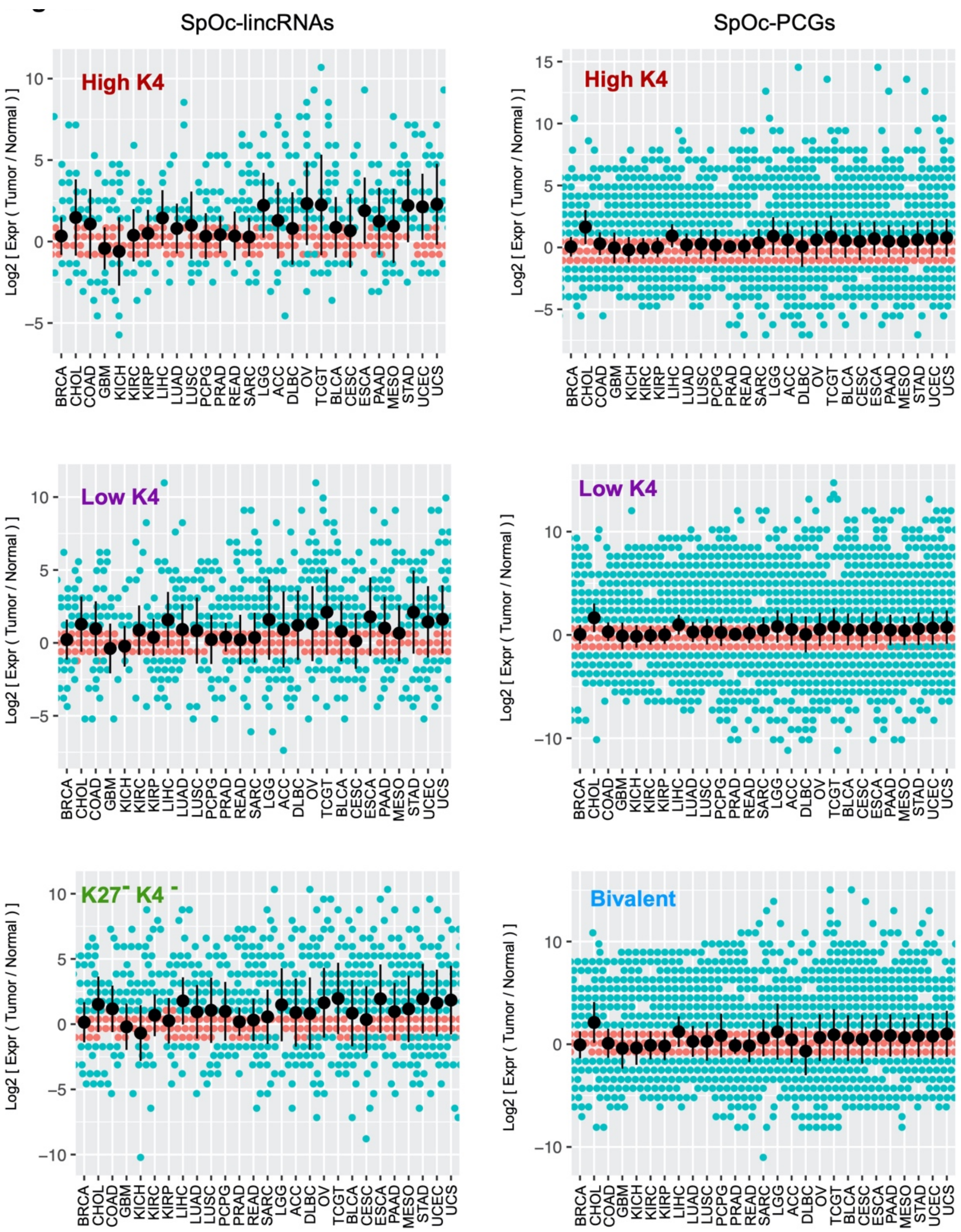


Figure S8. Directional regulation status of Sp transcripts in different cancers, Related to Figure 8. A-B) Fold differences of SpOc-lincRNAs (**A**) and SpOc-PCGs (**B**) from sperm derived chromatin clusters by comparing TCGA tumors with normal healthy tissues. The fold changes are calculated from normalized TPM values.

