Supplementary Materials

Expanded Methyl Sensitive Cut Counting Reveals Hypo-Methylation as an Epigenetic State that Highlights Functional Sequences of the Genome

Alejandro Colaneri[1], Nickolas G. Staffa, Jr.[1], David C. Fargo[2], Yuan Gao[3] , Tianyuan Wang[1], Shyamal D. Peddada[4] and Lutz Birnbaumer[1]#

from the

[1]Laboratory of Neurobiology, [2]Library and Information Services and [4]Biostatistics Branch, National Institute of Environmental Health, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, North Carolina 27709, and [3]Division of Genomics, Epigenomics and Bioinformatics, Lieber Institute for Brain Development & Neuroregeneration, and Stem Cell Program. Institute for Cell Engineering, Johns Hopkins University, Baltimore, MD 21205.

Running Title: Genome wide mapping of unmethylated regions

Keywords: Epigenetics – DNA methylation – Unmethylated regions – CpG islands - Methylome

*Supported by the Intramural Research Program of the National Institutes of Health (2009 NIH Director's Challenge Award and Project Z01 ES101643 to LB)

#to whom correspondence should be addressed at NIEHS, Building 101 Room F179, T.W. Alexander Dr., Research Triangle Park, NC 27709; e-mail: birnbau1@niehs.nih.gov.

This file contains the following information:

Acknowledgements

We are thankful for help received from members of the Biostatistics Branch (NIEHS). Dr. Pierre Bushel implemented, during the initial phases of the work, the counting of the number of CCGG sites in the mouse and human genomes and supervised the preparation of complete lists of these CCGG sites with their chromosomal coordinates and corresponding, up to 96-nt long, upstream and downstream flanking sequences in FASTA and text formats; Dr. Grace Kissling (Statistical Consulting Service) implemented the visualization, formatting and sorting of these lists using SAS's JMP software. We acknowledge Dr. Shuangshuang Dai and John Grovenstein for computational infrastructure support. We also acknowledge the special financial help during 2007 and 2008 from the Office of the Scientific Director (DIR, NIEHS) without which the first 10 CpG-tag libraries could not have been sequenced. We are also thankful to Dr. Wei-Chun Huang (Boston University) for an analysis of primary Illumina Genome Analyzer files and the preparation of frequency tables with the data from our first deep sequencing experiment. We are grateful to Dr. Jaya Kittur for careful reading and many suggestions that improved the manuscript. We thank Dr. Christoph Bock (Broad Institute of MIT and Harvard // Max Planck Institute for Informatics, Cambridge, MA) for providing us with the chromosomal coordinates of mouse Epi-CGIs.

Methods

Strategy

To identify unmethylated CpGs at the genome wide scale, we digested sheared genomic DNA, separately with each of four CpG methylation-sensitive restriction enzymes (MSREs): AciI, HpaII, HinP1I and HpyCH4IV (4Enz) (Fig. S1A). An approximately equimolar amount of unmethylated lambda phage DNA was added prior

