Distinctive chromatin in human sperm packages genes for embryo development

Saher Sue Hammoud[1,2], David A. Nix[3], Haiying Zhang[1], Jahnvi Purwar[1], Douglas T. Carrell[2] & Bradley R. Cairns[1]

Because nucleosomes are widely replaced by protamine in mature human sperm, the epigenetic contributions of sperm chromatin to embryo development have been considered highly limited. Here we show that the retained nucleosomes are significantly enriched at loci of developmental importance, including imprinted gene clusters, microRNA clusters, *HOX* gene clusters, and the promoters of stand-alone developmental transcription and signalling factors. Notably, histone modifications localize to particular developmental loci. Dimethylated lysine 4 on histone H3 (H3K4me2) is enriched at certain developmental promoters, whereas large blocks of H3K4me3 localize to a subset of developmental promoters, regions in *HOX* clusters, certain noncoding RNAs, and generally to paternally expressed imprinted loci, but not paternally repressed loci. Notably, trimethylated H3K27 (H3K27me3) is significantly enriched at developmental promoters that are repressed in early embryos, including many bivalent (H3K4me3/H3K27me3) promoters in embryonic stem cells. Furthermore, developmental promoters are generally DNA hypomethylated in sperm, but acquire methylation during differentiation. Taken together, epigenetic marking in sperm is extensive, and correlated with developmental regulators.

During spermiogenesis canonical histones are largely exchanged for protamines[1,2], small basic proteins that form tightly packed DNA structures important for normal sperm function[3]. We find about 4% of the haploid genome retained in nucleosomes (Supplementary Fig. 1a). The rare retained nucleosomes in sperm consist of either canonical or histone variant proteins, including a testes-specific histone H2B (TH2B) with an unknown specialized function[4,5]. Their presence may simply be due to inefficient protamine replacement, leading to a low random distribution genome-wide with no impact in the embryo. Alternatively, these retained nucleosomes, along with attendant modifications, might be enriched at particular genes/loci. This latter possibility would raise the possibility for programmatic retention for an epigenetic function in the embryo. To address these questions, we localized the nucleosomes retained in mature sperm from fertile donors using high-resolution genomic approaches.

Developmental loci bear nucleosomes

To address donor variability, we examined nucleosome retention in a single donor (D1) and/or a pool of four donors (donor pool). Sperm chromatin was separated into protamine-bound and histone-bound fractions. In brief, mononucleosomes were isolated (>95% yield) by sequential MNase digestion and sedimentation (Supplementary Fig. 1b–e). This mononucleosome pool was used for chromatin immunoprecipitation (ChIP; to select modified nucleosomes), or the DNA was isolated from the mononucleosome pool to represent all nucleosomes. Purified DNA was subjected to high-throughput sequencing (Illumina GAII), or alternatively, was labelled and hybridized to a high-density promoter-tiling array (9 kilobase (kb) tiled; Supplementary Fig. 2, schematic).

Our initial array approach examined three replicas of D1 (pairwise average $R^2 = 0.85$). Notably, Gene Ontology analysis revealed nucleosomes significantly enriched at promoters that guide embryonic development—primarily developmental transcription factors and signalling molecules (Gene Ontology term false discovery rate

(FDR) < 0.01; Box 1 and Supplementary Table 1; for all extended Gene Ontology categories see Supplementary Tables and Supplementary Data Set 1). To conduct genome-wide profiling, we performed high-throughput sequencing of nucleosomes from D1 or the donor pool. Regions significantly enriched for histone relative to the input control (sheared total sperm DNA) were identified using a 300-base-pair (bp) window metric[6]. For display, we depict the normalized difference score and FDR window scores (Fig. 1a, FDR transformation ($-10 \log_{10}$ (q-value FDR)), 20 = 0.01, 25 = 0.003, 30 = 0.001, and 40 = 0.0001). Histone-enriched loci for one individual (D1) were well correlated with a donor pool ($r = 0.7$). Globally, 76% of the top 9,841 histone-enriched regions (FDR 40 cutoff) intersect genic regions, whereas the expected intersection given random distribution is 36% ($P < 0.001$).

Interestingly, sequencing of D1 or the donor pool revealed significant (FDR < 0.001) histone retention at many loci important for embryo development, including embryonic transcription factors and signalling pathway components (Box 1, Supplementary Tables 2 and 3). We show this enrichment at *HOX* loci (Fig. 1, Supplementary Fig. 3), but also observe this at stand-alone developmental transcription factors (Supplementary Fig. 4) and signalling factors (Supplementary Fig. 5). An FDR of 60 yields 4,556 genes, of which 1,683 are grouped with developmental Gene Ontology categories (2,848 total developmental genes). The magnitude of nucleosome enrichment at developmental loci is modest, with high significance provided by a moderate average increase at a large number of loci. Histones are also significantly enriched at the promoters of microRNAs (miRNAs) ($P < 0.05$; Supplementary Fig. 6) and at the class of imprinted genes ($P < 0.0001$; Fig. 2), addressed in detail later. Selected loci were tested and confirmed by quantitative PCR (qPCR; Supplementary Fig. 7a–e). Outside of these enriched regions, we observe sequencing reads at low levels distributed genome-wide (for example, Figs 1a and 2a), an observation consistent with low levels of nucleosomes genome-wide, although contributions from non-nucleosomal contamination cannot be ruled out.

[1]Howard Hughes Medical Institute, Department of Oncological Sciences, and Huntsman Cancer Institute, [2]IVF and Andrology Laboratories, Departments of Surgery, Obstetrics and Gynecology, and Physiology, [3]Research Informatics and Bioinformatics Core Facility, Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, Utah 84112, USA.

473

Box 1 | Developmental genes are associated with particular chromatin attributes in human sperm

GoMiner was used to identify enriched categories, and all categories displayed have an FDR < 0.01. The top five general categories are listed, after omitting nearly identical/redundant classes. An expanded gene ontology table with the unfiltered top 30–60 categories, the total genes, number of changed genes, enrichment, and FDR are provided in the Supplementary Information.

