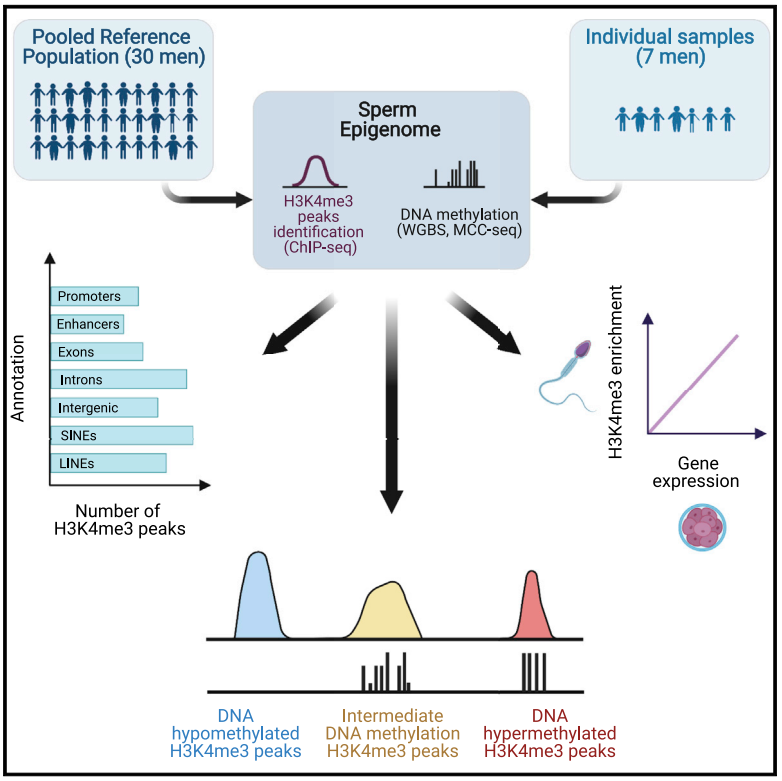
Cell Reports

Whole-genome sequencing of H3K4me3 and DNA methylation in human sperm reveals regions of overlap linked to fertility and development

Graphical abstract



Highlights

- Deep sequencing identifies regions bearing H3K4me3 in human sperm
- H3K4me3 is widely distributed in the genome and is enriched at promoters and SINEs
- Sperm H3K4me3 marks developmental genes and correlates with embryonic gene expression
- In addition to DNA hypomethylated loci, H3K4me3 also overlaps hypermethylated regions

Authors

Romain Lambrot, Donovan Chan, Xiaojian Shao, ..., Jacquetta Trasler, Vanessa Dumeaux, Sarah Kimmins

Correspondence

vanessa.dumeaux@concordia.ca (V.D.), sarah.kimmins@mcgill.ca (S.K.)

In brief

Lambrot et al. use H3K4me3 ChIP-seq and WGBS to profile the sperm epigenome of Canadian men. Deep sequencing reveals regions enriched for H3K4me3 that overlap with DNA methylation, suggesting a functional relationship in sperm.

Lambrot et al., 2021, Cell Reports *36*, 109418
July 20, 2021
https://doi.org/10.1016/j.celrep.2021.109418

CellPress

Cell Reports

CellPress
OPEN ACCESS

Resource

Whole-genome sequencing of H3K4me3 and DNA methylation in human sperm reveals regions of overlap linked to fertility and development

Romain Lambrot,[1] Donovan Chan,[2] Xiaojian Shao,[3] Mahmoud Aarabi,[4,5] Tony Kwan,[6] Guillaume Bourque,[6] Sergey Moskovtsev,[7,8] Clifford Librach,[7,8] Jacquetta Trasler,[2,6,9,10] Vanessa Dumeaux,[11,12,*] and Sarah Kimmins[1,9,13,*]

[1]Department of Animal Science, McGill University, Montreal, QC, Canada
[2]Research Institute of the McGill University Health Centre, Montreal, QC, Canada
[3]Digital Technologies Research Centre, National Research Council, Ottawa, ON, Canada
[4]Department of Laboratory Medicine and Pathobiology, North York General Hospital, University of Toronto, Toronto, ON, Canada
[5]Department of Obstetrics, Gynecology and Reproductive Sciences, University of Pittsburgh, Pittsburgh, PA, USA
[6]Human Genetics, McGill University, Montreal, QC, Canada
[7]CReATe Fertility Centre, Toronto, ON, Canada
[8]Department of Obstetrics and Gynecology, University of Toronto, Toronto, ON, Canada
[9]Department of Pharmacology and Therapeutics, McGill University, Montreal, QC, Canada
[10]Department of Pediatrics, McGill University, Montreal, QC, Canada
[11]PERFORM Center, Concordia University, Montreal, QC, Canada
[12]Department of Biology, Concordia University, Montreal, QC, Canada
[13]Lead contact
*Correspondence: vanessa.dumeaux@concordia.ca (V.D.), sarah.kimmins@mcgill.ca (S.K.)
https://doi.org/10.1016/j.celrep.2021.109418

SUMMARY

The paternal environment has been linked to infertility and negative outcomes. Such effects may be transmitted via sperm through histone modifications. To date, in-depth profiling of the sperm chromatin in men has been limited. Here, we use deep sequencing to characterize the sperm profiles of histone H3 lysine 4 tri-methylation (H3K4me3) and DNA methylation in a representative reference population of 37 men. Our analysis reveals that H3K4me3 is localized throughout the genome and at genes for fertility and development. Remarkably, enrichment is also found at regions that escape epigenetic reprogramming in primordial germ cells, embryonic enhancers, and short-interspersed nuclear elements (SINEs). There is significant overlap in H3K4me3 and DNA methylation throughout the genome, suggesting a potential interplay between these marks previously reported to be mutually exclusive in sperm. Comparisons made between H3K4me3 marked regions in sperm and the embryonic transcriptome suggest an influence of paternal chromatin on embryonic gene expression.

INTRODUCTION

Currently one in six couples experience infertility (Smith et al., 2011), with a male factor attributed as causal in up to half of those cases (Whitman-Elia and Baxley, 2001). In the past 40 years, sperm counts have declined by 50% (Levine et al., 2017), suggesting that an increased reliance on assisted reproductive technologies (ARTs) is imminent. Dependence on ART to achieve pregnancy has inherent health risks for women and babies and negative socio-economic consequences (Chen and Heilbronn, 2017). Although the factors contributing to the decline in sperm quality and quantity are poorly understood, environment-epigenome interactions are likely. Exposure to toxicants, diet, and elevated BMIs have been associated with alterations in the sperm epigenome and reduced fertility in clinical populations of men (Aarabi et al., 2015; Donkin and Barrès, 2018; Wu et al., 2017). However, the focus of those studies has been predominantly at the level of DNA methylation, and little is known regarding the relationship between chromatin and DNA methylation in sperm.

For the purposes of this study, we considered epigenetic inheritance to refer to the biochemical layer of information that exists beyond the DNA sequence. This includes histone and DNA modification, in sperm that have the potential to influence gene expression in the next generation. The sperm epigenome is unique in comparison with that of the somatic cell. In the final stages of spermiogenesis, 85%–90% of histones are replaced by sperm-specific proteins: protamines (Balhorn, 2007; Brykczynska et al., 2010; Hammoud et al., 2009; Jung et al., 2017). The sites of histone retention in the sperm genome from men have yet to be unequivocally established. The first studies in mice and men determined that histones were preferentially retained at CpG-rich regions and that those regions predominantly lacked DNA methylation (Erkek et al., 2013; Hammoud et al.,

2009). In contrast, other studies in sperm found that histones localized to gene-poor regions (Carone et al., 2014), specifically at distal intergenic sequences and retrotransposons (Samans et al., 2014). These opposing findings may be the consequence of experimental and data analysis differences (Carone et al., 2014; Royo et al., 2016; Samans et al., 2014) and require resolution. In mice, our studies and those from others demonstrated the presence of the gene-activating histone H3 lysine 4 di- and tri-methylation (H3K4me2/3) throughout the genome with strong enrichment at gene promoters (Jung et al., 2017; Lambrot et al., 2019; Lismer et al., 2020; Siklenka et al., 2015; Yamaguchi et al., 2018). In mice, histone modifications and DNA methylation have been implicated in the paternal transmission of environmentally driven phenotypes to their offspring (Lambrot et al., 2013; Lismer et al., 2021; Ly et al., 2017; Radford et al., 2014). Genetic alteration of histone methylation H3K4me2 and H3K4me3 profiles in sperm leads to changes in embryonic gene expression and birth defects that are transmitted across generations (Lismer et al., 2020; Siklenka et al., 2015). Paternal H3K4me3 is transmitted to the embryo, and diet-induced alterations in sperm H3K4me3 profiles are associated with disturbed embryonic gene expression and birth defects in offspring (Lismer et al., 2021). These mouse models indicate that H3K4me3 is responsive to the environment and is linked to embryonic development and offspring health. Whether similar functions and H3K4me3 profiles are conserved in men is not known, and there are limitations in existing data on H3K4me3 because it is derived from only a few men (Hammoud et al., 2009, 2011; Oikawa et al., 2020).

The DNA methylation pattern of sperm is unique in comparison with that of somatic cells; CpG islands located at promoters are mostly unmethylated, with the remainder of the genome being hyper-methylated, with the exception of specific repetitive elements (Molaro et al., 2011). We recently performed deep whole-genome bisulfite sequencing (WGBS) using sperm from 30 men and identified greater complexity in the sperm methylome than has previously been reported (Chan et al., 2019). We newly identified CpGs that were modified by perturbations to folate metabolism and with folate supplementation; most altered CpGs were found at sites that demonstrated intermediate methylation levels (i.e., 20%–80%), suggesting that these regions may be "dynamic," that is, responsive, in general, to the environment (Chan et al., 2019). In addition, those data suggested that, in contrast to previous data on DNA methylation in sperm, there may be an overlap with DNA methylation and H3K4me3. Previously, it has been reported in mice and men that regions enriched for H3K4me3 in sperm are mutually exclusive with DNA methylation (Erkek et al., 2013; Hammoud et al., 2009). We wanted to revisit that assumption because there are limitations in prior studies in the coverage and depth of the sperm methylome. For example, reduced-representation bisulfite sequencing (RRBS) and the 450K array are biased toward CpG regions, which tend to be poor in DNA methylation (Aarabi et al., 2015; Dere et al., 2016; Guo et al., 2014; Krausz et al., 2012), and prior WGBS has been low coverage (Molaro et al., 2011). Given the limitations of these previous approaches and prior reports on DNA methylation and H3K4me3 interplay in adipocytes and cancer cells (Dunican et al., 2020; Matsumura et al., 2015), a similar relationship may have been overlooked in sperm.

In this study, we investigated the location of H3K4me3 in sperm and its relationship to sperm DNAmethylation profiles in the same samples. High-depth sequencing newly revealed that H3K4me3 is present throughout the genome in sperm; is enriched at promoters, human embryonic stem cell (hESC) enhancers, and short-interspersed nuclear elements (SINEs); and is associated with gene expression in the embryo. We confirmed that H3K4me3 and DNA methylation are mutually exclusive at many promoters. However, within genomic regions bearing intermediate and hypermethylated DNA, there is an overlap with H3K4me3 peaks, including at genes implicated in reproductive processes important for fertility, such as gamete generation, but also in cell cycle and cell metabolic processes.

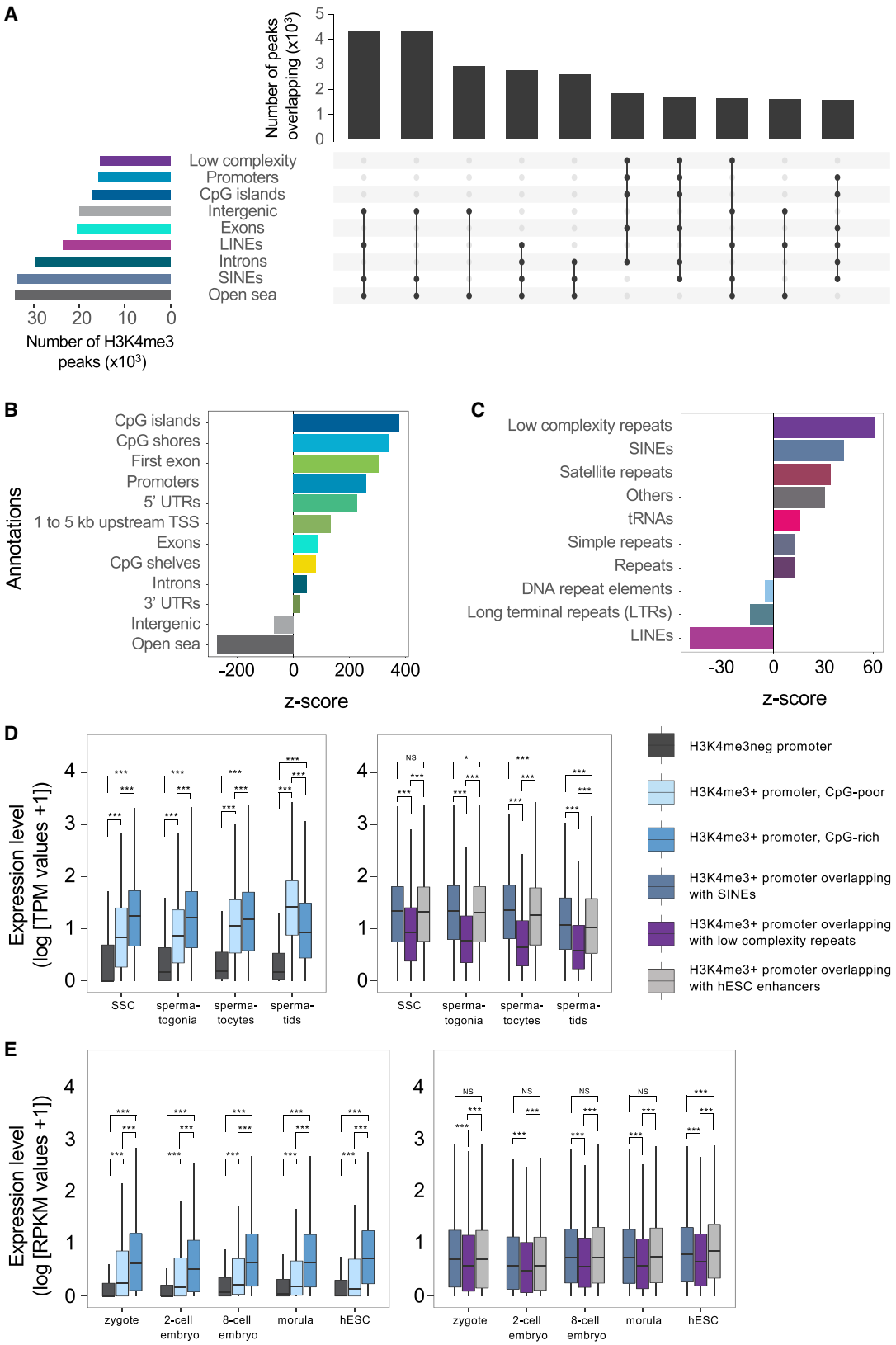
RESULTS

Histone H3K4me3 in sperm is enriched in genes and regulatory regions

To determine the H3K4me3-enrichment profile in sperm, we performed chromatin immunoprecipitation (ChIP), followed by deep sequencing on a purified, pooled sample of sperm from a reference population of 30 Canadian men, and seven individual samples (Tables S1 and S2; STAR Methods). ChIP sequencing (ChIP-seq) for H3K4me3 in the pooled sample yielded a robust dataset with more than 153 million reads; of which, 95% mapped to the human reference genome (Table S3). Using the peak-caller MACS2 (Zhang et al., 2008) with broad peak settings, we obtained a sperm profile of >50,000 peaks distributed throughout the genome (Figures 1A and S1). The individual sperm samples had between 29 and 53 million reads, and the pooled sample from four men used by Hammoud et al., 2009, had 18 million reads (Table S3). Our individual samples and the pooled sample from Hammoud et al. (2009) showed a maximum number of 21,000 peaks (Figure S1). The improvement in peak calling in our pooled sample versus that of the individual samples can be attributed to deeper sequencing, which yielded an increased signal-to-noise ratio, allowing us to detect H3K4me3 peaks in previously unreported functional regions of the genome (Figure 1) and to define peak boundaries more accurately (Figure S1). To compare the similarity in H3K4me3 profiles between the pooled sample and the individual samples, a Spearman correlation-coefficient analysis was used (Figure S2). There was a strong correlation between the reference population and the seven individual samples (Figure S2). This was based on the ChIP-seq signal of each sample expressed as reads per kilobase of exon per million (RPKM) fragments mapped, at the peaks identified in the pooled sample. This analysis indicates that the regions enriched for H3K4me3 in the pooled sample are also present in the individuals, a similarity confirmed by comparing the H3K4me3 profiles locally on genome-browser tracks (Figure S3A) and genome-wide on heatmaps (Figure S3B).

In sperm, H3K4me3 peaks were broad and ranged from 0.2 to 50 kb, with some peaks, because of their breadth, spanning several genomic compartments (Figure 1A). For example, some H3K4me3 peaks occurring in promoter regions

CellPress
OPEN ACCESS



(legend on next page)

overlapped with introns, exons, and regions inclusive of long-interspersed nuclear elements (LINEs) and SINEs (Figure 1A). Enrichment analysis for specific annotations showed that H3K4me3 peaks were significantly enriched in locations near the transcription start site (TSS), the promoter (1 kb upstream of the TSS), the first exon, and the 5′ untranslated region (5′ UTR) (Figure 1B). Most gene promoters have a high CpG content (Saxonov et al., 2006), and in mice, histones have previously been reported to be preferentially retained in sperm at CpGs (Erkek et al., 2013). This analysis confirms that conservation between mice and men, because H3K4me3 peaks were enriched at CpG dense regions, such as CpG islands and shores, but were not enriched at the CpG-poor open sea (Figures 1A and 1B). Notably, fewer H3K4me3 peaks were detected in intergenic regions than would be expected (Z score < 0), indicating a preferential genic retention of H3K4me3 in sperm (Figure 1B). Previously unreported is that H3K4me3 peaks were significantly enriched at some repetitive elements, in particular, SINEs and low-complexity repeats (Figures 1A–1C).

H3K4me3-bearing promoters that contain SINEs, low-complexity repeats, or embryonic enhancers are associated with differential embryonic processes

Whether H3K4me3 distribution in sperm is associated with a specific promoter architecture to drive gene expression programs particular to sperm or embryo development has not been explored in depth. Genomic repeats may influence gene expression at both transcriptional and post-transcriptional levels through *cis* and *trans* mechanisms and participate in the regulation of diverse biological and pathological processes (Lu et al., 2020). Here, we further explore potential *cis*-regulatory differences conferred by CpG content, repeat subclasses, or putative enhancers contained in promoters marked by H3K4me3 in sperm that might preferentially affect spermatogenic and embryonic gene expression. H3K4me3 peaks were first split by whether they overlapped a promoter or not. Gene promoters were further categorized by CpG content (poor versus rich), SINEs (with or without low complexity repeats), low-complexity repeats (with or without SINEs), and hESC enhancers (Figures 1D and 1E). We then analyzed the top gene functions of those categories (Figure S4; for full list, Table S4) and examined their expression in spermatogenic cells and at distinct embryonic stages (Figures 1D and 1E).

