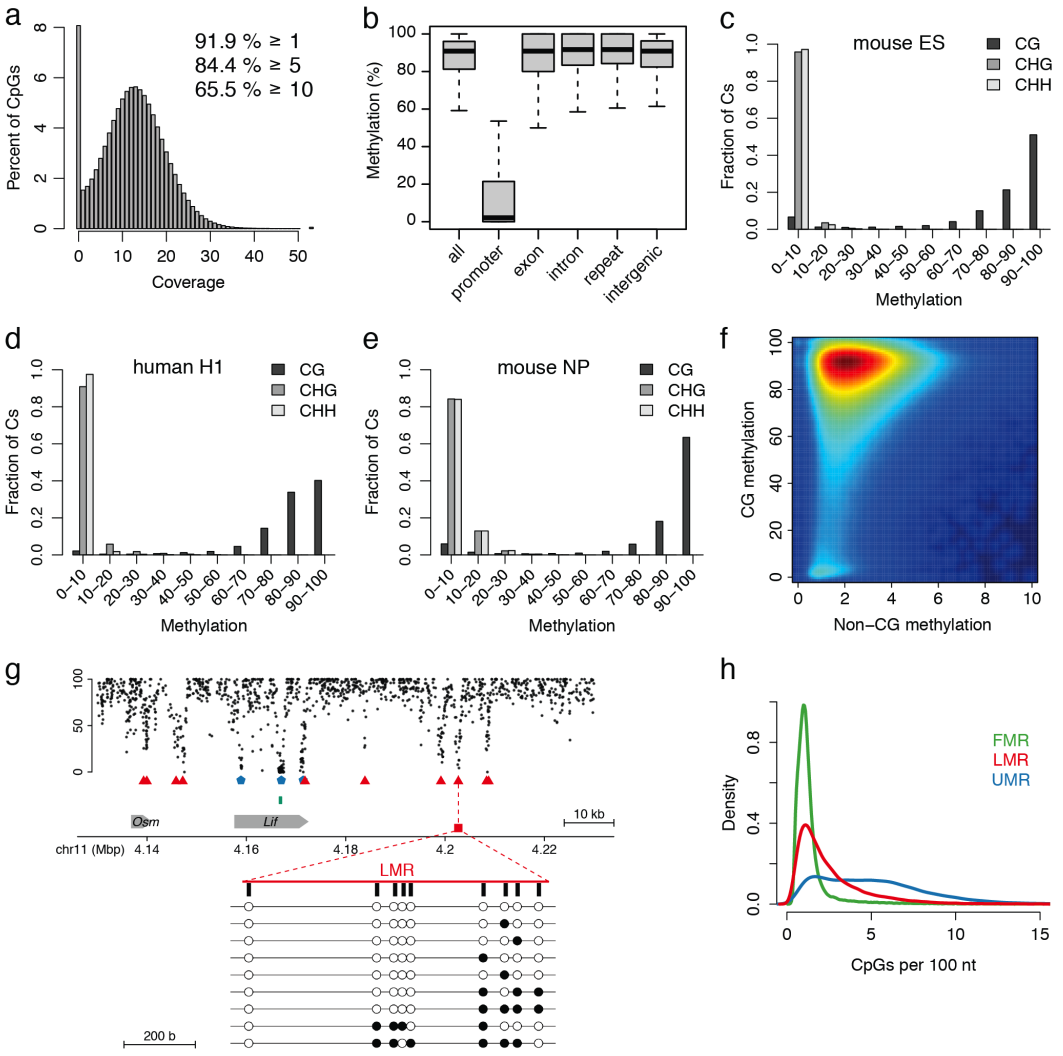
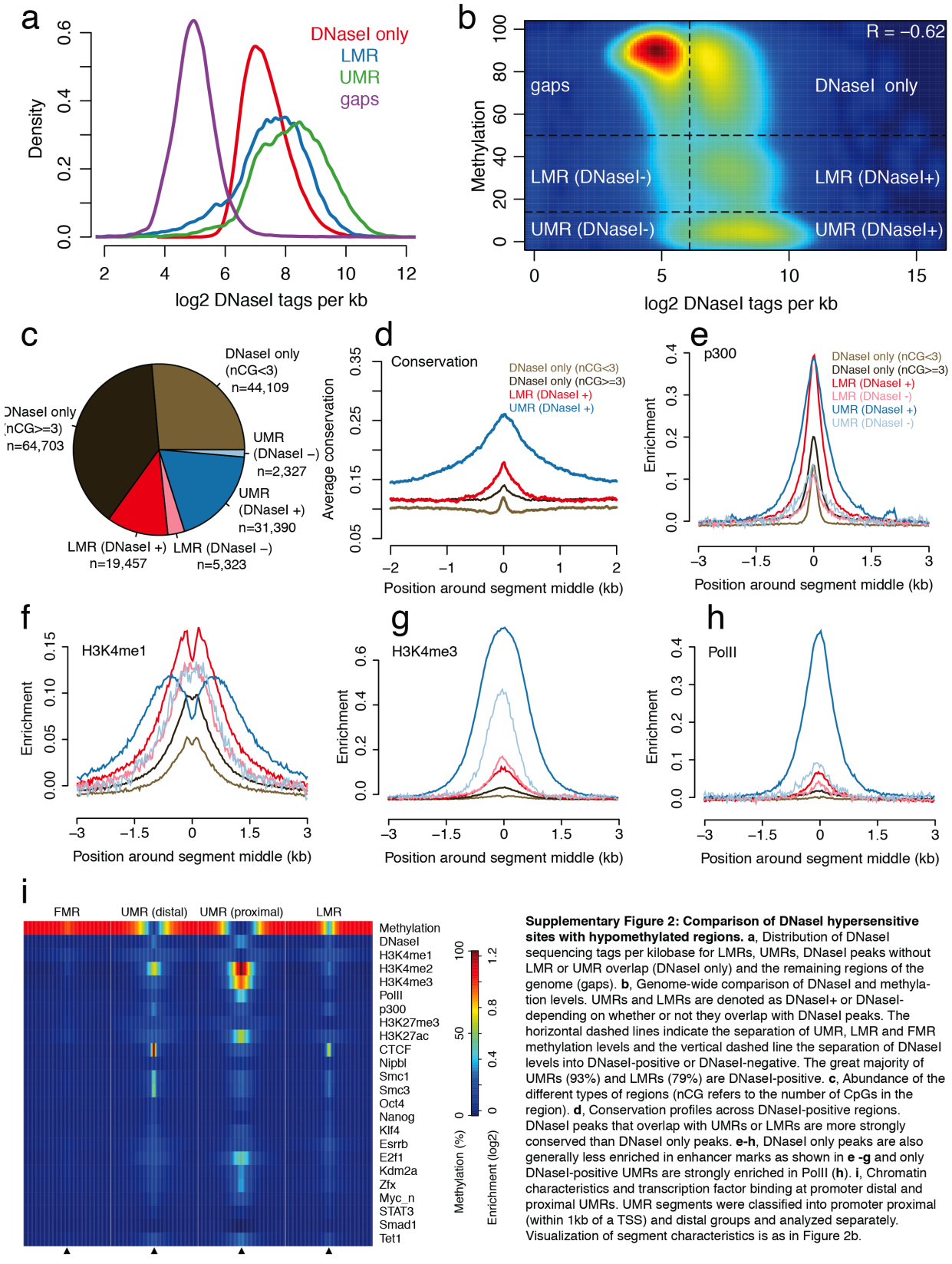
SUPPLEMENTARY INFORMATION

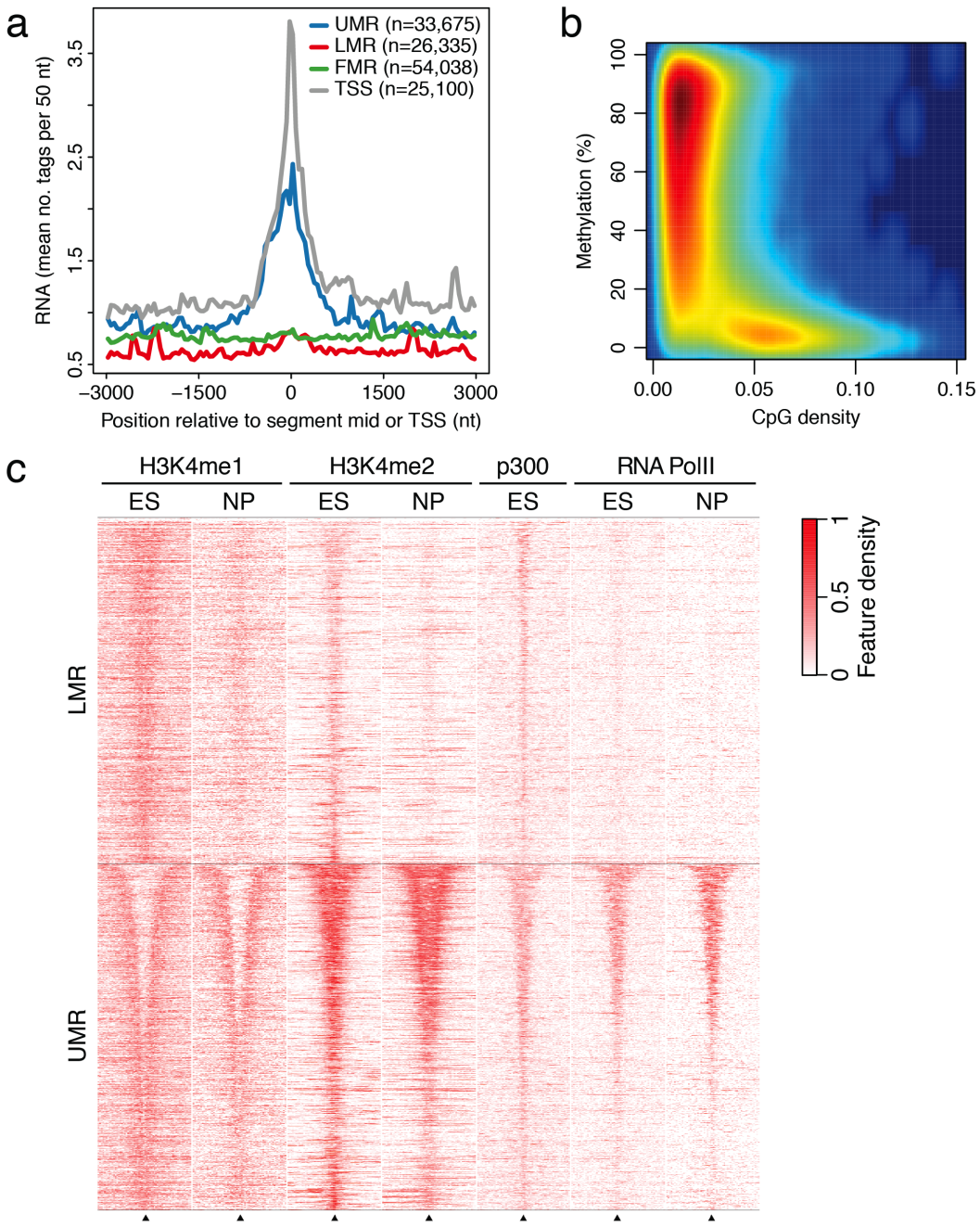
doi:10.1038/nature10716



Supplementary Figure 1: Features of the mouse ES methylome. a, Distribution of CpG coverage in the generated BisSeq dataset. **b**, Methylation levels of CpGs within specific genomic annotations illustrating that promoters are largely unmethylated. **c-f**, Comparison of CG and non-CG methylation in mouse ES, human H1 and mouse NP cells. **c**, Distribution of CG, CHG and CHH methylation in mouse ES. **d**, Same as in (**c**) for human H1 cells. **e**, Same as in (**c**) for mouse NP. **f**, Genome-wide comparison of CG versus non-CG methylation levels in consecutive windows of 1 kb. Non-CG methylation above 2% generally occurs only in regions with high CG methylation. **g**, Representative genomic region. Computational segmentation identifies: UMRs (blue pentagons), LMRs (red triangles) and FMRs (unmarked). Each dot represents one CpG.CpG islands are marked in green. Included is an example of an independently verified LMR downstream of the *LIF* gene. **h**, CpG density for each group: most UMRs are CpG rich, while LMRs and FMRs are CpG poor.