to the shearing step to be used as an internal standard (Fig. S1B). The newly created ends (unmethylated CpGs) were ligated to a dsDNA adaptor (Adaptor A) with a CAGCAG EcoP15I recognition sequence next to the ligation site. The ligated mixture was digested with EcoP15I. The 27-bp dsDNA fragments (adaptorA-CpG tags) excised by the EcoP15I restriction enzyme was isolated (Fig. S1C). These tags were ligated to a second dsDNA adaptor (adaptor B), creating fragments that were amplified by low-cycle-number PCR using forward and reverse primers complementary to adaptors A and B (Fig. S1D). We applied this method to a sample of DNA extracted from mouse liver. Four independent libraries were made, one for each methyl sensitive enzyme. The libraries were pooled to produce a single 4Enz CpG tag library and sequenced using an Illumina Genome Analyzer, and, unless stated otherwise shall discuss in detail results from one such experiment, referred to as Slxa3. We used 5 channels of the Analyzer's 8-channel flow cells, and obtained from each an average of 6 million reads mapped to unique sites. The data, consisting of 36 nt-long reads spanned not only the sequences of the CpG tags generated by EcoP15I, but also extended into the Adaptor B sequence. Adaptor B was found to start between 24 and 29 nt from the CAGCAG, owing to an inherent slippage exhibited by the endonuclease activity of EcoP15I. To eliminate slippage-based errors, all reads were trimmed to 25nt. The data returned from each channel were aligned separately against the mouse reference genome (build mm9) using MOM allowing one mismatch (1). MOM returned the following information: the sequence of the read; whether the read mapped to a single or to multiple sites; if single, the chromosome number and mm9 *Mus musculus* reference genome coordinate; and the sense of the tag (forward or reverse) identified by the read. Except when noted otherwise, the analyses presented here were performed on results for which the forward and reverse tag information was collapsed into one number for each identified CpG.

Preparation of Genomic DNA from Liver Nuclei

Liver nuclei were prepared according to (2). Briefly, adult male mice were euthanized in a CO_2 chamber; their livers were removed and placed on ice. Unless indicated otherwise all subsequent steps were performed at 4^{o}C. After dissection of liver lobes and removal of connective tissue and blood vessels, 2-3 g were washed in ice cold homogenizing medium (HM, 0.32M sucrose, 3mM $MgCl_2$), minced with scissors and homogenized in 4 volumes of HM using a 40 ml Wheaton Dounce homogenizer (*Fisher #06-435C*) and 25 strokes of the loose B pestle. The homogenate was filtered through 2 layers of gauze, taken to 20 ml with HM and centrifuged for 10 min at 700xg in a 30-ml Corex glass centrifuge tube (*Corning #8445*) using the Sorvall SS34 rotor. After discarding the supernatant, the pellet was gently resuspended in the same tube in 10 ml ultracentrifugation medium (UM, 2.4M sucrose, 1 mM $MgCl_2$) using the loose pestle of the Dounce homogenizer. The mixtures were centrifuged for 60 min at 20,000 rpm (50,000 xg) in two tubes of the Beckman SW40.1 Ti rotor. The pellets were each resuspended in 1.0ml of 0.25M sucrose, 1 mM $MgCl_2$, transferred to a 1.5 ml Eppendorf tube. Nuclei were collected by centrifugation in a Sorvall Biofuge Fresco table top centrifuge at 4^{o}C for 10 min at 700 xg and stored frozen at -70^{o}C until used as source of genomic DNA.

DNA was extracted from each nuclear pellet by combining with 40 ml of LB lysis buffer at room temperature (LB, 0.8M guanidinium-HCl, 30mM EDTA, 5% Tween-20, 0.5% Triton X-100, 30mM Tris-HCl, pH 8.0). The nuclear lysate was then incubated first with 200 µg/ml RNAse A (*Sigma #R4642*) for 30 min at 50^{o}C, and then with 20µg/ml proteinase K (*Invitrogen Fungal 25530*) for 180 min at 55^{o}C. The genomic DNA in the resulting mixtures was bound to an anionic exchange resin (Qiagen Genomic Tip 500) that had been prewashed with 10 ml buffer QBT (0.75M NaCl, 50 mM MOPS-NaOH, pH