Genes that had promoters not bearing H3K4me3 (7,443 genes) had low expression (Figure 1D) and were predominantly not associated with spermatogenesis and early embryo development (Figure S4). For example, H3K4me3-negative promoters included genes implicated in nervous system function, such as sensory stimulus detection and processing. In contrast, gene promoters enriched in H3K4me3 but CpG-poor (1,565 genes) were highly expressed, particularly during meiosis and spermiogenesis (Figure 1D). These genes were associated with spermatogenesis and, more specifically, with gamete formation, sperm function, and fertility (Figure S4). These included key spermatogenic factors, such as *DDX4* and *FOXO1*; genes implicated in meiosis, such as *MAEL* and *MEI4*; and spermiogenesis genes, such as *CYLC2*, *EQTN*, and *CAV1/2*. Interestingly, we previously demonstrated, in mice, that genes gaining H3K4me2 during spermiogenesis conserved that mark in sperm (Lambrot et al., 2019). Similarly, our analysis shows that human sperm retains H3K4me3 at genes expressed during spermatogenesis (Figure 1D). Genes regulated through promoters that are CpG-rich bearing H3K4me3 (14,875) were the vast majority and are implicated in a wide variety of cellular processes, including gene regulation, cell cycle, cellular transport, and chromosome organization (Figures 1D and S4).

About 72% of the gene promoters enriched for H3K4me3 in sperm contained a SINE. We found that they were significantly more expressed in spermatogenic cells compared with any other promoter categories, suggesting a new role for SINEs in regulation of gene expression during spermatogenesis (Figure 1D). These highly expressed genes were involved in basic cellular processes (Figure S4). In contrast to SINE-bearing promoters, those containing low-complexity repeats were the least expressed in spermatogenic cell types (Figure 1D) and included genes implicated in the nervous system and developmental processes (Figure S4), suggesting that these genes were marked for embryonic expression.

Next, we explored the connection between promoters marked by H3K4me3 in sperm to embryonic gene expression, in relation to promoter CpG content, low-complexity repeats, and SINEs. To do so, we compared human embryo gene-expression data from single-cell RNA sequencing (RNA-seq) and hESCs (Yan et al., 2013) to H3K4me3 enrichment at promoters in sperm. The genes marked in sperm by H3K4me3 with a CpG-poor promoter had greater expression than genes not marked by

Figure 1. H3K4me3 is enriched at CpG-rich regions, promoters, and SINEs in human sperm

ChIP-seq for H3K4me3 was performed on a sample prepared from the sperm of 30 men, and peaks were called using MACS2.

(A) Location of the H3K4me3 peaks based on genomic and CpG annotations and overlap with repetitive elements and low-complexity DNA sequences. Annotations were obtained from the Bioconductor package *annotatr* (Cavalcante and Sartor, 2017) and RepeatMasker (http://www.repeatmasker.org/). The overlap of peaks between different annotations is illustrated by connecting nodes, and the number of overlapping peaks is displayed on the bar graph shown above them. The top-10 overlaps are indicated.

(B) Enrichment in H3K4me3 peaks for each specific annotation. Positive and negative enrichments are indicated by Z scores were determined with the Bioconductor package *regioneR* (Gel et al., 2016). For all annotations displayed, $p < 0.001$ by permutation tests.

(C) Enrichment in H3K4me3 peaks for each repeat annotation; $p < 0.001$ by permutation tests.

(D and E) Boxplots showing the level of expression of genes with promoters bearing no H3K4me3 in sperm, marked by H3K4me3 at promoters located in CpG-poor or -rich regions, and promoters overlapping with SINEs, low-complexity repeats, or hESCs enhancers during spermatogenesis (D) and development (E). The level of transcription was visualized based on the transcript abundance expressed as transcripts per million (TPM) for spermatogenesis in spermatogonial stem cells (SSCs), spermatogonia (mitotic phase), spermatocytes (meiosis), and spermatids (spermiogenesis) or reads per kilobase of exon per million (RPKM) fragments mapped for development. The p values were determined by pairwise Wilcoxon rank-sum tests with a Benjamini-Hochberg correction for multiple testing. NS indicates a non-significant adjusted p value. *adjusted $p < 0.05$, ***adjusted $p < 0.001$. See also Figures S1–S4 and Table S1.

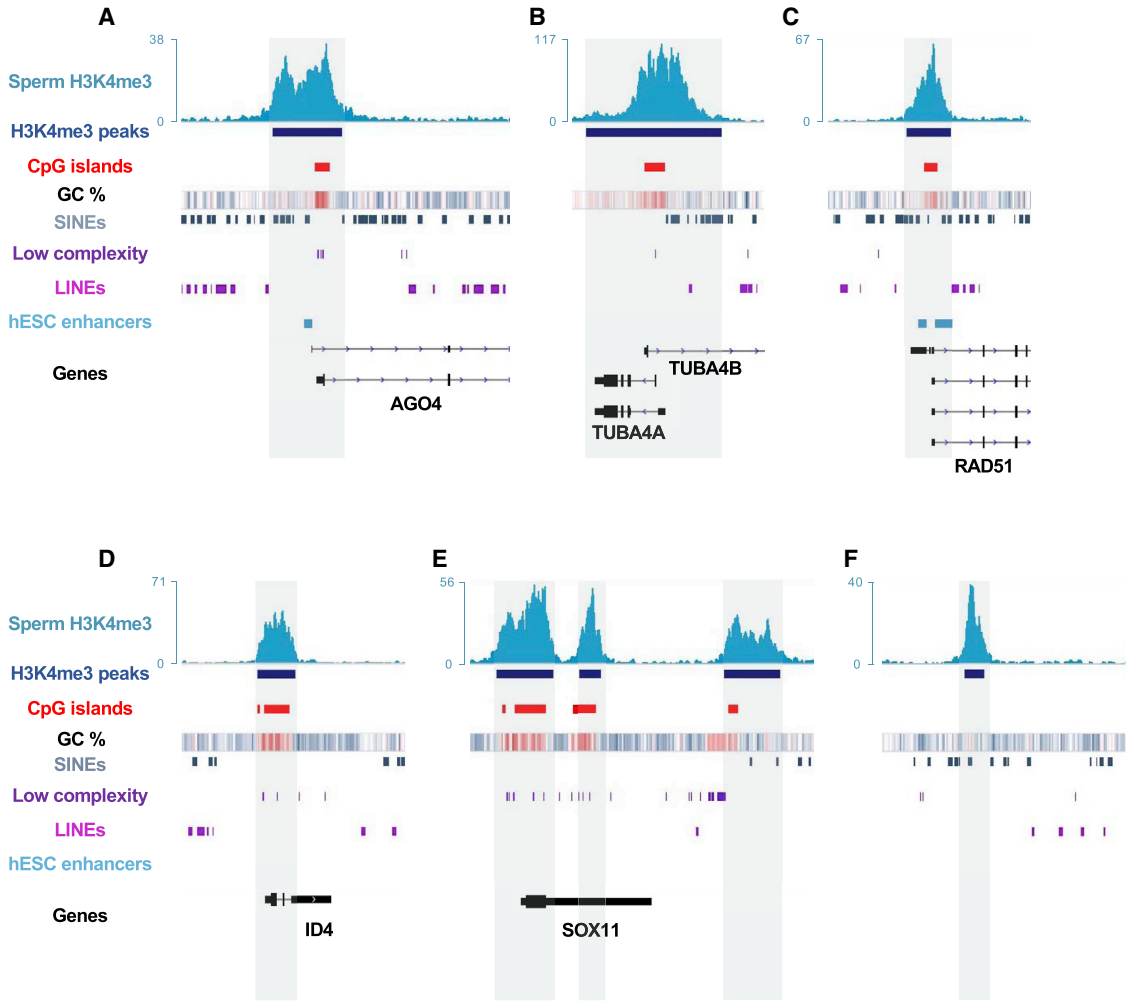


Figure 2. Overlap between sperm H3K4me3 peaks and CpG-rich regions, repetitive elements, and hESC enhancers

(A–F) Genome-browser snapshots showing the reference-population ChIP-seq track (blue), H3K4me3 peaks called by MACS2 (dark blue boxes), a CpG island location track (red boxes, UCSC), a GC percentage heatmap (blue, low GC percentage; white, 50% GC; red, high GC percentage), the location of SINEs (dark gray boxes), low-complexity repeats (dark purple boxes), the location of LINEs (violet boxes) (obtained from RepeatMasker), and the location of hESC enhancers (light blue boxes) (Barakat et al., 2018). The overlap of H3K4me3 with each of these elements is illustrated for the member of the Argonaute family of genes implicated in RNA-mediated gene silencing *AGO4* (A), for two genes coding for tubulin isoforms *TUBA4A* and *TUBA4B* (B), for the gene assisting in double-strand breaks repair *RAD51* (C), for the developmental genes *ID4* (D) and *SOX11* (E), and, finally, for a random intergenic region (F).

H3K4me3, but less expression than that of their CpG-rich counterparts during embryo development, confirming that they are, for the most part, specialized in spermatogenesis (Figure 1E). Similarly, we analyzed the expression profile of H3K4me3 peaks overlapping with low-complexity repeats and SINEs during development. We found those genes to be highly expressed during early embryonic development (Figure 1E). Because SINEs can operate as enhancers for neurodevelopment (Sasaki et al., 2008), and H3K4me3 shows low but significant enrichment at functional enhancers (Hu et al., 2017; Shen et al., 2016), we compared the level of expression of the 11,520 SINE-enriched H3K4me3^{+} genes with that of the 6,827 H3K4me3^{+} genes with an hESC functional enhancer (Barakat et al., 2018) in their promoter (Figure 1E). Both groups of genes had a very similar expression patterns during embryo development and were involved in comparable, essential "housekeeping" functions, including gene expression and cellular metabolic processes (Figure S4; for full list, Table S4). This suggests that there is an overlap between SINEs and hESC enhancers marked by H3K4me3 in sperm. Indeed, 73% of the hESC-enhancer-enriched H3K4me3^{+} promoters also contain a SINE. The genome location of H3K4me3-enriched promoters, in relation to hESC enhancers, CpG content, low-complexity repeats, LINEs, and SINEs, is exemplified in genome-browser snapshots in Figure 2. This H3K4me3 programming at SINE-enriched promoters in sperm could have a role in the specification of enhancers and the control of enhancer activity during embryonic development. Consistent with those findings, a recent study demonstrated that gene promoters harboring low-complexity repeats and SINEs were sequentially activated during early embryonic development

(Lu et al., 2020). Genes enriched in SINEs are first activated during zygotic genome activation and are highly expressed in ESCs, whereas genes enriched in low-complexity repeats tend to encode developmental and tissue-specific transcription factors that are preferentially targeted by PRC2 for transcriptional poising (Lu et al., 2020; Mendenhall et al., 2010). The activity of enhancers is tightly regulated by the balance between H3K4me1 and H3K4me3, but unlike what has been defined as a hallmark of enhancers, it is becoming increasingly clear that highly active enhancers tend to display H3K4me3 rather than H3K4me1 (Henriques et al., 2018; Shen et al., 2016).

Promoters containing low-complexity repeats and marked by H3K4me3 peaks in sperm were significantly less expressed than SINE-enriched H3K4me3^{+} promoters during spermatogenesis (Figure 1D). In contrast, the signature difference detected between these two types of promoters, in terms of their overall expression levels during embryonic development, was less important albeit significant (Figure 1E). Gene ontology (GO) analysis revealed distinct differences in processes associated with promoter annotation. For example, H3K4me3^{+} gene promoters containing only low-complexity repeats were highly enriched for developmental processes, whereas promoters overlapping with SINEs were only enriched for basic cellular processes (Figure S4; for full list, Table S4). These differences may associate with the temporal and spatial expression of genes during developmental stages or for specific tissues.

Finally, although H3K4me3 marked many LINEs and long-terminal repeat (LTR) elements, that was less than what would be expected by chance (Z score < 0), reflecting their large presence in the genome (Figures 1A–1C and 2). Taken together, these results indicate a preferential distribution of sperm H3K4me3 in repeats at SINEs and low-complexity elements, regions that are relevant for early embryo development.

H3K4me3 and DNA methylation co-occur at functional genomic regions in human sperm

Previously, it has been reported in mice and men that regions enriched for H3K4me3 in sperm are predominantly mutually exclusive with DNA methylation (Erkek et al., 2013; Hammoud et al., 2009; Molaro et al., 2011). We wanted to revisit that assumption by integrating our two robust datasets generated from the same pooled sperm sample. To determine the degree of overlap and potential for interaction in gene regulation between H3K4me3 and DNA methylation, we used our WGBS data (Chan et al., 2019) with deep coverage derived from the same reference population used for the H3K4me3 ChIP-seq. We confirmed by Spearman correlation analysis that the DNA methylation levels identified in the pooled sample were representative of the profiles of individual men obtained using our sperm-targeted methylC-capture sequencing (MCC-seq) (Chan et al., 2019) (Figure S5). This high degree of conservation of the sperm DNA-methylation levels between the pooled and individual samples is visualized in Figure S6.

Next, we determined the average DNA methylation levels within regions bearing H3K4me3 peaks and categorized those as being hypomethylated (<20% methylation), intermediate for methylation (20%–80% methylation), or hypermethylated (<80% methylation) (Figure 3A; STAR Methods). As previously shown (Molaro et al., 2011), most of the H3K4me3 peaks were associated with hypomethylated DNA (23,023 peaks, <20% methylation) (Figures 3A and 3B). Interestingly, we additionally detected >16,000 peaks marking regions with an average methylation >80% and, thus, hypermethylated. and >8,000 peaks with intermediate DNA methylation average levels (Figures 3A and 3B). The proportion of hypo-, intermediate-, and hyper-methylation within H3K4me3 peaks for annotated genomic regions was then evaluated (Figure 3C). Generally, genic and CpG-rich regions, such as promoters and exons, and CpG islands and shores, respectively, marked by H3K4me3 were predominantly DNA hypomethylated. In contrast, non-LTR retrotransposons, including the subtypes SINEs and LINEs, and non-coding RNA were predominantly in H3K4me3 peak regions bearing higher levels of DNA methylation (Figure 3C). Intergenic regions and other interspersed repetitive elements marked by H3K4me3 were commonly hypermethylated or bore intermediate methylation levels (Figure 3C). Genome browser tracks shown in Figure 4 show DNA methylation profiles in relation to CpG content and H3K4me3 peaks at various annotated DNA elements (Figure 4). For example, most DNA hypomethylated regions marked by H3K4me3 in the *HOXA* developmental gene cluster were enriched at CpG islands in the proximal promoter (Figure 4A). Chromatin modifier *ARID5B* and spermatogonia marker *ZBTB16*/*PLZF* displayed an identical pattern and presented additional hypermethylated H3K4me3 peak regions in their gene body (Figures 4B–4C). However, some DNA-hypermethylated H3K4me3 peak regions were also located in gene promoters. such as in the spermiogenesis gene *CYLC2* (Figures 3C and 4D). Finally, the H3K4me3-marked regions of intermediate DNA methylation were divided into large regions displaying a mix of fully methylated, intermediate, and/or unmethylated CpGs (Figures 3B and 4A), and smaller regions composed mostly of CpGs with intermediate methylation (Figures 3B and 4B).

Exploring the functional relationship of sperm H3K4me3 and DNA methylation in gene regulation during spermatogenesis and embryogenesis

Based on studies in mice, we predicted that sperm would retain H3K4me3 and DNA methylation at spermatogenic genes (Brykczynska et al., 2010; Lambrot et al., 2019; Lismer et al., 2021). To gain insight into a potential functional relationship between H3K4me3 and DNA methylation in relation to genes expressed in spermatogenesis and embryo development, we investigated the gene promoters designated as having either hypo-, intermediate-, or hypermethylated DNA, coincident with H3K4me3 peaks in gene promoters. There was an over-representation of promoters for genes expressed in spermatogonia that contained H3K4me3 peak regions with an average DNA hypomethylation (Figure 5A). Similarly, meiotic-specific genes promoters were DNA hypomethylated within peaks of H3K4me3, whereas genes expressed in spermatocytes and spermatids bore more DNA methylation in regions with H3K4me3 peaks. This analysis on the co-existence of DNA methylation within spermatogenic gene promoters with H3K4me3 was further confirmed by a gene ontology enrichment analysis (Figure 5B; for full list, Table S4). Interestingly, DNA hypermethylated regions bearing H3K4me3 located in gene promoters were confirmed to be

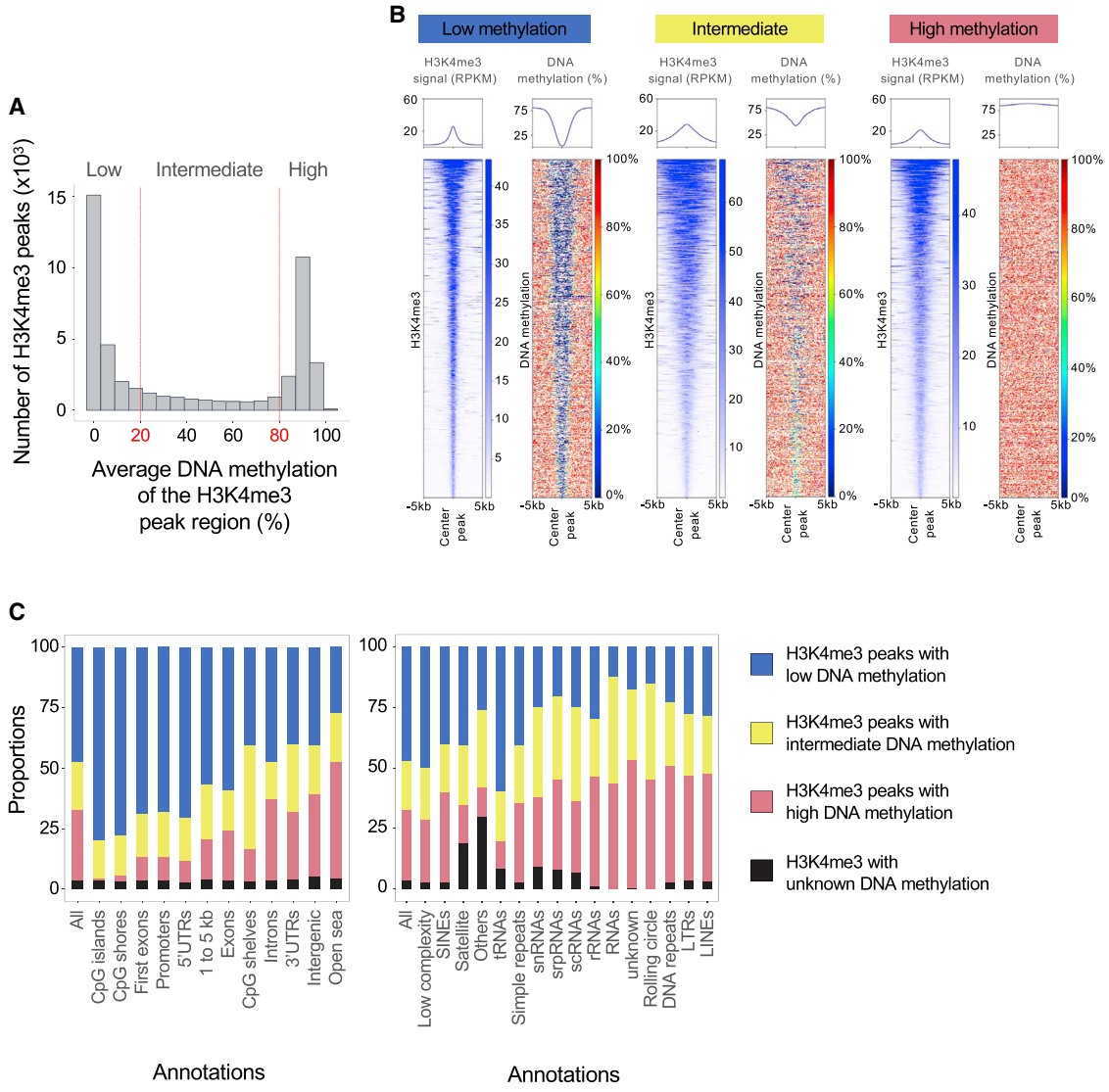


Figure 3. DNA methylation levels in regions with H3K4me3 peaks in human sperm

WGBS was performed on the sperm of the same 30 men used for H3K4me3 ChIP-seq, and the DNA methylation levels of the CpGs located within H3K4me3 peaks were assessed. The average DNA methylation level of the regions marked by peaks was then calculated.