Nucleosomes, Array D1
(1) Sequence-specific DNA binding; (2) multicellular organismal development; (3) regulation of transcription; (4) developmental process; (5) regulation of metabolic process.

Nucleosomes, Illumina GAII pooled donors
(1) Transcription factor activity; (2) cell fate commitment; (3) WNT receptor signalling; (4) neuron development; (5) embryonic development.

H3K4me2, Array D1
(1) Multicellular organismal development; (2) developmental process; (3) sequence-specific DNA binding; (4) anatomical structure development; (5) system development.

H3K4me3, Array D1
(1) mRNA processing; (2) RNA binding; (3) cell cycle; (4) transcription; (5) RNA splicing.

H3K4me3, Illumina GAII pooled donors
(1) RNA splicing; (2) translation; (3) cell cycle; (4) RNA metabolic process; (5) transcription.

H3K27me3, Illumina GAII pooled donors
(1) WNT receptor signalling; (2) embryonic organ development and morphogenesis; (3) cell fate commitment; (4) neuron differentiation; (5) sequence-specific DNA binding.

DNA hypomethylated promoters D1 and D2
(1) Embryonic development; (2) multicellular organismal development; (3) system development; (4) RNA biosynthetic process; (5) transcription factor activity.

DNA methylated promoters omitting CpG islands, array
(1) Transcription; (2) RNA biosynthetic process; (3) regulation of transcription; (4) embryonic development; (5) embryo morphogenesis.

Protamine occupancy (two replicas, $R^2 = 0.89$, arrays only) yielded 7,151 enriched regions (>2.5-fold), but failed to identify any enriched Gene Ontology term categories, although a few segments of the Y chromosome were notably enriched (including the testis-specific *TSPY* genes, data not shown). Regions of histone enrichment did not exclude protamine, consistent with a nucleosome-protamine mixture existing even at histone-enriched loci. However, as protamine fragments averaged ~750 bp, protamine depletion would have to be extensive (regions >2 kb) to be apparent on our arrays. Taken together, nucleosomes are significantly enriched in sperm at genes important for embryonic development, with transcription factors the most enriched class.

Localization of modified nucleosomes

Because histones replace protamines genome-wide at fertilization[7,8], unmodified histones retained in sperm would seem insufficient to influence gene regulation in embryos. Therefore, we examined three further chromatin properties in sperm: (1) histone variants, (2) histone modifications, and (3) DNA methylation. ChIP combined with promoter microarray analysis (termed ChIP-chip) of TH2B (two replicas, $R^2 = 0.93$) shows 0.3% of gene promoters with relatively high levels of TH2B (>twofold enrichment). Gene Ontology analysis showed significant (FDR < 0.06) enrichment at genes

important for sperm biology, capacitation and fertilization (Supplementary Table 4), but not at developmental categories. ChIP sequencing (ChIP-seq) analysis with H2A.Z nucleosomes (at standard conditions, 150–250 mM salt) did not show significant enriched Gene Ontology categories, with high enrichment limited to pericentric heterochromatin (Supplementary Fig. 8), consistent with prior immunostaining[9].

Modified nucleosomes were localized by performing ChIP on mononucleosomes, followed by either array analysis or sequencing (Supplementary Fig. 2, schematic). We normalized the data set for each modification to the data set derived from input mononucleosomes, determined enriched regions (array >twofold; sequencing FDR 40), found the nearest neighbouring gene, and performed Gene Ontology analysis. In somatic cells, H3K4me2 is correlated with euchromatic regions. In sperm, H3K4me2 was enriched at many promoters, and at significant levels at promoters for developmental transcription factors (two replicas $R^2 = 0.94$; Gene Ontology term FDR < 0.06; Box 1 and Supplementary Table 5). In somatic cells, H3K4me3 is localized to: (1) the transcription start sites (TSS) of active genes, (2) genes bearing 'poised' RNA polymerase II (Pol II), and (3) the proximal promoter of inactive developmental regulators in embryonic stem (ES) cells—promoters that also bear the silencing mark H3K27me3 (refs 10, 11), and thus termed bivalent. Mature sperm are transcriptionally inert, and Pol II protein levels are barely detectable (data not shown), so the high H3K4me3 levels we observed in sperm chromatin (Supplementary Fig. 1f) seemed surprising. H3K4me3 was localized by both ChIP-chip (three replicas, $R^2 = 0.96$) and ChIP-seq. The raw data sets were similar ($r = 0.7$) and the thresholded data sets were very similar (array twofold; sequencing, FDR 40; 96% intersection, $P < 0.001$). With both data sets, simple inspection showed small peaks at many 5′ gene ends, with high levels and broader blocks at a subset of genes (that is, *HOX* loci; Fig. 1 and Supplementary Fig. 3). Gene Ontology term analyses with either data set yielded genes that are important for changing nuclear architecture, RNA metabolism, spermatogenesis, and also selected transcription factors important for embryonic development (FDR < 0.01, Box 1, Supplementary Tables 6 and 7 and Supplementary Fig. 9). H3K4me3 at genes related to nuclear architecture and spermatogenesis can presumably be attributed to their prior activation during gametogenesis. RNA metabolism occurs both in gametogenesis and the early embryo, so attribution to a prior program as opposed to a potential poising for a future program cannot be unambiguously attributed. However, several transcription and signalling factors of importance in embryo development exhibited high levels and a broad distribution of H3K4me3, including *EVX1/2*, *ID1*, *STAT3*, *KLF5*, *FGF9*, *SOX7/9*, certain *HOX* genes, and certain noncoding RNAs (Fig. 1 and Supplementary Figs 3 and 6).

Interestingly, ChIP-seq analysis showed significant levels of H3K27me3 at developmental promoters in sperm (Box 1, Fig. 1b, Supplementary Table 8 and Supplementary Figs 3 and 4), and overlapped significantly with H3K27me3-occupied genes in ES cells ($P < 0.01$), which are silent before differentiation. Furthermore, bivalent genes (bearing H3K4me3 and H3K27me3) in ES cells had a significant overlap with bivalent genes in sperm (FDR < 0.001 for each mark). Of the 1,999 genes identified as bivalent in ES cells, 861 were bivalent in sperm ($P < 0.01$; Supplementary Table 9). Also notable but not explored further were many blocks of high H3K4me3 or H3K27me3 in regions lacking annotation (Fig. 1a, oval). Furthermore, H3K9me3 was not detected at the small set developmental promoters tested, but was high at pericentric regions (qPCR only, Supplementary Fig. 7d). Taken together, our results demonstrate extensive histone modification patterns in sperm, and significant similarities to patterns observed in ES cells.