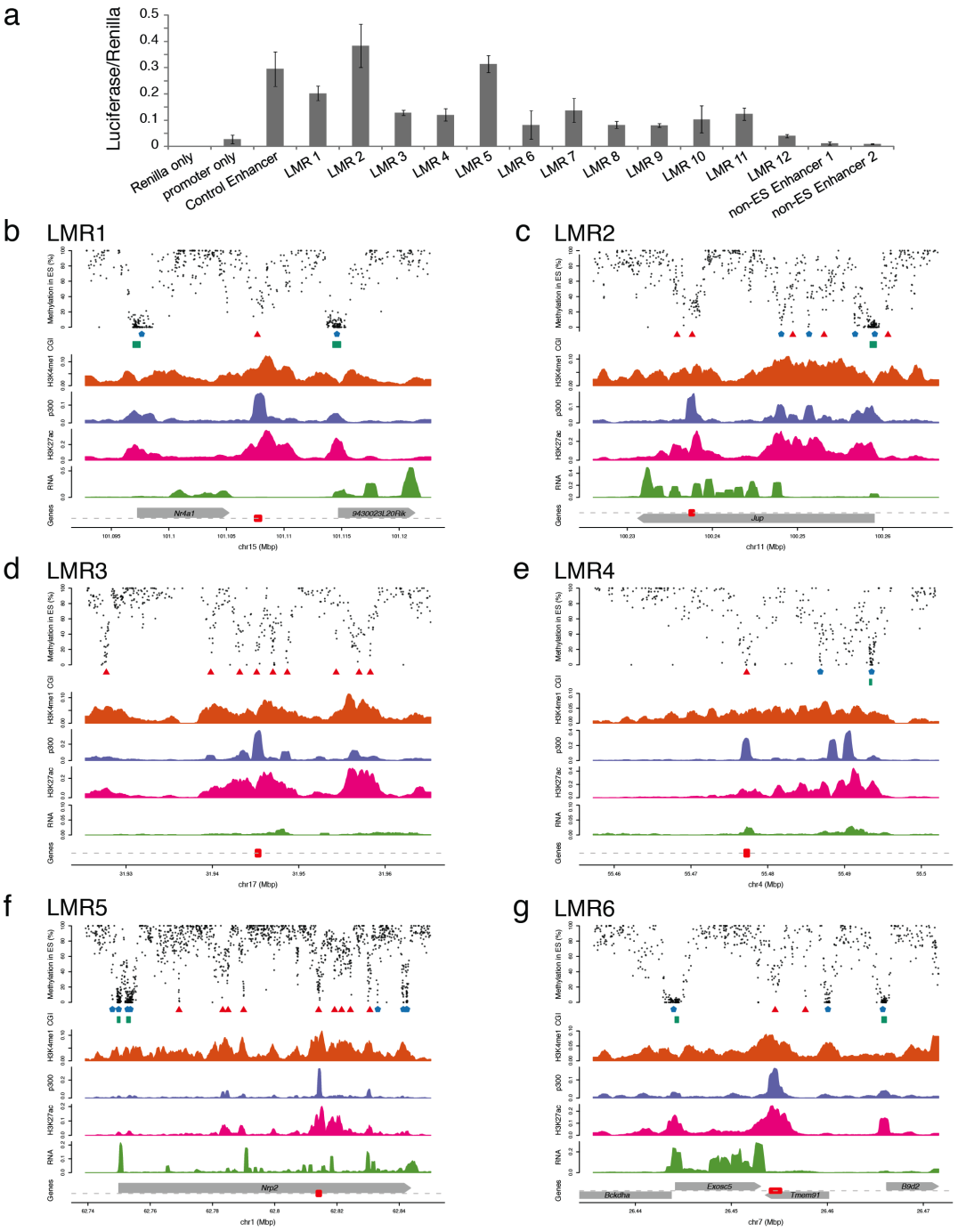


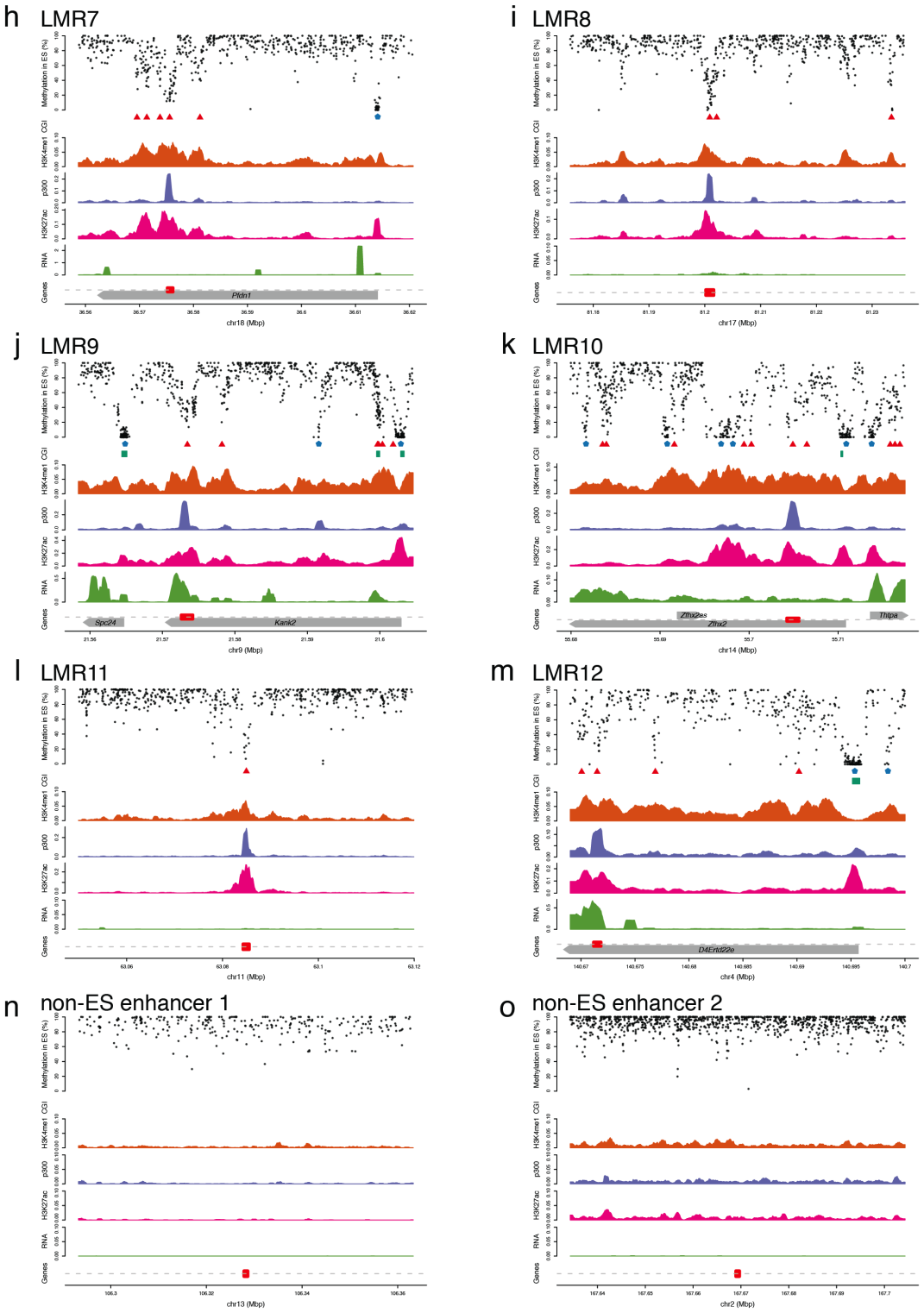
Supplementary Figure 2: Comparison of DNaseI hypersensitive sites with hypomethylated regions. **a**, Distribution of DNaseI sequencing tags per kilobase for LMRs, UMRs, DNaseI peaks without LMR or UMR overlap (DNaseI only) and the remaining regions of the genome (gaps). **b**, Genome-wide comparison of DNaseI and methylation levels. UMRs and LMRs are denoted as DNaseI+ or DNaseI- depending on whether or not they overlap with DNaseI peaks. The horizontal dashed lines indicate the separation of UMR, LMR and FMR methylation levels and the vertical dashed line the separation of DNaseI levels into DNaseI-positive or DNaseI-negative. The great majority of UMRs (93%) and LMRs (79%) are DNaseI-positive. **c**, Abundance of the different types of regions (nCG refers to the number of CpGs in the region). **d**, Conservation profiles across DNaseI-positive regions. DNaseI peaks that overlap with UMRs or LMRs are more strongly conserved than DNaseI only peaks. **e-h**, DNaseI only peaks are also generally less enriched in enhancer marks as shown in **e -g** and only DNaseI-positive UMRs are strongly enriched in PolII (**h**). **i**, Chromatin characteristics and transcription factor binding at promoter distal and proximal UMRs. UMR segments were classified into promoter proximal (within 1kb of a TSS) and distal groups and analyzed separately. Visualization of segment characteristics is as in Figure 2b.



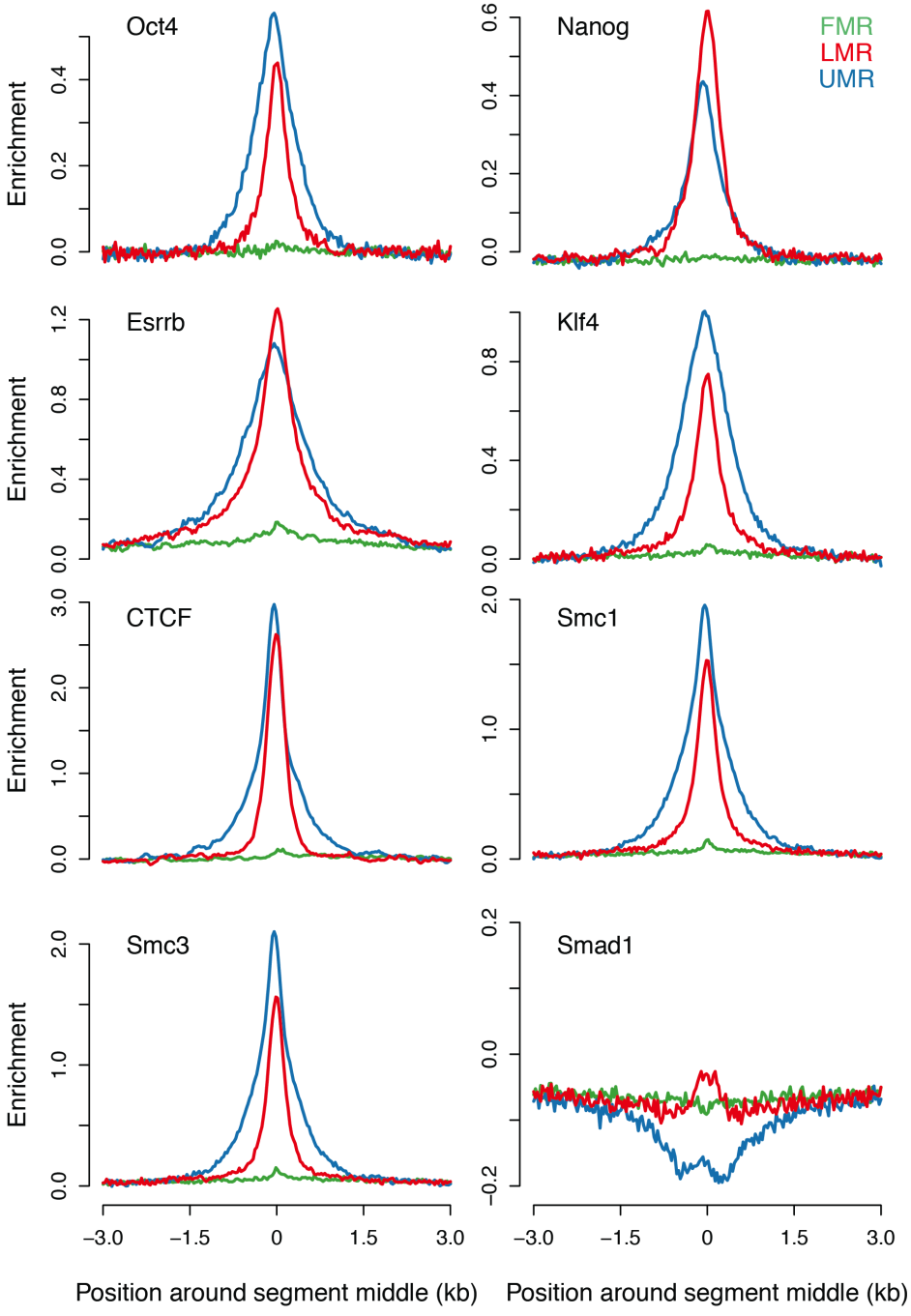
Supplementary Figure 3: LMRs are not promoters. a, RNA profiles at methylation segments and transcript start sites. Total RNA from ES was depleted for ribosomal RNA (see material and methods) and sequenced. The average number of RNA reads aligned to the genome per 50 nucleotides is shown for methylation segments (UMRs, LMRs and FMRs) as well as for known transcript start sites (TSS). **b**, CpG density versus methylation at DHS. The figure shows two populations of DHS, one with high CpG density and very low methylation levels, which correspond to UMRs including CpG islands, and one with low CpG density and methylation levels between 10% and 100%. A subset of these with methylation levels below 50% corresponds to LMRs. Given that the large majority of unmethylated regions are DNaseI positive (Supplementary Figure 4b), this strongly argues that LMRs and UMRs form two distinct classes. Only DHS with at least 3 CpGs and 10 reads overlapping CpGs were used for the analysis. **c**, Features of individual LMR and UMR loci. Enrichments for H3K4me1, H3K4me2, p300 and RNA PolII in ES and NP around a set of 1000 randomly selected LMRs and UMRs sorted by segment length. Scaled densities of ChIP-seq tags are shown (from 0 = white to 1 = red for each antibody) in regions of 6kb centered on segment midpoints (indicated by triangles).