(A) Number of H3K4me3 peaks with low (hypomethylated, ≤20% methylation), intermediate (20% ≤ methylation ≤ 80%), or high DNA methylation (hypermethylated, ≥80% methylation).

(B) Heatmaps showing the H3K4me3 signal (in RPKM) and the corresponding DNA methylation levels (in %) around the center of the peaks (±5 kb) for the hypomethylated (23,023 peaks), intermediate (8,087), or hypermethylated (16,777) H3K4me3 peaks. Each line represents a H3K4me3 peak. The overall H3K4me3 signal and DNA methylation for all peaks are summarized on top of the heatmaps.

(C) Proportion of hypomethylated, intermediate, and hypermethylated DNA within H3K4me3 peaks for genomic, CpG, RNAs, and repeats annotations.

See also Figures S5 and S6.

highly enriched for genes implicated in spermatid development and function. This suggests that DNA hypermethylation at some spermatogenic gene promoters functions to prevent their expression in the embryo.

In murine sperm, H3K4me3 localizes to genes active in embryonic development, and we recently demonstrated it is transmitted on paternal chromatin and is associated with embryonic gene expression (Lismer et al., 2021). At developmental genes in sperm, H3K4me3 often co-localizes with H3K27me3, and this is referred to as bivalency. Genes that are bivalent for H3K4me3/H3K27me3 are frequently poised for expression in later stages of embryogenesis (Bernstein et al., 2006). To gain insight into the functional potential of H3K4me3 in embryogenesis, we investigated the relationship of H3K4me3 and H3K27me3 patterns at DNA methylated regions occurring within H3K4me3 peaks in eight-cell embryos, the inner-cell mass, hESCs, and fetal organs (Figure 6).

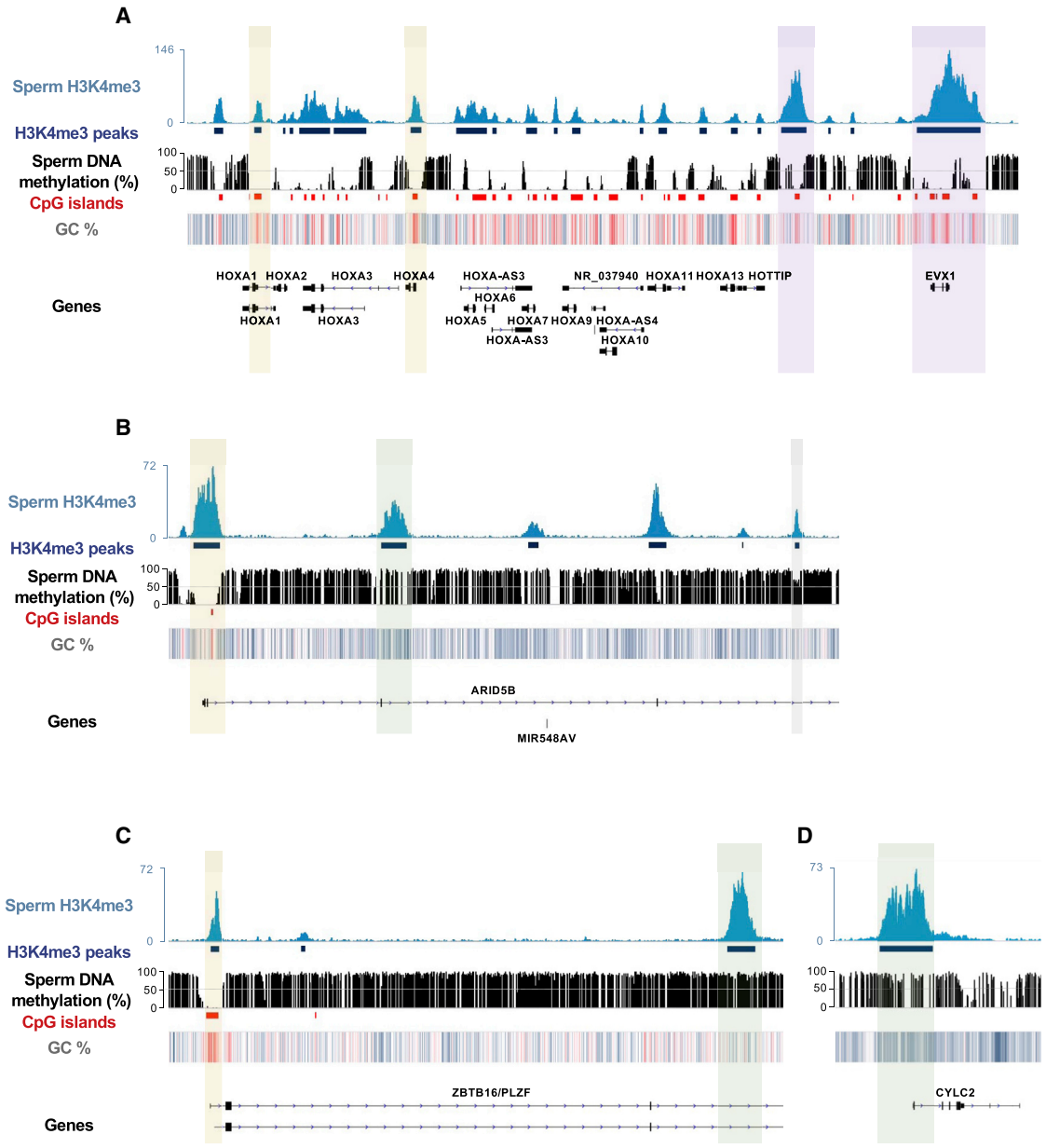


Figure 4. H3K4me3 and DNA methylation profiles in human sperm

(A–C) The overlap between H3K4me3 peaks and DNA methylation is displayed in genome-browser snapshots showing the ChIP-seq track (blue), H3K4me3 peaks called by MACS2 (dark blue), the WGBS track (black), a CpG islands location track (red, UCSC), and a GC percentage heatmap (blue, low GC %; white, 50% GC; red, high GC %) for the *HOXA* developmental gene cluster (A), for the chromatin modifier *ARID5B* (B), for the spermatogonial marker *ZBTB16*/*PLZF* (C), and for the spermiogenesis gene *CYLC2* (C). H3K4me3 peaks located at non-DNA methylated CpG islands (yellow), at regions including various degrees of methylation (purple), at CpG-poor DNA methylated regions (green), and at regions of intermediate DNA methylation (gray) are highlighted.

See also Figure S6.

Strikingly, the DNA hypermethylated H3K4me3 peak regions present in sperm were not detected in the embryo, hESC, and fetal organs (Figure 6), and those regions were also devoid of H3K27me3. Perhaps, those regions losing their H3K4me3 activation signal are programmed for repression in the next generation. Most of those hypermethylated H3K4me3 peaks were located in non-coding regions, such as intergenic regions and introns (Figure 3C). Nevertheless, the 1,294 promoters marked by DNA hypermethylated H3K4me3 peaks in sperm were indeed less expressed during embryonic development than the promoters marked by H3K4me3 peaks with a low or intermediate DNA methylation (Figure S7A). Notably, 431 gene promoters marked in sperm by DNA hypermethylated H3K4me3 peaks were associated with very low to no expression (RPKM $\leq$ 1) in

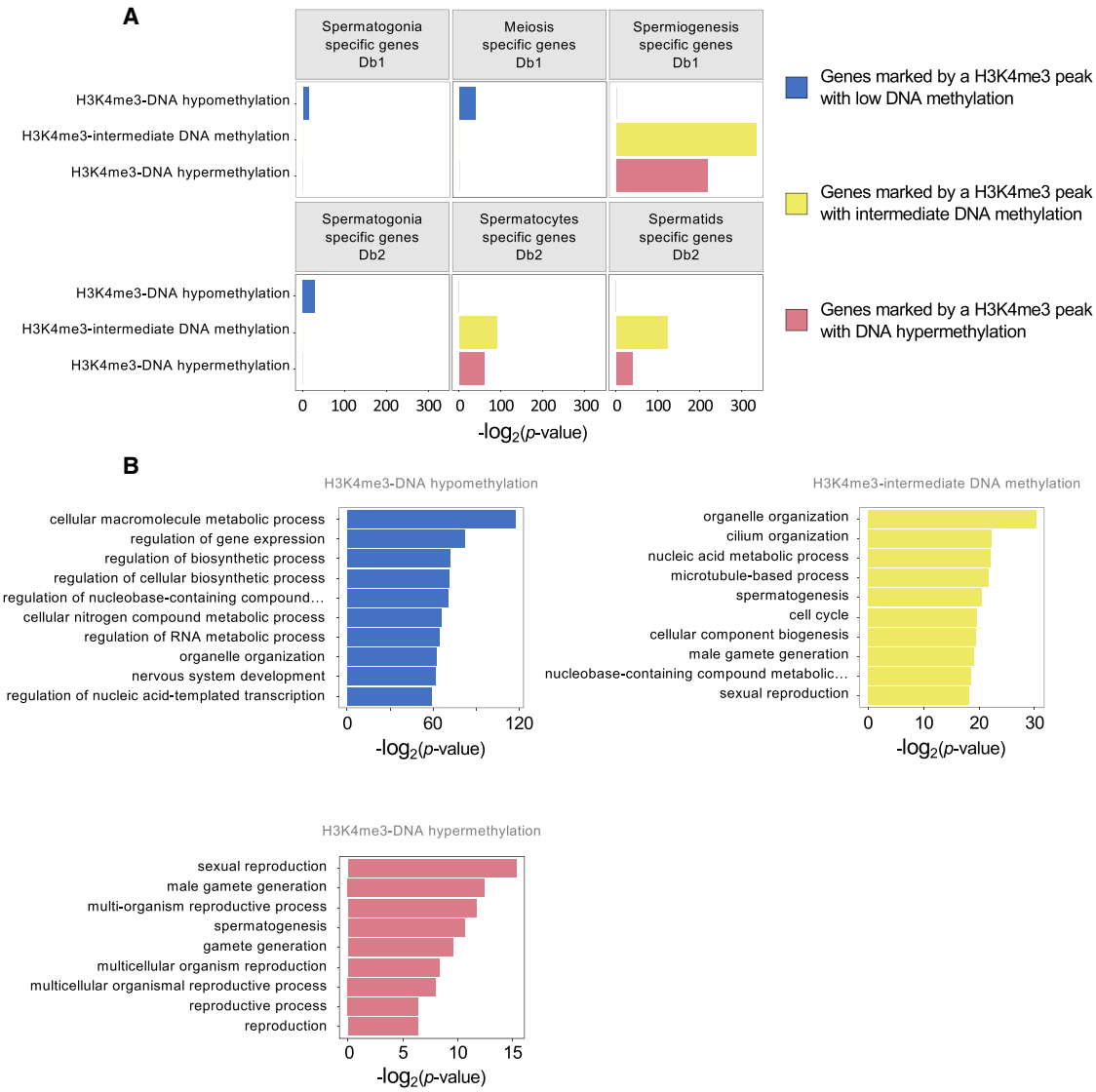


Figure 5. DNA methylation levels vary within H3K4me3 peaks at gene promoters with functions in spermatogenesis, embryogenesis, and basic cellular processes

(A) Over-representation of H3K4me3 peaks with hypo-, intermediate-, and hyper- average DNA methylation in the promoter of genes implicated in spermatogenesis. The genes relevant to early spermatogenesis, meiosis, and spermiogenesis were selected from two databases: Db1 (Jan et al., 2017) and Db2 (Wang et al., 2018). Benjamini-Hochberg-adjusted hypergeometric minimum-likelihood p values are indicated.

(B) Gene ontology analysis of the H3K4me3 peaks located in gene promoters (1 kb upstream of the TSS). The enrichment in a specific biological process was measured by PANTHER over-representation test (GO biological process complete). The p value was determined by Fisher's exact test with a Bonferroni correction for multiple testing. The top-10 processes with the lowest p value are indicated after filtering for non-redundant processes with an enrichment $\geq$1.25. Full list of GO-enriched terms is provided in Table S4.

the eight-cell embryo. Gene ontology analysis of those hypermethylated 431 promoters showed they were enriched in reproductive processes (Figure S7B). Taken together, this indicates that the combination of the gene-activating H3K4me3 with DNA hypermethylation at those promoters, may be functionally dominated by DNA methylation to repress those genes in the next generation. However, there were exceptions because some promoters (146 genes) that were DNA hypermethylated within H3K4me3 peaks and were linked with essential cellular functions, displayed medium to high expression levels in the embryo (RPKM $\geq$ 10) (Figure S7C).

The H3K4me3 peak regions with an intermediate DNA methylation showed a reduced signal in embryos, and fetal tissue with narrower peaks showing bivalency (Figure 6). These sperm H3K4me3 peak regions with intermediate DNA methylation were mostly located at non-protein coding regions (Figure 3C). Strikingly, similar to hypermethylated regions, the genes with promoters characterized by intermediate DNA methylation at sperm

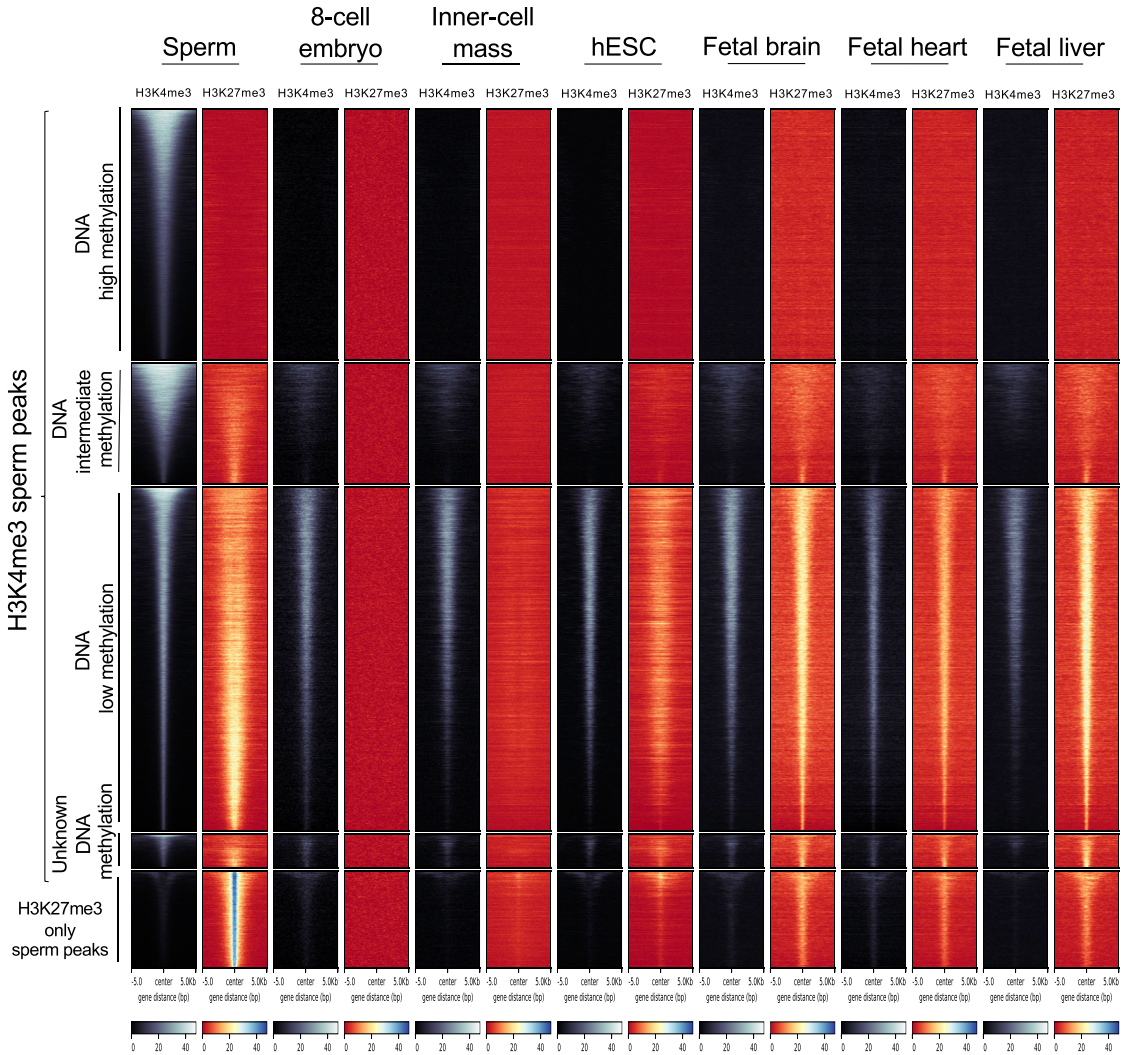


Figure 6. H3K4me3/H3K27me3 bivalency at H3K4me3 sperm peak regions with a hypo-, intermediate, or hyper- DNA methylation

Heatmaps showing the H3K4me3 and H3K27me3 signal (in RPKM) around the center of the H3K4me3 peaks (±5 kb) in human sperm (reference population of 30 men for H3K4me3 and published data for H3K27me3 (Hammoud et al., 2009)), eight-cell human embryo and inner cell mass (Xia et al., 2019), hESC (Grandy et al., 2015), and human fetal organs (Yan et al., 2016). Each lane represents an H3K4me3 peak region identified in the reference population with a DNA hypo- (23,023), intermediate- (8,087), hyper- (16,777), or unknown (with no CpG or CpGs that had less than 10× coverage in WGBS) methylation, or a H3K27me3 sperm peak region not marked by H3K4me3. See also Figure S7.

H3K4me3 peaks that were not expressed in the eight-cell embryo were spermiogenic genes (Figure S7B), whereas promoters enriched for critical processes, such as cell cycle, intercellular transport, and RNA metabolism, were expressed (Figure S7C).

Hypomethylated DNA regions within H3K4me3 peaks were implicated in essential functions, such as transcription, intracellular metabolism, and development (Figure 5B). Interestingly, those regions marked in sperm by DNA hypomethylated H3K4me3 peaks conserved their H3K4me3 signal during development and were not bivalent in eight cell embryos or the inner cells mass, but showed bivalency in hESC and fetal tissues (Figure 6). Given that these regions were found at >12,000 gene promoters (Figure 3C), this concurs with our findings that sperm H3K4me3 is associated with early embryonic gene expression (Figures 1E and S7A). The gain of bivalency in fetal tissues might indicate that these genes are poised for later expression or that bivalency serves as a means of gene silencing in regions lacking DNA methylation.

Based on this analysis, it is the predominantly H3K4me3 regions with hypomethylated DNA that may be transmitted to human embryos.