DNA methylation profiles

DNA methylation profiles examined two fertile donors (D2 and D4) using a methylated DNA immunoprecipitation (MeDIP) procedure
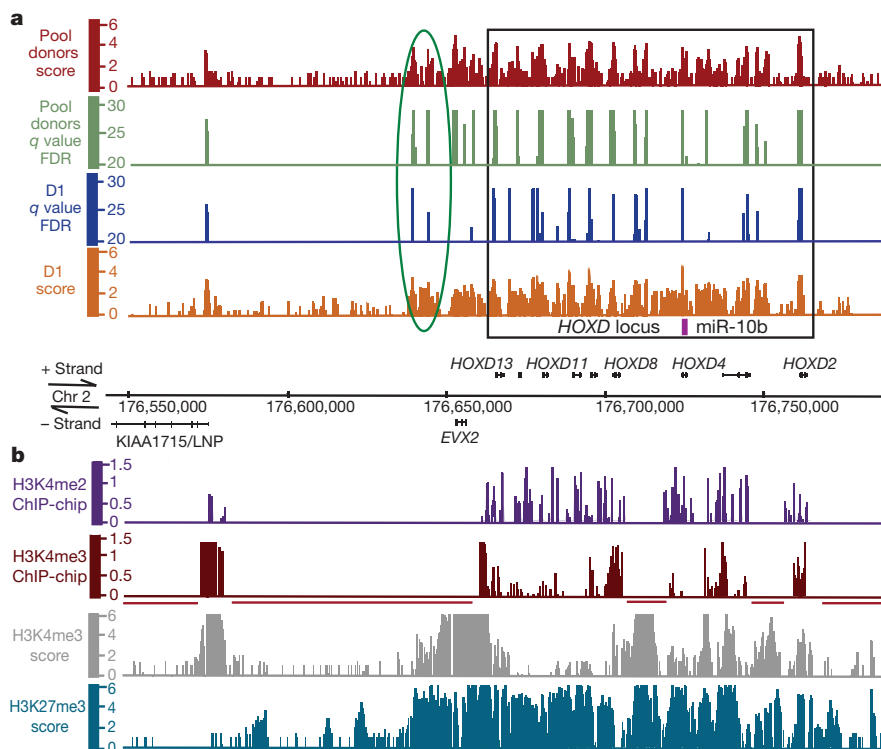
Figure 1 | Profiling of nucleosomes and their modifications at *HOXD*. For high-throughput sequencing, we show the mapped sequencing reads from D1 or a donor pool (red or orange bars, respectively; normalized difference score), and their significance (green or blue bars; FDR of 20 is <1% and FDR of 30 is <0.1%). **a**, The *HOXD* locus (black box) and an uncharacterized flanking locus (green oval). **b**, Profiling of nucleosome modifications at *HOXD* (in part **a**). The *y* axis is signal intensity ($\log_2$, for ChIP-chip), or the normalized difference score for sequencing. The regions not tiled on the array are underlined in red. Chr, chromosome.

and promoter arrays (individual replicates average D2 $R^2 = 0.97$ and D4 $R^2 = 0.89$). Their methylation patterns were highly similar (pairwise $R^2 = 0.86$), and extensive qPCR validated our array threshold (Supplementary Fig. 7e). Gene Ontology analysis of genes with pronounced DNA hypomethylation yielded transcription and signalling factors that guide embryo development (FDR < 0.05; Box 1 and Supplementary Table 10) including *HOX* loci (Fig. 3, blue bars, and Supplementary Figs 4 and 10). Hypomethylation also overlapped very significantly with histone-enriched promoters ($P < 0.02$; Supplementary Table 11). Bisulphite sequencing verified the MeDIP results, revealing extensive hypomethylation at developmental promoters in sperm (Supplementary Fig. 10b, c).

Notably, DNA-hypomethylated promoters in mature sperm overlap greatly with developmental promoters bound by the self-renewal network of transcription factors in human ES cells (for example, OCT4 (also known as POU5F1), SOX2, NANOG, KLF4 and FOXD3 proteins[12]; intersection of OCT4 protein occupancy and DNA hypomethylation, $P < 0.01$). In ES cells, these proteins promote self-renewal and also work with repressive polycomb complexes (PRC2; containing core component SUZ12) to help repress a large set of developmental regulators (including *HOX* genes) to prevent differentiation[10,13–20]. However, the hypomethylation of developmental genes in sperm is extensive (Fig. 3 and Supplementary Fig. 4). In fact, when CpG islands are omitted from the data sets, Gene Ontology term analysis of hypomethylated promoters still yields developmental genes (Box 1 and Supplementary Table 12). Notably, many of these developmental genes become methylated after differentiation; differential analysis of sperm and primary human fibroblasts (MeDIP, two replicas $R^2 = 0.86$) showed that many promoters occupied by PRC2 in human ES cells acquire methylation in fibroblasts (FDR < 0.01, Supplementary Tables 13 and 14; *HOXD* illustrated in Fig. 3, Supplementary Figs 4 and 5). Furthermore, the promoters driving several key members of the self-renewal network are themselves markedly hypermethylated in sperm

(*OCT4*, *NANOG* and *FOXD3*, bisulphite sequencing in Supplementary Fig. 10c), whereas their developmental target genes are hypomethylated (bisulphite sequencing in Supplementary Fig. 10b), consistent with recent studies in mice[21–24].