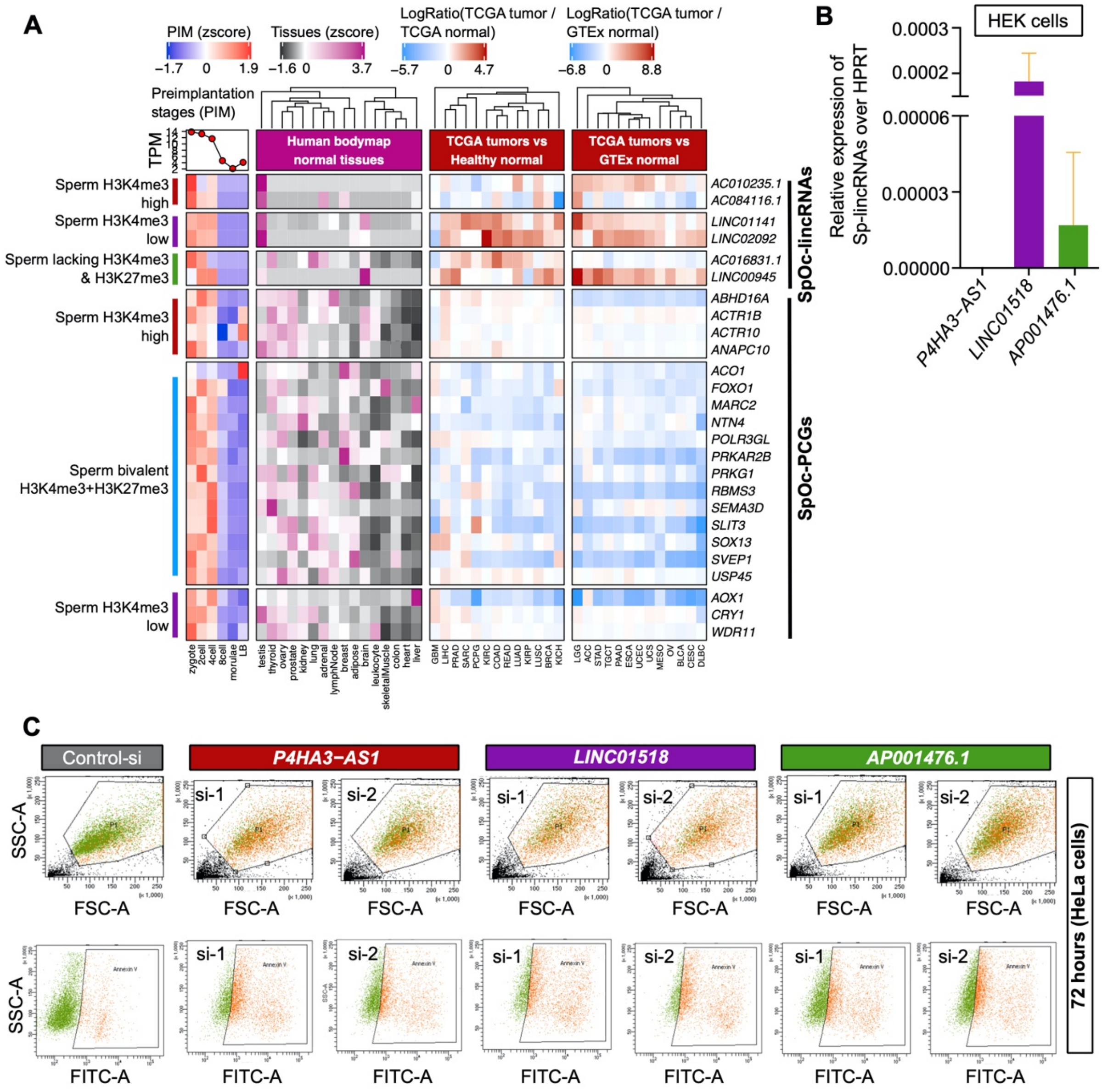


Figure S9. SpOc transcripts show tumor-specific activation, Related to Figure 9. A) Status of SpOc transcripts in germ cells, preimplantation stage embryos, the human body map tissues, TCGA tumors compared with the corresponding healthy samples or TCGA tumors compared with GTEx health samples. Z-score in the plots is derived from the normalized TPM expression values. The log fold change is calculated by comparing the expression of tumors with the health samples expression. **B**) Relative expression of Sp transcripts in HEK cell line. Data from barplot is represented as mean ± SD. **C**) Scatter plots showing the Hela cells stained with Alexa Fluor 488 annexin V conjugate, which is green Fluorescein (FITC) dye after 72hrs of transfection with siRNA for Sp lincRNAs and control siRNA samples. Single cells (green dots) were gated from debris (black dots) using a forward scatter area (FSC-A) versus side scatter area (SSC-A) gate (P1). Red colored dots represent Annexin V positive staining cells. Two different siRNAs were used for each gene (si1 and si2).

Table S4. siRNA sequence used in this study, Related to Figure 9 and Figure S9.

Oligo Name	Sequence (5' - 3')
AP001476.1_si1	CGAAGUCUCCUUGUCAAGA
AP001476.1_si2	CCGCUAUUUGGGCGGUGUU
P4HA3-AS1_si1	CCACACAAUAGUGAGAUAU
P4HA3-AS1_si2	GUGAGAUGUUAGGCCAUUA
LINC01518_si1	GGGAUCAUGGCCAAAGUUG
LINC01518_si2	GGCAGUGACGGAACAGUAC

Transparent Methods

Data collection

We collected 17,705 RNA-seq and ChIP-seq datasets from different studies, GEO, ENCODE, dbGAP, GTEx and TCGA. The detailed list of samples and their sources are provided in **Table S1**. Datasets are chosen after quality check using FastQC (RNA-seq and ChIP-seq), SAMstat(Lassmann et al., 2011) (RNA-seq and ChIP-seq) and deepTools(Ramirez et al., 2016) fingerprints (ChIP-seq vs input). Transcriptome data for different cancers are downloaded as raw read counts using Bioconductor package TCGAbiolinks(Colaprico et al., 2016; Mounir et al., 2019; Silva et al., 2016) and the raw read count data for tissues used for comparison with TCGA patients are downloaded from GTEx consortium.