H3K4me3 in sperm localizes to enhancers and regions escaping epigenetic reprogramming

Given that more than 6,000 H3K4me3^{+} genes are enriched in hESC enhancers in sperm, we wondered whether an association between sperm H3K4me3 and active enhancers was found throughout the genome. By intersecting our ChIP-Seq data

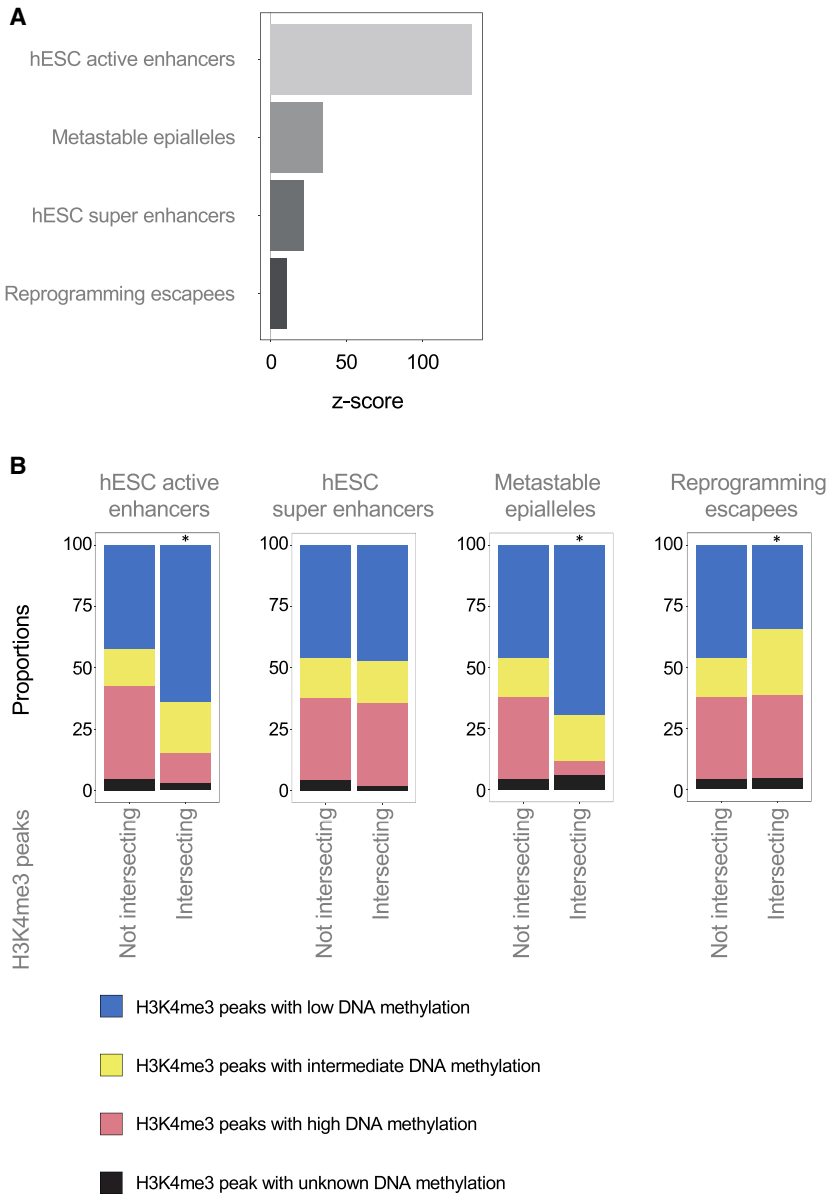


Figure 7. H3K4me3 is enriched at hESC enhancers and at regions that escape reprogramming

(A) Enrichment in H3K4me3 peaks at identified hESC enhancers and super-enhancers (Barakat et al., 2018), at metastable epialleles (Kessler et al., 2018), and in regions escaping reprogramming through loss of DNA methylation in human primordial germ cells (PGCs) (Tang et al., 2015). Positive enrichments are indicated by *Z* scores determined by the Bioconductor package *regioneR*. For all regions displayed, $p < 0.001$ by permutation tests.

(B) Proportion of hypomethylated, intermediate, and hypermethylated H3K4me3 peaks intersecting or not with the regions of interest. Changes in the proportion of peaks not intersecting the regions of interest and peaks intersecting were tested by χ^2 test, and an asterisk indicates $p < 0.001$.

with functionally characterized hESC enhancers, we determined that H3K4me3 is localized preferentially at enhancers that are defined as active and as super enhancers (Barakat et al., 2018) (Figure 7A). Interestingly, the H3K4me3 peaks found at enhancers, were mostly associated with hypo- and intermediate methylation (Figure 7B). We extended this analysis to probe regions that have been implicated in epigenetic inheritance, such as metastable epi-alleles (Kessler et al., 2018), and those previously determined in human primordial germ cells to escape epigenome reprogramming (Tang et al., 2015) (Figure 7B). We found that H3K4me3 peaks were enriched at those regions and that the DNA methylation levels tended to be hypomethylated in metastable epialleles, and in regions escaping epigenome reprogramming, there is a mix of hypo-, intermediate-, and hyper-DNA methylation (Figure 7B). These findings hint at a possible role for H3K4me3 in sperm from men in intergenerational inheritance, such as we have shown for mice (Lismer et al., 2020, 2021).

DISCUSSION

Sperm was historically viewed as a vehicle for the transmission of the paternal genome to the oocyte, but recent studies on the sperm epigenome and environmental interactions show that its role in heritability is more complex. Studies in model organisms revealed that epigenetic information is carried by sperm in the form of chromatin, DNA methylation, and noncoding RNA. Despite the intensive focus on DNA methylation and chromatin in sperm, they are infrequently studied in the same samples. Research in model organisms shows that there is an interplay between these epigenetic systems, yet whether the same is true for men is unclear. In this study, we found significant regions of overlap between sperm DNA methylation and H3K4me3. Retro-transposons are usually enriched in the gene-silencing modification H3K9me3 and DNA methylation to protect genome integrity (review in Zamudio and Bourc'his, 2010). However, in sperm, there is significant enrichment within SINEs for H3K4me3, which, in some regions, coincided with DNA hypermethylation, whereas, in others, with low DNA methylation. The permissive epigenetic state in sperm is particularly interesting, given that retro-transposons can influence the expression pattern of nearby genes (Sasaki et al., 2008; Zamudio and Bourc'his, 2010). We also observed an overlap of SINEs with hESC enhancers, marked by H3K4me3. This observation highlights a potential role of sperm epigenetic programming at SINEs in the specification of regulatory regions for the next generation.

As has been demonstrated in mice, data from recent studies in human embryos, combined with our study, suggest that histones in sperm from men also function in gene regulation and development. Here, we report that H3K4me3 has widespread accumulation at gene promoters in sperm and correlates with pre-implantation embryonic gene expression. In line with paternal H3K4me3 influencing embryonic gene expression is the overlap between H3K4me3 in human sperm and in pre-implantation embryos at DNA-hypomethylated regions and its transmission to the embryo in mouse models (Lismer et al., 2020, 2021). In addition to regions with low DNA methylation, our integration of embryo and sperm H3K4me3 datasets revealed that sites of overlap exist at H3K4me3-enriched regions with intermediate levels of DNA methylation. Notably, these H3K4me3 regions, particularly those presenting individual CpGs with DNA methylation levels between 20% and 80%, may be responsive to environmental cues, as was determined by MCC-seq analysis of sperm from men who were folate supplemented (Chan et al., 2019). Likewise, H3K4me3 in mice responds to paternal folate (Lismer et al., 2021) levels and may be conserved from sperm to the embryo to influence intergenerational gene expression. In support of this idea, in a transgenic mouse model, we determined that sperm with altered histone H3K4me2/me3 profiles at the transcriptional start sites of genes implicated in cellular metabolism, spermatogenesis, and embryo development had dire consequences for fertility and was associated with transgenerational transmission of abnormal developmental phenotypes (Lismer et al., 2020; Siklenka et al., 2015). In addition, in mice, paternal folate deficiency lead to DNA methylation and histone-methylation changes in sperm associated with altered gene expression in the embryo and the placenta and increased embryo loss and birth defects (Lambrot et al., 2013; Lismer et al., 2021). Remarkably, in mice, sperm H3K4me3 is retained in the embryo on paternal chromatin and is strongly correlated with pre-implantation gene expression (Lismer et al., 2021). Whether H3K4me3 in sperm from men serves a similar function in sensing paternal environmental exposures and transmits those epigenome alterations to the embryo to affect child development and disease remains unknown.

Limitations of the study

In this study, we analyzed the sperm epigenome of a representative population of Canadian men using the gold standard of DNA methylation analysis, WGBS, and deep ChIP-seq for H3K4me3, following ENCODE guidelines (Landt et al., 2012). The ChIP-seq was validated on seven individual men, and the WGBS was validated with a custom sperm MCC-seq approach. Limitations of this approach are that MCC-seq does not cover the entire genome (Chan et al., 2019). However, we can say with confidence that the regions that are covered in the capture validated those in the WGBS. There remains the possibility that, by pooling sperm, we are blind to some regions in which there may be strong individual variability in H3K4me3 and DNA methylation. In future studies, the degree of sperm-to-sperm heterogeneity can be addressed once single-cell analysis for chromatin is sensitive enough to profile sperm, despite the low level of histone retention.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHODS DETAILS
  - Sperm preparation
  - Native ChIP
  - Library preparation and sequencing
  - ChIP-seq data processing
  - H3K4me3 peak calling
  - H3K4me3 peak annotation and enrichment analyses
  - DNA methylation
  - MethylC-capture-sequencing
  - MCC-seq data processing
  - Correlation of ChIP signal with gene expression in spermatogenesis and development
  - Gene ontology analysis
  - Heatmaps
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2021.109418.

ACKNOWLEDGMENTS

We thank the team at the McGill University and Genome Quebec Innovation Centre for performing the sequencing and library preparations. The project was funded by the Canadian Institutes of Health Research (grants 358654 and 350129 to S.K., EP1-120608 to G.B., and CEE-151618 and FDN-1482425 to J.T.).

AUTHOR CONTRIBUTIONS

S.K. conceived, funded (Canadian Institutes of Health Research, grants 358654 and 350129), and guided the project. R.L. and D.C. performed the experiments. M.A., S.M., and C.L. recruited participants and collected sperm samples. J.T, T.K., and G.B. provided platform support for the MCC-seq (EP1-120608 and CEE-151618; CIHR FDN-148425). D.C. and J.T. reviewed and edited the manuscript. R.L., V.D., D.C., and X.S. performed the analysis; S.K., R.L., and V.D. interpreted the findings and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 29, 2020
Revised: February 12, 2021
Accepted: June 28, 2021
Published: July 20, 2021

REFERENCES

Aarabi, M., San Gabriel, M.C., Chan, D., Behan, N.A., Caron, M., Pastinen, T., Bourque, G., MacFarlane, A.J., Zini, A., and Trasler, J. (2015). High-dose folic

acid supplementation alters the human sperm methylome and is influenced by the MTHFR C677T polymorphism. Hum. Mol. Genet. *24*, 6301–6313.

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A., and Mason, C.E. (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. *13*, R87.

Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE blacklist: identification of problematic regions of the genome. Sci. Rep. *9*, 9354.

Balhorn, R. (2007). The protamine family of sperm nuclear proteins. Genome Biol. *8*, 227.

Barakat, T.S., Halbritter, F., Zhang, M., Rendeiro, A.F., Perenthaler, E., Bock, C., and Chambers, I. (2018). Functional dissection of the enhancer repertoire in human embryonic stem cells. Cell Stem Cell *23*, 276–288.e8.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bourgey, M., Dali, R., Eveleigh, R., Chen, K.C., Letourneau, L., Fillon, J., Michaud, M., Caron, M., Sandoval, J., Lefebvre, F., et al. (2019). GenPipes: an open-source framework for distributed and scalable genomic analyses. Gigascience *8*, giz037.

Brykczynska, U., Hisano, M., Erkek, S., Ramos, L., Oakeley, E.J., Roloff, T.C., Beisel, C., Schübeler, 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.

Carone, B.R., Hung, J.H., Hainer, S.J., Chou, M.T., Carone, D.M., Weng, Z., Fazzio, T.G., and Rando, O.J. (2014). High-resolution mapping of chromatin packaging in mouse embryonic stem cells and sperm. Dev. Cell *30*, 11–22.

Cavalcante, R.G., and Sartor, M.A. (2017). annotatr: genomic regions in context. Bioinformatics *33*, 2381–2383.

Chan, D., Shao, X., Dumargne, M.C., Aarabi, M., Simon, M.M., Kwan, T., Bailey, J.L., Robaire, B., Kimmins, S., San Gabriel, M.C., et al. (2019). Customized methylC-capture sequencing to evaluate variation in the human sperm DNA methylome representative of altered folate metabolism. Environ. Health Perspect. *127*, 87002.

Chen, M., and Heilbronn, L.K. (2017). The health outcomes of human offspring conceived by assisted reproductive technologies (ART). J. Dev. Orig. Health Dis. *8*, 388–402.

Dere, E., Huse, S., Hwang, K., Sigman, M., and Boekelheide, K. (2016). Intra- and inter-individual differences in human sperm DNA methylation. Andrology *4*, 832–842.

Donkin, I., and Barrès, R. (2018). Sperm epigenetics and influence of environmental factors. Mol. Metab. *14*, 1–11.

Dunican, D.S., Mjoseng, H.K., Duthie, L., Flyamer, I.M., Bickmore, W.A., and Meehan, R.R. (2020). Bivalent promoter hypermethylation in cancer is linked to the H327me3/H3K4me3 ratio in embryonic stem cells. BMC Biol. *18*, 25.

Erkek, S., Hisano, M., Liang, C.Y., Gill, M., Murr, R., Dieker, J., Schübeler, D., van der Vlag, J., Stadler, M.B., and Peters, A.H. (2013). Molecular determinants of nucleosome retention at CpG-rich sequences in mouse spermatozoa. Nat. Struct. Mol. Biol. *20*, 868–875.

Gel, B., Díez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M.A., and Malinverni, R. (2016). regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics *32*, 289–291.

Goodrich, R.J., Anton, E., and Krawetz, S.A. (2013). Isolating mRNA and small noncoding RNAs from human sperm. Methods Mol. Biol. *927*, 385–396.

Grandy, R.A., Whitfield, T.W., Wu, H., Fitzgerald, M.P., VanOudenhove, J.J., Zaidi, S.K., Montecino, M.A., Lian, J.B., van Wijnen, A.J., Stein, J.L., and Stein, G.S. (2015). Genome-wide studies reveal that h3k4me3 modification in bivalent genes is dynamically regulated during the pluripotent cell cycle and stabilized upon differentiation. Mol. Cell. Biol. *36*, 615–627.

Guo, H., Zhu, P., Yan, L., Li, R., Hu, B., Lian, Y., Yan, J., Ren, X., Lin, S., Li, J., et al. (2014). The DNA methylation landscape of human early embryos. Nature *511*, 606–610.

Haeussler, M., Zweig, A.S., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Hinrichs, A.S., Gonzalez, J.N., et al. (2019). The UCSC Genome Browser database: 2019 update. Nucleic Acids Res. *47* (*D1*), D853–D858.

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.

Hammoud, S.S., Nix, D.A., Hammoud, A.O., Gibson, M., Cairns, B.R., and Carrell, D.T. (2011). Genome-wide analysis identifies changes in histone retention and epigenetic modifications at developmental and imprinted gene loci in the sperm of infertile men. Hum. Reprod. *26*, 2558–2569.

Henriques, T., Scruggs, B.S., Inouye, M.O., Muse, G.W., Williams, L.H., Burkholder, A.B., Lavender, C.A., Fargo, D.C., and Adelman, K. (2018). Widespread transcriptional pausing and elongation control at enhancers. Genes Dev. *32*, 26–41.

Hisano, M., Erkek, S., Dessus-Babus, S., Ramos, L., Stadler, M.B., and Peters, A.H. (2013). Genome-wide chromatin analysis in mature mouse and human spermatozoa. Nat. Protoc. *8*, 2449–2470.

Hu, D., Gao, X., Cao, K., Morgan, M.A., Mas, G., Smith, E.R., Volk, A.G., Bartom, E.T., Crispino, J.D., Di Croce, L., and Shilatifard, A. (2017). Not All H3K4 Methylations Are Created Equal: Mll2/COMPASS Dependency in Primordial Germ Cell Specification. Mol. Cell *65*, 460–475.e6.

Jan, S.Z., Vormer, T.L., Jongejan, A., Röling, 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.

Kessler, N.J., Waterland, R.A., Prentice, A.M., and Silver, M.J. (2018). Establishment of environmentally sensitive DNA methylation states in the very early human embryo. Sci. Adv. *4*, eaat2624.

Krausz, C., Sandoval, J., Sayols, S., Chianese, C., Giachini, C., Heyn, H., and Esteller, M. (2012). Novel insights into DNA methylation features in spermatozoa: stability and peculiarities. PLoS ONE *7*, e44479.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572.

Lambrot, R., Xu, C., Saint-Phar, S., Chountalos, G., Cohen, T., Paquet, M., Suderman, M., Hallett, M., and Kimmins, S. (2013). Low paternal dietary folate alters the mouse sperm epigenome and is associated with negative pregnancy outcomes. Nat. Commun. *4*, 2889.

Lambrot, R., Siklenka, K., Lafleur, C., and Kimmins, S. (2019). The genomic distribution of histone H3K4me2 in spermatogonia is highly conserved in sperm. Biol. Reprod. *100*, 1661–1672.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. *22*, 1813–1831.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Levine, H., Jørgensen, N., Martino-Andrade, A., Mendiola, J., Weksler-Derri, D., Mindlis, I., Pinotti, R., and Swan, S.H. (2017). Temporal trends in sperm count: a systematic review and meta-regression analysis. Hum. Reprod. Update *23*, 646–659.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lismer, A., Siklenka, K., Lafleur, C., Dumeaux, V., and Kimmins, S. (2020). Sperm histone H3 lysine 4 trimethylation is altered in a genetic mouse model

of transgenerational epigenetic inheritance. Nucleic Acids Res. *48*, 11380–11393.

Lismer, A., Dumeaux, V., Lafleur, C., Lambrot, R., Brind'Amour, J., Lorincz, M.C., and Kimmins, S. (2021). Histone H3 lysine 4 trimethylation in sperm is transmitted to the embryo and associated with diet-induced phenotypes in the offspring. Dev. Cell *56*, 671–686.e6.

Lu, J.Y., Shao, W., Chang, L., Yin, Y., Li, T., Zhang, H., Hong, Y., Percharde, M., Guo, L., Wu, Z., et al. (2020). Genomic repeats categorize genes with distinct functions for orchestrated regulation. Cell Rep. *30*, 3296–3311.e5.

Ly, L., Chan, D., Aarabi, M., Landry, M., Behan, N.A., MacFarlane, A.J., and Trasler, J. (2017). Intergenerational impact of paternal lifetime exposures to both folic acid deficiency and supplementation on reproductive outcomes and imprinted gene methylation. Mol. Hum. Reprod. *23*, 461–477.

Matsumura, Y., Nakaki, R., Inagaki, T., Yoshida, A., Kano, Y., Kimura, H., Tanaka, T., Tsutsumi, S., Nakao, M., Doi, T., et al. (2015). H3K4/H3K9me3 bivalent chromatin domains targeted by lineage-specific DNA methylation pauses adipocyte differentiation. Mol. Cell *60*, 584–596.