Attributes of *HOX* clusters and miRNAs

Nucleosome enrichment was clear across *HOX* loci and proximal flanking regions, but falls off precipitously outside (*HOXD*, Fig. 1a; *HOXA*, Supplementary Fig. 3a). Histone-enriched *HOXD* regions with a single donor (D1) were largely shared with the donor pool (Fig. 1a; D1 versus donor pool, $r = 0.7$). Notably, retained nucleosomes have regional covalent modifications. For example, distinct and very large (5–20 kb) blocks of H3K4me3 are clearly observed at all *HOX* loci, and also at certain imprinted genes (addressed later). At *HOXD*, high H3K4me3 extends for ~20 kb, encompassing all of *EVX2* and extending to the 3′ region of *HOXD13* (Fig. 1b). Remarkably, a similar profile is observed at the related *HOXA* locus (Supplementary Fig. 3a). At *HOXD* a second block of H3K4me3 is observed in the region between *HOXD4* and *HOXD8* (Fig. 1b), a region that encodes several noncoding RNAs expressed during development. This region represents a marked difference from the chromatin status in ES cells; in ES cells *HOXD8–D11* are all bivalent. The distribution of H3K4me2 (determined from two replicas of D1) is clearly different from H3K4me3 at *HOX* loci (Fig. 1b and Supplementary Fig. 3). For example, at *HOXD*, H3K4me2 is enriched in *HOXD8–D11*, a region deficient in H3K4me3 (Fig. 1b). Notably, high H3K27me3 encompasses all *HOX* loci and their proximal flanking regions. In contrast, high levels of H3K9me (a mark of heterochromatin; Supplementary Fig. 7d) or H2A.Z were not detected at the *HOX* loci tested.

Histones are enriched at many miRNAs, especially miRNA clusters (Supplementary Fig. 6). For example, 16 of the 29 miRNA clusters on autosomes were significantly enriched ($P < 0.05$). Clusters include those bearing *let7e*, *mir-17*, *mir-15a*, *mir-96*, *mir-135b* and *mir-10a/*
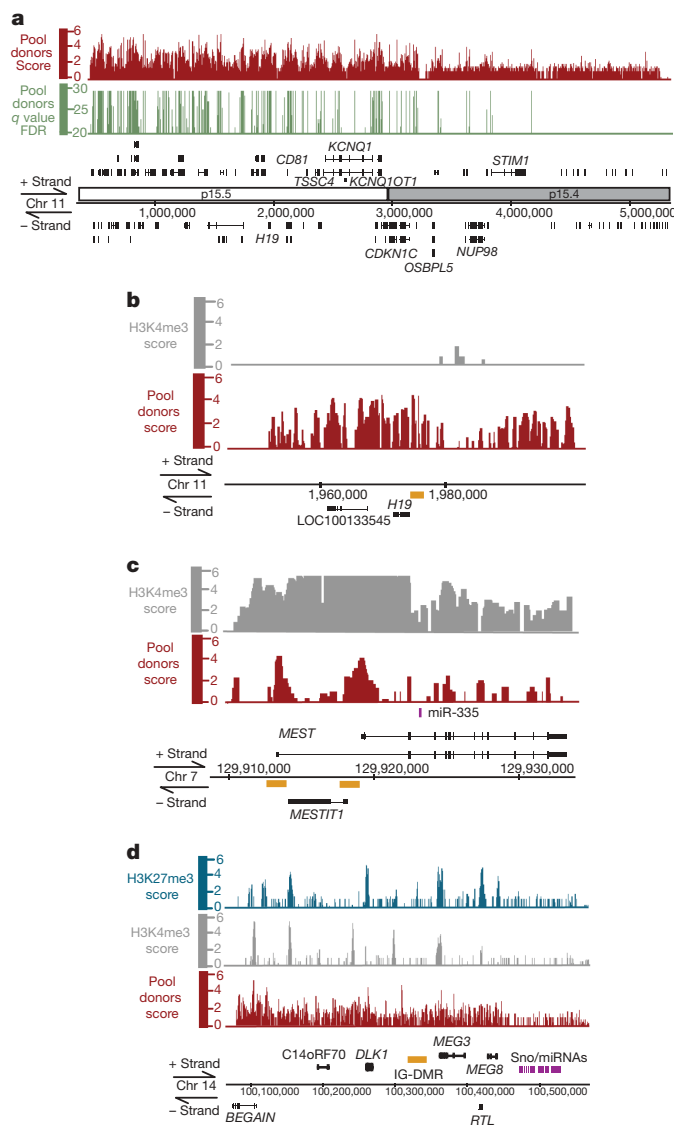
Figure 2 | Nucleosome enrichment at imprinted gene clusters, with high H3K4me3 at paternally expressed noncoding RNAs, and paternally demethylated regions. a, Histone enrichment at the 11p15.5 imprinted cluster (ending near *OSBPL5*), but not in the adjacent region. **b, c**, An expanded view of the DMRs (yellow rectangles) of *H19* (paternally methylated) (**b**) and *MEST* (paternally demethylated) (**c**). **d**, Moderate H3K4me3 at the promoters of the paternally expressed genes *BEGAIN*, *DLK1* and *RTL*, and the lack of H3K4me3 at the methylated intergenic-differentially methylated region (IG-DMR) of *MEG3* in sperm. Notably, both H3K4me3 and H3K27me3 reside at the promoter of *MEG3*, which later acquires DNA methylation in the embryo. Sno, small nucleolar.

hypomethylated (Supplementary Fig. 10d). Furthermore, 7 of the 12 miRNAs on autosomes that are occupied by OCT4, NANOG and SOX2 in human ES cells[17] are also significantly occupied by histone (from pooled sequencing data). However, we do not at present understand the logic for their modification status; certain miRNA clusters have high histone and bivalent status, whereas others lack either modification (Supplementary Fig. 6).