Processing transcriptome dataset (RNA-seq)

All downloaded RNA-seq samples were first subjected to quality check with FastQC before and after adaptor removal using cutadapt (if necessary). Samples were considered for further analysis only when it passes FastQC quality check. We used HISAT2 (v2.1.0)(Kim et al., 2019) to align reads to hg38 (human) or mm10 (mouse) with parameters *'--dta --avoid-pseudogene --known-splicesite- --qc-filter'* and *'--rna-strandness'* if the sequencing samples are paired-end. The known splice sites are generated from Ensembl annotation(Yates et al., 2019) GRCh38.90 (human) or GRCm38.90 (mouse) using 'hisat2_extract_splice_sites.py' and the genomes are indexed using 'hisat2-build'. The generated aligned SAM files are sorted and converted into BAM files with the help of SAMtools (v 1.5)(Li et al., 2009). Alignment quality and statistics are obtained using SAMstat. For assigning reads to individual transcripts (genes or RNAs) from Ensembl annotation GRCh38.90 (human) or GRCm38.90 (mouse), featureCounts(Liao et al., 2014) from Subread Bioconductor package was employed using parameters *'-Q 30 -t exon --ignoreDup -J –minOverlap -C'* and *'-B -s'* for paired-end and strand-specific samples. Obtained raw read counts for individual transcripts are normalized as Transcripts Per Kilobase Million (TPM) by using library size and length of exons. If necessary, batch effects were removed from normalized reads using 'removeBatchEffect' from Bioconductor package.
According to Hammoud SS et al 2014, both the sperm donors were similar in age (D1-32 and D2-37), both white-Caucasian, comparable semen parameters (progressive motility ~65%, sperm count ~80x10^6). Oocytes samples were obtained from Yan L et al 2013 in which the couples who had more than 20 oocytes derived from the same IVF cycle were sequenced. Embryos that were produced by routine fertilization were cultured individually. The women had an average age of 30 years (25–35 years). Detailed information on the studies associated with entire dataset used in the current analysis can be found in **Table S1**.