Mendenhall, E.M., Koche, R.P., Truong, T., Zhou, V.W., Issac, B., Chi, A.S., Ku, M., and Bernstein, B.E. (2010). GC-rich sequence elements recruit PRC2 in mammalian ES cells. PLoS Genet. *6*, e1001244.

Molaro, A., Hodges, E., Fang, F., Song, Q., McCombie, W.R., Hannon, G.J., and Smith, A.D. (2011). Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. Cell *146*, 1029–1041.

Oikawa, M., Simeone, A., Hormanseder, E., Teperek, M., Gaggioli, V., O'Doherty, A., Falk, E., Sporniak, M., D'Santos, C., Franklin, V.N.R., et al. (2020). Epigenetic homogeneity in histone methylation underlies sperm programming for embryonic transcription. Nat. Commun. *11*, 3491.

Radford, E.J., Ito, M., Shi, H., Corish, J.A., Yamazawa, K., Isganaitis, E., Seisenberger, S., Hore, T.A., Reik, W., Erkek, S., et al. (2014). In utero effects. In utero undernourishment perturbs the adult sperm methylome and intergenerational metabolism. Science *345*, 1255903.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44* (*W1*), W160–W165.

Royo, H., Stadler, M.B., and Peters, A.H.F.M. (2016). Alternative computational analysis shows no evidence for nucleosome enrichment at repetitive sequences in mammalian spermatozoa. Dev. Cell *37*, 98–104.

Samans, B., Yang, Y., Krebs, S., Sarode, G.V., Blum, H., Reichenbach, M., Wolf, E., Steger, K., Dansranjavin, T., and Schagdarsurengin, U. (2014). Uniformity of nucleosome preservation pattern in mammalian sperm and its connection to repetitive DNA elements. Dev. Cell *30*, 23–35.

Sasaki, T., Nishihara, H., Hirakawa, M., Fujimura, K., Tanaka, M., Kokubo, N., Kimura-Yoshida, C., Matsuo, I., Sumiyama, K., Saitou, N., et al. (2008). Possible involvement of SINEs in mammalian-specific brain formation. Proc. Natl. Acad. Sci. USA *105*, 4220–4225.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. USA *103*, 1412–1417.

Shen, H., Xu, W., Guo, R., Rong, B., Gu, L., Wang, Z., He, C., Zheng, L., Hu, X., Hu, Z., et al. (2016). Suppression of enhancer overactivation by a RACK7-histone demethylase complex. Cell *165*, 331–342.

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.

Smith, J.F., Eisenberg, M.L., Millstein, S.G., Nachtigall, R.D., Sadetsky, N., Cedars, M.I., and Katz, P.P.; Infertility Outcomes Program Project Group (2011). Fertility treatments and outcomes among couples seeking fertility care: data from a prospective fertility cohort in the United States. Fertil. Steril. *95*, 79–84.

Tang, W.W., Dietmann, S., Irie, N., Leitch, H.G., Floros, V.I., Bradshaw, C.R., Hackett, J.A., Chinnery, P.F., and Surani, M.A. (2015). A unique gene regulatory network resets the human germline epigenome for development. Cell *161*, 1453–1467.

Wang, M., Liu, X., Chang, G., Chen, Y., An, G., Yan, L., Gao, S., Xu, Y., Cui, Y., Dong, J., et al. (2018). Single-cell RNA sequencing analysis reveals sequential cell fate transition during human spermatogenesis. Cell Stem Cell *23*, 599–614.e4.

Whitman-Elia, G.F., and Baxley, E.G. (2001). A primary care approach to the infertile couple. J. Am. Board Fam. Pract. *14*, 33–45.

Wu, H., Estill, M.S., Shershebnev, A., Suvorov, A., Krawetz, S.A., Whitcomb, B.W., Dinnie, H., Rahil, T., Sites, C.K., and Pilsner, J.R. (2017). Preconception urinary phthalate concentrations and sperm DNA methylation profiles among men undergoing IVF treatment: a cross-sectional study. Hum. Reprod. *32*, 2159–2169.

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.

Yamaguchi, K., Hada, M., Fukuda, Y., Inoue, E., Makino, Y., Katou, Y., Shirahige, K., and Okada, Y. (2018). Re-evaluating the localization of sperm-retained histones revealed the modification-dependent accumulation in specific genome regions. Cell Rep. *23*, 3920–3932.

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.

Yan, L., Guo, H., Hu, B., Li, R., Yong, J., Zhao, Y., Zhi, X., Fan, X., Guo, F., Wang, X., et al. (2016). Epigenomic landscape of human fetal brain, heart, and liver. J. Biol. Chem. *291*, 4386–4398.

Zamudio, N., and Bourc'his, D. (2010). Transposable elements in the mammalian germline: a comfortable niche or a deadly trap? Heredity *105*, 92–104.

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

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit monoclonal anti-H3K4me3	Cell Signaling Technology	9751S; RRID: AB_2616028
Biological samples		
Human semen samples	CReAte fertility center (Toronto); McGill University Reproductive Centre and OVO clinic (Montreal); Ottawa (Ottawa Fertility Clinic).	https://www.createivf.com; https://muhc.ca/reproductivecentre; https://www.cliniqueovo.com; https://conceive.ca/
Chemicals, peptides, and recombinant proteins		
Puresperm 100	Nicadon	PS100-100
Puresperm buffer	Nicadon	PSB-100
MNase (nuclease S7)	Roche	10107921001
cOmplete Protease Inhibitor Cocktail	Roche	4693116001
Critical Commercial Assays		
Kapa HTP Library Preparation Kit	Roche	KK8230
Protein A Dynabeads	Thermo Fisher	10002D
Agencourt AMPure XP beads	Beckman Coulter	A63880
Deposited data		
Human sperm H3K4me3 ChIP-seq	This study	GEO: GSE156108
Human sperm WGBS	(Chan et al., 2019)	EGA: EGAS00001003617
Human sperm MCC-seq	This study	GEO: GSE156108
Human sperm H3K4me3 & H3K27me3 ChIP-seq	(Hammoud et al., 2009)	GEO: GSE15594
hESCs H3K4me3 ChIP-seq	(Grandy et al., 2015)	GEO:GSE55502
hESCs enhancers and super-enhancers (ChIP-STARR-seq)	(Barakat et al., 2018)	GEO: GSE99631
DNA methylation reprogramming escapees (BS-Seq)	(Tang et al., 2015)	NCBI SRA: SRP057098
Metastable epialleles (Bioinformatics analysis of published data)	(Kessler et al., 2018)	GEO: GSM1091961, GSM1091962, GSM1091963, GSM1091964, GSE49828, NCBI SRA: SRX388732, SRX175349, SRX263877, SRX388741, SRX388745, SRX109155, SRX263886, SRX263887, and SRX388747
Spermatogenesis RNA-seq Database 1	(Jan et al., 2017)	NCBI SRA: SRP069329
Spermatogenesis RNA-seq Database 2	(Wang et al., 2018)	GEO: GSE106487
Embryo development scRNA-seq	(Yan et al., 2013)	GEO: GSE36552
Human embryo H3K4me3/H3K27me3 ChIP-seq	(Xia et al., 2019)	GEO: GSE124718
Human fetus H3K4me3/H3K27me3 ChIP-seq	(Yan et al., 2016)	GEO: GSE63634
Software and algorithms		
Trimmomatic version 0.36	(Bolger et al., 2014)	http://www.usadellab.org/cms/?page=trimmomatic
Bowtie 2 version 2.3.4.1	(Langmead and Salzberg, 2012)	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
SAMtools version 1.9	(Li et al., 2009)	http://www.htslib.org/doc/
deepTools2 version 3.1.1	(Ramírez et al., 2016)	https://deeptools.readthedocs.io/en/develop/
MACS2 version 2.1.2	(Zhang et al., 2008)	https://pypi.org/project/MACS2/
annotatr	(Cavalcante and Sartor, 2017)	http://bioconductor.org/packages/release/bioc/html/annotatr.html
RepeatMasker	N/A	http://www.repeatmasker.org
regioneR	(Gel et al., 2016)	http://bioconductor.org/packages/release/bioc/html/regioneR.html

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
methylKit	(Akalin et al., 2012)	https://bioconductor.org/packages/release/bioc/html/methylKit.html
GenPipes	(Bourgey et al., 2019)	https://bitbucket.org/mugqic/genpipes/src/master/
Bismarck version 0.18.1	(Krueger and Andrews, 2011)	https://github.com/FelixKrueger/Bismark
ggplot2	N/A	https://ggplot2.tidyverse.org/
PANTHER	N/A	http://www.pantherdb.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Sarah Kimmins, email: sarah.kimmins@mcgill.ca, Phone: 514-398-7658.

Materials availability

This study did not generate new unique reagents.

Data and code availability

The accession number for sperm H3K4me3 ChIP-Seq and MCC-seq data reported in this paper is GEO: GSE156108 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE156108).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

This study was approved by the McGill Research Ethics Board and informed consent was obtained from all participants. Study Participants representative of a reference population in Canada were recruited in three Canadian cities: Toronto, Montreal, and Ottawa. Information pertaining to the participants age, fertility status, semen analysis etc., is presented in Table S1. Semen samples were collected by masturbation after at least 3 days of abstinence. Following semen liquefaction (20-30min at room temperature), an aliquot was used for sperm count and the remaining sample was immediately frozen. Seven additional men were recruited at the CReATe fertility center for comparison and validation of the ChIP-seq data generated on the reference population (Table S2). These samples were identically processed.

METHODS DETAILS

Sperm preparation

The sperm samples from the 30 men (Table S1) were thawed and an equal number of sperm cells from each sample were combined to establish the pooled reference population sample. This sample was purified to ensure that it was free from somatic cell contamination using a density gradient method (Goodrich et al., 2013) then washed with PureSperm buffer (Nicadon). After centrifugation for 15 min at 4°C at 300 g, the pellet was resuspended in 1 mL of PureSperm buffer and the sample was overlaid onto a 50% PureSperm gradient. This gradient was centrifuged for 20 min at 200 g at room temperature and the resulting sperm pellet was transferred to a new tube and resuspended in PBS. After further centrifugation at 4000 g for 10 min, two additional washes in PBS were followed by snap-freezing the pellet in liquid nitrogen. The 7 individual sperm samples (Table S2) were processed separately using the same procedure as the pooled samples. Sample purity and cell number were determined by cell counts using a hemocytometer. No somatic cell contamination was observed in any of the individual samples following cleaning and purification.

Native ChIP

ChIP (Chromatin ImmunoPrecipitation) was performed as previously described (Hisano et al., 2013; Siklenka et al., 2015). Briefly, 12 million spermatozoa were resuspended in 300 μL of buffer 1 (15 mM Tris-HCl, 60 mM KCl, 5 mM MgCl2 and 0.1 mM EGTA) containing 0.3 M sucrose and 10 mM DTT. The sample was then split into 6 tubes (2 million cells/tube) and 50μL of buffer 1 was supplemented with 0.5% NP-40 and 1% sodium deoxycholate and added to each tube. After 30 min incubation on ice, 100μL of MNase buffer (85 mM Tris-HCl, pH 7.5, 3 mM MgCl2 and 2 mM CaCl2) containing 0.3 M sucrose and MNase (30 units of MNase for every 2 million sperm, Roche Nuclease S7) was added to each tube. The tubes were immediately placed at 37°C in a thermomixer for exactly 5 min. The MNase treatment was stopped by adding 2 μL of 0.5 M EDTA and placing the tubes on ice for 10-20 min. The tubes were then centrifuged at 17,000 g for 10 min at room temperature to remove the debris and protamines. The supernatants were pooled into one 1.5 mL tube and 1X protease inhibitors (Roche) added to the chromatin. The immunoprecipitation was carried

out overnight at 4°C with an antibody targeting H3K4me3 (Cell Signaling Technology) using Dynabeads coupled with protein A (ThermoFisher Scientific). The mononucleosomal fraction (147bp) was size selected using Agencourt XP AMPure beads (Beckman Coulter).

Library preparation and sequencing

ChIP-seq libraries were prepared using the Kapa HTP Library Preparation Kit (Roche) as per the manufacturer's recommendations except that the size selection was performed after the PCR step. Two libraries were prepared for the reference sample to allow for a better quality of sequencing. The quality of the libraries was assessed using a LabChip GX instrument (PerkinElmer) before being sequenced on an Illumina HiSeq 2000 or HiSeq 4000 (individual samples) system (SR-100).

ChIP-seq data processing

Raw reads were trimmed from the 3′ end to have a phred score of at least 30. Illumina sequencing adapters were removed from the reads. Trimming and clipping were performed using Trimmomatic version 0.36 (Bolger et al., 2014). Each read was mapped against the Feb. 2009 assembly of the human genome (hg19, GRCh37 Genome Reference Consortium Human Reference 37 (GCA_000001405.1) downloaded from UCSC genome browser (Haeussler et al., 2019) using Bowtie 2 version 2.3.4.1 (Langmead and Salzberg, 2012). Bowtie2 searched for distinct, valid alignments for each read and reported the best one according to mapping quality metrics, thus preventing any redundant use of sequencing reads (Royo et al., 2016). To further strengthen our confidence in reported alignments, we excluded reads that exhibited more than 3 mismatches. SAMtools version 1.9 (Li et al., 2009) was then used to sort and convert SAM files. Coverage tracks were generated from aligned reads using *deepTools2* (Ramírez et al., 2016). The coverage was calculated as the number of reads extended to 150bp fragment size per 25 bp bin and normalized using Reads Per Kilobase per Million mapped reads (RPKM) not located in chromosome X. Sperm samples and data from single individuals were processed similarly.

H3K4me3 peak calling

Peaks were called for each pooled or individual samples using MACS2 version 2.1.2 (Zhang et al., 2008) using the following parameters '–broad–broad-cutoff 0.000001–bdg –nomodel'. Visual assessment of tracks confirmed that these parameters allowed for the most accurate identification of the broad H3K4me3 peaks observed in sperm for our reference population. We compared the similarity of H3K4me3 signals across samples by calculating the pairwise Spearman correlation coefficient between the ChIP-seq signals calculated in RPKM at each H3K4me3 peak identified in the reference population using the multiBigwigSummary and plotCorrelation functions from the deepTools2 suite (Ramírez et al., 2016).

H3K4me3 peak annotation and enrichment analyses

The genomic annotation of peaks was performed using the R/Bioconductor package *annotatr* package (Cavalcante and Sartor, 2017) using data from the TxDb.Hsapiens.UCSC.hg19.knownGene and org.hs.eg.db packages. To examine if H3K4me3 occurred in repetitive and low complexity regions and to classify these regions we used the consensus annotations provided by RepeatMasker (http://www.repeatmasker.org). Each H3K4me3 peak is annotated to genic and intergenic (including repeat) annotations when region boundaries overlap by at least 1bp. CpG shores were defined as 2Kb upstream/downstream of the ends of the CpG islands, less the CpG islands. CpG shelves were defined as another 2Kb upstream/downstream of the farthest upstream/downstream limits of the CpG shores, less the CpG islands and CpG shores. The remaining genomic regions make up the open sea annotation. Promoters were defined as the 1Kb upstream of the TSS.

The enrichment of the H3K4me3 peaks at specific genomic locations and curated regions of interest (i.e., reprogramming escapees, hESC enhancers and super-enhancers, metastable epiallele) was assessed by evaluating whether the observed overlap between peaks and regions of interest occurred significantly more or less frequently than chance. Briefly, the same number of regions of the same width are randomly chosen along the non-masked parts of the genome. For this analysis, we set the number of permutations of randomly chosen sets of regions at 1000 and significance of enrichment for a specific annotation was determined when permuted $p < 0.001$. The result of the permutation test implemented in the *regioneR* package (Gel et al., 2016) provides a z-score, which estimates if there is a positive enrichment of the peaks at a given region, i.e., more peaks observed in this region than expected by chance, or a negative enrichment, i.e., less peaks observed than expected by chance. The sperm enhancers and super-enhancers coordinates were obtained from Barakat et al. (2018). The locations of the regions escaping the epigenetic reprogramming happening in human primordial germ cells were retrieved from Tang et al. (2015). The metastable epialleles were obtained from Kessler et al. (2018).

DNA methylation

Whole-genome bisulfite sequencing was performed on a pooled sperm sample from the same 30 men as used for ChIP-Seq (Table S1) (Chan et al., 2019). We used the DNA methylation levels calculated by Chan et al., 2019 for each CpG that had a minimum of 10 reads coverage. We then summarized the DNA methylation information for each genomic region covered by an H3K4me3 peaks

using the methylKit R package (Akalin et al., 2012). Regions of DNA methylation within H3K4me3 peaks were defined as hypo, intermediate and hypermethylated when presenting an average methylation $\leq$ 20%, between 20 to 80%, and $\geq$ 80% methylated respectively.

MethylC-capture-sequencing

Targeted bisulfite sequencing was performed as previously described (Chan et al., 2019). The WGBS library was constructed using the KAPA® High Throughput Library Preparation kit (Roche/KAPA® Biosystems). Briefly, 1 μg of the sperm DNA was spiked with 0.1% (w/w) unmethylated λ and pUC19 DNA (Promega). DNA was sonicated (S220 Focused-ultrasonicator, Covaris) and fragment sizes of 300–400 bp were controlled on a Bioanalyzer DNA 1000 LabChip® (Agilent). Following fragmentation, DNA-end repair of double-stranded DNA breaks, 3′ end adenylation, adaptor ligation, and clean-up steps were conducted according to KAPA® Biosystems' protocols. The sample was then bisulfite converted using the EpiTect® Fast DNA bisulfite kit (QIAGEN) following the manufacturer's protocol. The resulting bisulfite DNA was amplified with 9–12 PCR cycles using the KAPA® HiFi HotStart Uracil+Uracil+ DNA Polymerase kit (Roche/KAPA® Biosystems). The final WGBS library was purified using Agencout® AMPure® Beads (Beckman Coulter), validated on Bioanalyzer High Sensitivity DNA LabChip® kits (Agilent) and quantified by PicoGreen® (ThermoFisher). The MCC-Seq protocol developed and optimized by Roche NimbleGen® was then applied. Briefly, the SeqCap® Epi Enrichment System protocol (Roche NimbleGen®) was used to capture the regions of interest. Equal amounts of multiplexed libraries (84 ng of each) were combined to obtain 1 μg of total input library, which was hybridized to the capture panel at 47°C for 72 h. Washing, recovery, and PCR amplification of the captured libraries, as well as final purification were conducted as recommended by the manufacturer. Bioanalyzer High Sensitivity DNA LabChip® kits (Agilent) were used to determine quality, concentration, and size distribution of the final captured libraries before being sequenced on an Illumina NovaSeq 6000 system (PE-100).

MCC-seq data processing

Targeted sperm panel MCC-Seq HiSeq reads were aligned using the Genpipes pipeline (Bourgey et al., 2019). Specifically, the MCC-Seq paired-end fastq reads were first trimmed for quality (phred33 > = 30), length (n > 50bp) and Illumina adapters using Trimmomatic (version 0.36). Then the trimmed reads were aligned, per sequencing lane, to the bisulfite-converted hg19/GRCh37 reference genome using Bismark (version 0.18.1) (Krueger and Andrews, 2011) with Bowtie 2 (version 318 2.3.1) in pair-end mode using the non-directional protocol setting and other default parameters. Lane bam files were merged and then de-duplicated using Picard (version 2.9.0). Methylation calls were obtained using Bismark to record counts of methylated and unmethylated cytosines at each cytosine position in the genome. A methylation level of each CpG was calculated by the number of methylated reads over the total number of sequenced reads. CpGs that were found to be overlapping with SNPs (dbSNP 137), the DAC Blacklisted Regions or Duke Excluded Regions (generated by the Encyclopedia of DNA elements - ENCODE project) (Amemiya et al., 2019) were removed. We compared the similarity of methylation levels across samples by calculating the pairwise Spearman correlation coefficient between the methylation levels at CpGs identified in the reference population and in all 7 individual samples (~2 millions CpGs) averaged across 100bp windows using the multiBigwigSummary and plotCorrelation functions from the deepTools2 suite (Ramírez et al., 2016).

