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.

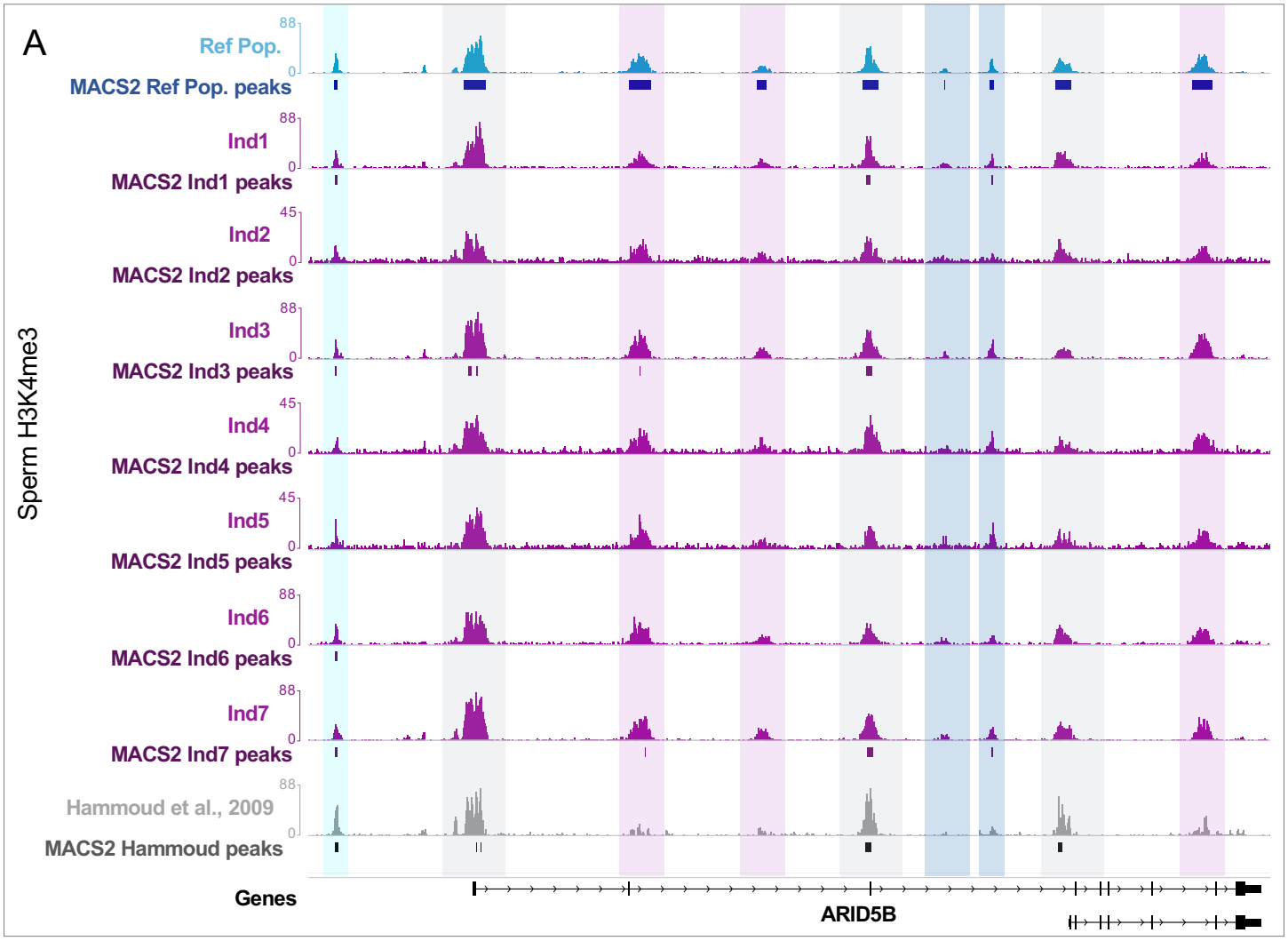
Patient ID	Age	Source	Count (M/mL)	DFI (%)	Motility (%)
Ind1	31	Toronto	109	10	61
Ind2	33	Toronto	35	20	69
Ind3	44	Toronto	161	10	55
Ind4	33	Toronto	103	14	33
Ind5	36	Toronto	123	20	47
Ind6	38	Toronto	94	21	70
Ind7	40	Toronto	153	10	75

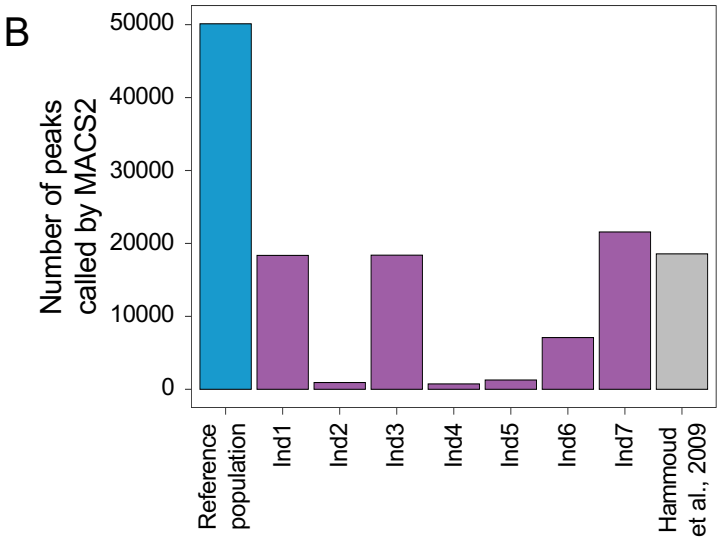
M/mL: million spermatozoa per mL semen.