Processing chromatin immunoprecipitation data (ChIP-seq)

After FastQC quality check, raw reads from ChIP-seq were aligned using BWA (v 0.7.17) aligner(Li and Durbin, 2009) and obtained SAM files are aligned and converted into BAM using SAMtools (v 1.5). De-duplication of reads was done using Picard MarkDuplicates (v 2.18.11)(McKenna et al., 2010) and the known blacklisted regions are removed using BEDTools '*intersect*'(Quinlan and Hall, 2010). Additionally, we removed low quality reads having mapping quality less than 30 using *bamutils* from NGSUtils(Breese and Liu, 2013). Finally, the BAM files with no duplicate reads and higher mapping quality were used for downstream analysis. To check quality of enrichment in ChIP-seq, we compared read coverage of ChIP-seq against the input/control samples using ‘plotFingerprint’ from deepTools package (v 2.5.1). Only samples with higher enrichment were used in bamCompare (deepTools) to obtain base-by-base resolution of genome-wide fold enrichment (log2ratio/enrichment) by comparing ChIP-seq with input samples *(--scaleFactorsMethod readCount --ratio log2 --pseudocount 1*). These base-by-base enrichment values are plotted for promoters of genes using’ plotHeatmap’ from deepTools. Promoters of Sp and SpOc transcripts are clustered by chromatin profiles using k-means clustering and tested up to 10 clusters to find optimal number of clusters. We optimized the cluster numbers by considering only non-repetitive combination (H3K27me3 and H3K4me3) of clusters. The protein coding genes (PCGs) from individual clusters are used for further functional enrichment analysis. We used GeneSCF(Subhash and Kanduri, 2016) to derive enriched biological process from individual clusters by using gene ontology from human and mouse. Detailed information on the studies associated with entire dataset used in the current analysis can be found in **Table S1**.

Motif analysis of Sp transcript promoters

Promoter sequences of Sp and SpOc transcripts are extracted by considering ± 250 base-pairs (bp) from transcription start site (TSS). The FASTA sequences of these promoters were extracted using ‘*faidx*’ from SAMtools (v 1.5). To find transcription factor motif sequences or binding sites on this promoter, we used ‘*findMotifs.pl*’ (*-mset vertebrates -homer2*) from HOMER (v 4.10) package(Duttke et al., 2019). HOMER matches provided sequences with known transcription factor motifs predicted from published ChIP-seq datasets and ranks by its significance. We considered motif to be enriched only if it has p-value < 0.01. Also, top enriched motifs from Sp and SpOc transcript promoters are selected by highest number of target genes having particular motif binding site.

TCGA data access and processing

The raw read counts for TCGA patient data and its corresponding controls were downloaded using TCGAbiolinks. Obtained raw reads are normalized to TPM and calculated average expression of patients and normal sample groups per gene.