7.0, 15% v/v isopropanol, 0.15% v/v Triton X-100) at a ratio of 1 tip per 10 ml of RNAse A and proteinase K treated nuclear lysate. Each tip was washed twice with 15 ml buffer QC (1.0 M NaCl, 50 mM MOPS-NaOH, pH 7.0, 15% v/v isopropanol). The genomic DNA was eluted with 15 ml of buffer QF (1.25M NaCl, 50 mM Tris-HCl, pH 8.5, 15% v/v isopropanol). The DNA solutions were concentrated by ultrafiltration to a volume of 1.0 ml using two Amicon Ultra-10 ultrafiltration devices centrifuged at 4,000xg for 20 min in the SH-3000 rotor of a Sorval Legend-RT centrifuge. The 1.0-ml sample from each ultrafiltration devise was dialyzed by 4 ultrafiltration cycles (dilution with 15 ml TE (Tris-HCl, pH8.0, 1mM EDTA) followed by concentration to 1.0 ml). The final retentate in each filtration device was taken to 1.0 with TE. The DNA was rinsed off the membrane and transferred a 1.5 ml Eppendorf tubes. A typical yield of genomic DNA in this procedure determined by absorbance at 260nm, was 1 mg from 2-3g liver, with a 260/280 ratio of 1.8-1.9.

Addition of *lambda* Phage DNA to Genomic DNA and Fragmentation

Genomic DNA was diluted to a concentration of 300-500 ng/µl with 10 mM Tris-HCl, pH 7.5, and supplemented with an equimolar concentration of unmethylated *lambda* DNA (Promega *cat #D1521*) using the equivalence numbers Mr[haploid mouse genome] = 1.799 x 10^12 and Mr[*lambda*]= 32 x10^6. Genomic plus lambda DNA mixtures were split into 200 µl aliquots and fragmented by hydrodynamic shearing using the standard tip (*HSH 204004*) of a Gene Machine Hydroshear apparatus (*Harvard Apparatus*) set at speed code 12 for 80 cycles at a retraction speed setting of 23. An analysis of such sheared DNA by polyacrylamide gel electrophoresis is shown in Fig. S1B.

Digestion with Restriction Enzymes

Twenty-µg aliquots of sheared genomic DNA (1,500-3,500 bp fragments) were digested with 250U of HpaII, 250U HpyCh4IV, 250U HinP1I or 100U AciI each, for 16 h at 37^{o}C at a final volume of 250 µl of the digestion buffer supplied with the enzyme, followed by inactivation of the enzyme by incubation with proteinase K at 16µg/ml for 60 min at 55^{o}C. Subsequent steps were at room temperature and were identical for each of the digests. After addition of 2.50 ml of Qiagen's isopropanol-, $NaClO_4$- and high salt-containing PN buffer, the digested DNA fragments were purified by adsorption to the silica membrane of four Qiaquick spin columns. Samples were applied under centrifugal force in a table-top Eppendorf model 5415D centrifuge at 13,200 rpm for 2 min. After washing each spin column with 700 µl of buffer PE, the residual buffer was removed by centrifugation using an empty holding microfuge tube. The DNA samples were then eluted into a fresh Eppendorf tube by addition of 100 µl of EB buffer (EB, Tris-HCl, pH 8.0) warmed to 60^{o}C, incubating for 5 min at room temperature and collecting the eluate by centrifugation at 13,200 rpm. The eluates were combined and the DNA recovered (typically 15-17 µg) was quantified by UV absorption at 260nm. To this were added 3 µg (142 pmol) of adaptor A in 50 µl 10 mM Tris-HCl, pH 7.5 and the mixtures were concentrated to a volume of 20 µl by centrifugation in a Millipore Microcon YM-3 concentrating devise at 12,200 rpm at 4^{o}C for 60 min.

Ligation to Adaptor A and Generation of CpG Tag Fragments

The DNA fragments recovered after adsorption to and elution from the silica matrix mixed with adaptor A were subjected to ligation by incubation for 16 hours at 16^{o}C in a final volume of 30 µl containing 50 mM Tris-HCl, pH 7.5, 10 mM $MgCl_2$, 1 mM ATP, 1 mM DTT (DTT, dithiothreitol), 5% w/v polyethyleneglycol 8,000, and 10 units of T4

DNA ligase (*Invitrogen #15224*). The reaction was terminated by heating at 60^{o}C for 30 min.