Supplementary Figure 4

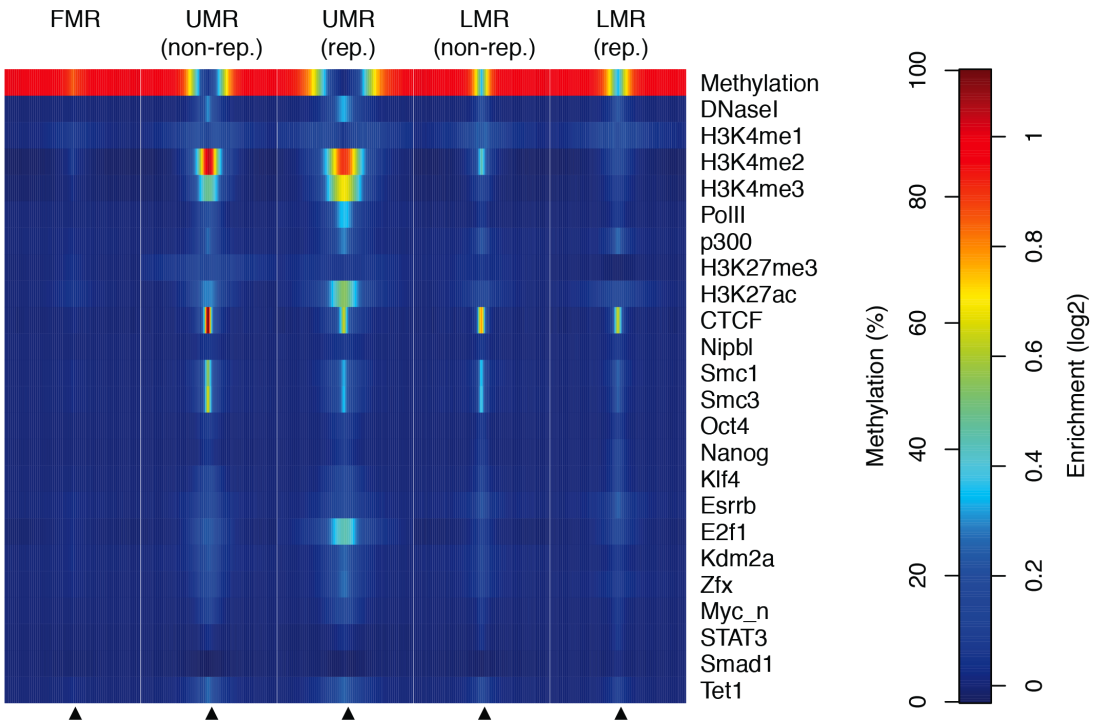




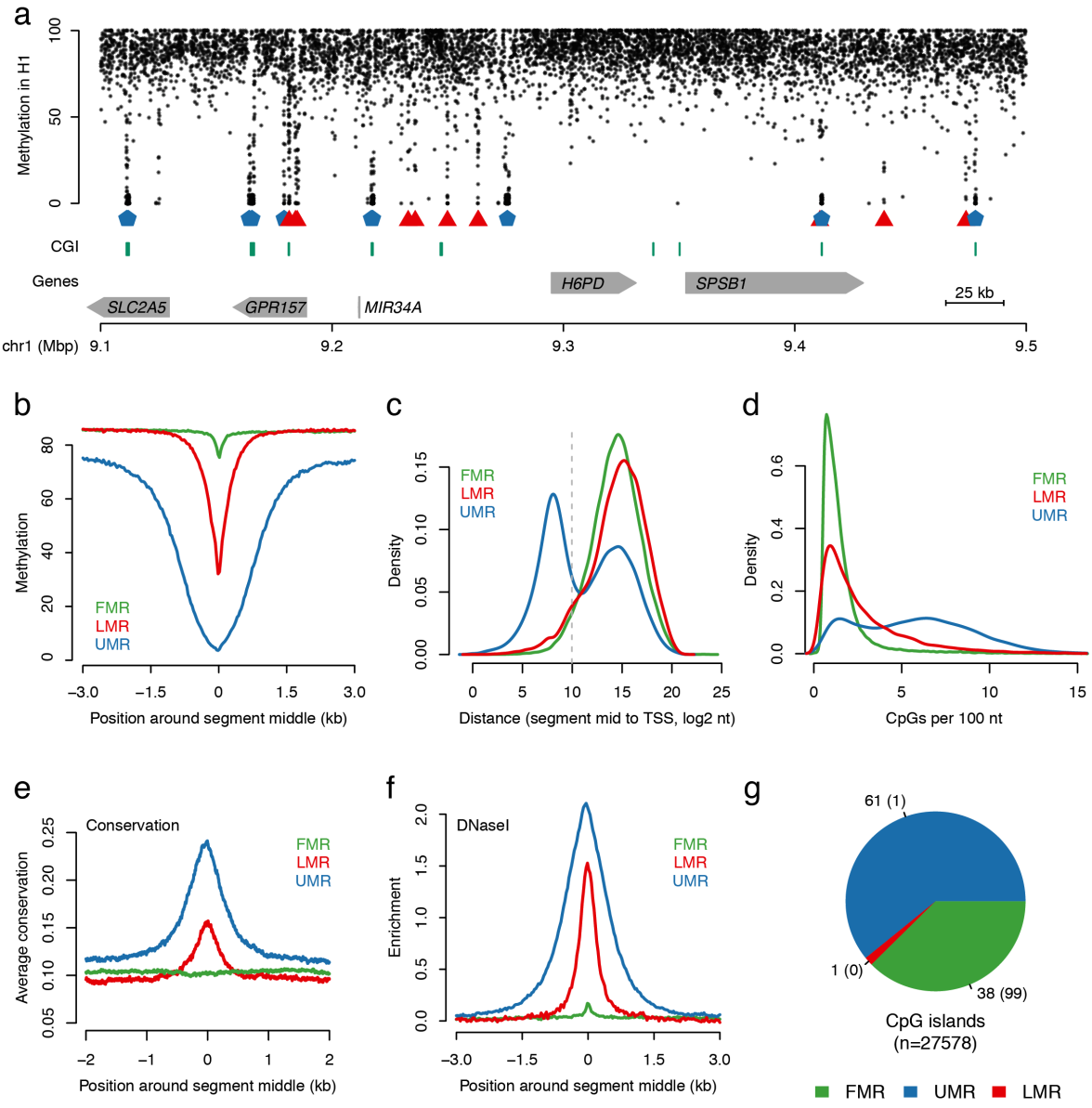
Supplementary Figure 4: LMRs show enhancer activity in transgenic reporter assays. a, 12 different LMRs plus a positive control and two non-stem cells but embryo-specific enhancers (non-ES enhancers), were assayed for enhancer activity after transfection in ES. The values are the mean relative luminescence of three independent transfections, error bars reflect standard deviation). **b-o**, Example chromosomal profiles of LMRs tested in enhancer assays. Methylation profiles for the 12 LMRs (**b-m**) and the two non-ES enhancer regions (**n, o**) shown in (**a**). In addition, tag densities for ChIP (H3K4me1, p300 and H3K27ac) and RNA are shown as a running mean over 1kb windows. A red box at the bottom of each panel indicates the location of the tested fragments.



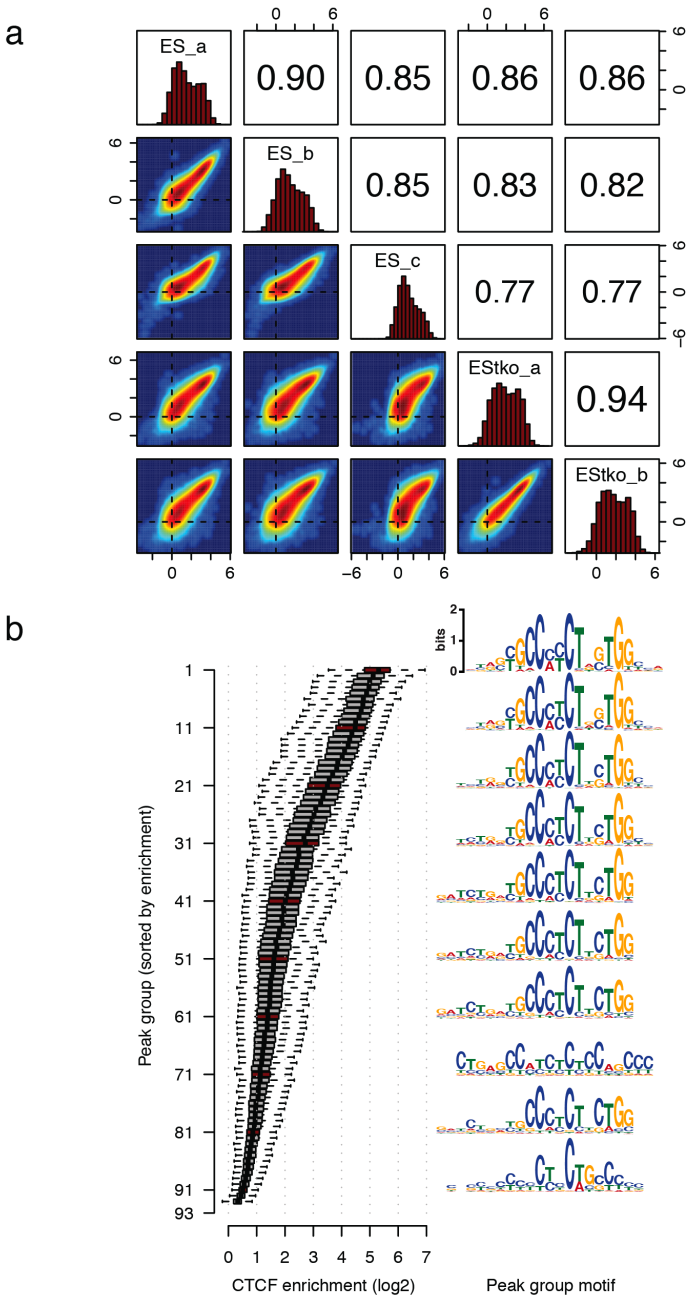
Supplementary Figure 5: Composite profiles for transcription factor enrichment at LMRs, FMRs and UMRs in ES. Both UMRs and LMRs show enrichment for stem cell-specific factors: Oct4, Nanog, Esrrb and Klf4, as well as for CTCF and members of the cohesin complex: Smc1 and Smc3. Smad1 is not enriched in LMRs or UMRs.



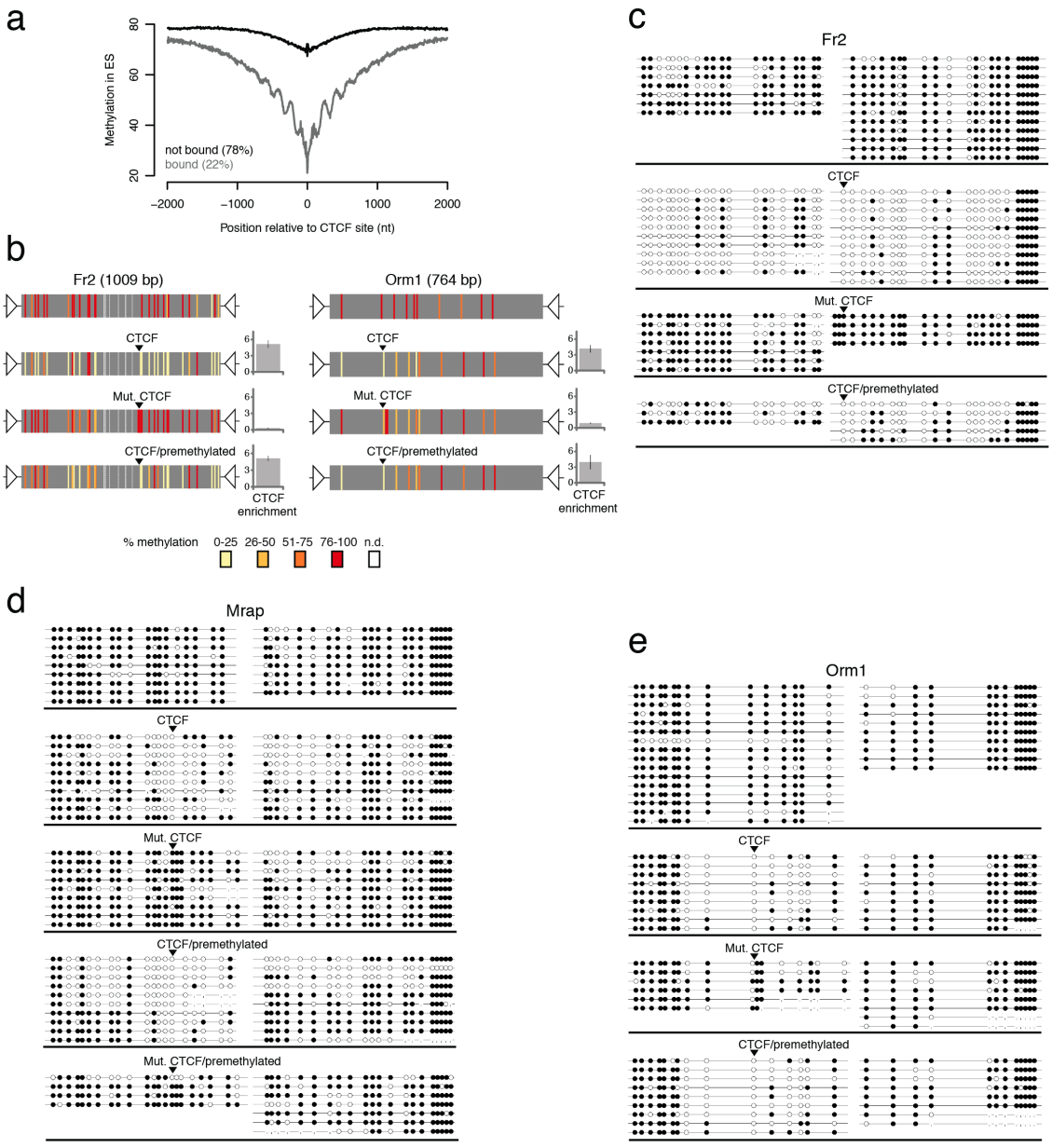
Supplementary Figure 6: Chromatin characteristics and transcription factor binding are similar at segments with and without repeat overlaps. LMR and UMR segments were classified into groups with (at least 1 bp) and without overlaps with repetitive sequences (using RepeatMasker annotation from UCSC) and analyzed separately. Visualization of segment characteristics is as in Figure 2b.