Attributes of primary and secondary imprinted genes

Nucleosomes are significantly enriched at most imprinted genes in sperm, but at both paternally and maternally expressed loci. However, we observe marked specificity of H3K4me3 localization, with high and broad levels present at genes and noncoding RNAs that are paternally expressed. Locus 11p15.5 (Fig. 2a) is a large imprinted cluster with *IGF2*, *H19* and *KCNQ1* and several miRNAs. Here, increased levels of histone are present throughout the imprinted region (up to *OSBPL5*), but not in the large adjacent region lacking imprinted genes (Fig. 2a). Notably, the paternally silenced *H19* locus upstream of *KCNQ1* has a methylated DMR (Supplementary Fig. 10a) that lacks H3K4me3 (Fig. 2b). In contrast, *MEST* (a paternally expressed gene) has high H3K4me3 that extends from its promoter and first exon (containing the demethylated differentially methylated region (DMR); Fig. 2c and Supplementary Fig. 10a) through the second exon. The antisense noncoding RNA *MESTIT1* (also paternally expressed) is transcribed from the first intron, and is also very high in H3K4me3 (Fig. 2c). Furthermore, the promoter region of the paternally expressed antisense noncoding RNA *KCNQ1OT1* displays H3K4me3 (Fig. 2a and data not shown), and the DMR is DNA demethylated (Supplementary Fig. 10a). Several other examples of paternally expressed loci with blocks of H3K4me3 are provided in Supplementary Fig. 11, including *PEG3*, the noncoding RNAs *AIRN* (antisense to *IGF2R*) and *GNASAS* (antisense to *GNAS*). In contrast, genes flanking *KCNQ1* that are repressed by the noncoding RNA *KCNQOT1* (such as *OSBPL5*, *TSSC4* and *CD81*; Fig. 2a, expanded in Supplementary Fig. 11) contain histone, but lack H3K4me3. Notably, several paternally silenced genes (bearing DNA methylation) bore moderate (2–3-fold) enrichment of H3K9me3, a mark absent at paternally expressed genes (Supplementary Fig. 7d).

The 14q32.33 region (*DLK-DIO3*) is complex and interesting; paternally expressed genes such as *DLK1* and *RTL1* have moderate levels of H3K4me3 in their promoters, and the imprinting control locus (IG-DMR) lacks H3K4me3 (Fig. 2d) and is DNA methylated[26–28]. Notably, the promoter of *MEG3* (also known as *GTL2*; just downstream of the IG-DMR) lacks DNA methylation in sperm, but acquires DNA methylation in the embryo[26–28], termed secondary imprinting. Notably, the *MEG3* promoter region that later acquires DNA methylation initially bears both H3K4me3 and H3K27me3 in sperm; it is bivalent. One interpretation is that for mature sperm and early embryos, H3K4me3 prevents DNA methylation while H3K27me3 promotes silencing, with subsequent H3K4me removal enabling tissue-specific DNA methylation and secondary imprinting. Furthermore, our examination of the X chromosome inactivation centre showed an apparent bivalent status (and DNA hypomethylation) at the TSS of the *XIST* noncoding RNA, but not at *TSIX*,
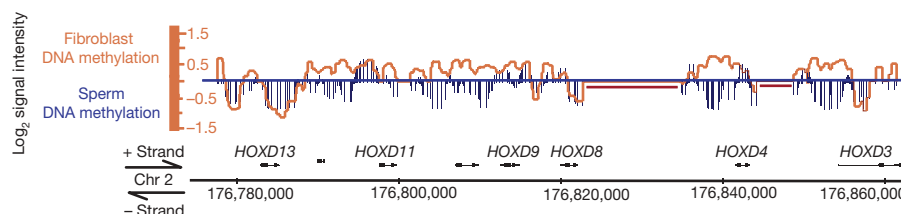
b, as well as the stand-alone miRNAs *mir-153-1*, *mir-488* and *mir-760*. Notably, many histone-occupied miRNAs are associated with embryonic development[25] ($P < 0.01$), and their promoters were largely

Figure 3 | Developmental promoters in sperm lack DNA methylation, but acquire methylation during development. DNA methylation of the *HOXD* locus in the mature sperm (blue bars) or primary fibroblasts (orange line overlay). The *y* axis is the signal intensity ($\log_2$) and the *x* axis is the annotated physical map (HG17). The regions not tiled on the array are underlined in red.

476

although future studies are required to determine whether these marks influence the regulation of this locus in the embryo (Supplementary Figs 6 and 10d; note that sequence reads on the X chromosome are half that on autosomes, as it is only present in 50% of sperm).

Modifications and expression timing

Transcriptome analysis has been performed in 4-cell and 8-cell human embryos, with 29 or 65 messenger RNAs identified as enriched, respectively[29]. Notably, genes in sperm bearing H3K4me3 but not H3K27me3 correlated with genes expressed at the 4-cell stage (14 out of 24, $P = 0.059$). Also, genes bearing high H3K4me2 were significantly enriched at genes expressed in the 4–8-cell stage (23 out of 49, $P < 0.02$; only 49 tiled on our array). In contrast, no significant correlation was observed with H3K27me3, which instead associates with transcription factors required for differentiation and organogenesis (discussed earlier). Furthermore, we verified by qPCR the presence of H3K4me2 or H3K4me3 at a subset of these stage-specific gene promoters (Supplementary Fig. 12). Thus, these findings reveal correlations of H3K4me2/3 enrichment, but not H3K27 enrichment, with early expression.

Conclusion

We provide several lines of evidence that the parental genome is packaged and covalently modified in a manner consistent with influencing embryo development. Previous analyses of DNA methylation in sperm identified hypomethylated promoters[23,24,30,31], showed similarities to the pattern in ES cells[24,31], and overlap between PRC2 and CpG islands[15,17,21,22]. We add that hypomethylated developmental promoters in human sperm overlap significantly with developmental promoters (in ES cells) occupied by the self-renewal network. Also, the promoters that acquire methylation in fibroblasts are primarily developmental transcription factors that are bound by PRC2 in human ES cells, consistent with recent work linking PRC2 to DNA methylation in development and neuronal differentiation in mice[21,32,33]. Thus, components of the self-renewal network emerge as candidates for helping to direct DNA hypomethylation in the germ line, and also to guide DNA hypermethylation to particular loci during differentiation, possibly to help 'lock in' differentiation decisions, although this remains to be tested.