Digestion with EcoP15I and Isolation of Tagged CpGs

DNA fragments from the previous step with adaptor A at each of their ends (30 µl containing about 12µg DNA) were incubated for 16 hr at 37^{o}C in a final volume of 100 µl containing NEB buffer 3 (1 x NEB buffer 3, 100 mM NaCl, 50 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM DTT, pH 7.9), 0.1 mg/ml BSA (BSA, bovine serum albumin), 10 mM ATP, 10 µM sinefungin (also *adenosylornithine*; *Axxora #A-914550*) and 100 units EcoP15I (*New England Biolabs #R0646*). The reaction was terminated by addition of 2 µl proteinase K (20 µg/ml) and incubation at 55^{o}C for 1 hour followed by addition of 5x DNA loading buffer (1x DNA loading buffer, 0.05 w/v xylene-cyanol, 0.08 w/v bromophenol blue, 1 mM EDTA, 5% glycerol). One fourth of the reaction mixture was subjected to 10% polyacrylamide gel electrophoresis (PAGE) using a precast 10cm x10 cm x 1.0 mm gel slab made in TBE (TBE, 90 mM Tris-borate, 2 mM EDTA, pH 8.0, *Invitrogen #EC6275*) and electrophoresed at 200 V at room temperature until the bromophenol blue dye reached the bottom of the gel.

Fig. S1C depicts the PAGE analysis of sheared DNA mixed with ^{32}P-labeled adaptor A2 before and after incubation with DNA ligase and after incubation of the ligated mixture with EcoP15I. The CpG dinucleotides with their 3'-flanking 25-27 nt tags (tagged CpGs) attached at their 5' end to adaptor A were extracted from the polyacrylamide gel by cutting out the bands with the tagged CpGs and processing each area separately as follows: The polyacrylamide fragments were flash frozen on dry ice, placed into a 0.5-ml Eppendorf tube into which an approximately 1-mm wide hole had been made at the bottom. The tube was placed into a 1.5 ml Eppendorf holding tube and the gel was extruded through the hole by centrifugation at 13,000. The DNA fragments were then extracted by soaking with 300 µl H_2O at 37^{o}C for 60 min with occasional shaking. 200µl of the extracted DNA were removed from the mixture after centrifuging in an Eppendorf Model 5415D microfuge for 15 min at maximum speed. The DNA was precipitated by addition 4 µl glyco-blue (Ambion), 300 µl of 5M ammonium acetate, and 1.5 ml absolute ethanol followed by incubation at 4^{o}C over night, collected by centrifugation at 4^{o}C. The precipitate was washed once with 80% ethanol and resuspended in 15 µl of Tris-HCl 25 mM, pH 7.4 (37ºC for one hour). The molar concentration of the purified and concentrated tagged CpG fragments was determined by fluorescence in a Nanodrop ND3300 fluorospectrometer in the presence of PicoGreen reagent (*Invitrogen #P-11496*).

Ligation of CpG Tags to Adaptor B

The tagged CpG fragments with adaptor A at one end and a random two nucleotide 5' overhang at the other, were ligated to adaptor B (with a degenerate cohesive 5'overhang) by incubation in a final volume of 25 µl under the same conditions as used for ligation of adaptor A, with the exception that the concentration of adaptor B was 0.25 µM.

Amplification of CpGs Tags Flanked by Adaptors A and B (Fig. S1D, left panel)

Aliquots of the ligation mixture (1µl) were diluted 10 and 50 times with H_20 and subjected to amplification by PCR using 10, 12 and 14 cycles. The PCR reaction contained in 50 µl (added in the indicated sequence): 34 µl H_2O µl, 10 µl of 5 x HF reaction buffer (Invitrogen), 1 µl with 10 mM of each dNTP, 1 µl 10 µM primer A, 1 µl 10 µM primer B, 1.5 µl dimethyl sulfoxide, 1 µl diluted ligation mixture and 0.5 µl Phusion Hot Start High Fidelity DNA Polymerase (*New England Biolabs #F540*). The PCR

reaction sequences were: one incubation at 98°C for 30 sec, 10, 12 or 14 cycles of 10 sec at 98°C (denaturation), 30 sec at 55°C (annealing) and 60 sec at 72°C (elongation), followed by a final incubation at 72°C for 5 min. Conditions were chosen to preserve proportionality as a function of number of cycles and minimized appearance of high molecular weight artifacts and used to amplify the tagged CpGs flanked by adaptors A and B in 24 parallel amplification reactions, each yielding between 25 and 50 ng DNA.