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

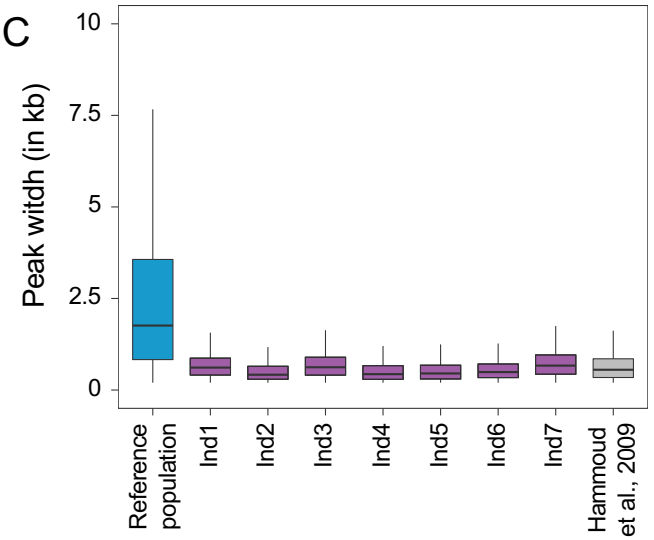

C
Peak witdh (in kb)
10
7.5
5
2.5
0
Reference population
Ind1
Ind2
Ind3
Ind4
Ind5
Ind6
Ind7
Hammoud et al., 2009

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.