RNA extraction, cDNA synthesis and qRT-PCR

Total RNA was isolated from cells using Relia RNA isolationkit (Promega) following manufacturer’s directions. cDNA synthesis was carried out using ImProm-II Reverse Transcriptase (Promega). RT-qPCR analysis was done using Power SYBR Green PCR master mix (Applied Biosystems, Warrington, UK). Differences in expression were calculated using the ΔΔCt method using HPRT as control gene. The primers used for qRT-PCR are 5’AAGAGCGGCAAGAGGACAG3’ and 5’ GGACTATGCAGTTCCTTCCTG3’ for *P4HA3-AS1*; 5’ ACTGAAGCTGGTGGCTGTG3’ and 5’ CATGGACTCGAGAGCTGACA3’ for *LINC01518* and 5’ ACTGCCACTCGGACTGTCTC3’ and 5’ GAAAGGCACTCTGTGGTCGT3’ for *AP001476.1.* Two Custom designed small interfering RNAs (siRNAs) for each lncRNA were used for transfections along with control siRNA (Invitrogen).

MTT assay and Cell cycle analysis

Percentage of cell proliferation was analyzed using MTT assay after 48 h of post transfection according to the manufacturer’s protocol using CellTiter-Glo® 3D Cell Viability Assay (G9681, Promega Madison, USA). Proliferation capacity of HeLa cells as measured based on the Ultra-Glo Recombinant Luciferase, which generates a stable luminescent signal measured using Clariostar Plus Microplate Reader (BMG labtech). The error bars were calculated based on two independent transfections. The cell cycle profiling was performed using NucleoCounter NC-3000 platform (Chemometec, Denmark). The cells were fixed using absolute ethanol after transfection and were stained with DAPI solution provided by the manufacturer and analyzed according to manufacturer’s instructions. All siRNA sequences used in this study are listed in **Table S4**.

Apoptosis assay

Hela cells are stained with Alexa Fluor 488 annexin V conjugate, which is green Fluorescein (FITC) dye after 48hrs and 72hrs of transfection with siRNA for Sp-lincRNAs and control siRNA samples. Single cells were gated from debris using a forward scatter area (FSC-A) versus side scatter area (SSC-A) gate (P1). We measured Annexin V positively stained cells in cells transfected with two different siRNAs for each Sp-lincRNAs (si1 and si2) along with corresponding control siRNAs. All siRNA sequences used in this study are listed in **Table S4**.

Supplemental References

1. Breese, M.R., and Liu, Y. (2013). NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. Bioinformatics *29*, 494-496.
2. Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I*., et al.* (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res *44*, e71.
3. Duttke, S.H., Chang, M.W., Heinz, S., and Benner, C. (2019). Identification and dynamic quantification of regulatory elements using total RNA. Genome Res *29*, 1836-1846.
4. Friedman, C.E., Nguyen, Q., Lukowski, S.W., Helfer, A., Chiu, H.S., Miklas, J., Levy, S., Suo, S., Han, J.J., Osteil, P*., et al.* (2018). Single-Cell Transcriptomic Analysis of Cardiac Differentiation from Human PSCs Reveals HOPX-Dependent Cardiomyocyte Maturation. Cell Stem Cell *23*, 586-598 e588.
5. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol *37*, 907-915.
6. Lassmann, T., Hayashizaki, Y., and Daub, C.O. (2011). SAMStat: monitoring biases in next generation sequencing data. Bioinformatics *27*, 130-131.
7. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.
8. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.
9. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923-930.
10. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M*., et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res *20*, 1297-1303.
11. Mounir, M., Lucchetta, M., Silva, T.C., Olsen, C., Bontempi, G., Chen, X., Noushmehr, H., Colaprico, A., and Papaleo, E. (2019). New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. PLoS Comput Biol *15*, e1006701.
12. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.
13. Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dundar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res *44*, W160-165.
14. Silva, T.C., Colaprico, A., Olsen, C., D'Angelo, F., Bontempi, G., Ceccarelli, M., and Noushmehr, H. (2016). TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. F1000Res *5*, 1542.
15. Subhash, S., and Kanduri, C. (2016). GeneSCF: a real-time based functional enrichment tool with support for multiple organisms. BMC Bioinformatics *17*, 365.
16. Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R*., et al.* (2019). Ensembl 2020. Nucleic Acids Res.