The central findings of our work involve the significant enrichment of modified nucleosomes in the sperm genome at genes for embryo development, and a specificity to their modification patterns that might be instructive for the regulation of developmental genes, noncoding RNAs and imprinted loci. For example, histone retention and modification were clear at *HOX* loci and most of the targets of the self-renewal network in ES cells. One key concept in ES cell chromatin is the prevalence of developmental promoters with a bivalent status—bearing both H3K27me3 and H3K4me3 (ref. 10). Many promoters bivalent in ES cells are also bivalent in sperm, although some bear only H3K27me3 in sperm. Notably, H3K27me3 covers essentially all of the four *HOX* loci in sperm, whereas H3K4me3 is present in large blocks at only a subset of locations in *HOX* loci. Our work also provides correlations between H3K4me, but not H3K27me, and early expression in the embryo. In contrast, protamine-enriched loci did not show any significant Gene Ontology categories. However, there were certain segments of the Y chromosome with protamine enrichment, including the testis-specific *TSPY* genes, although the significance is not known.

We also find histones enriched at imprinted gene clusters, and a notable correlation between H3K4me3 and paternally expressed noncoding RNAs and genes; loci that lack DNA methylation in sperm. In contrast, maternally expressed noncoding RNAs/genes, and especially paternally methylated regions, lack H3K4me3 and (for the selected genes tested) contain moderate H3K9me3. Consistent with these observations, recent structural and *in vitro* data show that H3K4 methylation deters DNA methylation by DNMT3A2 and DNMT3L in mice[34]. However, experiments in model organisms are needed to address whether the modification patterns we report influence

imprinting patterns *in vivo*. Taken together, we reveal chromatin features in sperm that may contribute to totipotency, developmental decisions and imprinting patterns, and open new questions about whether ageing and lifestyle affects chromatin in a manner that impacts fertility or embryo development.

METHODS SUMMARY

Biological samples. Sperm samples were obtained from four men of known fertility attending the University of Utah Andrology laboratory, consented for research. Samples were collected after 2–5 days abstinence and subjected to a density gradient (to purify viable, motile, mature sperm) and treated with somatic cell lysis buffer (0.1% SDS, 0.5% Triton X-100 in DEPC H_2O) for 20 min on ice to eliminate white blood cell contamination. Samples were centrifuged at 10,000g for 3 min, and the sperm pellet was resuspended in PBS and used immediately for chromatin preparation. Clontech human fibroblast cells (Lonza cc-2251) were cultured (37 °C and 5% CO_2) in DMEM containing 10% FBS and supplemented with penicillin and streptomyocin.

Chromatin immunoprecipitation. Standard ChIP methods were used[35], but we omitted crosslinking and used the following salt concentrations in the numbered buffers[35]: (1) 150 mM NaCl, (2) 250 mM NaCl, (3) 200 mM LiCl, and (4) 150 mM NaCl (the PBS wash). Antibodies used were: anti-H3K27me3 (Upstate 07-449), H3K4me3 (Abcam 8580), H3K4me2 (Abcam 32356), TH2B (Upstate 07-680), H2A.Z (Abcam 4174) and H3K9me3 (Abcam 8898). For each, 4 μl of antibody was coupled to 100 μl of Dynabeads (Invitrogen). After ChIP, samples for sequencing were not amplified, whereas for arrays the DNA was amplified (WGA, Sigma) before hybridization.

Methylation profiling using MeDIP. MeDIP procedures for sperm and primary human fibroblasts (Clontech) were performed as described previously[30].

Sequencing. Sequencing used the Illumina GAII (Illumina Inc.) with standard protocols. Read numbers are final mapped microsatellite filtered reads (26–36 bases). Nucleosomes from D1: 19,658,110, D2–D4: 18,842,467, D1–4: 25,933,196 with equal contribution from each donor (random sub-sampling). Input, human sperm DNA: 17,991,622, H3K4me3: 13,337,105, H3K27me3:10,344,413, and H2A.Z: 5,449,000. All genomics data sets have been deposited in the Gene Expression Omnibus (GEO) under the SuperSeries GSE15594.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Ward, W. S. & Coffey, D. S. DNA packaging and organization in mammalian spermatozoa: comparison with somatic cells. *Biol. Reprod.* **44**, 569–574 (1991).
2. Wykes, S. M. & Krawetz, S. A. The structural organization of sperm chromatin. *J. Biol. Chem.* **278**, 29471–29477 (2003).
3. Balhorn, R., Brewer, L. & Corzett, M. DNA condensation by protamine and arginine-rich peptides: analysis of toroid stability using single DNA molecules. *Mol. Reprod. Dev.* **56**, 230–234 (2000).
4. Gatewood, J. M., Cook, G. R., Balhorn, R., Schmid, C. W. & Bradbury, E. M. Isolation of four core histones from human sperm chromatin representing a minor subset of somatic histones. *J. Biol. Chem.* **265**, 20662–20666 (1990).
5. Kimmins, S. & Sassone-Corsi, P. Chromatin remodelling and epigenetic features of germ cells. *Nature* **434**, 583–589 (2005).
6. Nix, D. A., Courdy, S. J. & Boucher, K. M. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics* **9**, 523 (2008).
7. Reik, W., Santos, F. & Dean, W. Mammalian epigenomics: reprogramming the genome for development and therapy. *Theriogenology* **59**, 21–32 (2003).
8. Santos, F., Hendrich, B., Reik, W. & Dean, W. Dynamic reprogramming of DNA methylation in the early mouse embryo. *Dev. Biol.* **241**, 172–182 (2002).
9. Rangasamy, D., Berven, L., Ridgway, P. & Tremethick, D. J. Pericentric heterochromatin becomes enriched with H2A.Z during early mammalian development. *EMBO J.* **22**, 1599–1607 (2003).
10. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
11. Bernstein, B. E. *et al.* Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc. Natl Acad. Sci. USA* **99**, 8695–8700 (2002).
12. Boyer, L. A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–956 (2005).
13. Muller, J. & Kassis, J. A. Polycomb response elements and targeting of Polycomb group proteins in *Drosophila*. *Curr. Opin. Genet. Dev.* **16**, 476–484 (2006).
14. Schwartz, Y. B. *et al.* Genome-wide analysis of Polycomb targets in *Drosophila melanogaster*. *Nature Genet.* **38**, 700–705 (2006).
15. Tanay, A., O'Donnell, A. H., Damelin, M. & Bestor, T. H. Hyperconserved CpG domains underlie Polycomb-binding sites. *Proc. Natl Acad. Sci. USA* **104**, 5521–5526 (2007).