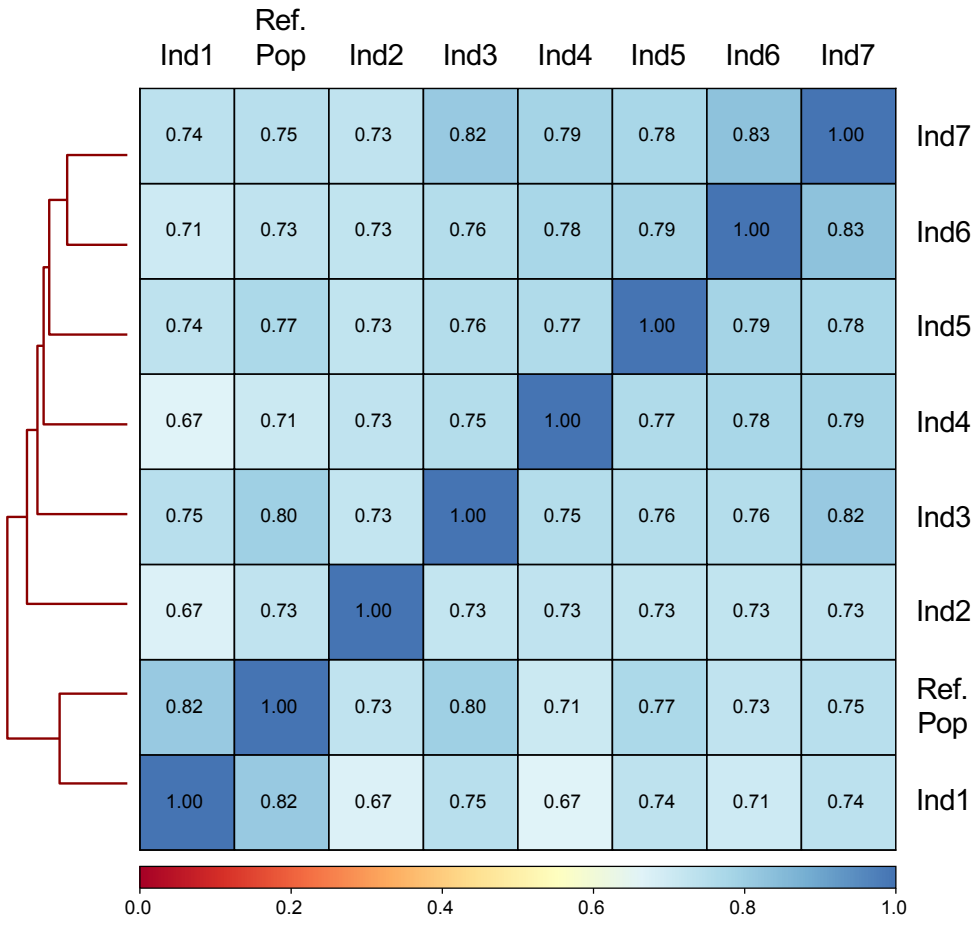
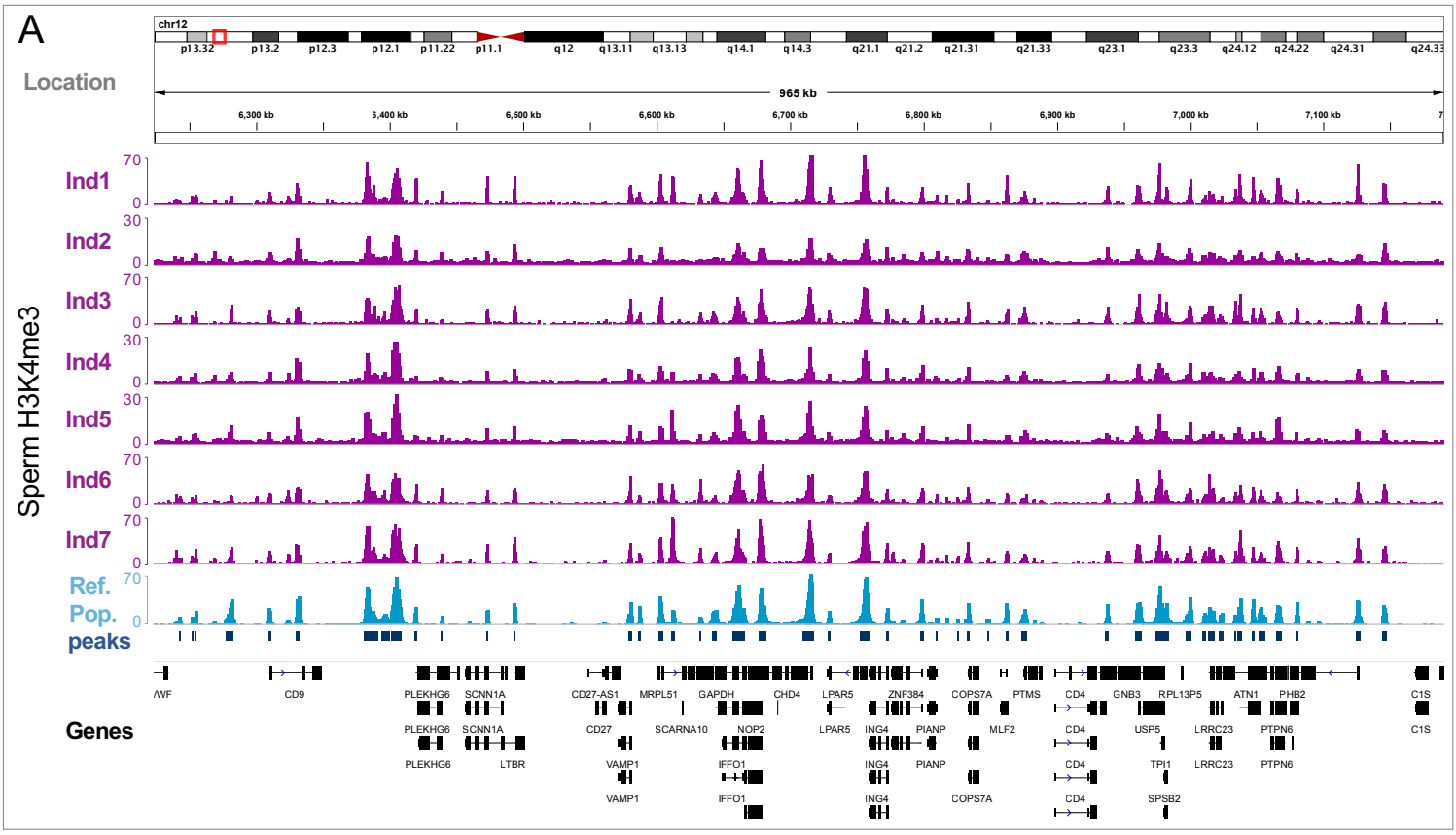

Ind1
Ref. Pop
Ind2
Ind3
Ind4
Ind5
Ind6
Ind7
0.74 0.75 0.73 0.82 0.79 0.78 0.83 1.00 Ind7
0.71 0.73 0.73 0.76 0.78 0.79 1.00 0.83 Ind6
0.74 0.77 0.73 0.76 0.77 1.00 0.79 0.78 Ind5
0.67 0.71 0.73 0.75 1.00 0.77 0.78 0.79 Ind4
0.75 0.80 0.73 1.00 0.75 0.76 0.76 0.82 Ind3
0.67 0.73 1.00 0.73 0.73 0.73 0.73 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