Correlation of ChIP signal with gene expression in spermatogenesis and development

The statistical comparison of the expression levels of genes marked by H3K4me3 in sperm during spermatogenesis and development (Figures 1D, 1E, and S7A) was performed in R using a pairwise Wilcoxon rank sum tests with a Benjamini-Hochberg correction for multiple testing. The RPKM values were obtained from the published material (Wang et al., 2018; Yan et al., 2013). The over-representation of hypo, intermediate and hypermethylated H3K4me3 peaks in the promoter of genes implicated in spermatogenesis was assessed by Fisher's test with a Benjamini-Hochberg adjustment to correct for multiple testing (Figure 5A). Genes relevant to early spermatogenesis, meiosis, and spermiogenesis were selected in 2 published databases: Database 1 (Db1) (Jan et al., 2017) and 2 (Db2) (Wang et al., 2018).

Gene ontology analysis

Gene ontology (GO) enrichment analyses were carried out using PANTHER Overrepresentation Test (http://www.pantherdb.org/). The annotation dataset used was the GO biological process complete. Functional enrichment was evaluated by Fisher's exact test and adjusted using the Bonferroni or Benjamini-Hochberg correction.

Heatmaps

Tag density heatmaps were plotted using the *computeMatrix* and *plotHeatmap* functions from the *deepTools2* suite (Ramírez et al., 2016). The datasets used were our H3K4me3 pooled reference population, individual samples, H3K4me3 and H3K27me3 ChIP-seq data from human embryo (Xia et al., 2019), fetal organ (Yan et al., 2016) and hESC (Grandy et al., 2015), and H3K27me3 ChIP-seq data from human sperm (Hammoud et al., 2009).

QUANTIFICATION AND STATISTICAL ANALYSIS

Information about statistical tests and softwares used for each analysis is detailed in the relevant STAR Methods sections and figure legends. Analyses were performed on the regions enriched or not in H3K4me3 in the pooled sample of the reference population. We confirmed the representativity of these regions in independent individual samples by calculating the pairwise Spearman correlation coefficient between the reference population and 7 individual samples using the deepTools2 suite (Ramírez et al., 2016). The level of significance for all statistical tests used in this study was set at $p < 0.05$ with an exception for the permutation and chi-square tests, with a significance set at $p < 0.001$.

Cell Reports, Volume *36*

Supplemental information

Whole-genome sequencing of H3K4me3 and DNA methylation in human sperm reveals regions of overlap linked to fertility and development

Romain Lambrot, Donovan Chan, Xiaojian Shao, Mahmoud Aarabi, Tony Kwan, Guillaume Bourque, Sergey Moskovtsev, Clifford Librach, Jacquetta Trasler, Vanessa Dumeaux, and Sarah Kimmins

Table S1. List of 30 men included in the reference population and their health and sperm parameters. Related to Figure 1 and STAR methods.

Patient ID	Age	Source	Male Infertility	Smoking	MTHFR 677	Count (M/mL)	DFI (%)	Motility (%)
T1	28	Toronto	no	no	CC	75	18	73
T2	37	Toronto	no	yes	TT	51	8	49
T3	43	Toronto	no	no	TT	98	12	64
T4	37	Toronto	no	no	TT	84	14	61
T5	39	Toronto	no	yes	CC	64	16	50
T6	43	Toronto	no	no	TT	49	12	41
T7	31	Toronto	no	no	TT	82	31	66
T8	51	Toronto	no	no	CC	51	8	45
T9	34	Toronto	no	yes	TT	66	12	58
T10	38	Toronto	no	yes	CC	38	15	41
T11	36	Toronto	no	no	CC	44	22	48
T12	27	Toronto	no	no	CT	unknown	unknown	unknown
T13	35	Toronto	no	yes	CC	116	10	64
T14	42	Toronto	no	no	CC	45	14	68
T15	50	Toronto	no	no	TT	51	29	46
T16	34	Toronto	no	no	TT	60	8	63
T17	38	Toronto	no	yes	TT	45	12	44
T18	61	Toronto	no	no	CC	140	17	46
T19	46	Toronto	no	no	CC	128	9	51
T20	29	Toronto	no	no	CC	49	6	67
T21	55	Toronto	no	no	CC	85	10	63
M1	49	Montreal	yes	no	TT	119	28	41.6
M2	41	Montreal	yes	no	TT	108	8	26
M3	45	Montreal	yes	no	TT	94	19	49
M4	53	Montreal	yes	no	TT	498	26	76
M5	36	Montreal	yes	no	TT	119	17	22.8
M6	29	Montreal	yes	no	TT	138	9	10.3
O1	55	Ottawa	unknown	unknown	CT	330	unknown	27
O2	50	Ottawa	unknown	unknown	CC	4	unknown	30
O3	50	Ottawa	unknown	unknown	CC	92	unknown	9
All	41 ± 9					101 ± 96	15 ± 7	48 ± 16
Toronto	40 ± 9					71 ± 30	14 ± 7	55 ± 10
Montreal	42 ± 9					179 ± 157	18 ± 8	38 ± 23
Ottawa	52 ± 3					142 ± 169	na	35 ± 24

MTHFR: Methylenetetrahydrofolate Reductase; MTHFR C677T SNP (rs1801133), M/mL: million spermatozoa per mL semen; na: not applicable.

Table S2. List of 7 men and their sperm parameters. Related to STAR methods.

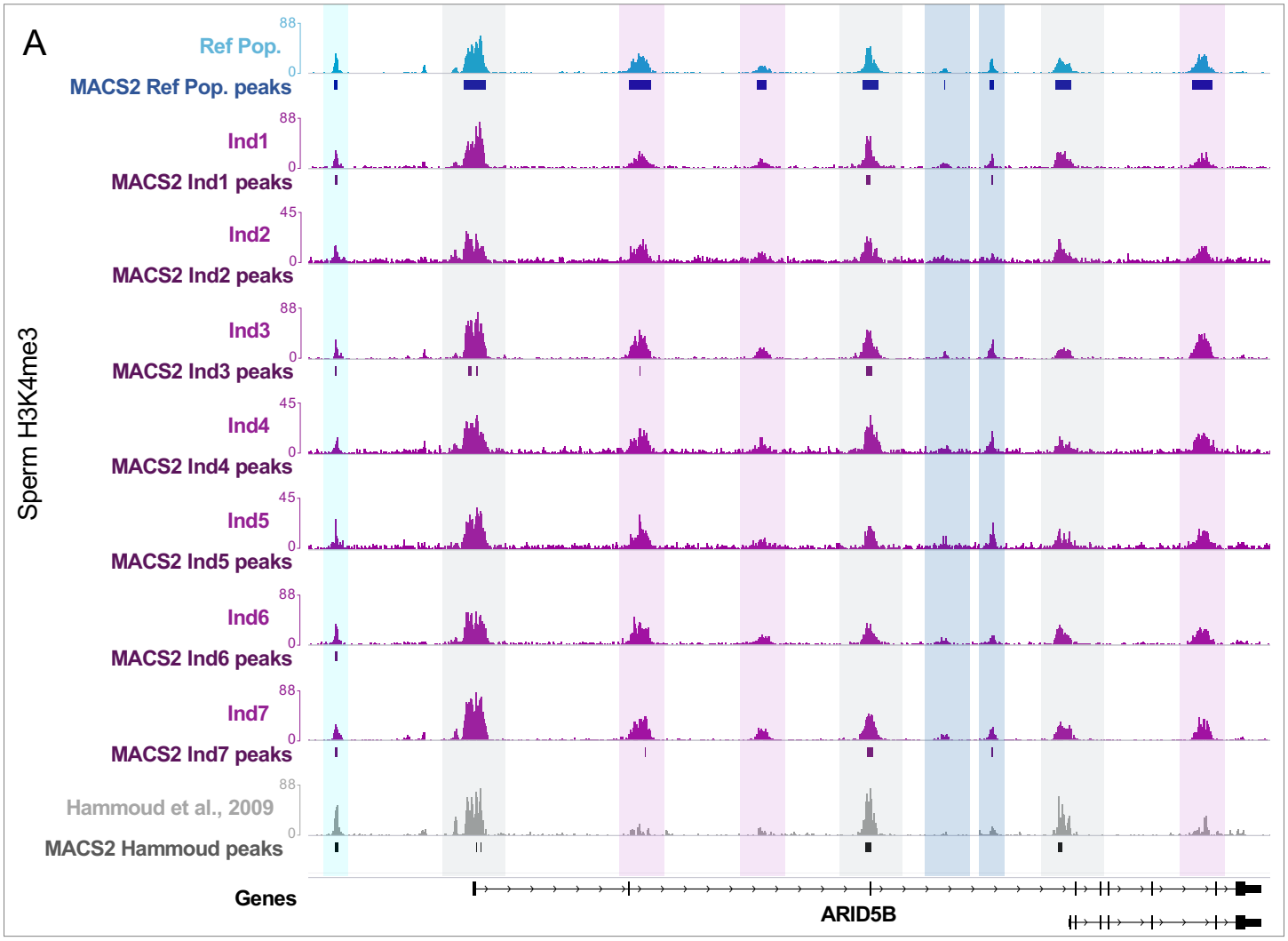
Patient ID	Age	Source	Count (M/mL)	DFI (%)	Motility (%)
Ind1	31	Toronto	109	10	61
Ind2	33	Toronto	35	20	69
Ind3	44	Toronto	161	10	55
Ind4	33	Toronto	103	14	33
Ind5	36	Toronto	123	20	47
Ind6	38	Toronto	94	21	70
Ind7	40	Toronto	153	10	75

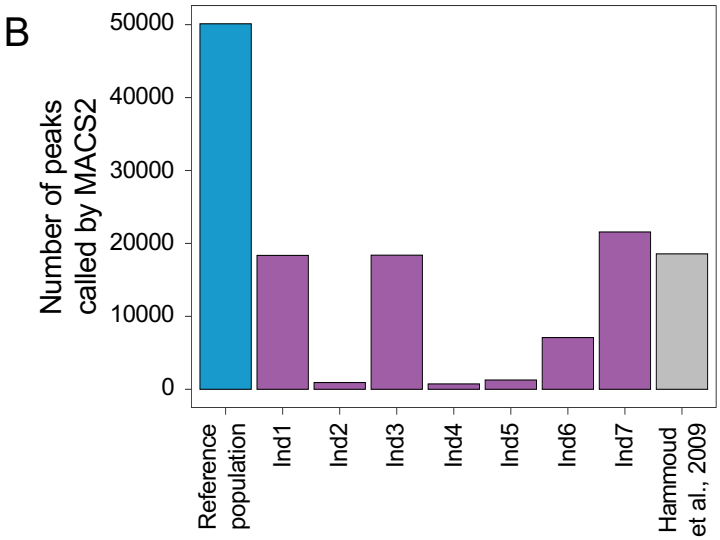
M/mL: million spermatozoa per mL semen.

Table S3. Sequencing and alignment statistics of the human sperm H3K4me3 ChIP-seq. Related to STAR methods.

Sample	Replicate	Source	Number of men	Number of Raw Reads	Duplicates (%)	Reads After Trimming	Mappable Reads	Mapping Efficiency	Genome assembly
Human sperm H3K4me3 ChIP-seq	library 1	This study	30	72,218,577	29.4	-	-	-	hg19
	library2			82,918,355	30.6	-	-	-	hg19
	Merged			155,136,932	30.0	153,384,915	145,690,862	94.98%	hg19
Human sperm H3K4me3 ChIP-seq	Ind1	This study	1	32,765,454	23.7	32,736,129	32,182,888	98.31%	hg19
	Ind2	This study	1	53,551,639	21.4	53,431,175	52,261,032	97.81%	hg19
	Ind3	This study	1	33,913,972	22.4	33,159,777	32,732,015	98.71%	hg19
	Ind4	This study	1	36,153,909	19.6	34,899,033	34,204,542	98.01%	hg19
	Ind5	This study	1	31,128,166	19.3	30,308,242	29,896,049	98.64%	hg19
	Ind6	This study	1	29,946,466	19.9	29,278,746	28,915,689	98.76%	hg19
	Ind7	This study	1	37,856,626	22.9	37,189,782	36,821,603	99.01%	hg19
Human sperm H3K4me3 ChIP-seq	Lane 1	(Hammoud et al., 2009)	4	5,393,483	-	-	-	-	hg19
	Lane 2			5,329,868	-	-	-	-	hg19
	Lane 3			7,623,257	-	-	-	-	hg19
	Merged			18,346,608	13.2	13,824,350	13,159,750	95.19%	hg19

Figure S1.


A
Sperm H3K4me3
Ref Pop.
MACS2 Ref Pop. peaks
Ind1
MACS2 Ind1 peaks
Ind2
MACS2 Ind2 peaks
Ind3
MACS2 Ind3 peaks
Ind4
MACS2 Ind4 peaks
Ind5
MACS2 Ind5 peaks
Ind6
MACS2 Ind6 peaks
Ind7
MACS2 Ind7 peaks
Hammoud et al., 2009
MACS2 Hammoud peaks
Genes
ARID5B


B
Number of peaks called by MACS2
50000
40000
30000
20000
10000
0
Reference population
Ind1
Ind2
Ind3
Ind4
Ind5
Ind6
Ind7
Hammoud et al., 2009

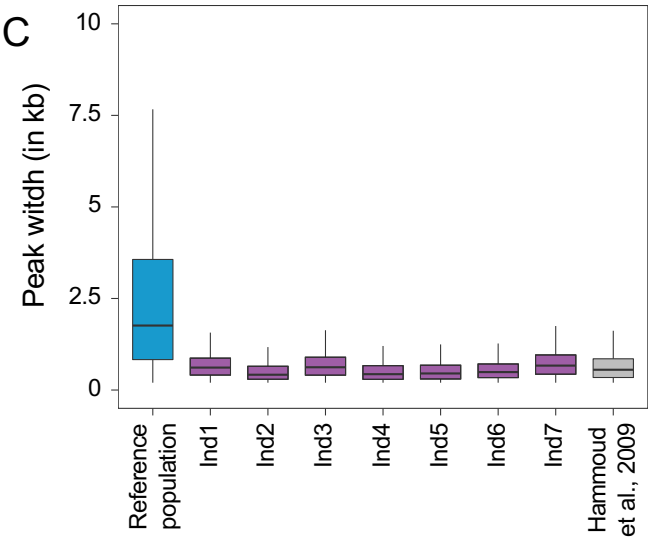

C
Peak witdh (in kb)
10
7.5
5
2.5
0
Reference population
Ind1
Ind2
Ind3
Ind4
Ind5
Ind6
Ind7
Hammoud et al., 2009

Figure S1. Improved identification of sperm H3K4me3 peaks by MACS2 using a pooled sample from a reference population for ChIP and sequenced at high-depth. Related to Figure 1. (**A**) The improved detection of H3K4me3 peaks and their actual size is illustrated by a genome browser snapshot for the reference population of 30 men (blue track; dark blue boxes underneath the track: H3K4me3 peaks called), 7 individual men (Ind1-7, purple tracks; dark purple boxes underneath: peaks called), and Hammoud et al. data (grey track; dark grey boxes underneath: peaks called) at the chromatin modifier, *ARID5B* locus. The turquoise window highlights the narrow peaks called identically for most ChIP-seq samples. The blue (narrow or low peaks) and pink (broad peaks) windows pinpoint the peaks not identified in most samples but the reference population. The grey windows display the improved characterization of width of the peaks in the reference population sample in comparison to individual and published datasets. (**B**) Number of H3K4me3 peaks called by MACS2 for the pooled sperm of the reference population of 30 men, 7 individual men (Ind1-7), and the pooled sperm of 4 men (Hammoud et al., 2009). (**C**) Boxplots showing the increased width of sperm H3K4me3 peaks identified by MACS2 in the reference population vs individual men and a published database.

Figure S2.