16. Cao, R. *et al.* Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* **298**, 1039–1043 (2002).

17. Lee, T. I. *et al.* Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**, 301–313 (2006).

18. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).

19. Kopp, J. L., Ormsbee, B. D., Desler, M. & Rizzino, A. Small increases in the level of Sox2 trigger the differentiation of mouse embryonic stem cells. *Stem Cells* **26**, 903–911 (2008).

20. Wernig, M. *et al. In vitro* reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* **448**, 318–324 (2007).

21. Mohn, F. *et al.* Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Mol. Cell* **30**, 755–766 (2008).

22. Illingworth, R. *et al.* A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol.* **6**, e22 (2008).

23. Down, T. A. *et al.* A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature Biotechnol.* **26**, 779–785 (2008).

24. Farthing, C. R. *et al.* Global mapping of DNA methylation in mouse promoters reveals epigenetic reprogramming of pluripotency genes. *PLoS Genet.* **4**, e1000116 (2008).

25. Landgraf, P. *et al.* A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129**, 1401–1414 (2007).

26. Glazov, E. A., McWilliam, S., Barris, W. C. & Dalrymple, B. P. Origin, evolution, and biological role of miRNA cluster in DLK-DIO3 genomic region in placental mammals. *Mol. Biol. Evol.* **25**, 939–948 (2008).

27. Takada, S. *et al. Delta-like* and *Gtl2* are reciprocally expressed, differentially methylated linked imprinted genes on mouse chromosome 12. *Curr. Biol.* **10**, 1135–1138 (2000).

28. da Rocha, S. T., Edwards, C. A., Ito, M., Ogata, T. & Ferguson-Smith, A. C. Genomic imprinting at the mammalian *Dlk1-Dio3* domain. *Trends Genet.* **24**, 306–316 (2008).

29. Li, S. S., Liu, Y. H., Tseng, C. N. & Singh, S. Analysis of gene expression in single human oocytes and preimplantation embryos. *Biochem. Biophys. Res. Commun.* **340**, 48–53 (2006).

30. Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genet.* **39**, 457–466 (2007).

31. Fouse, S. D. *et al.* Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation. *Cell Stem Cell* **2**, 160–169 (2008).

32. Vire, E. *et al.* The Polycomb group protein EZH2 directly controls DNA methylation. *Nature* **439**, 871–874 (2006).

33. Schlesinger, Y. *et al.* Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nature Genet.* **39**, 232–236 (2007).

34. Ooi, S. K. *et al.* DNMT3L connects unmethylated lysine 4 of histone H3 to *de novo* methylation of DNA. *Nature* **448**, 714–717 (2007).

35. Gordon, M. *et al.* Genome-wide dynamics of SAPHIRE, an essential complex for gene activation and chromatin boundaries. *Mol. Cell. Biol.* **27**, 4058–4069 (2007).

METHODS

Partitioning of histone- and protamine-associated DNA. Chromatin was prepared from 40 million sperm as described previously[36] in the absence of crosslinking reagent, treated with sequential and increasing MNase (10–160 U), and centrifuged to sediment protamine-associated DNA, releasing mononucleosomes. The pooled mononucleosomes were used for ChIP, or the DNA was extracted and gel purified (~140–155 bp) for sequencing and array analysis.

ChIP and preparation for genomics methods. All ChIPs for sequencing were performed using the same pool of mononucleosomes from pooled donors. For arrays, a single pool was used from D1. ChIP methods were as described previously[35] but were performed without a crosslinking agent and slight modifications to the salt levels (250 mM NaCl, 200 mM LiCl), and the TE wash was replaced with a 150 mM PBS wash. ChIP methods used anti-H3K27me3 (Upstate 07-449), H3K4me3 (Abcam 8580), H3K4me2 (Abcam 32356), TH2B (Upstate 07-680), or H2A.Z (Abcam 4174) antibodies. For each, 4 μl of antibody was coupled to 100 μl of Dynabeads (Invitrogen). After the ChIP procedure, the DNA was amplified (WGA, Sigma) before hybridization to arrays, whereas samples used for Solexa were not amplified. For sequencing, DNA lengths corresponding to mononucleosomes with adapters (220–280 bp) were gel purified after the addition of the Illumina adaptors. This size selection was also performed for the nucleosomal DNA from pooled donors not subjected to ChIP.

Methylation profiling using MeDIP. This procedure was described previously[30]. In brief, sonicated sperm DNA was obtained from two different donors and sonicated fibroblast DNA was obtained from Clontech primary human fibroblasts (Lonza CC-2251) (4 μg, 300–1,000-bp fragments). Immunoprecipated DNA was washed, subjected to whole genome amplification (Sigma Aldrich). Amplified DNA (6 μg) was labelled with Cy5, and input DNA (6 μg) was labelled with Cy3 (Bio labs) by standard methods. Samples were hybridized to Agilent expanded promoter arrays, treated according to standard Agilent conditions, and scanned in an Agilent scanner.

Computational analytical methods. The software used in this analysis are open source and available from the TIMAT2 (http://timat2.sourceforge.net) and USeq (http://useq.sourceforge.net) project websites. Human annotation and genomic sequence (May 2004, NCBI Build 35, HG17 and March 2006, NCBI Build 36.1, HG18) were obtained from the UCSC Genome Bioinformatic website.