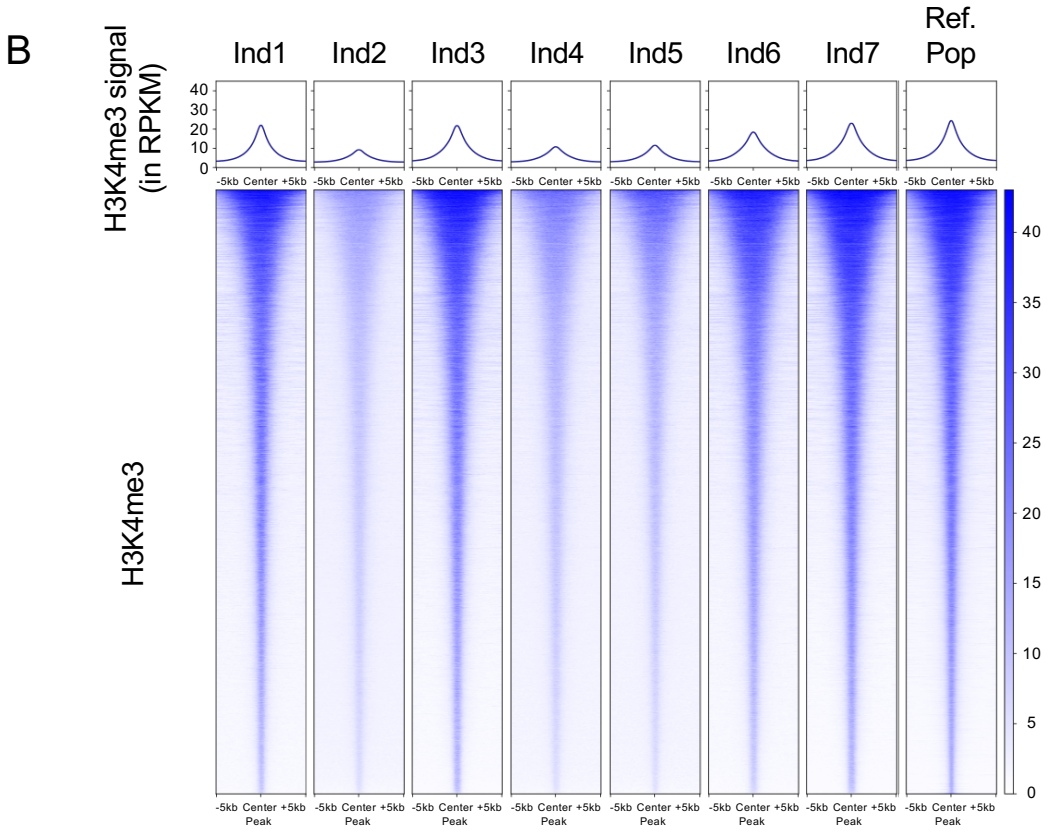
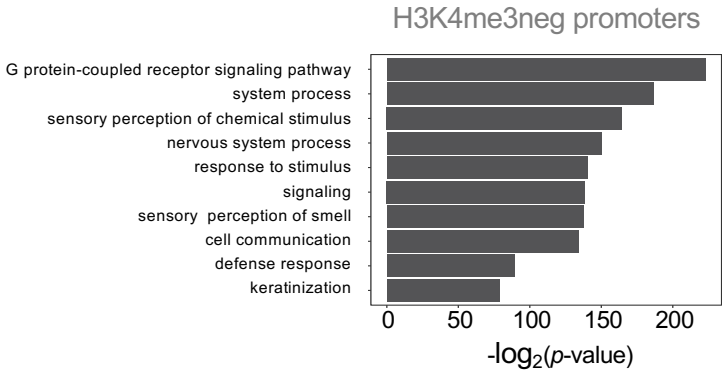
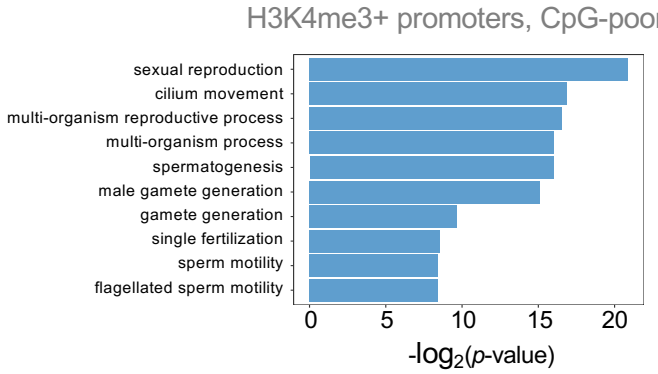
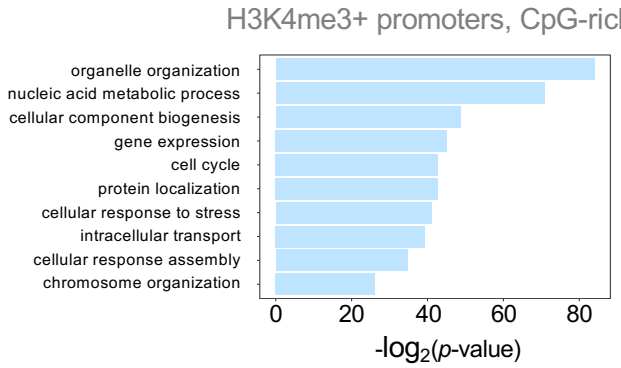

B
Ind1
Ind2
Ind3
Ind4
Ind5
Ind6
Ind7
Ref. Pop
H3K4me3 signal (in RPKM)
H3K4me3
-5kb Center +5kb
Peak

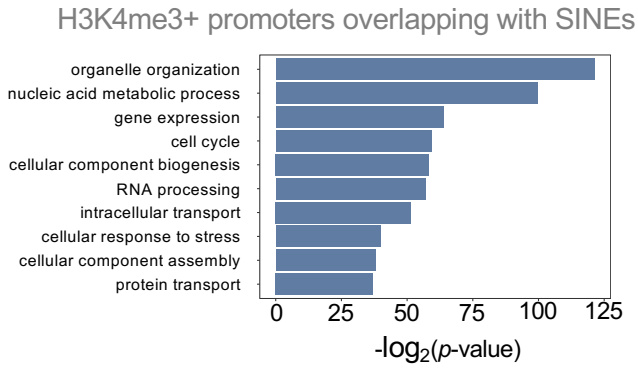
Figure S3. Comparison of the H3K4me3 ChIP-seq signal between the reference population and individual sperm samples. Related to Figure 1. **(A)** The high degree of conservation of H3K4me3 peaks between pooled and individual sperm samples is displayed in a genome browser snapshot for the reference population 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.