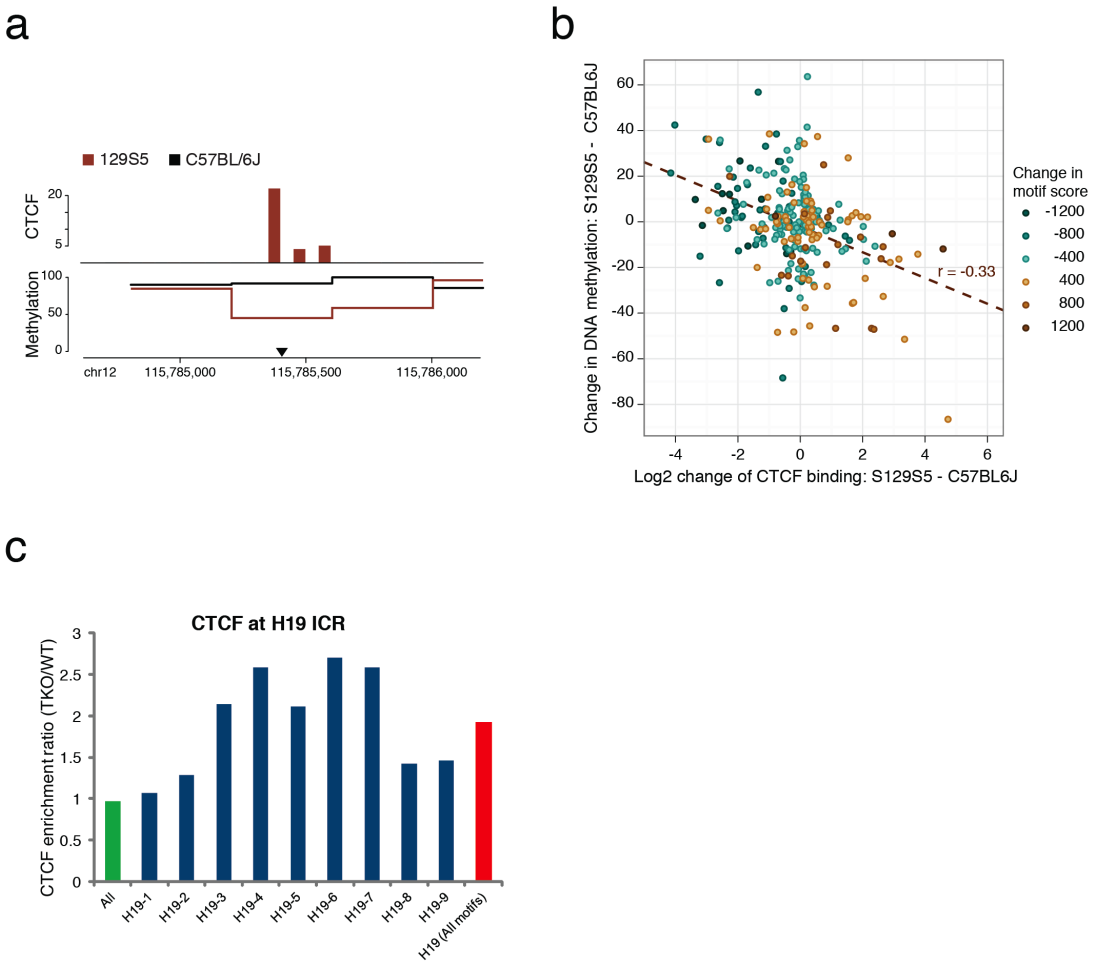
Supplementary Figure 7: LMRs are conserved in human ES cells. a, Chromosomal profile of a selected genomic region. Each dot represents location and methylation level of one CpG. Similarly to mouse ES cells, unbiased segmentation identified UMRs, LMRs, and FMRs. CpG islands (CGI) are shown in green. **b**, Composite profile of CpG methylation in UMRs, LMRs and FMRs aligned at their midpoints, illustrating similar methylation levels and sizes as in mouse ES cells. **c**, As in mouse ES cells, most human UMRs overlap start sites, while FMRs and LMRs are distal to promoters. **d**, Distribution of the CpG density for each group indicating that human LMRs are mostly CpG poor. **e**, Evolutionary conservation of UMRs, LMRs and FMRs aligned at segment midpoints. Both human LMRs and UMRs are conserved. **f**, Enrichment of DNaseI tags within the three groups illustrating that LMRs and UMRs show comparable hypersensitivity to DNaseI cleavage. **g**, Pie chart illustrating the small overlap of LMRs with CpG islands.



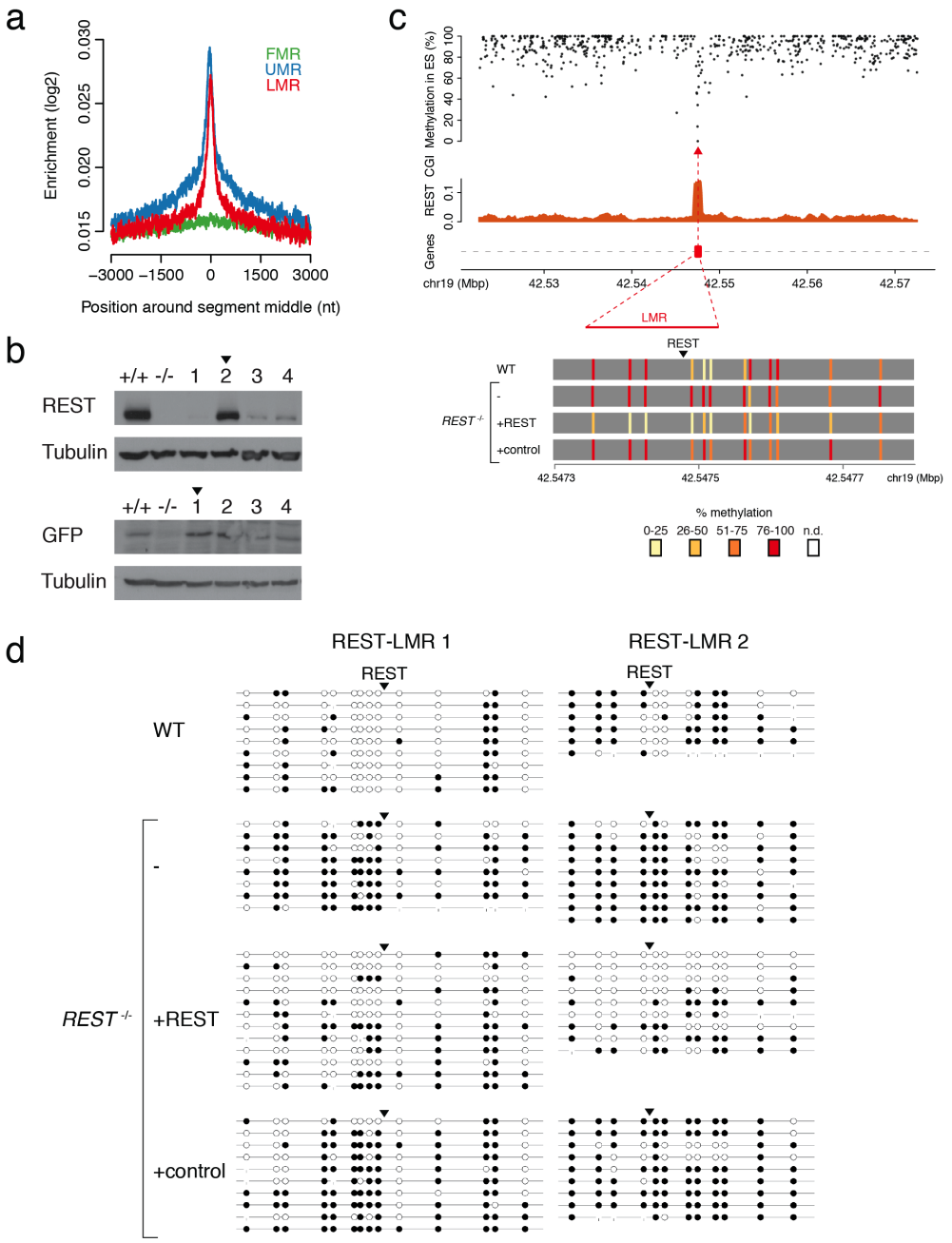
Supplementary Figure 8: CTCF ChIP-Seq in mouse wild-type ES and EStko cells. a, Peak enrichments for each pair of individual replicate samples were compared: Pearson's correlation coefficients are shown in the panels above the diagonal, and two-dimensional density plots are shown below the diagonal, with colors from blue (low) to red (high) indicating the local density of peaks. The overall similarity of CTCF enrichments between ES and EStko cells indicates that CTCF binding is globally unaltered in EStko cells. Histograms of enrichments in each sample are shown in the diagonal panels. **b**, CTCF peaks were sorted according to their average enrichment and classified into 93 groups of 1000 peaks each. For each group, the enrichments in individual replicates are shown as a box plot in the left panel (whiskers corresponding to minimum and maximum, the box to 25th and 75th percentiles and the black line to the median). Every 10th group of peaks (shown in dark red color) was selected to identify a sequence motif (see material and methods). The resulting logos (right panel) indicate that CTCF binding sites can be found in the large majority of peaks and even in peaks with average or low enrichments.



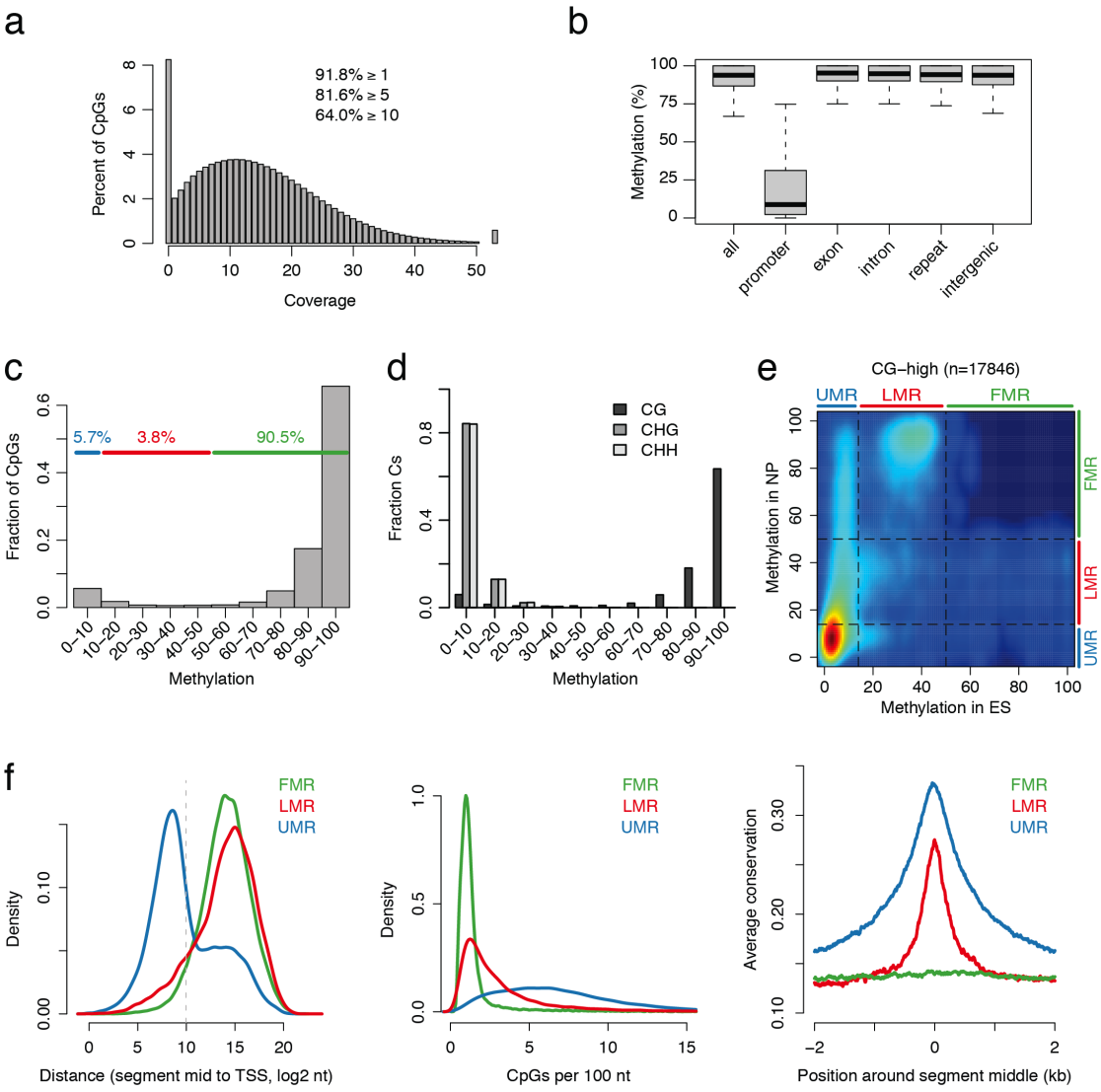
Supplementary Figure 9: Transgenic assay to test the effect of CTCF binding on DNA methylation. a, Composite profile of methylation levels around CTCF binding sites illustrating LMR-like methylation at the site of binding and increased methylation in its vicinity. Percentage of bound and unbound sites is given in legend. **b**, DNA fragments corresponding to the prokaryotic fragment FR2 or the Orm1 promoter are methylated when inserted via RMCE. As for the Mrap fragment in Figure 3b, introduction of a CTCF motif leads to CTCF binding and local demethylation in an otherwise methylated fragment, while insertion of a mutated motif does not affect methylation and does not lead to CTCF binding. Insertion of an in vitro methylated transgene with CTCF motif similarly leads to a demethylated state and CTCF binding. Bars show log2 CTCF enrichment as measured by ChIP. Error bars represent standard deviation of 3 independent experiments. **c-e**, Single bisulfite sequences for all 3 inserted fragments.