Purification of the PCR-amplified Library of tagged CpGs

The DNA in the 24 PCR reactions was pooled. One-ml aliquots of the combined products were precipitated by addition of 10 volumes of 3M sodium acetate, 4 µl Dr. Gentle carrier reagent (Ambion), 1 µl glycoblue and 1 ml isopropanol. After 10-15 min at room temperature, the precipitated DNA was collected by centrifugation at 4°C at maximum speed of the Sorval "Fresco" Biofuge table top centrifuge for 10 min. The precipitate was washed with 500 µl 80% ethanol and resuspended in 15 µl 25 mM Tris-HCl, pH 7.5. The DNA fragments obtained from 24 PCR reactions (30 µl) were then applied to a 5-mm wide well of a 10% polyacrylamide gel slab of 1 mm thickness (*Invitrogen #EC6275*), electrophoresed as above and extracted as above. The extract was concentrated in a Microcon YM-3 device to a final volume of 50 µl, added to 500 µl Qiagen denaturing PN buffer and purified further by adsorption to the silica membrane of a Qiagen Spin column in PN buffer. After washing as above, the DNA was eluted with 50 µl of Qiagen EB buffer (Fig. S1D, right panel).

Quality Analysis of the Amplified Library

1. The concentration and spectrum of the library of tagged CpGs prepared as described above was determined using a Nanodrop 1100 spectro-photometer of 1 mm light path length (*Nanodrop, Inc*). Typically the DNA had a 260nm/280nm absorbance ratio >1.85 and a concentration >20ng/µl, corresponding to >300 nM (molecular weight of tagged CpGs with 25-27 nt of 3' flanking sequence (with respect to the CAGCAG EcoP15I binding site embedded in adaptor A) and flanking adaptors A and B (95 bp) = 62.7 KDa).

2. The integrity of amplified CpG tags was determined by cloning into plasmid pCR-Blunt II-TOPO, purchased covalently bound to topoisomerase from *Invitrogen(cat #2800*), and sequencing the inserts of plasmids recovered from >200 transformed colonies. Equal quantities of the amplified libraries obtained from genomic DNA digested with HpaII, HinP1I, HpyCH4IV and AciI (Fig. S1D) were combined and sequenced.

Massive parallel sequencing of the PCR-amplified CpG tag library

Aliquots of the libraries synthesized as described above were sequenced without further treatment by Illumina, Inc. using the Solexa technology of the Illumina Genome Analyzer - 1 gigabase sequencing service (*http://www.Illumina.com*) and a custom CpG tag-sequencing primer.

Bisulfite Sequencing Analysis

The methodology of Olek et al. (3) in which the DNA to be treated is embedded in agarose, was used as described. Alternatively, an in–column treatment of DNA with sodium bisulfite/hydroquinone was used to convert cytosine to uracil using reagents purchased in kit form from Zymo Research Corp, following the instructions supplied with the kit (EZ DNA Methylation-Gold cat # D5005-6). PCR primers were designed with the aid of the BiSearch Web Server's Primer Design and Search Tool at http://bisearch.enzim.hu/

Materials.

Oligonucleotides were purchased from Integrated DNA Technologies (IDT).