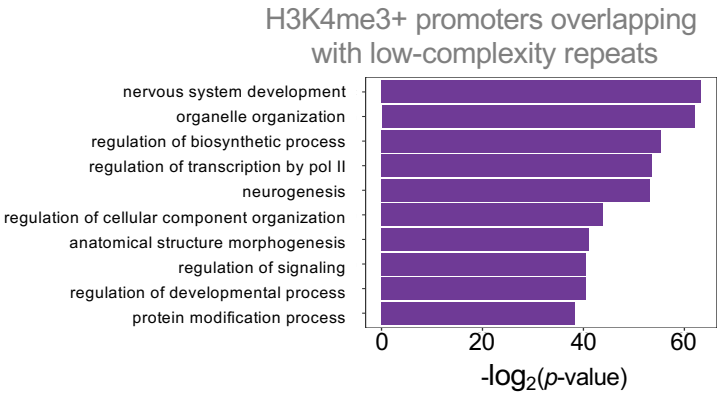
Figure S4.

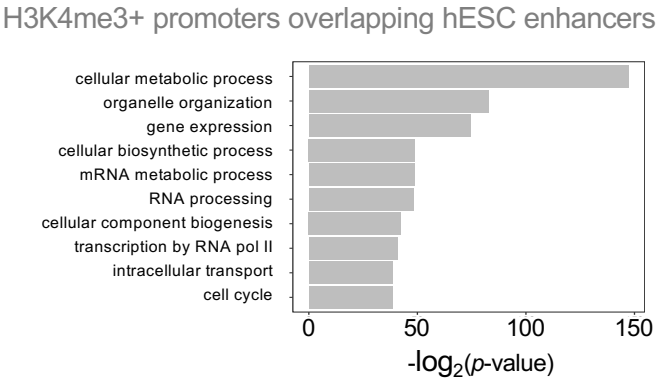

H3K4me3neg promoters
G protein-coupled receptor signaling pathway
system process
sensory perception of chemical stimulus
nervous system process
response to stimulus
signaling
sensory perception of smell
cell communication
defense response
keratinization
0
50
100
150
200
-log2(p-value)

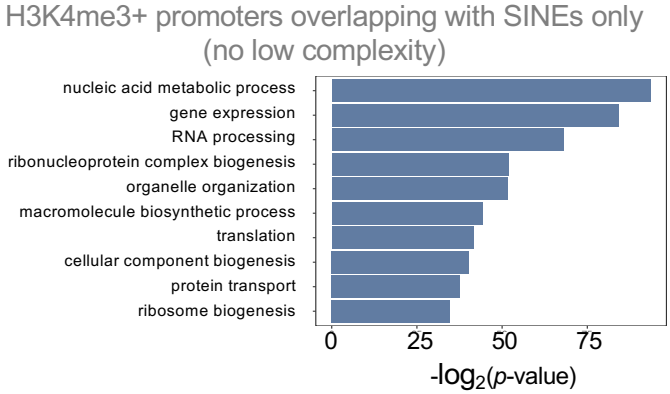

H3K4me3+ promoters, CpG-poor
sexual reproduction
cilium movement
multi-organism reproductive process
multi-organism process
spermatogenesis
male gamete generation
gamete generation
single fertilization
sperm motility
flagellated sperm motility
0
5
10
15
20
-log2(p-value)


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)

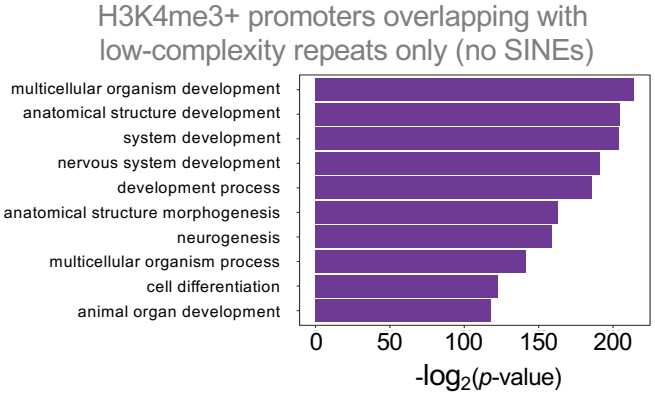

H3K4me3+ promoters overlapping with
low-complexity repeats only (no SINEs)
multicellular organism development
anatomical structure development
system development
nervous system development
development process
anatomical structure morphogenesis
neurogenesis
multicellular organism process
cell differentiation
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.