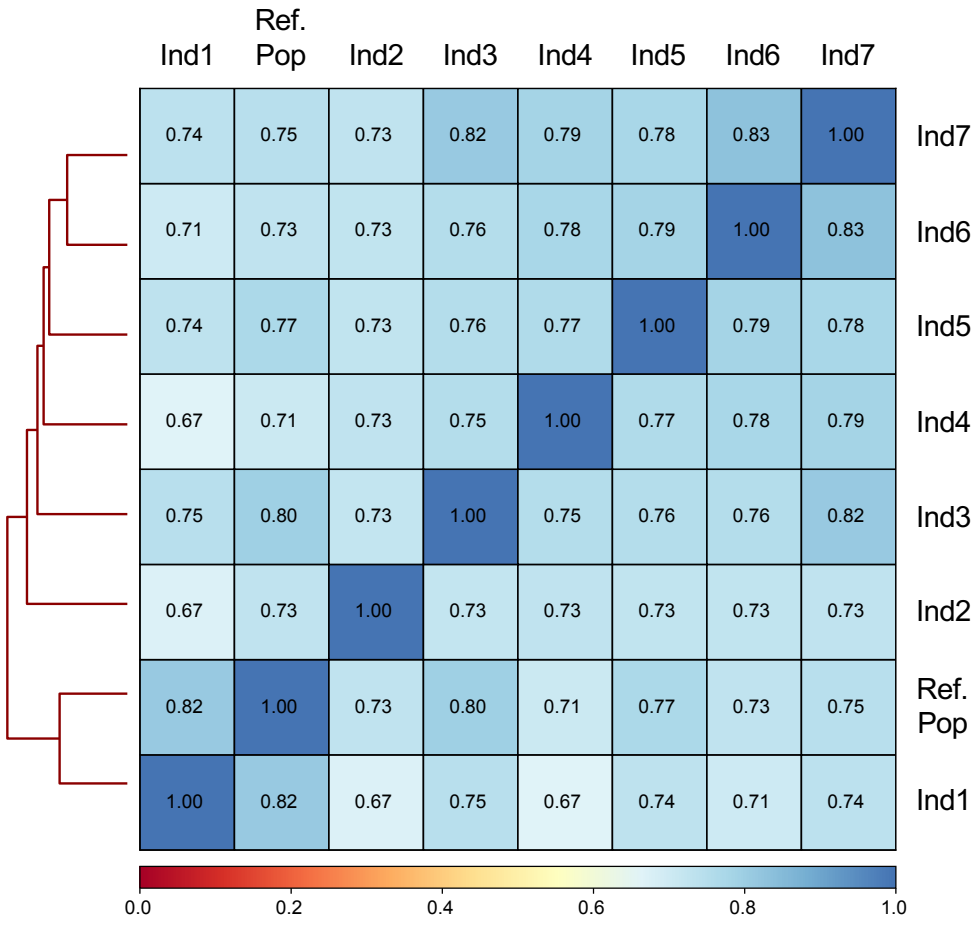
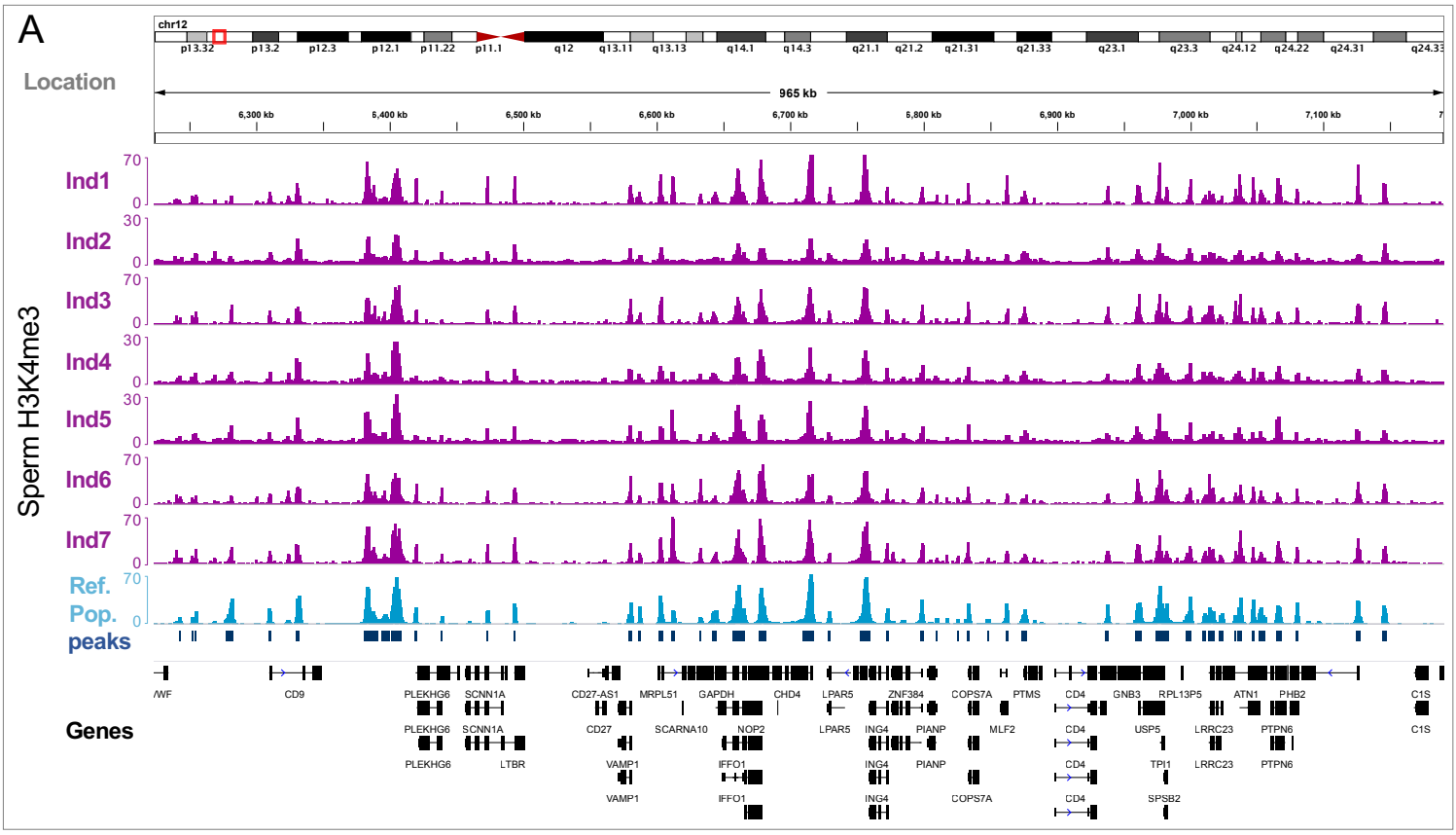

Ind1
Ref. Pop
Ind2
Ind3
Ind4
Ind5
Ind6
Ind7
0.74 0.75 0.73 0.82 0.79 0.78 0.83 1.00 Ind7
0.71 0.73 0.73 0.76 0.78 0.79 1.00 0.83 Ind6
0.74 0.77 0.73 0.76 0.77 1.00 0.79 0.78 Ind5
0.67 0.71 0.73 0.75 1.00 0.77 0.78 0.79 Ind4
0.75 0.80 0.73 1.00 0.75 0.76 0.76 0.82 Ind3
0.67 0.73 1.00 0.73 0.73 0.73 0.73 0.73 Ind2
0.82 1.00 0.73 0.80 0.71 0.77 0.73 0.75 Ref. Pop
1.00 0.82 0.67 0.75 0.67 0.74 0.71 0.74 Ind1
0.0 0.2 0.4 0.6 0.8 1.0

Figure S2. There is a high degree of correlation between the sperm H3K4me3 ChIP-seq signal in individual men and the pooled reference population. Related to Figure 1. The heat map shows the Spearman correlation coefficient of pairwise comparison between the sample prepared from the pooled sperm of 30 men (Ref.Pop., for Reference Population) and 7 individual men (Ind1-7). The Spearman correlation coefficient between two samples was calculated based on the ChIP-seq signal calculated in RPKM (Reads Per Kilobase of exon per Million fragments mapped) at each H3K4me3 peak identified in the reference population.

Figure S3.


A
Location
chr12
965 kb
Sperm H3K4me3
Ind1
Ind2
Ind3
Ind4
Ind5
Ind6
Ind7
Ref. Pop. peaks
Genes

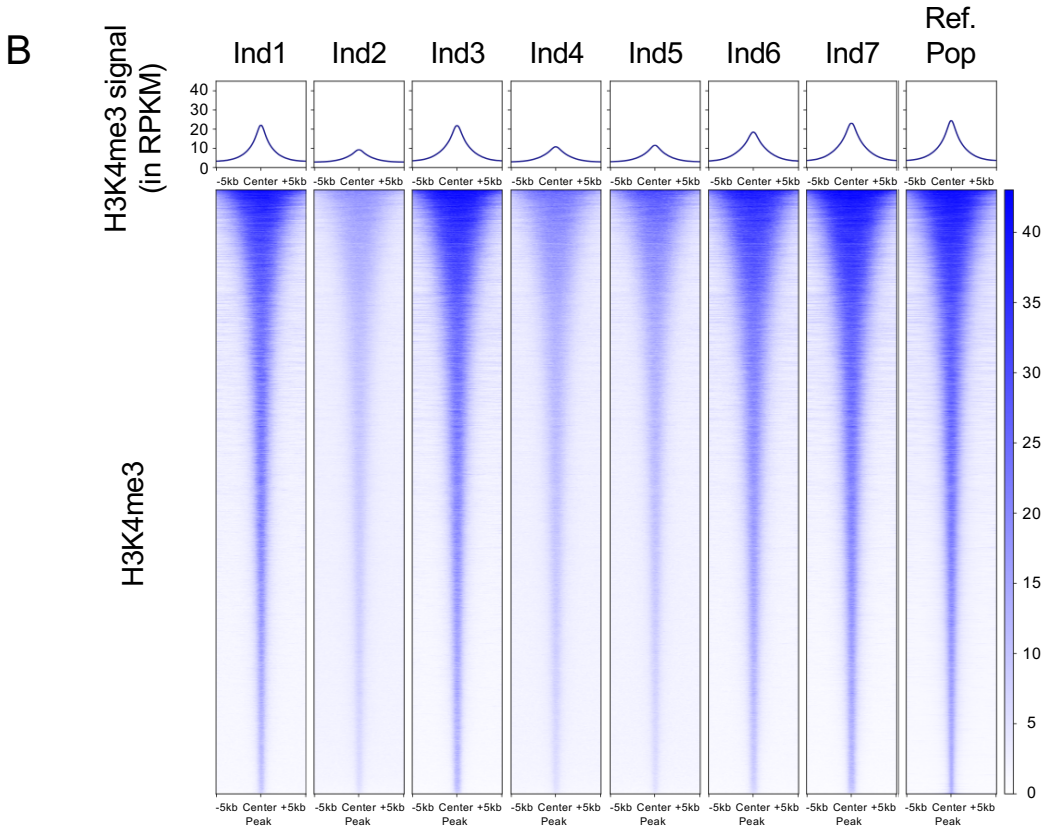
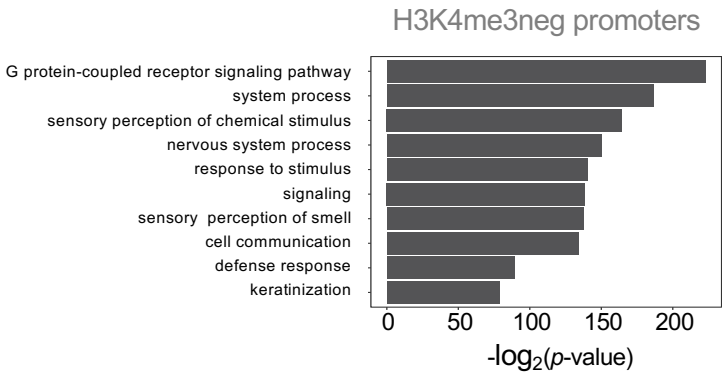
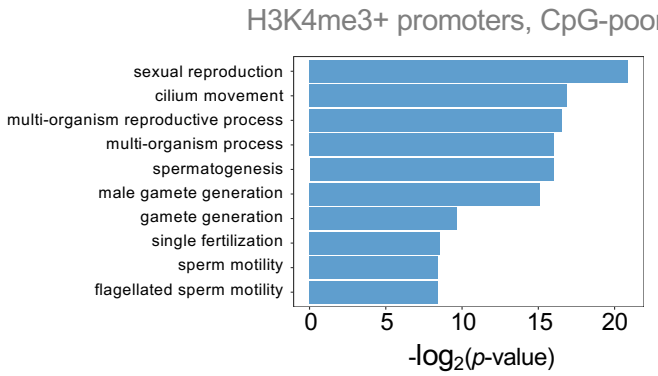
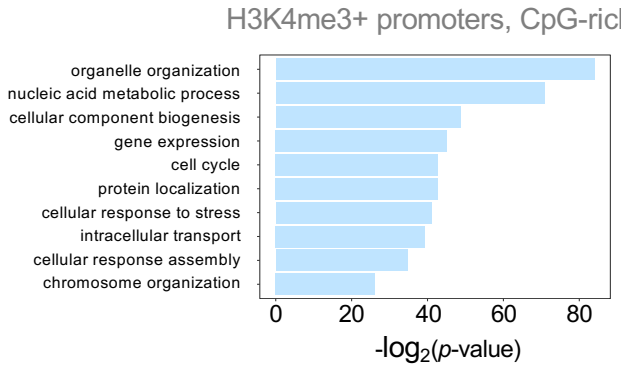

B
Ind1
Ind2
Ind3
Ind4
Ind5
Ind6
Ind7
Ref. Pop
H3K4me3 signal (in RPKM)
H3K4me3
-5kb Center +5kb
Peak

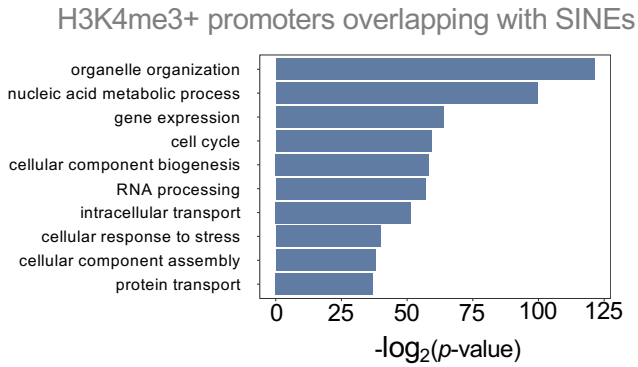
Figure S3. Comparison of the H3K4me3 ChIP-seq signal between the reference population and individual sperm samples. Related to Figure 1. **(A)** The high degree of conservation of H3K4me3 peaks between pooled and individual sperm samples is displayed in a genome browser snapshot for the reference population of 30 men (blue track; dark blue boxes indicate H3K4me3 peaks called by MACS2) and 7 individual men (Ind1-7, purple tracks) at a locus located on chromosome 12 (965kb). (**B**) Heat maps showing the H3K4me3 signal (in RPKM) around the center of the peaks (±5 Kb) for the reference population of 30 men and 7 individual men (Ind1-7). Each line represents one of the 50,117 H3K4me3 peaks identified in the reference population. The overall H3K4me3 signal is summarized on top of the heat maps.

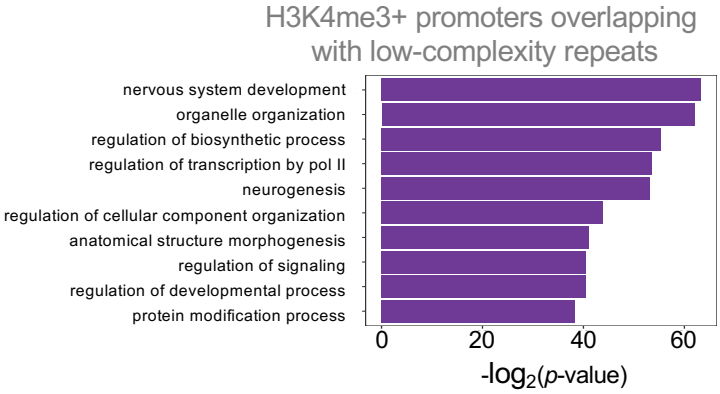
Figure S4.

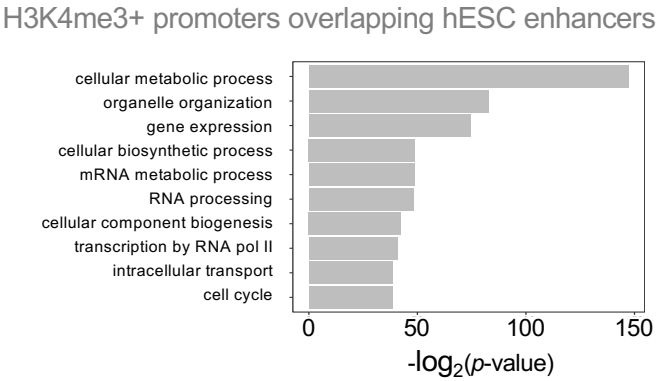

H3K4me3neg promoters
G protein-coupled receptor signaling pathway
system process
sensory perception of chemical stimulus
nervous system process
response to stimulus
signaling
sensory perception of smell
cell communication
defense response
keratinization
0
50
100
150
200
-log2(p-value)

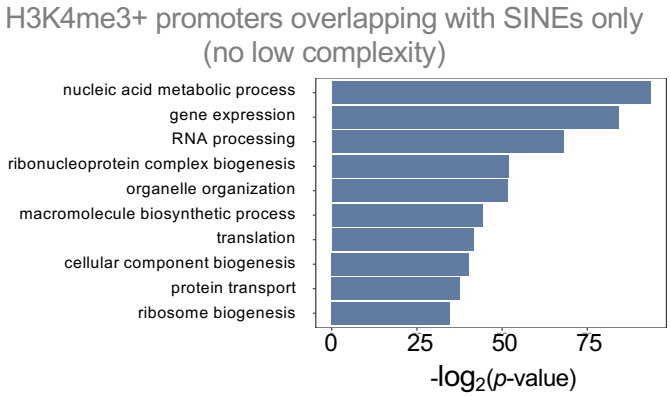

H3K4me3+ promoters, CpG-poor
sexual reproduction
cilium movement
multi-organism reproductive process
multi-organism process
spermatogenesis
male gamete generation
gamete generation
single fertilization
sperm motility
flagellated sperm motility
0
5
10
15
20
-log2(p-value)


H3K4me3+ promoters, CpG-rich
organelle organization
nucleic acid metabolic process
cellular component biogenesis
gene expression
cell cycle
protein localization
cellular response to stress
intracellular transport
cellular response assembly
chromosome organization
0
20
40
60
80
-log2(p-value)


H3K4me3+ promoters overlapping with SINEs
organelle organization
nucleic acid metabolic process
gene expression
cell cycle
cellular component biogenesis
RNA processing
intracellular transport
cellular response to stress
cellular component assembly
protein transport
0
25
50
75
100
125
-log2(p-value)


H3K4me3+ promoters overlapping
with low-complexity repeats
nervous system development
organelle organization
regulation of biosynthetic process
regulation of transcription by pol II
neurogenesis
regulation of cellular component organization
anatomical structure morphogenesis
regulation of signaling
regulation of developmental process
protein modification process
0
20
40
60
-log2(p-value)


H3K4me3+ promoters overlapping hESC enhancers
cellular metabolic process
organelle organization
gene expression
cellular biosynthetic process
mRNA metabolic process
RNA processing
cellular component biogenesis
transcription by RNA pol II
intracellular transport
cell cycle
0
50
100
150
-log2(p-value)


H3K4me3+ promoters overlapping with SINEs only
(no low complexity)
nucleic acid metabolic process
gene expression
RNA processing
ribonucleoprotein complex biogenesis
organelle organization
macromolecule biosynthetic process
translation
cellular component biogenesis
protein transport
ribosome biogenesis
0
25
50
75
-log2(p-value)

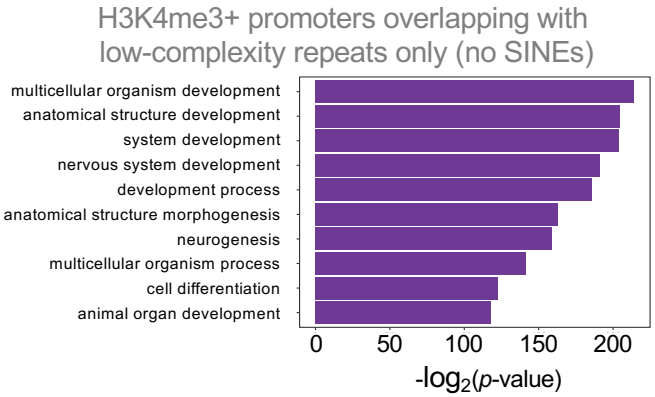

H3K4me3+ promoters overlapping with
low-complexity repeats only (no SINEs)
multicellular organism development
anatomical structure development
system development
nervous system development
development process
anatomical structure morphogenesis
neurogenesis
multicellular organism process
cell differentiation
animal organ development
0
50
100
150
200
-log2(p-value)

Figure S4. Gene ontology analysis of the different types of genes enriched or not in H3K4me3 in human sperm. Related to Figure 1. The enrichment in a specific biological process was measured by PANTHER Over-representation Test (GO biological process complete). The *p* value was determined by Fisher's Exact test with Bonferroni correction for multiple testing. Top10 processes with the lowest *p* value are indicated after filtering for non-redundant processes with an enrichment ≥1.25. Full list of gene ontology enriched terms provided in Table S4.

Figure S5.