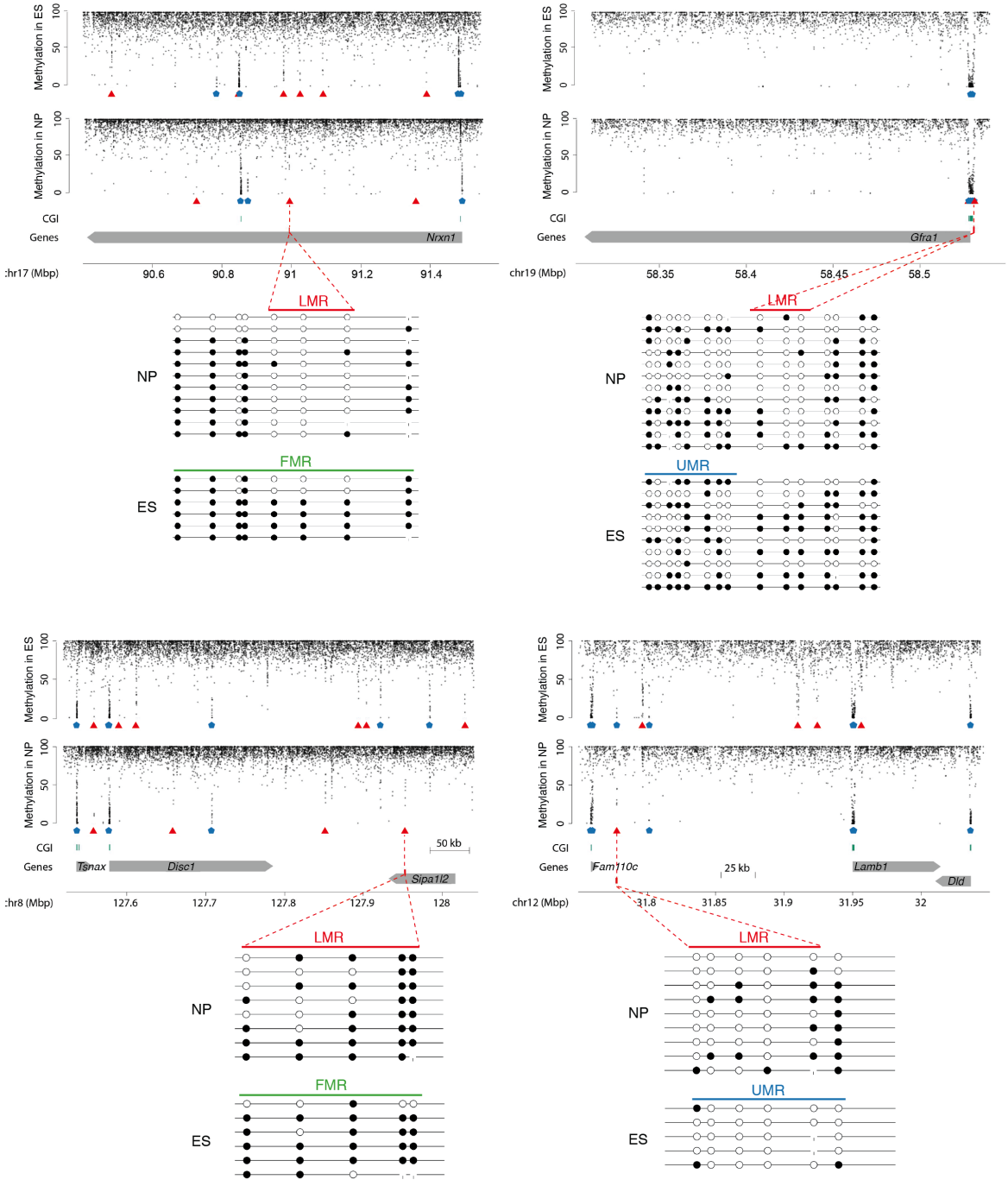
Supplementary Figure 10: Comparison of allele-specific CTCF binding and DNA methylation. a, An additional example of a CTCF binding site illustrating allele specific CTCF binding and DNA methylation for one heterozygous locus. **b**, For CTCF binding sites overlapping with heterozygous SNPs, ChIP-seq tags and DNA methylation were quantified separately for both alleles in 400bp windows centered on CTCF sites. The plot shows the differences between 129S5 and C57BL6 alleles, and the difference of motif score (indicated by the fill color). Increased CTCF binding correlates with reduced methylation (Spearman correlation coefficient r = -0.33) and increased motif scores. **c**, Comparison of CTCF binding at the H19 ICR. The y-axis represents the ratio of CTCF binding (linear scale) between EStko and WT ES cells. “All” represents all CTCF motifs in the genome. H19-1 to H19-9 represent CTCT binding at individual motifs contained in the H19 ICR. “H19 (all motifs)” represents average CTCF binding at all 9 motifs. Note the ~2 fold increase in CTCF binding in EStko cells.



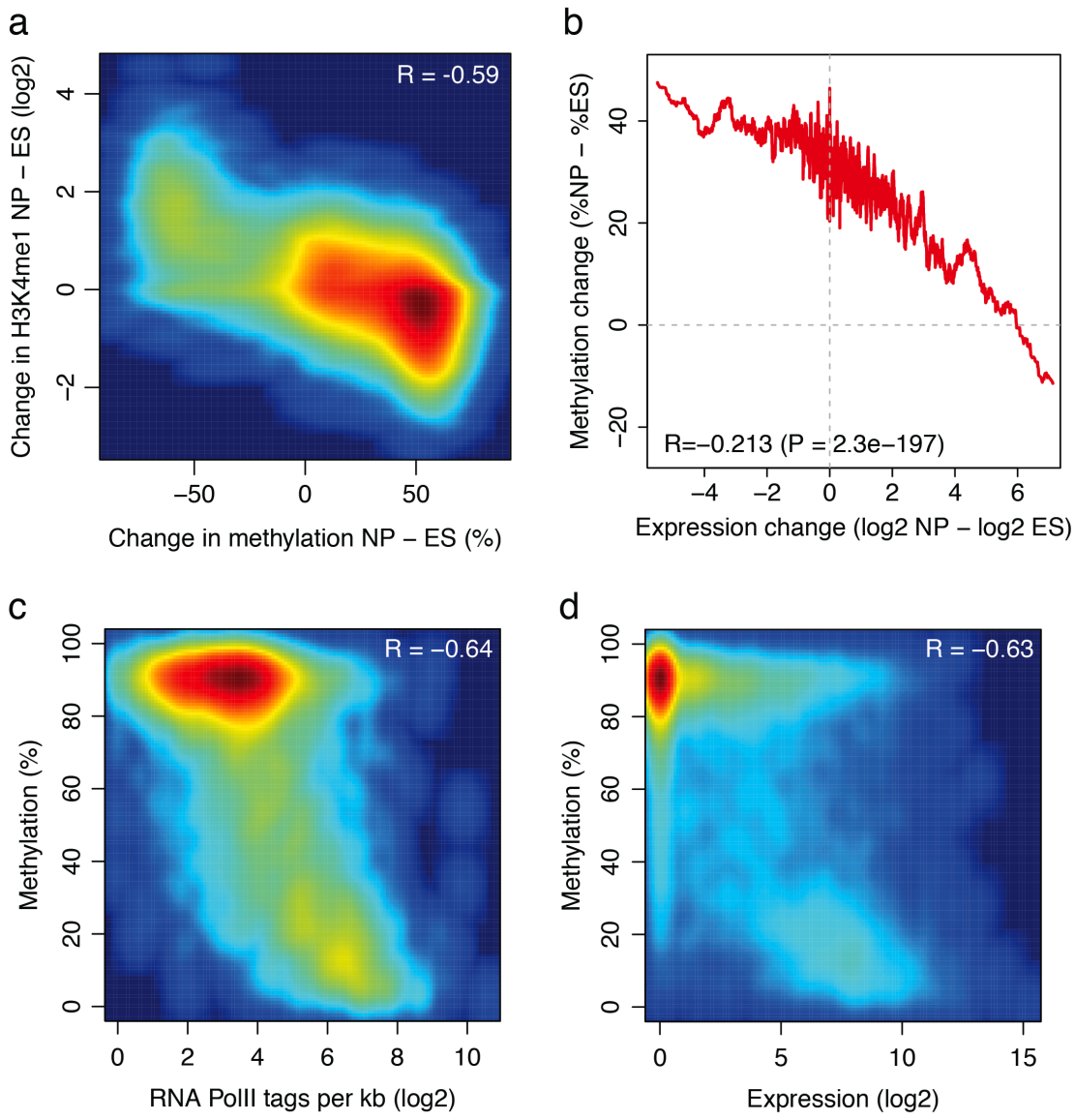
Supplementary Figure 11: REST binding leads to LMR formation. a, Composite profile showing that REST is enriched at LMRs. **b**, Western Blot showing protein levels for REST and GFP in 4 blasticidin resistant clones, as well as negative (-/-) and positive (+/+) control cells. Black triangles indicate the clones that were used for bisulfite sequencing. Tubulin serves as a loading control. **c**, Bisulfite profile of another genomic region containing a REST binding site and coinciding LMR formation. This LMR regains methylation in absence of REST (*REST*$^{-/-}$ ES cells), while reintroduction of REST reestablishes it. **d**, Single bisulfite sequences for the REST-occupied LMRs tested here and in Figure 4.



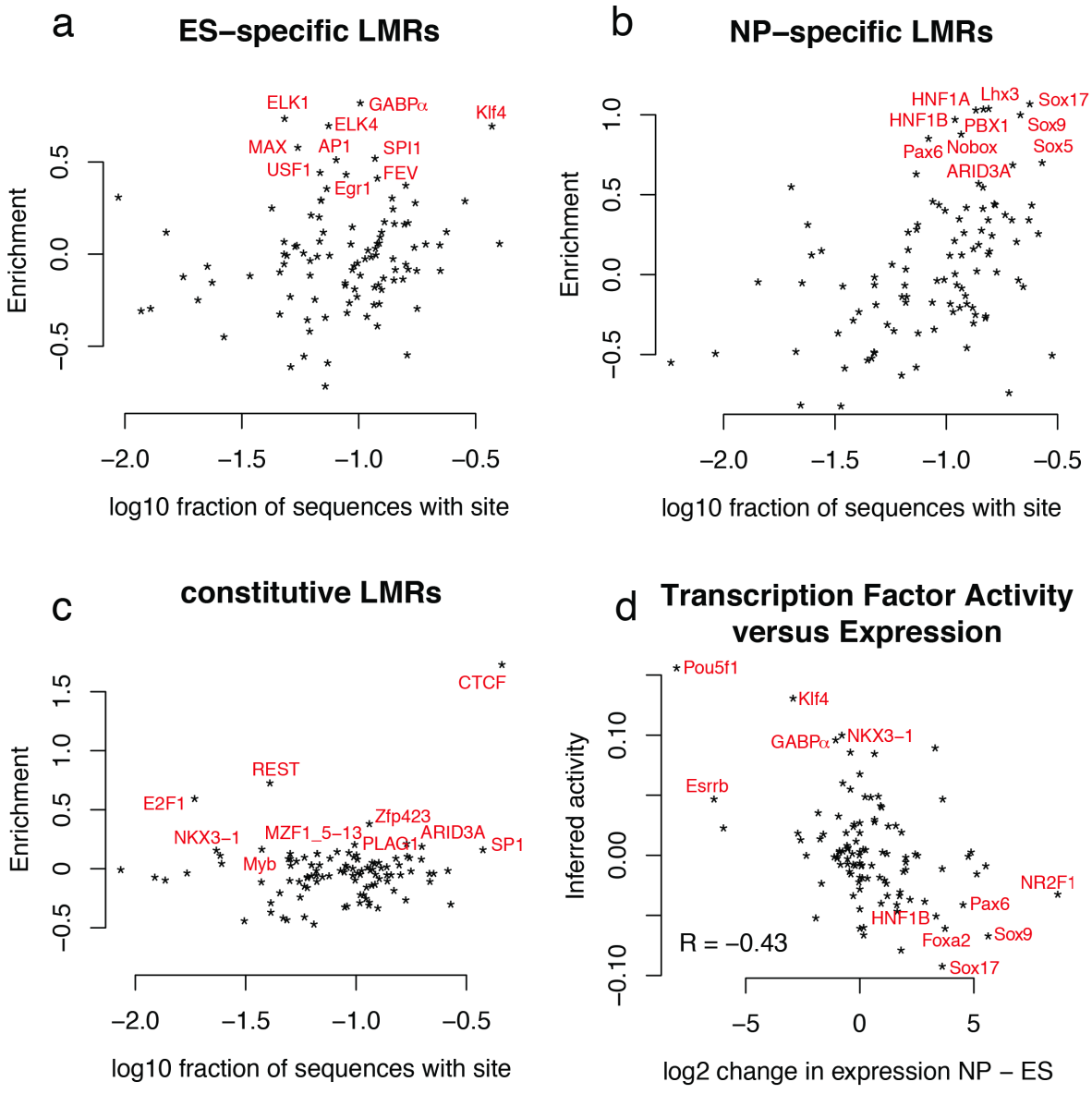
Supplementary Figure 12: Features of the mouse NP methylome. a, Distribution of coverage in the generated BisSeq dataset of NP illustrating that on average each cytosine was sequenced 13 times. CpGs with coverage higher than 50 are shown as a separate box on the right. **b**, Methylation status of CpGs within specific genomic annotations illustrating that promoters are largely devoid of methylation, while the rest of the genome appears methylated. **c**, Histogram showing the percentage methylation for individual CpGs (coverage of at least 10) revealing a binary distribution of methylation states. In addition 3.8% of all cytosines show methylation between 10 and 50%. **d**, Distribution of CG, CHG and CHH methylation in NP. **e**, 2D-density plot showing the methylation changes from ES to NP in all regions classified as CG-high. **f**, Distances to TSS (left panel), CpG density (middle panel) and sequence conservation (right panel).