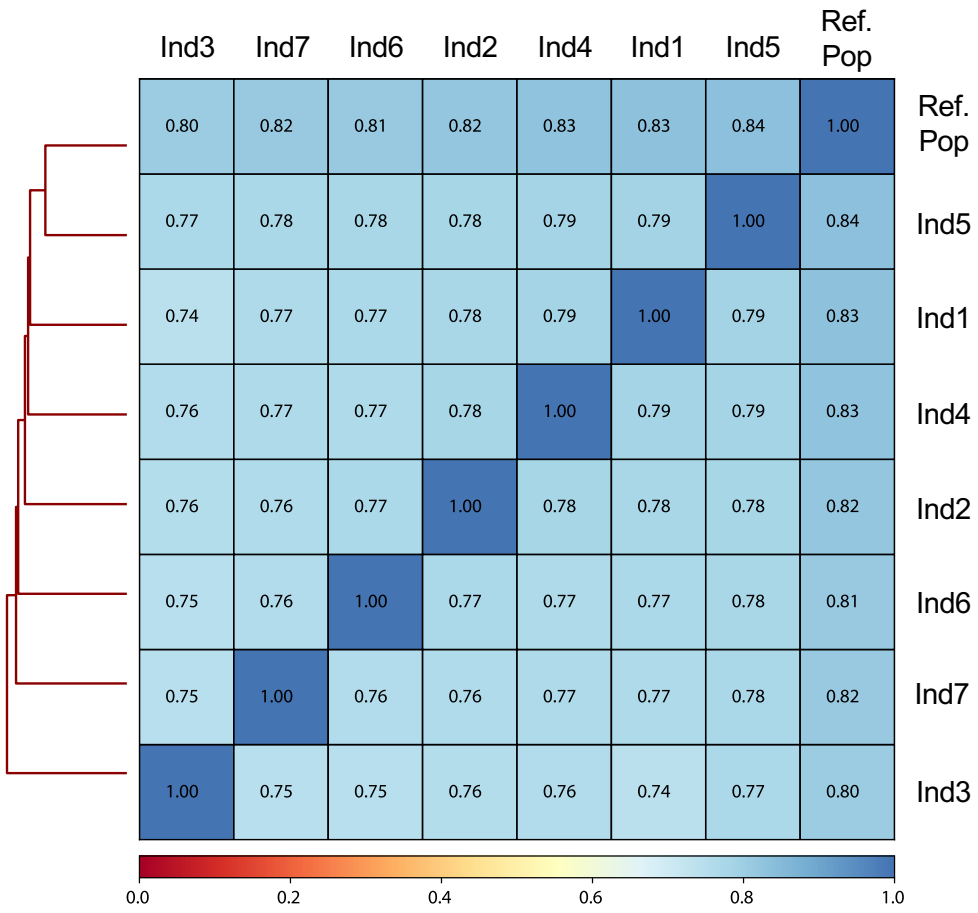

Ind3
Ind7
Ind6
Ind2
Ind4
Ind1
Ind5
Ref. Pop
0.80 0.82 0.81 0.82 0.83 0.83 0.84 1.00
0.77 0.78 0.78 0.78 0.79 0.79 1.00 0.84
0.74 0.77 0.77 0.78 0.79 1.00 0.79 0.83
0.76 0.77 0.77 0.78 1.00 0.79 0.79 0.83
0.76 0.76 0.77 1.00 0.78 0.78 0.78 0.82
0.75 0.76 1.00 0.77 0.77 0.77 0.78 0.81
0.75 1.00 0.76 0.76 0.77 0.77 0.78 0.82
1.00 0.75 0.75 0.76 0.76 0.74 0.77 0.80
Ref. Pop
Ind5
Ind1
Ind4
Ind2
Ind6
Ind7
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.