Adaptors and PCR primers:

Adaptor A:

5'-ACAG GTT CAG AGT TCT ACA GTC CGA **CAG CAG** C
CAA GTC TCA AGA TGT CAG GCT GTC GTC GGC -5'

Adaptor A2 (Fig S1C):

5′- CTA ATA CGA CTC ACT ATA GGG AGA **CAG CAG** C
A_{24}GAT TAT GCT GAG TGA TAT CCC TCT GTC GTC GGC -5′

Adaptor B:

5'-NN TCG TAT GCC GTC TTC TGC TTG
AGC ATA CGG CAG AAG ACG AAC -5'

Double stranded adaptors were prepared from single stranded Oligonucleotides by heating 5 µM of each strand in 10 mM Tris-HCl, pH 7.5, to 65^{o}C for 10 min followed by slow cooling to room temperature for 20 min and purification by electrophoresis through 10% polyacrylamide in TBE buffer. The annealed adaptors were extracted, concentrated and quantified as described for CpGs tags.

PCR Primer A:

5'-aatga tacgg cgacc accg ACAG GTT CAG AGT TCT ACA GTC CGA

PCR Primer B:

5'-CAA GCA GAA GAC GGC ATA CGA

CpG tag-sequencing primer:

5'-ccg ACAG GTT CAG AGT TCT ACA GTC CGA CAG CAG

Receiver Operating Characteristic (ROC) decision strategy applied to MSCC scores of CpGs whose methylation status was validated by bifulfite sequencing. The classification problem was divided in two parts: 1. Separation of hypomethylated sites (<75% methylation) from heavily methylated sites (>75% methylation). 2. Separationn of mostly unmethylated sites (<25% methylation) from the rest (> 25% methylation). A control data panel was created consisting of 358 sites, representing all 5 XCGX patterns, whose methylation states were measured by bisulfite sequencing. These CpGs mapped to 36 different genomic regions covering many different levels of methylation. Fig. 3A shows the trade-off between the number of hypomethylated CpGs that were correctly categorized (true positive rate) and the number of heavily methylated CpGs that were measured as hypomethylated (false positive rate) at different MSCC score cutoffs. True positive and false positive rates (TPR and FPR respectively) were calculated according the formulas: TPR=(nº CpG (<75% methylation & MSCC score> cutoff))/(nº CpG (<75% methylation)), and FPR=(nº CpG (>75% methylation & MSCC > cutoff))/(nº CpG (>75% methylation)).

Informatics

The mm9 Mus musculus reference genome was downloaded from the UC Santa Cruz Genome data base (4). To list for any given chromosomal interval, the number: of

CpGs, of 4Enz CpGs and of unique addressable CpGs we used a script in C (**intersection_tool.c**) that searches two tables and returns the physical location intersection (using chromosome, start and stop). The number of unique addressable CpGs were obtained using intersection-tool.c that looked up the number of addressable CpGs having at least one of its flanking 25mers (tags) being represented only once in the genome (unique flanking sequences).

The physical locations and sequences near target restriction sites in the all-CpG-all-hits frequency table were extracted from mm9 raw sequence files (by chromosome) using awk by searching individual files as single long strings awk '{printf("%s", $0 (NR==1 ? "" : ""))}' infile > outfile. Followed by for-loop-step restriction site substring lookup and corresponding sequence and position information extraction such as: awk '{ for (i = 1; i <=length($1); ++i) if (toupper (substr($1,i,4)) == "ACGT") print “ACGT”, i, substr($1,i+1,25); else if… }' infile > outfile.

Reads were aligned to the mm9 genome using MOM (1) and hit frequencies were integrated into a all-CpG-all-hit frequency table by intersection of physical location. The 5’ start location of hit reads were extracted from the MOM outfile in a strand specific manner and the physical locations were compiled using sort | uniq -c. The locations were then integrated using intersection_tool.c. Additionally, locations that did not return MOM hits were labeled as having 0 hits. The physical location of RepeatMasker, Refseq and CpG Island elements (4) were also integrated into the all-CpG-all-hit- frequency table by physical location from flat files using the intersection_tool.c script with a similar sort | uniq -c compilation workflow.