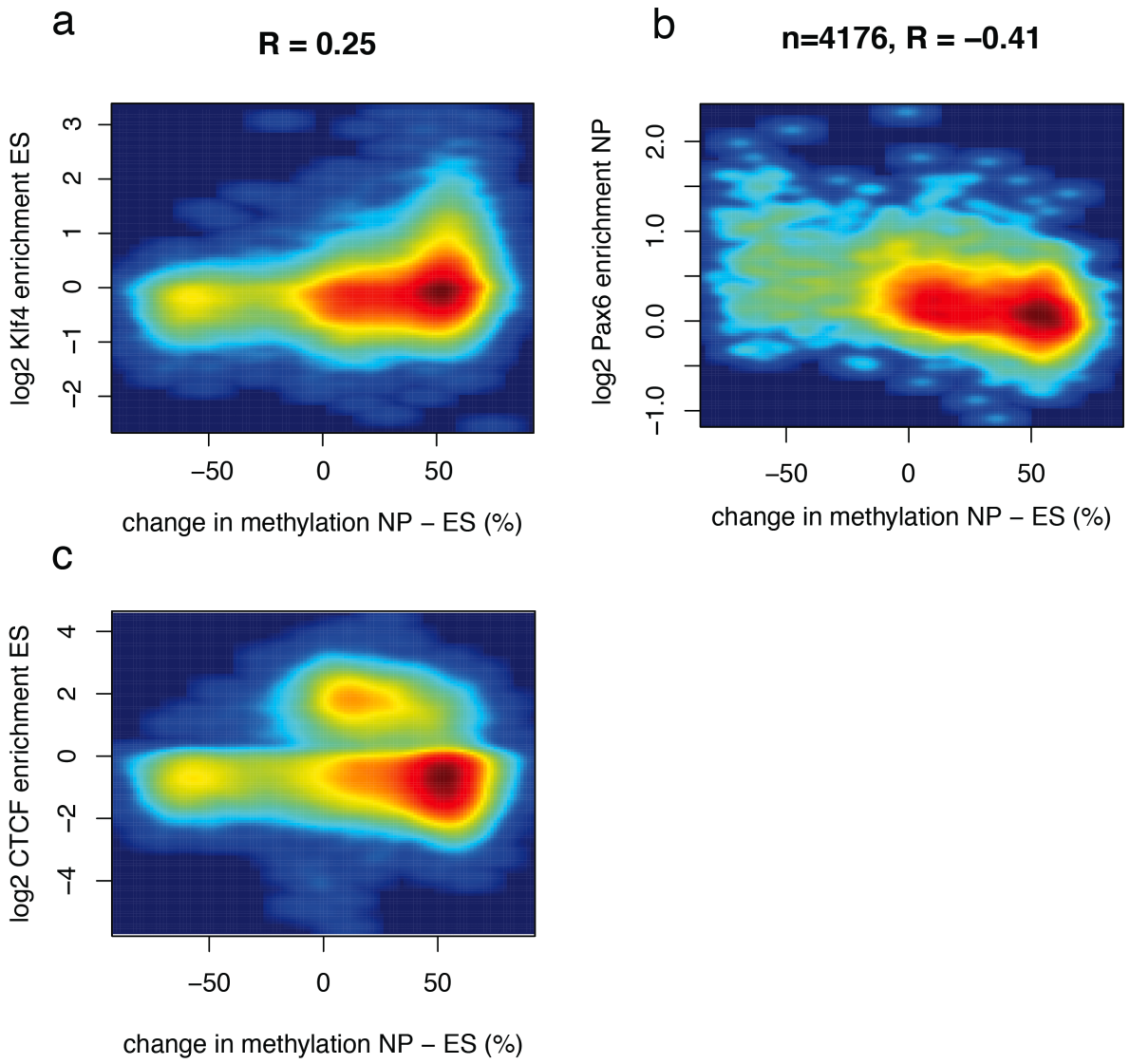
Supplementary Figure 13: Bisulfite sequencing showing NP-specific LMRs. All 4 LMRs are located near genes important for neuronal differentiation (*Disc1*, *Lamb1*, *Nrxn1*, and *Gfra1*). Display as in Supplementary Figure 1g. Mbp, million base pairs. kb, kilobases.



Supplementary Figure 14: Comparison of DNA methylation with H3K4me1, expression and RNA PolII binding. a, Changes in methylation correlate with changes in H3K4me1 (R=-0.59). While many ES-specific LMRs loose H3K4me1, most NP-specific LMRs gain H3K4me1 from ES to NP, providing further evidence that LMR formation follows established chromatin characteristics of active enhancers. **b**, Expression changes of genes correlate significantly (P=2.3e-197) with changes in methylation of neighboring LMRs. Each gene was associated with the LMR closest to its TSS in ES and/or in NP, and gene expression changes between ES and NP were correlated to the methylation change in the associated segment. The plot shows running means over 101 values (R= -0,213). **c-d** Gene expression and RNA PolII binding versus methylation levels in non-CpG island promoters. RNA Pol II bound promoters (**c**) and expressed genes (**d**) tend to have lower methylation levels, as evidenced by the strong negative correlation (R=-0.64 and R=-0.63, respectively). Methylation levels were determined in windows of 2kb centered on TSS and only promoters containing at least 3 CpGs and 10 reads overlapping CpGs were used for the analysis.



Supplementary Figure 15: Motif enrichments and inferred transcription factor activities. **a–c**, Enrichment of binding sites for any of the 126 transcription factors with known binding preference versus the fraction of regions with at least one predicted site in ES-specific LMRs (**a**), NP-specific LMRs (**b**) and constitutive LMRs (**c**). The 10 transcription factors with the highest site enrichment are indicated in red. **d**, Transcription factor expression changes from ES to NP versus their inferred activities. The transcription factors with largest activities and expression changes are indicated in red. Pearson's correlation coefficient is shown at the bottom left corner.



Supplementary Figure 16: Transcription factor binding versus methylation changes in LMRs. a, Klf4 binding in ES cells versus change in methylation from ES to NP. **b**, as in (**a**) but for Pax6 binding in NP. For Pax6, only those LMRs covered with at least three probes of the array were used for the analysis. **c**, Change in CTCF binding versus change in methylation. Only LMRs that are bound either in ES or NP are shown. Most regions do not change in binding nor methylation state.

Supplementary Material and Methods:

Experimental procedures:

Cell Culture and Differentiation. Wild-type embryonic stem cells 159-2 were derived form blastocysts (3.5 PC) of mixed 129-C57Bl/6 background and cultivated on feeder cells or 0.2% gelatin coated dishes (37°C, 7% CO2). Differentiation was performed essentially as previously described [5,7]. In brief, ES cells were deprived of feeder cells during 3 to 4 passages, then $4.3x10^6$ cells were used for formation of cellular aggregates. These were cultivated in non-adherent bacterial dishes for 8 days. These represent the neuronal progenitors (NP).

ChIP-Seq and ChIP-chip. Chromatin immunoprecipitation (ChIP) assay for CTCF was performed according to the *Upstate protocol*. ChIP assay for REST was performed as previously described [51]. ChIP assays for Pax6, PolII and monomethylated H3K4 (H3K4me1) were performed as previously described [52]. The antibodies used were: anti-monomethyl-H3K4 (Abcam #ab8895), anti-CTCF (SantaCruz #15914), anti-PolII(N20) (SantaCruz #sc-899), anti-Pax6 (Covance PRB-278P). ChIP-real time PCR was performed using SYBR Green chemistry (ABI) and 1/80 of ChIP or 20ng of input chromatin per PCR reaction. H3K4me1, CTCF and PolII ChIP-Seq libraries for Illumina sequencing were prepared with the Illumina ChIP-Seq DNA Sample Prep Kit (Cat# IP-102-1001) according to Illumina's instructions and sequenced on the Genome Analyzer II following the manufacturer's protocols. Pax6 ChIP samples were amplified using the WGA2 kit (Sigma) and hybridized to a custom tiling microarray [7] (NimbleGen Systems Inc.).

mRNA-Seq. RNA from ES cells and NP of two independent biological replicates each was used for cDNA preparation using oligo-dT primers followed by sequencing on an Illumina GA II analyzer.

Strand specific RNA-Seq. Total RNA was isolated using Trizol (Invitrogen). Two micrograms of total RNA were depleted from ribosomal RNA (rRNA) using the Ribo-Zero rRNA removal kit (Epicentre). The rRNA depleted samples were used to construct the strand specific RNA-Seq libraries following the Illumina pre-release version of the Directional mRNA-Seq Library Preparation Guide.

BisSeq library preparation. The protocol was adapted from Illumina Genomic DNA Sample Preparation Guide and Paired-End Sample Preparation Guide. Briefly, One to five µg of input DNA were fragmented by sonication to 50-500 bp with a Bioruptor (Diagenode, Sparta, NJ). DNA fragments were end repaired by incubation at 20°C for 30 minutes with 400µM dNTP, 15 units of T4 DNA polymerase (NEB #M0203S), 5 units of DNA Polymerase I Lg. Frag. (Klenow) (NEB #M0210S), 50 units of T4 PNK (NEB #M0201S), 1x T4 DNA ligase buffer containing 10mM ATP (NEB). 3' ends of DNA fragments were adenylated by incubation at 37°C for 30 minutes with 200µM dATP, 1xNEB Buffer 2, 15 units Klenow Fragment (3´→5´ exo–) (NEB # M0212L). Adapter sequences were reproduced based on Illumina adapter sequences (Oligonucleotide sequences © 2006-2008 Illumina, Inc. All rights reserved). For single end sequencing: 5' P- GATXGGAAGAGXTXGTATGXXGTXTTXTGXTTG and
5' AXAXTXTTTXXXTAXAXGAXGXTXTTXXGATXT, and for paired end sequencing:
5' P-GATXGGAAGAGXGGTTXAGXAGGAATGXXGAG and
5' AXAXTXTTTXXXTAXAXGAXGXTXTTXXGATXT where X is a methylated cytosine. Adapters were ordered as single stranded oligos (Microsynth AG), resuspended in annealing buffer (10mM Tris pH7.5, 50mM NaCl, 1mM EDTA), annealed by heating at 95°C for 10 minutes and cooling down slowly. Annealed adapters were ligated to the DNA fragments as per manufacturer's instructions for genomic DNA library construction. Adapter-ligated DNA of 140-210 bp (for single end sequencing) or 340-410 (for paired end sequencing) was isolated by 2% agarose gel electrophoresis. Gel purified DNA was then converted with sodium bisulfite using

the Imprint® DNA Modification Kit (Sigma-Aldrich) as per manufacturer's instructions. One third of the bisulfite-converted, adapter-ligated DNA molecules were enriched by 7 cycles of PCR with the following reaction composition: 2.5 U of uracil-insensitive *PfuTurboCx* Hotstart DNA polymerase (Stratagene), 5 µl 10X *PfuTurbo* reaction buffer, 25 µM dNTPs, 0.5µM of Illumina PCR primers. The thermocycling parameters were: 95°C 2 min, 98°C 30 sec, then 7 cycles of 98°C 15 sec, 65°C 30 sec and 72°C 3 min, ending with one 72°C 5 min step. The reaction products were purified using the MinElute PCR purification kit (Qiagen, Valencia, CA), run on 2% agarose gel electrophoresis to separate the library from adapter-adapter ligation products, and purified from the gel using the MinElute gel purification kit (Qiagen, Valencia, CA). Quality of the libraries and template size distribution were assessed by running an aliquot of the library on an Agilent 2100 Bioanalyzer (Agilent Technologies).

Whole Genome (WG) Sequencing library preparation. The protocol was adapted from Illumina Genomic DNA Sample Preparation Guide and Paired-End Sample Preparation Guide. Briefly, One to five µg of input DNA were fragmented by sonication to 50-500 bp with a Bioruptor (Diagenode, Sparta, NJ). DNA fragments were end repaired by incubation at 20°C for 30 minutes with 400µM dNTP, 15 units of T4 DNA polymerase (NEB #M0203S), 5 units of DNA Polymerase I Lg. Frag. (Klenow) (NEB #M0210S), 50 units of T4 PNK (NEB #M0201S), 1x T4 DNA ligase buffer containing 10mM ATP (NEB). 3' ends of DNA fragments were adenylated by incubation at 37°C for 30 minutes with 200µM dATP, 1xNEB Buffer 2, 15 units Klenow Fragment (3´→5´ exo–) (NEB # M0212L). Adapter sequences were reproduced based on Illumina adapter sequences (Oligonucleotide sequences © 2006-2008 Illumina, Inc. All rights reserved). For paired end sequencing:

5' P-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG and

5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT. Adapters were ordered as single

stranded oligos (Microsynth AG), resuspended in annealing buffer (10mM Tris pH7.5, 50mM NaCl, 1mM EDTA), annealed by heating at 95°C for 10 minutes and cooling down slowly. Annealed adapters were ligated to the DNA fragments as per manufacturer's instructions for genomic DNA library construction. Adapter-ligated DNA of 340-410 (for paired end sequencing) was isolated by 2% agarose gel electrophoresis. One third of the adapter-ligated DNA molecules were enriched by 7 cycles of PCR with the following reaction composition: 2.5 U of *PfuTurboCx* Hotstart DNA polymerase (Stratagene), 5 µl 10X *PfuTurbo* reaction buffer, 25 µM dNTPs, 0.5µM of Illumina PCR primers. The thermocycling parameters were: 95°C 2 min, 98°C 30 sec, then 7 cycles of 98°C 15 sec, 65°C 30 sec and 72°C 3 min, ending with one 72°C 5 min step. The reaction products were purified using the MinElute PCR purification kit (Qiagen, Valencia, CA), separated by 2% agarose gel electrophoresis to separate the library from adapter-adapter ligation products, and purified from the gel using the MinElute gel purification kit (Qiagen, Valencia, CA). Quality of the libraries and template size distribution were assessed by running an aliquot of the library on an Agilent 2100 Bioanalyzer (Agilent Technologies).

High-throughput sequencing. BisSeq, ChIP-Seq, RNA-seq and whole genome (WG) DNA libraries were sequenced using the Illumina Genome Analyzer II (GA II) and the Illumina HiSeq 2000 as per manufacturer's instructions. Sequencing of BisSeq and WG libraries was performed up to 100 cycles to yield longer sequences that are more amenable for unambiguous mapping to the mouse genome reference sequence. Image analysis and base calling were performed with the standard Illumina pipeline (RTA v1.9-v1.12, Casava v1.7, OLB v1.8), performing automated matrix and phasing calculations on the PhiX control that was run in the fifth (GAII) or the eighth (HiSeq) lane of each flowcell.

Bisulfite Sanger Sequencing. 1-5 µg genomic DNA was fragmented by sonication and bisulfite converted with the Imprint® DNA Modification Kit (Sigma-Aldrich) or the EpiTec Bisulfite Kit

(QIAGEN). Primers were designed using MethPrimer [53] to amplify specific regions of the genome following bisulfite conversion. Regions of interest were amplified by PCR, separated by gel electrophoresis, gel purified, and cloned by TOPO-TA cloning (Invitrogen). Primers for PCR amplification are available upon request. Prior to sequencing, plasmids forming individual clones were amplified using the QIAprep Spin Miniprep Kit (QIAGEN) according to the manufacturer's protocol. Sanger sequencing of multiple clones for each amplicon was performed to identify the methylation status of cytosines. BiQ Analyzer tools were then used to compare sequences from individual clones to the original sequence, perform quality controls, eliminate sequences that might generate from the same clone and draw diagrams in a standardized manner.