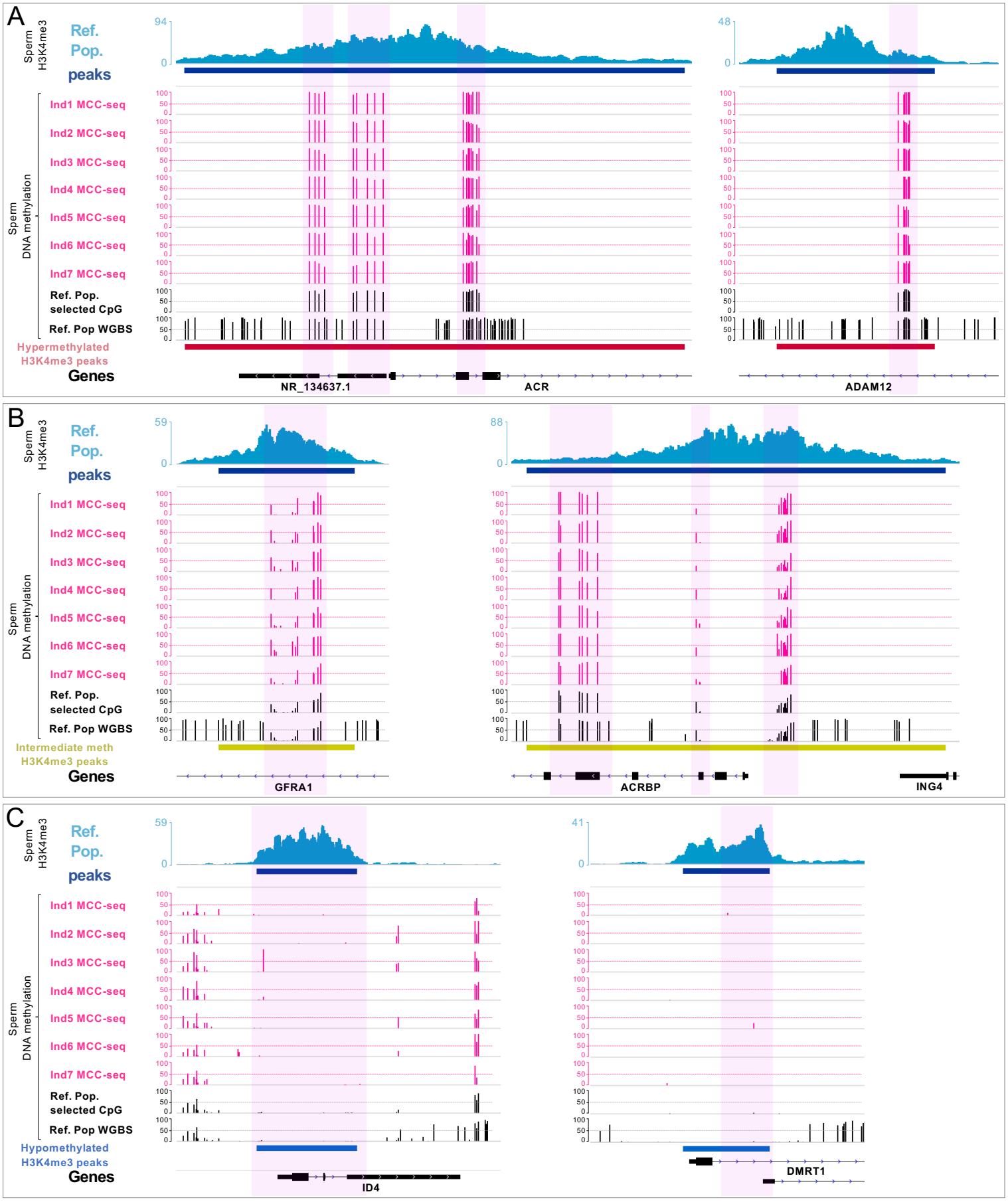
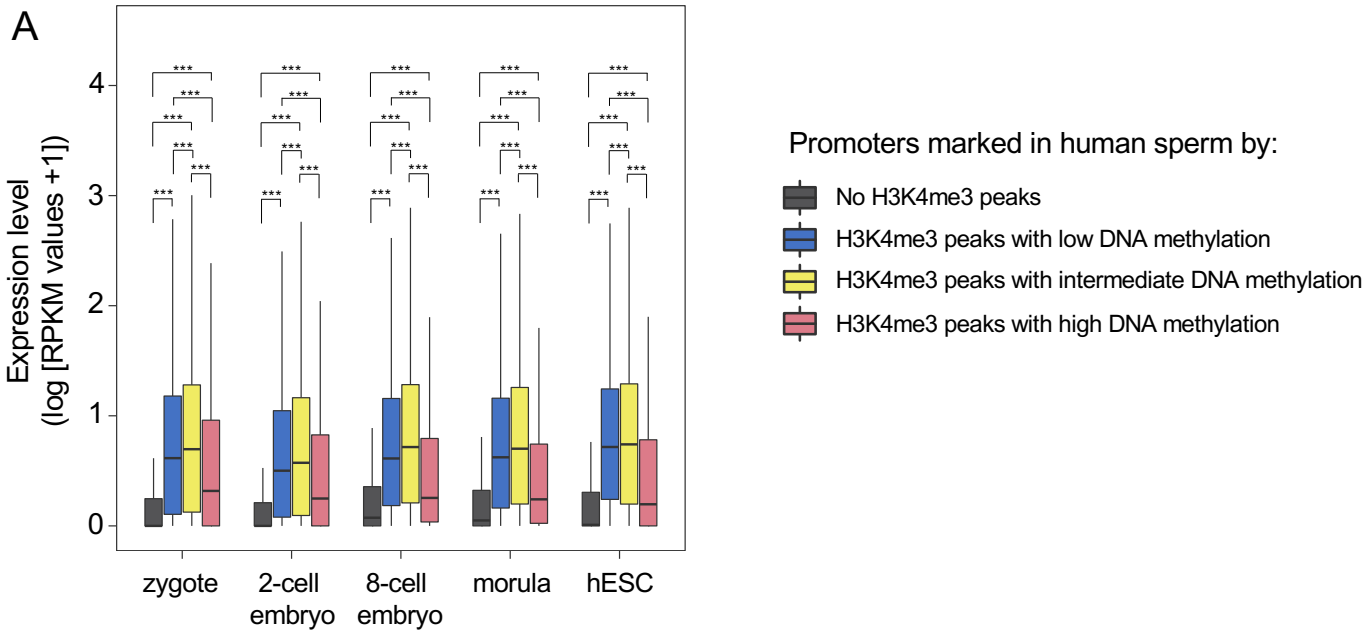
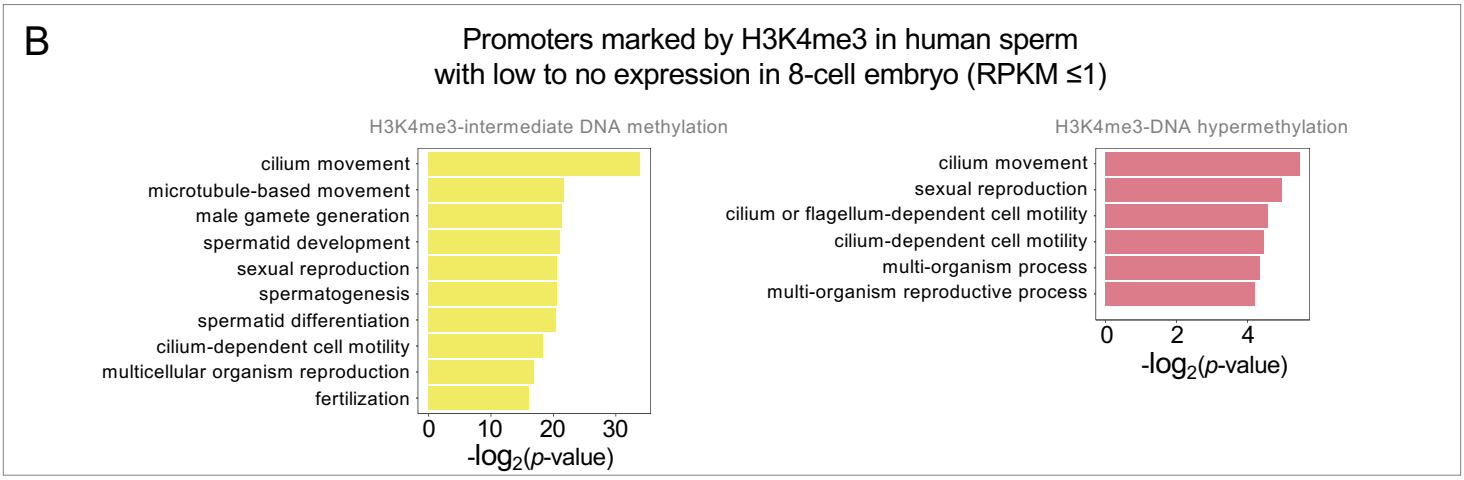