Low-level ChIP-chip analysis. Processing of the Agilent microarray promoter data was performed in three basic steps: data normalization, sliding window summaries, and enriched region identification. For each data set, the median unadjusted signal intensities from the Cy3 and Cy5 channels were extracted. Probes were then mapped to the HG17 or HG18 builds. Biological replicas were quantile normalized and median scaled to 100 (ref. 37). This normalization was applied to the treatment (ChIP samples) and control (whole genomic input DNA for the MeDIP and protamine data sets or DNA derived from mononucleosomes) replicas separately (see later for replica-averaged R^2). Probe level 'Oligo' summaries were calculated by taking the $\log_2$ ratio (mean treatment replicas/mean control replicas). 'Window' level summaries were generated by identifying windows of a particular size (100 bp for data sets derived from mononucleosomes, 675 bp for MeDIP and protamine data sets) containing a minimum number of oligonucleotide start positions (one for the data sets derived from mononucleosomes, three for the MeDIP and protamine data sets), and calculating an all pair (treatment versus control) relative difference pseudo median. This window summary score was assigned to the centre position of the window 'Pse' or represented as heat map 'PseHM' data. Extended regions of high-scoring windows, called 'intervals', were identified by merging windows that exceed a set threshold and are located within 250 bp of one another. Intervals were then ranked by their best window score. Relative difference pseudo median scores were converted to $\log_2$ ratio values.

The average R^2 values for microarray data were as follows: 0.85 for the three D1 MNase replicas; 0.89 for the three Protamine replicas; 0.96 for the two H3C replicas; 0.94 for the two H3K4me2 replicas; 0.93 for the two TH2B replicas; 0.96 for the three H3K4me3 replicas; and 0.93 for the two H3K27me3 replicas. The average MeDIP R^2 values for the three replicas of each donor were as follows: D2 average $R^2 = 0.97$ and D4 = 0.89, and the correlation between D2 versus D4 was 0.87. The average R^2 for the two primary human fibroblast MeDIP replicas was 0.86.

Low-level Chip-seq analysis. The DNA samples derived from mononucleosomes, and the sonicated control input genomic DNA were prepared for sequencing using Illumina's ChIP-seq kit. The 26-bp and 36-bp reads were generated using Illumina's Genome Analyser II and their standard software pipeline. Reads were mapped to the March 2006 NCBI Build 36.1 human genome using the pipeline's eland_extended aligner.

The USeq package[6] was used to identify regions of histone enrichment relative to input control. This entailed selecting reads that mapped with an alignment score ≥13 ($-10\log_{10}(0.05)$), shifting their centre position 73 bp 3' to accommodate the 146-bp mononucleosome fragment length, and using a sliding window of 300 bp to score each region in the genome for significant histone enrichment. Significance was determined by calculating a binomial P value for each 300-bp window and controlled for multiple testing by applying Storey's q value FDR estimation[38,39].

Read numbers. Note the sperm genome has only 4% of the genome in nucleosomes. For nucleosome enrichment D1 had 19,658,110 reads, and the pool of three additional donors had 18,842,467 reads. The raw correlation for D1 versus the donor pool was $r = 0.7$. For all the analysis containing pool donors (D1, and a pooled sample of three additional individuals D2, D3 and D4) we used 25,933,196 mapped filtered reads with equal contribution from each donor (random subsampling). A total of 17,991,622 reads were generated from control input human sperm DNA, 3,337,105 reads from the H3K4me3 sample, 10,344,413 reads for H3K27me3, and 5,449,000 reads for H2Az. The raw unfiltered reads (fastq format) are deposited at GEO under the superseries GSE15594, which encompasses the Subseries entries GSE15690 for ChIP-seq and GSE15701 for ChIP-chip data.

To assess histone enrichment consistency, the QCSeqs application in the USeq package[6] was used to correlate the read counts between the D1 and pooled sample by calculating a Pearson correlation on the basis of the number of mapped reads falling within 500-bp windowed regions stepped every 250 bp across all chromosomes. Only windows with five or more reads in either of the samples were included in the correlation.

To create lists of candidate histone enriched regions, q-value thresholds of 20 (0.01) and 30 (0.001) ($-10\log_{10}(q\text{value})$) were selected. Overlapping windows that pass a given threshold were merged and scores from the best window assigned to the enriched region. The normalized window score was then used to rank and sort the regions.

A modification was made to score gene promoters and miRNAs for significant histone enrichment. The first step was to define regions for scoring. For gene promoters, the start of the first exon was used to define its hypothetical promoter by selecting a region 9 kb upstream and 2 kb downstream. For miRNAs, the centre position of each was expanded ±300 bp. These defined regions were scored for significant enrichment using the window statistics above.

High-level ChIP-chip and ChIP-seq analysis. Intersect regions. To identify regions of significant intersection between enriched region lists from various data sets, the USeq IntersectRegions application was used. This application counts the number of intersections between two lists of genomic coordinates that occur within a minimum 'max gap' distance. To estimate confidence in the intersections, a thousand 'random' data sets are generated that were matched to the chromosome and size of the original regions, and randomly picked from the interrogated regions on the array or sequenced regions in the genome. These randomized data sets were used to calculate a P value for the intersection and fold enrichment (fraction real intersection/fraction average random data set intersection) over random. Initial pilots that imposed a fraction GC match when picking random regions showed little difference with non-GC-matched random data sets and were thus subsequently dropped.

Find neighbouring genes (FNG). Genes associating with histones or histone modifications were determined using the FNG application in the USeq package. The gene lists were uploaded in GoMiner (http://discover.nci.nih.gov/gominer/htgm.jsp) to identify over represented Gene Ontology terms.

Intersect lists. To determine whether the 4- and 8-cell transcripts identified in early human embryo correlated with any of our histone modifications we used The IntersectLists USeq application which uses random permutation to calculate the significance of intersection between two lists of genes.

Aggregate plots. The USeq AggregatePlots application was used to compare the degree of enrichment and distribution of histone reads surrounding the TSS of developmental and non developmental genes. The gene classes were derived on the basis of Gene Ontology term categories.

36. Zalenskaya, I. A., Bradbury, E. M. & Zalensky, A. O. Chromatin structure of telomere domain in human sperm. *Biochem. Biophys. Res. Commun.* **279**, 213–218 (2000).

37. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).

38. Dudoit, S., Gilbert, H. N. & van der Laan, M. J. Resampling-based empirical Bayes multiple testing procedures for controlling generalized tail probability and expected value error rates: focus on the false discovery rate and simulation study. *Biom. J.* **50**, 716–744 (2008).

39. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).