Homologous recombination. The homologous recombination strategy to create a recombination substrate was described previously [25]. Briefly, **pZRMCE** plasmid used for targeting mouse TC-1 ES cells (background 129S6/SvEvTac) in the β -globin locus was constructed in the pZERO multiple cloning site (MCS) and included a 2.4 kb Not I - Xho I fragment designated upstream arm (from positions -3700 to -1300 relative to the εy ATG start) and a 3.0 kb Kpn I - Not I downstream arm (positions +2332 to +5432) cloned 5' and 3', respectively, to the selection cassette which was flanked by inverted *lox p* sites. TC-1 ES cells were electroporated with 100 µg **pZRMCE** plasmid using a BioRad gene pulser (500 µF and 250 V/cm). Cells were selected with 150 µg/ml of hygromycin for 7-10 days after transfection. Clones were tested for successful recombination events by Southern blot analysis.

Recombinase mediated cassette exchange (RMCE). RMCE was performed as previously transcribed [25]. DNA fragments were cloned into a plasmid containing a multiple cloning site flanked by two inverted L1 Lox sites (kind gift from Matthew Lorincz). Promoter regions were amplified from TC-1 ES cells genomic DNA. We inserted 2 promoter fragments corresponding to the following genomic coordinates (mm9): Orm1 (chr4:63'005'184-63'005'948), Mrap

(chr16:90'738'245-90'738'944) and one E. Coli DNA fragment: Fr2. When indicated, wild-type CTCF motif: ATAGCGCCCCCTAGTGGCCA, CTCF mutated motif: ATAGCGCGCCGTAGTGGCCA were inserted in the fragments. *In vitro* methylation of the fragments before insertion, was performed as previously described [54]. The completion of the reaction was tested by digestion with methylation sensitive restriction enzyme (HpaII). Methylation-insensitive restriction enzyme MspI was used as a control. Insertions were performed as previously described [25] with slight modifications. TC-1 ES cells were selected under hygromycin (25 µg/ml, Roche) for 10 days. Next, $4x10^6$ cells were electroporated (Amaxa nucleofection, Amaxa) with 25 µg L1-promoter-1L plasmid and 15 µg of pIC-Cre (kind gift from Rémi Terranova). Selection with 3 µM Ganciclovir (Roche) was started two days after transfection and continued for 7-10 days. Clones were tested for successful insertion events by PCR.

REST overexpression. *REST* and *eGFP* cds were cloned into **pcDNA6-IRES-Blasticidin** vector (kind gift from Deborah Schmitz). We replaced the CMV promoter by a CAG promoter which showed better activity in ES cells. This resulted in 2 constructs: **pcDNA6-CAG-Rest-IRES-Blasticidin** and **pcDNA6-CAG-eGFP-IRES-Blasticidin**. More details on cloning process are available upon request. *REST* $^{-/-}$ ES cells were electroporated with both of these constructs separately and were subjected to Blasticidin selection for 2 weeks. Surviving clones were tested for REST and GFP overexpression by western blot. DNA was extracted from one positive clone for REST and one for GFP and subjected to bisulfite sequencing.

Western Blot Analysis. For detection of REST and GFP protein levels, total cell lysates from wildtype ES cells (+/+), REST knockout ES cells (-/-), wildtype 159-2 ES cells overexpressing GFP (+/+) and Blasticidin resistant clones were used for western blot analysis. The membrane was probed with mouse anti-REST (12C11, kind gift from David Anderson), anti-GFP (Abcam

#ab290) and rat anti-tubulin (tissue culture supernatant, cell line YL1/2, ECACC) as a loading control, in combination with appropriate secondary antibodies coupled to HRP.

Luciferase enhancer assays. Fragments corresponding to LMRs and non-ES enhancers were amplified from 100ng of 159-2 ES cell genomic DNA using Pfu-polymerase (Catalys) and 5uM of primers containing BamHI- or BglII-restriction sites (primer sequences available upon request). Resulting PCR products were gel-purified using Qiagen Gel purification kit, digested with either BamHI or BglII and subsequently purified with Qiagen PCR-cleanup kit. They were then cloned into the BamHI-site of the **pGL3-promoter-vector** (Promega, E1761), using Rapid DNA Ligation Kit (Roche) and transformed into chemical competent DH5α-cells. Colonies were checked by minipreps and the positive clones verified by sequencing.

4 µg of resulting constructs corresponding to LMRs, non-ES enhancers, positive control (**pGL3-control-vector**, Promega, E1741) or empty **pGL3-promoter-vector** (promoter only) were transfected in triplicate into $0.5x10^6$ 159-2 ES cells using Lipofectamine 2000 (Invitrogen). We also performed a mock control (renilla only). In all instances 0.4 µg of renilla luciferase plasmid driven by a CMV promoter were co-transfected as an internal control. Medium was replaced after 6 hours and cells were harvested after 48 hours. Reported luciferase expression levels are relative to internal renilla control.

Computational procedures:

Genomic Coordinates

The July 2007 M. musculus genome assembly (NCBI37/mm9) provided by NCBI (http://www.ncbi.nlm.nih.gov/genome/guide/mouse/) and the Mouse Genome Sequencing

Consortium (http://www.sanger.ac.uk/Projects/M_musculus/) was used as a basis for all analyses. Annotation of known RefSeq transcripts and repeat elements was obtained from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/mm9/database/refGene.txt.gz from Oct 18, 2009, and http://hgdownload.cse.ucsc.edu/goldenPath/mm9/database/chr*_rmsk.txt.gz from Jan 30, 2009). Five types of genomic regions were defined as follows: “Promoter” contains all bases within 1000 basepairs of a known RefSeq transcription start site (TSS). “Exon” are all non-promoter bases that overlap exons of RefSeq transcripts, “repeat” are non-promoter/non-exon bases that overlap repeat elements and “intron” are all non-promoter/non-exon/non-repeat bases that are flanked by two exons of a single transcript. All remaining bases were assigned to the “intergenic” region type.

ChIP-Seq and RNA-Seq: Read Filtering, Alignment and Weighting

Low-complexity reads were filtered out based on their dinucleotide entropy (removing reads with less than half the average entropy of a genomic sequence of the same length, typically accounting for <0.5% of the reads). Alignments to the mouse genome were performed by the software bowtie (version 0.9.9.1) [50] with parameters -v 2 -a -m 100, tracking up to 100 best alignment positions per query and allowing at most two mismatches. To track genomically untemplated hits (e.g., exon-exon junctions or missing parts in the current assembly), the reads were also mapped to an annotation database containing known mouse sequences (miRNA from ftp://ftp.sanger.ac.uk/pub/mirbase/sequences/13.0, rRNA, snRNA, snoRNA and RefSeq mRNA from GenBank http://www.ncbi.nlm.nih.gov/sites/entrez, downloaded on July 16, 2009, tRNA from http://lowelab.ucsc.edu/GtRNAdb/ and piRNA from NCBI accessions DQ539889 to DQ569912). In that case, all best hits with at most two mismatches were tracked. Each alignment was weighted by the inverse of the number of hits. In the cases where a read had

more hits to an individual sequence from the annotation database than to the whole genome, the former number of hits was selected to ensure that the total weight of a read does not exceed one. All quantifications were based on weighted alignments, and alignments from ChIP-Seq experiments were shifted 60 bases towards their 3'-end to account for an estimated fragment length of 120 basepairs.

BisSeq: Read Filtering, Alignment and Quantification of Methylation Levels

All C nucleotides in sequence reads from bisulfite converted samples were converted in silico to T nucleotides, and the converted reads were aligned to a similarly converted genome separately to each strand using the software bowtie (version 0.10.0.1) [50] with parameters --best --strata -v 3 --norc -a. Only reads with a unique alignment in this reduced alphabet base-space were retained, and C nucleotides from the original reads and genome were reintroduced. To eliminate effects caused by polymorphisms in our experimental system, C nucleotides that overlapped known SNPs between the reference C57BL/6J and the 129S5 strains were removed from further analysis based on the SNPs identified by the Mouse Genomes Project at Sanger Institute (downloaded from ftp://ftp-mouse.sanger.ac.uk/REL-1003/SNPs/20100301-all-snps.tab.gz). Percent methylation for individual Cs in the genome were calculated as the ratio of the number of alignments with C (methylated), over the number of alignments with either C (methylated) or T (unmethylated). For strand-independent analysis of CpG methylation, counts from the two Cs in a CpG and its reverse complement (position *i* on plus strand and position *i*+1 on minus strand) were combined.

Definition of Mappable Regions in the Mouse Genome

Genomic bases were defined as mappable, if they were contained in an oligonucleotide of the same length as a sequencing read that did not produce more than 100 hits in the whole genome when aligned with above alignment parameters.

Unsupervised Segmentation of CpG Methylome

Methylation levels were calculated as described above for individual CpGs (combining counts from both strands) and removing CpGs with a coverage of less than five or overlapping a known SNP between C57BL/6J and 129S5 mouse strains, resulting in a total of 17.6 and 17.0 million CpGs used as input for the segmentation in ES cells and NP, respectively (corresponding to 92.1% and 89.2% of all covered CpGs in ES cells and NP). The segmentation was performed using the R package RHmm (version 1.4.4, http://CRAN.R-project.org/package=RHmm) and a three state Hidden Markov Model (HMM) corresponding to fully methylated, low methylated and unmethylated CpGs. Emission probabilities of HMM states were modeled as Gaussian distributions with means 0.8, 0.28, 0.03 and variances 0.01, 0.005, 0.0026, and transition probabilities were ((0.98, 0.02, 1e-10), (0.21, 0.76, 0.03), (1e-10, 0.05, 0.95)), essentially preventing any direct state transitions from the fully methylated to the unmethylated state or vice versa without going through the low methylation state. Parameters were initially estimated by the Baum-Welch algorithm using ES cell methylation data from chromosome 19, and the means of fully and low methylated states were then manually lowered to reduce detection of low methylated states in randomly permuted data (see estimation of false discovery rate below). Using the maximum likelihood path through the model (Viterbi algorithm), initial state labels were assigned to individual CpGs for each chromosome separately. Consecutive CpGs with identical state labels were combined into the same segment, and segments with only one or two CpGs were combined with their upstream neighboring segment. Finally, UMRs were defined as unmethylated segments, FMRs as fully methylated segments, and LMRs as low methylated segments flanked by FMRs on both sides. Low methylated segments that were immediate neighbors of UMRs were removed from further analysis. Non-CpG bases in the genome were assigned to the segment that contained the nearest CpG, resulting in segments with genomic start and end coordinates corresponding to the midpoints between consecutive CpGs. Identical

parameters were used to segment mouse ES cells, mouse NP and human H1 cell methylomes. In order to estimate a false discovery rate (FDR) for the detection of the difficult to identify LMRs, methylation levels from CpGs in FMRs and LMRs were randomly permuted, and FDR was calculated as the number of LMRs detected in the randomized data, divided by the number of LMRs detected in the real data, resulting in an FDR of 6.7%, 6.3% and 8.6% in mouse ES cells, mouse NP and human H1 cells, respectively. Finally, segments were classified as either "TSS proximal" (have a RefSeq annotated TSS within 1 kb of the segment midpoint) or "TSS distal" (no TSS within 1kb), and as either "CG low" (GC content below 50% or CpG observed over expected ratio below 0.6) or "CG high" (others). The "CG high" class definition essentially corresponds to the CpG island definition according to Gardiner-Frommer [55], except for their additional length constraint of at least 200 base pairs. Segmentation results for ES cells and NP are available in Supplementary Tables 2 and 3, respectively.

Calculation of Segment Enrichments in Genomic Regions

Enrichment of segments in genomic regions (see above for definition) were calculated as the ratio of observed over expected number of bases of each segment type (FMR, LMR, UMR) in a region (e.g. exon, intron etc.), where the observed number is the number of bases in segments of a given type that overlap a region, and the expected number is the fraction of genomic bases in that region type, multiplied with the total number of bases in all segments of that type.

Calculation of Expected Segment Overlaps with CpG Islands

The expected numbers of CpG islands overlapping different segment types were calculated by multiplying the total number of CpG islands with the fractions of genomic bases per segment type. This corresponds to the null model assumption that CpG island locations are independent from the methylation segments and would create overlaps with at a frequency that is proportional to their genomic abundance.

Calculation of IP Enrichments and RNA-Seq Fold-Changes

IP enrichments of a genomic region were calculated as $e = \log2((n\_fg / N\_fg * \min(N\_fg, N\_bg) + p) / (n\_bg / N\_bg * \min(N\_fg, N\_bg) + p))$, where *n_fg* and *n_bg* are the summed weights of overlapping foreground and background (input chromatin) read alignments, respectively. *N_fg* and *N_bg* are the total number of aligned reads in foreground and background samples, and *p* is a pseudocount constant (*p*=8) used to regularize enrichments based on low counts that would otherwise be dominated by sampling noise. For comparison of different sequencing datasets, tag counts were normalized in an analogous fashion. Enrichment profiles around segment middles were calculated similarly by summing up all alignments at a given distance from the midpoint, and using a pseudocount constant *p* of 8 times the expected number of alignments per base pair (total number of alignments divided by the number of mappable bases in the genome). For the H3K9me2 ChIP-chip samples, enrichments for individual microarray probes were calculated as described [6]. To reduce signal bias at probes from regions with high GC or CpG content, only probes with less than 45 CpGs in a 1000 base pair window centered at the probe were retained (92.8% of all probes).

Expression fold changes based on RNA-Seq data were calculated similarly to the IP enrichments *e* described above, except for foreground (*fg*) and background (*bg*) corresponding to ES and NP samples, and *n* to the weighted sum of alignments to a representative RefSeq transcript for each gene (the one with median 3'UTR length from all RefSeq transcripts of a given gene).

Association of LMRs with gene expression

Each gene was assigned to the LMR nearest to the transcript start site of its representative RefSeq transcript (using LMRs in either ES cells or NP, disregarding UMRs). Expression fold-changes for all genes were calculated as described above and correlated to the methylation

changes of assigned LMR segments. Significance of the correlation was calculated using the cor.test function in R (www.r-project.org).