A
Sperm H3K4me3
Ref. Pop. peaks
94
0
48
0
Sperm DNA methylation
Ind1 MCC-seq
Ind2 MCC-seq
Ind3 MCC-seq
Ind4 MCC-seq
Ind5 MCC-seq
Ind6 MCC-seq
Ind7 MCC-seq
Ref. Pop. selected CpG
Ref. Pop WGBS
Hypermethylated H3K4me3 peaks
Genes
NR_134637.1
ACR
ADAM12
B
Sperm H3K4me3
Ref. Pop. peaks
59
0
88
0
Sperm DNA 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)

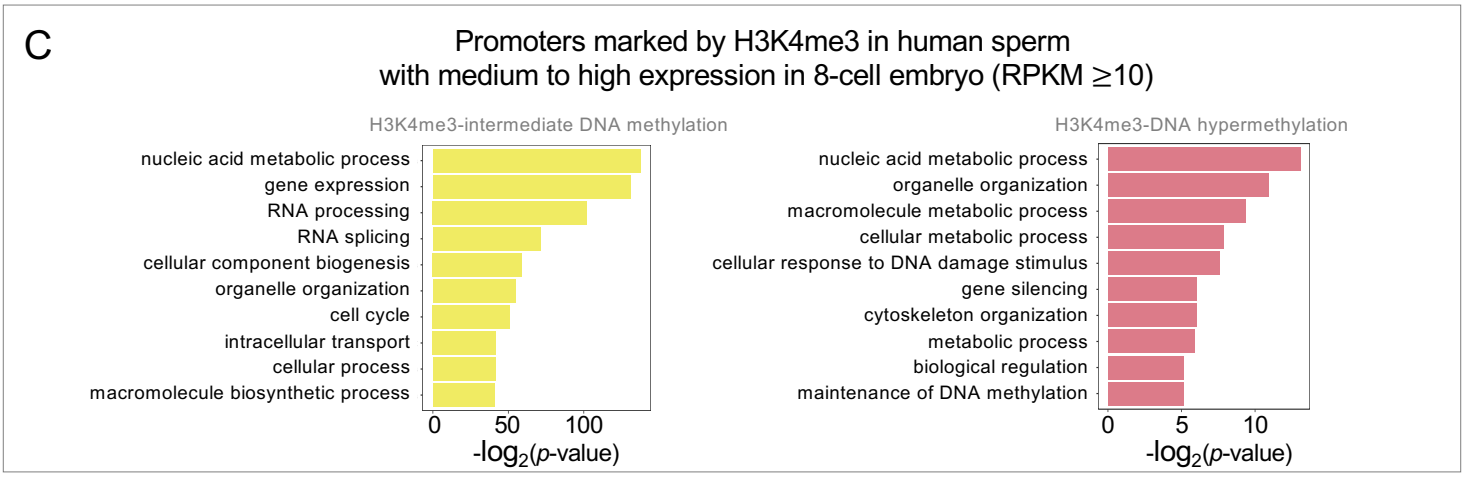

C
Promoters marked by H3K4me3 in human sperm
with medium to high expression in 8-cell embryo (RPKM ≥10)
H3K4me3-intermediate DNA methylation
nucleic acid metabolic process
gene expression
RNA processing
RNA splicing
cellular component biogenesis
organelle organization
cell cycle
intracellular transport
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.