Full Methods

Partitioning of Histone- and Protamine- Associated DNA: Chromatin was prepared from 40 million sperm as described previously[34] in the absence of crosslinking reagent, treated with sequential and increasing MNase (10U-160U), and centrifuged to sediment protamine-associated DNA, releasing mononucleosomes. The pooled mononucleosomes were used for chromatin immunopreciptation (below), or the DNA extracted and gel purified (~140-155 bp) for sequencing and array analysis.

Chromatin IP and Preparation for Genomics Methods: All Chromatin IPs (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 cross linking agent and slight modifications to the salt levels, 250 mM NaCl, 200 mM LiCl, and replaced the TE wash with 150 mM PBS wash. ChIP methods used anti-H3K27me3 (Upstate 07-449), H3K4me3 (Abcam 8580), H3K4me2 (Abcam 32356), or TH2B (Upstate 07-680), H2A.z (Abcam 4174) antibodies. For each, 4 µL of antibody was coupled to 100 µL of Dynabeads (Invitrogen). Following the ChIP procedure, the DNA was amplified (WGA, Sigma) prior to 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]. Briefly, sonicated sperm DNA was obtained from two different donors and sonicated fibroblast DNA was obtained from Clontech primary human fibroblasts (Lonza CC-2251) (4 ug, 300-1000 bp fragments). Immunoprecipated DNA was washed, subjected to whole genome amplification (Sigma, Aldrich). Amplified DNA (6 ug) was labeled with Cy5 and input DNA (6 ug) was labeled 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 softwares used in this analysis are open source and available from the TIMAT2 (http://timat2.sourceforge.net) and USeq (http://useq.sourceforge.net) project web sites. 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 web site.

Low Level ChIP-chip Analysis:
Processing of the Agilent microarray promoter data were performed in three basic steps: data normalization, sliding window summaries, and enriched region identification. For each dataset, 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[36]. This normalization was applied to the treatment (ChIP samples) and control (whole genomic input DNA for the MeDIP and protamine datasets or DNA derived from mononucleosomes) replicas separately (see below for replica averaged R^2). Probe level "Oligo" summaries were calculated by taking the log2 ratio (mean treatment replicas/ mean control

replicas). "Window" level summaries were generated by identifying windows of a particular size (100bp for datasets derived from mononucleosomes, 675 bp for MeDIP and protamine datasets) containing a minimum number of oligo start positions (1 for the datasets derived from mononucleosomes, 3 for the MeDIP and protamine datasets) and calculating an all pair (treatment vs. control) relative difference pseudo median. This window summary score was assigned to the center 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 bps of one another. Intervals were then ranked by their best window score. Relative difference pseudo median scores were converted to log2 ratio values.

The average R^2 for microarray data were as follows:
The average R^2 for the three D1 MNAse replicas were 0.85
The average R^2 for the three Protamine replicas were 0.89
The average R^2 for the two H3C replicas were 0.96
The average R^2 for the two H3K4me2 replicas were 0.94
The average R^2 for the two Th2B replicas were 0.93
The average R^2 for the three H3K4me3 replicas were 0.96
The average R^2 for the two H3K27me3 replicas were 0.93
The average MeDIP R^2 for the three replicas of each donor were the following: D2 average R^2 = 0.97 and D4= 0.89 and the correlation between D2 vs D4 was 0.87
The average R^2 for the two primary human fibroblast MeDIP replicas were 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. 26bp 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 center position 73bp 3' to accommodate the 146bp mononucleosome fragment length, and using a sliding window of 300bp to score each region in the genome for significant histone enrichment. Significance was determined by calculating a binomial p-value for each 300bp window and controlled for multiple testing by applying Storey's q-value FDR estimation[37, 38].

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 vs the donor pool was r = 0.7. For all the analysis containing pool donors (D1**,** and a pooled sample of 3 additional individuals D2, D3, D4) we used 25,933,196 mapped filtered reads with equal contribution from each donor (random subsampling). 17,991,622 reads were generated from control input human sperm DNA and 13,337,105 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 following 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 based on the number of mapped reads falling within 500bp windowed regions stepped every 250bp across all chromosomes. Only windows with 5 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-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 to the fore mentioned 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 9kb upstream and 2kb downstream. For miRNAs, the center position of each was expanded +/- 300bp. 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 datasets, 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" datasets 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 datasets were used to calculate a p-value for the intersection and fold enrichment (fraction real intersection / fraction average random dataset intersection) over random. Initial pilots that imposed a fraction GC match when picking random regions showed little difference with non GC matched random datasets and was thus subsequently dropped.

Find Neighboring 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 GO 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 transcription start sites (TSS) of developmental and non developmental genes. The gene classes were derived based on GO term categories.

Supplemental References

36. 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**, 185-93 (2003).
37. 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**, 716-44 (2008).
38. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-5 (2003).