Allele specific analysis

All reads from ChIP-Seq and whole genome (WG) DNA libraries generated in the lab were pooled and mapped to the reference genome C57BL/6J. For each known SNP between the reference C57BL/6J and the 129S5 strains, an allelic ratio was calculated as number of alignments with 129S5 allele over number of alignments with C57BL/6J or 129S5 allele. Only SNPs located on autosomes and covered by at least 15 reads were used as an input for segmentation. The segmentation was performed using the R package RHmm (version 1.4.4, http://CRAN.R-project.org/package=RHmm) and a three state Hidden Markov Model (HMM) corresponding to homozygous 129S5, heterozygous and homozygous C57BL/6J state. Emission probabilities of HMM states were modeled as Gaussian distributions with means 0.995, 0.423, 0.009 and variances 0.006, 0.026, 0.001, and transition probabilities were ((0.99, 1e-10, 1e-10), (1e-10, 0.99, 1e-10), (1e-10, 1e-10, 0.99)) and allowing any direct state transitions. Parameters were initially estimated by the Baum-Welch algorithm using data from chromosome 1, and the variance of homozygous C57BL/6J and heterozygous state were then manually increased to reduce detection of heterozygous state in randomly permuted data. Using the maximum likelihood path through the model (Viterbi algorithm), initial state labels were assigned to individual SNPs for each chromosome separately. Consecutive SNPs with identical state labels were combined into the same segment. Segments with only one or two SNPs were combined with their neighboring segments if they both were identical. The resulting ES cell genotype reconstruction is available in Supplementary Table 4.

In order to prevent mapping bias towards reference (C57BL/6J) allele, the second genome corresponding to 129S5 allele was generated by incorporating all single nucleotide

polymorphisms into the reference genome. The SNPs identified by the Mouse Genomes Project at Sanger Institute (downloaded from ftp://ftp-mouse.sanger.ac.uk/REL-1003/SNPs/20100301-all-snps.tab.gz) were injected using the R package Biostrings (version 2.20.1, http://www.bioconductor.org/packages/2.8/bioc/html/Biostrings.html). The analyses of allele specific DNA methylation or transcription factor binding were performed only in heterozygous regions. Estimation of DNA methylation and CTCF binding for C57BL/6J and 129S5 alleles were based on the alignments to the reference genome and 129S5 genome, respectively.

Peak Finding in ChIP-Seq Data

Genomic regions of increased ChIP-Seq read alignment densities were identified using macs (version 1.3.7.1) [56], using a pool of read alignments from all biological replicates and cellular stages (weights rounded to integers) as input, parameters --mfold=8 --gsize=2700000000 --tsize=36 --nomodel --shiftsize=60 and default values for all other parameters. IP enrichments (see below) of resulting peak candidates were calculated and peak candidates with enrichments lower than 2-fold above background (combining biological replicates by summing read counts prior to calculation of IP enrichments) were removed.

Binding Motif Identification and Site Prediction for CTCF

Motif weight matrices overrepresented in ChIP-enriched regions were identified using MEME version 4.3.0 [57] on the top 1000 enriched peaks with parameters -dna -mod oops -revcomp and -w 20, using a zero order Markov model as background estimated from all mappable regions in the genome that did not overlap any ChIP peaks, respectively. The obtained motif was used to scan the genome using MAST [58] with default parameters and the same background model as used for MEME. For subsequent analysis, only matches with a score greater than 1000 were retained, resulting in 1.0 match per peak and 391,862 predicted CTCF sites in the genome. Of these, 22% were overlapping ChIP peaks.

Prediction of CTCF Binding using Linear Models

IP enrichments for CTCF in ES cells were calculated as described above for windows of 200 base pairs centered on genomic sequences matching the identified weight matrix model, only retaining windows with at least four CpGs (n=89,167). Similarly, the average methylation levels in ES cells were calculated for the same 200 base pair windows. Two linear models were then fitted to the experimentally observed CTCF IP enrichments: The first one using the motif match score only and the second using motif match score and the average methylation level. Model performance was assessed using adjusted R-square values, evaluating the significance of the increase using analysis of variance (ANOVA). All model fitting was done in R (www.r-project.org) using the lm function.

Conservation Analysis

For conservation analysis, we downloaded the PhastCons tracks from UCSC (http://genome.ucsc.edu/). For mouse, we used phastCons11way (downloaded on Oct 27, 2009), which includes all euarchontoglires, for human phastCons44way (Nov 4, 2010), which includes all placental mammals. This measure is precalculated for each position in the mouse genome based on multigenome alignments between, in this case, all euarchontoglires for mouse and all placental mammals for human [59]. We displayed this measure as an average value for all regions within each class. The higher conservation of UMRs compared to LMRs is driven by the fact that these are mostly promoters and thus often include first exons.

Methylation Dynamics from ES to NP

To study the methylation dynamics between ES cells and NP, all hypomethylated segments (UMRs and LMRs) that resulted from the independent segmentation of the ES and NP methylomes were joined into one set of consolidated segments. If ES and NP segments

overlapped, a new segment was created that contained all the nucleotides of each segment. The methylation levels of these segments were recalculated using all contained CpGs. Although there were subtle changes in the shape of overlapping segments, such as a general decrease in length of UMRs in NP compared to ES cells (data not shown), the methylation changes in the set of consolidated segments agreed well with the changes observed on the separate set of ES and NP segments. The consolidated segments (many of which had changed in their exact size due to overlaps) were then reclassified depending on whether their methylation level was below 13.9 % (UMR), between 13.9% and 50% percent (LMR) or above 50% percent (FMR). The cut-offs of 13.9% and 50% were determined as the intersection points of the Gaussian emission probability distributions of the respective HMM states.

Motif Search and Site Prediction in LMRs

For motif search, we first selected the subset of segments that were LMRs in at least one of the two stages and, in order to focus on LMR to FMR dynamics as well as on distal elements, removed those segments that were UMRs in any of the two stages. We then masked from each segment all the nucleotides that overlapped with UCSC-annotated repeats since these cause strong biases in motif enrichments. Segments were then partitioned into ES-specific segments (LMR in ES cells, FMR in NP), NP-specific (FMR in ES cells, LMR in NP) and constitutive segments (LMR in both ES cells and NP).

Motivated by an initial unbiased analysis of k-mer enrichments, that indicated enrichments for several transcription factors with known binding motifs, we comprehensively investigated motif enrichments for factors with known binding preference. For this purpose, we downloaded the set of non-redundant vertebrate weight matrices from the Jaspar database [60], which contains weight matrices for 130 transcription factors. From this set, we retained only those factors that could be mapped to RefSeq transcripts (for assignment of expression levels, see below), either via gene

name matching or by manual annotation via the Uniprot identifiers of each factor as provided by Jaspar. This resulted in a reduced set of 126 transcription factors. We then scanned each segment with each of the 126 weight matrices using the countPWM function provided by the Biostrings R package (version 2.20.0, http://www.bioconductor.org/packages/2.8/bioc/html/Biostrings.html). A sequence was considered a match to a weight matrix if the posterior probability of it stemming from the weight matrix was larger than 0.999. This cut-off was chosen such that the average number of segments with a site was roughly 10 percent of the total number of segments. As a background model, we used a 0-order Markov model with equal probabilities for A, C, G and T (using a 0-order Markov model trained on the input data as well as a binning of segments by CpG density and using different 0-order Markov models for each bin resulted in very similar results). Enrichments for each set of segments (ES-specific, NP-specific and constitutive) were then determined by comparing the frequency of predicted sites for a particular factor in one set (the foreground set) compared to the union of the other two sets (the background). To be precise, for both background and foreground sets, the total number of predicted sites for each transcription factor was determined. The total number of predicted sites (summed over all transcription factors) was then rescaled such that the total number of predicted sites in both foreground and background was equal to the minimum of the total number of predicted sites in either set. Finally, to avoid spuriously high enrichments due to small counts (for transcription factors with only small numbers of predicted binding sites), a pseudocount of 20 was added to the number of predicted binding sites of each factor. The enrichment of binding sites for a particular factor was then calculated as the ratio of the rescaled number of sites in the foreground and background set (Supplementary Figure 15). We determined the significance of the enrichments as follows. For each transcription factor, we calculated the background probability of a motif occurrence as the number of predicted sites in the background set divided by the total number of predicted sites for all transcription factors in the background set. Using this probability, we

determined a p-value using a binomial model, with the number of successes corresponding to the observed number of matches in the foreground and the number of trials equaling the total number of predicted sites in the foreground.

Linear Model for the Inference of Transcription Factor Activity

In order to investigate to what extent changes in methylation could be explained by the activity of the 126 transcription factors with known weight matrices, we fitted a linear model to the methylation data. We modeled the change in methylation in LMRs (the response variable) as a linear sum of the activities (regression coefficients) of each factor times the number of predicted binding sites for each factor (the explanatory variables). As an additional explanatory variable, we added the CpG density of each segment.

Transcription factors with less than 10 sites in all segments (n=22) were removed from the model as their activities cannot be accurately inferred. The fit of the model was performed using the lm function in R. Since there were many more segments that are de-novo methylated than de-methylated from ES cells to NP, we performed a weighted linear regression. The weights were chosen such that the segments within each bin of methylation changes (10 bins of equal width, ranging from the lowest to the largest observed methylation change) had the same total weight and that the total sum of weights added up to the number of segments. With this model, we were able to explain ~20 % of the total variance (adjusted R-squared of 0.19 and 0.17 if the CpG density is taken out of the model).

Comparison of Hypomethylated Regions to DNaseI-hypersensitive Sites

We downloaded fastq files for DNaseI hypersensitivity samples for mouse ES cells (two replicates, wgEncodeUwDNaseIEscj7S129ME0) and human H1 cells (one replicate, wgEncodeUwDNaseISeqRawDataRep1H1es) from Encode at UCSC (http://genome.ucsc.edu/ENCODE/, Supplementary Table 1).

In both mouse and human, initial visual inspection of DNaseI hypersensitivity tracks revealed a very strong overlap of our inferred hypomethylated regions with DNaseI-enriched regions. To investigate the relationship of hypersensitive sites with hypomethylated regions in detail, we related DNaseI tag counts to methylation levels in a consolidated set of regions. This set was constructed as follows. We first selected all DNaseI peaks (fwgEncodeUwEscj7S129ME0HotspotsRep1.broadPeak for mouse and gEncodeUwSeqHotspotsRep1H1es.broadPeak for human) and selected all peaks with p-value <= 0.001 (n=159659, average length of 525 for mouse and n=156747, average length of 424 for human). Comparison of DNaseI levels within these peaks compared to DNaseI levels of randomly selected regions of the genome with the same length distribution revealed that at this p-value cutoff, the great majority of sites had DNaseI levels above background (data not shown). We then separated UMRs and LMRs into DNaseI positive and negative, depending on whether they overlapped with a DNaseI peak (overlap of at least one nucleotide). Finally, we created a set of background region (termed "gaps"), which correspond to all the regions that lie between any combination of DNaseI peaks, UMRs and LMRs. For each type of region, we determined DNaseI levels as the number of DNaseI tags per 1 kb and average methylation levels of CpGs in the respective region. In Supplementary Figure 2b, we plotted, for each region, its DNaseI level against its methylation level, with the exception of DNaseI-positive LMRs and UMRs, for which we plotted the methylation level of the LMR/UMR against the DNaseI level of the DNaseI peak it overlapped with (in order to optimally measure the signals in each respective assay).

Pax6 ChIP-chip Data Analysis

Nimblegen array intensity files were read and log2 enrichments (log2 bound/input ratios) for

each individual probe were calculated using the R package Ringo [61]. All arrays were loess-normalized using the normalizeWithinArrays function from the limma package [62]. To quantify Pax6 levels in LMRs, we intersected the probe coordinates with LMR coordinates and retained all LMRs with at least 3 assigned probes (resulting in 4176 (12 % of all) LMRs). Pax6 levels per LMR were then calculated as the mean log2 enrichment over all probes. Enrichments in LMRs showed high reproducibility (average pairwise Pearson's Correlation Coefficient of 0.92). For comparison with methylation levels, we averaged the log2 enrichment in each LMR over all three replicates.

References

5. Bibel, M., Richter, J., Lacroix, E. & Barde, Y.A. Generation of a defined and uniform population of CNS progenitors and neurons from mouse embryonic stem cells. *Nat Protoc* **2**, 1034-43 (2007).
6. Lienert, F. *et al.* Genomic prevalence of heterochromatic H3K9me2 and transcription do not discriminate pluripotent from terminally differentiated cells. *PLoS genetics* **7**, e1002090 (2011).
7. Mohn, F. *et al.* Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Mol Cell* **30**, 755-66 (2008).
25. Lienert, F. *et al.* Identification of genetic elements that autonomously determine DNA methylation states. *Nature genetics* **43**, 1091-7 (2011).
50. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
51. Koch, C.M. *et al.* The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome research* **17**, 691-707 (2007).
52. Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* **39**, 457-66 (2007).
53. Li, L.C. & Dahiya, R. MethPrimer: designing primers for methylation PCRs. *Bioinformatics* **18**, 1427-31 (2002).
54. Schubeler, D., Lorincz, M.C. & Groudine, M. Targeting silence: the use of site-specific recombination to introduce in vitro methylated DNA into the genome. *Sci STKE* **2001**, pl1 (2001).
55. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J Mol Biol* **196**, 261-82 (1987).
56. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
57. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**, 28-36 (1994).
58. Bailey, T.L. & Gribskov, M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**, 48-54 (1998).
59. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* **15**, 1034-50 (2005).
60. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W. & Lenhard, B. JASPAR: an

open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**, D91-4 (2004).
61. Toedling, J. *et al.* Ringo--an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC bioinformatics* **8**, 221 (2007).
62. Gentleman, R., Carey, S., Dudoit, R., Irizarry, R. & Huber. *Limma: linear models for microarray data*, (Springer, New York, 2005).