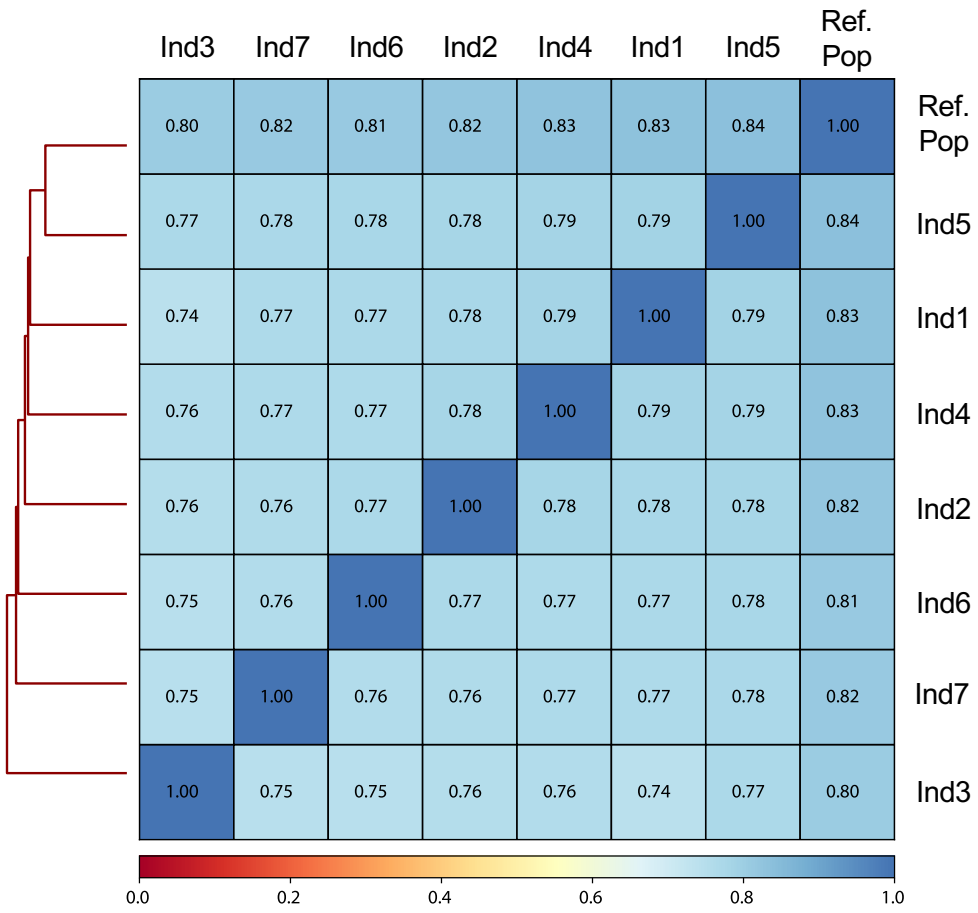

Ind3 Ind7 Ind6 Ind2 Ind4 Ind1 Ind5 Ref. Pop
0.80 0.82 0.81 0.82 0.83 0.83 0.84 1.00 Ref. Pop
0.77 0.78 0.78 0.78 0.79 0.79 1.00 0.84 Ind5
0.74 0.77 0.77 0.78 0.79 1.00 0.79 0.83 Ind1
0.76 0.77 0.77 0.78 1.00 0.79 0.79 0.83 Ind4
0.76 0.76 0.77 1.00 0.78 0.78 0.78 0.82 Ind2
0.75 0.76 1.00 0.77 0.77 0.77 0.78 0.81 Ind6
0.75 1.00 0.76 0.76 0.77 0.77 0.78 0.82 Ind7
1.00 0.75 0.75 0.76 0.76 0.74 0.77 0.80 Ind3
0.0 0.2 0.4 0.6 0.8 1.0

Figure S5. Correlation of the sperm DNA methylation levels between the pooled reference population and 7 individual men. Related to Figure 3. Heat map showing the Spearman correlation coefficient of pairwise comparison between the sample prepared from the pooled sperm of 30 men (Ref.Pop., for reference population) and 7 individual men (Ind1-7). The reference population data was obtained by WGBS (Chan, et al. 2020), whereas the 7 individual men data were determined by MethylC-Capture-seq (MCC-seq) on a reduced number of CpG in the genome. The spearman correlation coefficient between two samples was calculated based on the CpG methylation levels calculated by the software Bismark (version 0.18.1) (Krueger and Andrews 2011) for each CpG present in all datasets.

Figure S6.

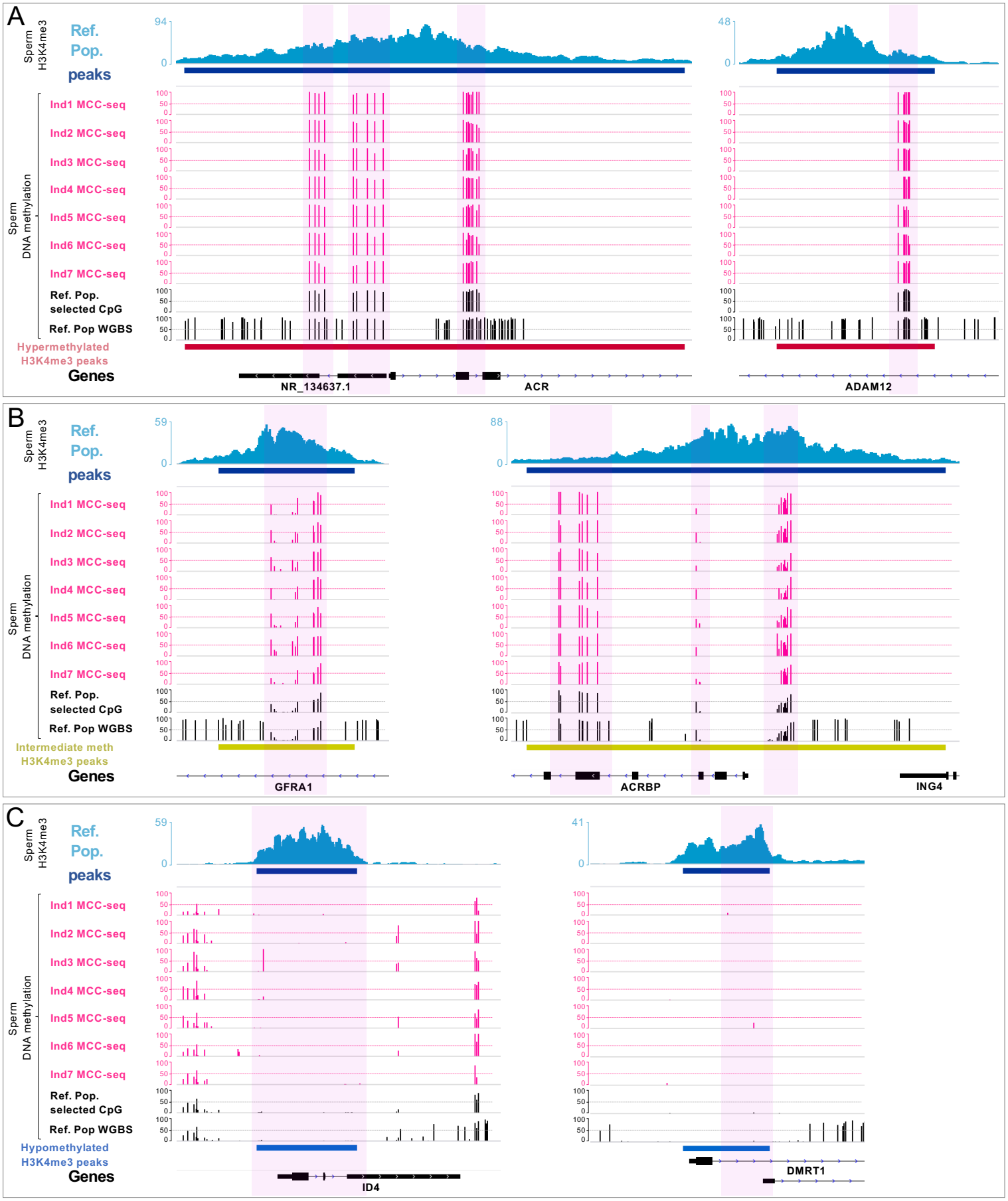
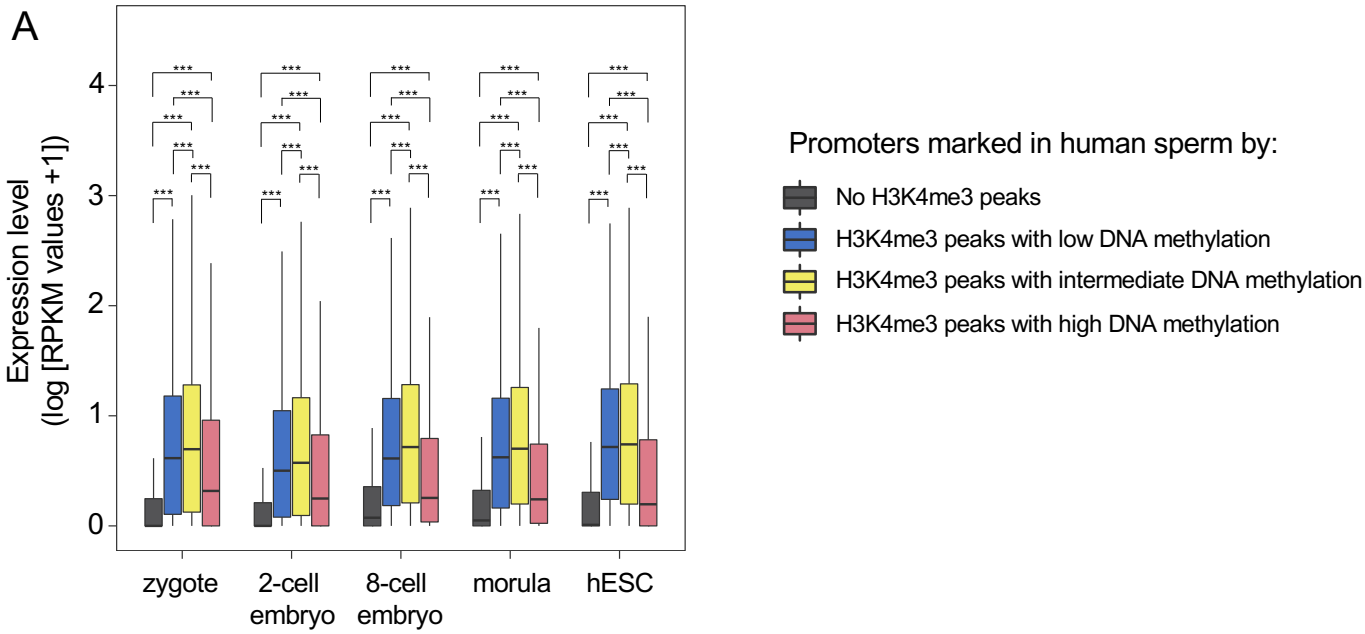
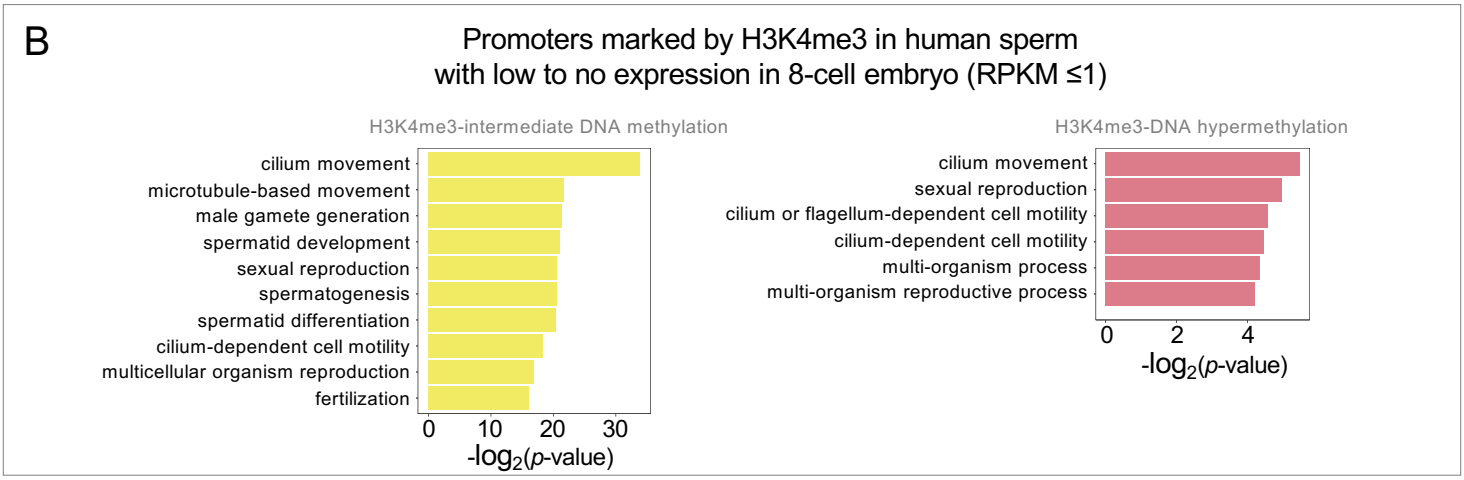

A
Sperm H3K4me3
Ref. Pop. peaks
94
0
48
0
Sperm DNA methylation
Ind1 MCC-seq
Ind2 MCC-seq
Ind3 MCC-seq
Ind4 MCC-seq
Ind5 MCC-seq
Ind6 MCC-seq
Ind7 MCC-seq
Ref. Pop. selected CpG
Ref. Pop WGBS
Hypermethylated H3K4me3 peaks
Genes
NR_134637.1
ACR
ADAM12
B
Sperm H3K4me3
Ref. Pop. peaks
59
0
88
0
Sperm DNA methylation
Ind1 MCC-seq
Ind2 MCC-seq
Ind3 MCC-seq
Ind4 MCC-seq
Ind5 MCC-seq
Ind6 MCC-seq
Ind7 MCC-seq
Ref. Pop. selected CpG
Ref. Pop WGBS
Intermediate meth H3K4me3 peaks
Genes
GFRA1
ACRBP
ING4
C
Sperm H3K4me3
Ref. Pop. peaks
59
0
41
0
Sperm DNA methylation
Ind1 MCC-seq
Ind2 MCC-seq
Ind3 MCC-seq
Ind4 MCC-seq
Ind5 MCC-seq
Ind6 MCC-seq
Ind7 MCC-seq
Ref. Pop. selected CpG
Ref. Pop WGBS
Hypomethylated H3K4me3 peaks
Genes
ID4
DMRT1

Figure S6. High conservation of the sperm DNA methylation levels between the pooled reference population and 7 individual men for the DNA hyper- and hypomethylated H3K4me3 peaks. Related to Figure 3 and 4. Genome browser snapshots showing the reference population ChIP-seq track (blue), H3K4me3 peaks called by MACS2 (dark blue boxes), the MCC-seq tracks of each individual man (pink), the WGBS track of the reference population (black), the same data but only for the CpG covered by the MCC-seq (Ref.Pop. Selected CpGs) for two DNA hypermethylated H3K4me3 peaks (**A**), two H3K4me3 peaks with intermediate DNA methylation (**B**), and two hypomethylated peaks (**C**).

Figure S7.


A
Expression level
(log [RPKM values +1])
0
1
2
3
4
zygote
2-cell embryo
8-cell embryo
morula
hESC
Promoters marked in human sperm by:
No H3K4me3 peaks
H3K4me3 peaks with low DNA methylation
H3K4me3 peaks with intermediate DNA methylation
H3K4me3 peaks with high DNA methylation


B
Promoters marked by H3K4me3 in human sperm
with low to no expression in 8-cell embryo (RPKM ≤1)
H3K4me3-intermediate DNA methylation
cilium movement
microtubule-based movement
male gamete generation
spermatid development
sexual reproduction
spermatogenesis
spermatid differentiation
cilium-dependent cell motility
multicellular organism reproduction
fertilization
0
10
20
30
-log2(p-value)
H3K4me3-DNA hypermethylation
cilium movement
sexual reproduction
cilium or flagellum-dependent cell motility
cilium-dependent cell motility
multi-organism process
multi-organism reproductive process
0
2
4
-log2(p-value)

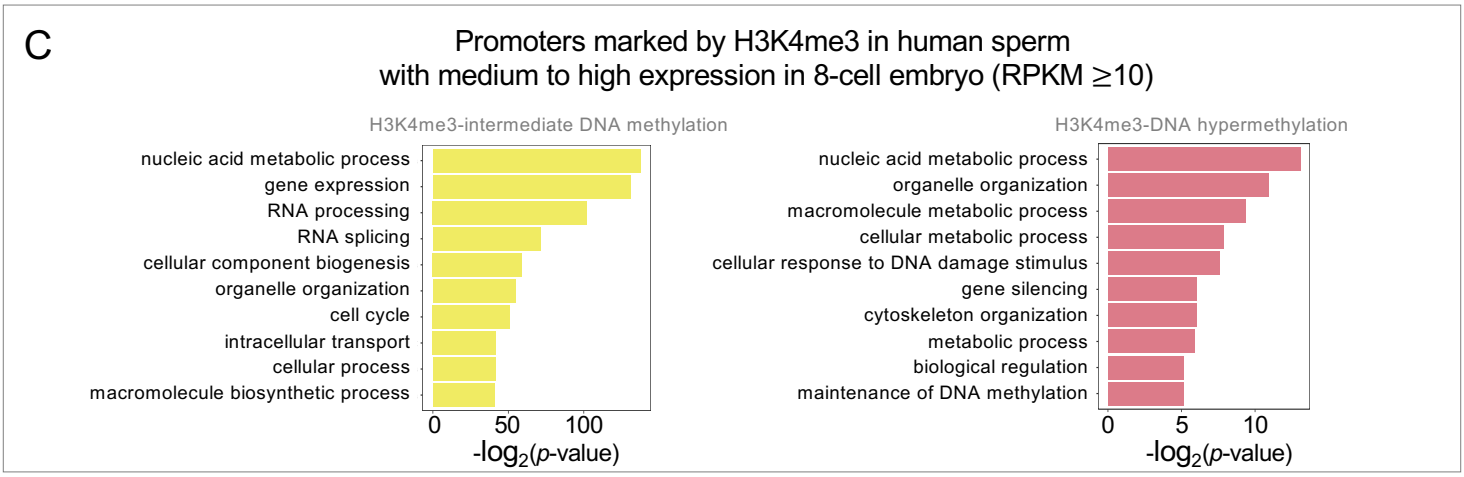

C
Promoters marked by H3K4me3 in human sperm
with medium to high expression in 8-cell embryo (RPKM ≥10)
H3K4me3-intermediate DNA methylation
nucleic acid metabolic process
gene expression
RNA processing
RNA splicing
cellular component biogenesis
organelle organization
cell cycle
intracellular transport
cellular process
macromolecule biosynthetic process
0
50
100
-log2(p-value)
H3K4me3-DNA hypermethylation
nucleic acid metabolic process
organelle organization
macromolecule metabolic process
cellular metabolic process
cellular response to DNA damage stimulus
gene silencing
cytoskeleton organization
metabolic process
biological regulation
maintenance of DNA methylation
0
5
10
-log2(p-value)

Figure S7. Expression levels in the embryo of the genes marked by H3K4me3 at their promoter in sperm that partially conserve or entirely lose this mark in the 8-cell embryo. Related to Figure 6. (**A**) Boxplot showing the level of expression of genes with promoters bearing no H3K4me3 in sperm, marked by H3K4me3 with low, intermediate or high DNA methylation during development. The level of transcription was visualized based on the transcript abundance expressed as RPKM. The adjusted *p* value was determined by pairwise Wilcoxon rank sum tests with a Benjamini-Hochberg correction for multiple testing; *** indicates an adjusted *p*-value <0.001. The enrichment in a specific biological process was measured by PANTHER Over-representation Test (GO biological process complete) for the genes marked by H3K4me3 sperm peaks with high and intermediated DNA methylation in sperm that had a low to no expression (**B**) or a medium to high expression (**C**) in 8-cell embryo. The *p* value was determined by Fisher's Exact test with Benjamini-Hochberg correction for multiple testing. Top10 processes with the lowest adjusted *p* value are indicated after filtering for non-redundant processes with an enrichment ≥1.25. Full list of gene ontology enriched terms provided in Table S4.