Tables with CpG islands (CGIs) listing the chromosomal intervals were extracted from the Mus musculus genome assembly following instructions outlined in Gardniner-Garden and Frommer (5, GG&F CGIs), in Takai and Jones (6, T&J CGIs), and in Hackenberg et al.(7, CpGClusters). The Java scripts used to extract GG&F and T&J CGIs are contained in **cpgilandfinder.zip**, The Java output was translated into a form that could be imported into SAS JMP software using the script **CreateJmp_Islands.pl**. CpGCluster CGIs were extracted using perl script **CpGcluster.pl** obtained from http://bioinfo2.ugr.es/CpGcluster/ (see also ref. 7). The chromosomal coordinates Epi-CGIs, a restricted subset of GG&F CGIs incorporating epigenetic data (8), were provided by Christoph Bock (Broad Institute of MIT and Harvard/Max Planck Institute for Informatics).

Gene regions: The information about genes and genomic regions in tables was derived from UCSC Genome Browser table refGene, for track RefSeq Genes, and table rmsk, for track RepeatMasker (4). For each gene we defined a 5Kb TSS region as starting 3Kb before the transcription start coordinate (txStart for + genes and txEnd for -) and continuing +2Kb into the gene. We compensated for UCSC’s zero-based counting for start coordinates. We also defined a 3Kb Three Prime Region (3'R) starting after the transcription end coordinates (txEnd for + genes and txStart for -). Sequence between a gene’s TSS and 3'R regions from another gene is considered integenic. TSS, exon, intron, 3’R, and Intergenic regions were added to the all-CpG-all-hit frequency table by creating tables of labeled begin and end records from the data contained in UCSC genome browser and sort/merging them with the existing all-CpG-all-hits frequency table using **Add_Regions_tab.pl**. When a read, an experimentally defined cluster of CpGs, or an UMR overlapped more than one genomic region, they were assigned according to the hierarchy TSS > 3'R >exon>intron> intergenic

Distribution of CpGs among genomic intervals such as repetitive elements, gene regions, CGIs and both the number of reads that identified each CpG and the number of channels in which each CpG was identified were extracted with Perl script **Get_channels_count.pl**. For each position, the Perl script Get_channels_count.pl

would read the records for the forward and reverse tags and sum the total number of reads. The number of channels that had reads was determined separately for both forward and reverse tags. The greater of these channel counts was reported for the position along with the total reads. For each CGI or gene region a total was kept of the number of positions that had reads in no channels, less than four channels, four channels, and five channels. Counts were also kept for CpGs that were not in any defined CGI.

Clusters of hypomethylated CpGs were created with perl script **four_enz_cluster.pl** This script is a modification of perl script CpGcluster.pl, by J.L. Oliver & M. Hackenberg obtained from http://bioinfo2.ugr.es/CpGcluster/ that allows the program to read positions from a file and accept a threshold value for the distance between CpGs that is necessary for a CpG to be included in the cluster. The input file for four_enz_cluster.pl is a table with headers Chrom, Location, Cut and Count, listing all CpGs identified in 5 channels, enzyme recognition site, and the number of identifying reads. The program generates two out files. One out file lists each CpG classified according to its position with respect to a cluster: start, end, middle, or lone (the CpG is located outside of any cluster, meaning that the nearest neighbors are at distances bigger that the maximum allowed for it to be part of a cluster). The second out file contains the different clusters listed in rows; each row defining chromosome number, and the beginning and ending coordinates of a cluster.

To calculate the enrichments of ORegAnno and MMC elements, in UMRs or subsets of UMRs, respect to the genome we used the ratio between (number of ORegAnno elements overlapping UMRs / Total number of ORegAnno elements) and (total size of UMRs / total size of genome). We used simulation to measure the significance of the fold enrichments. We developed a Perl script to simulate the distribution of the fold enrichment of randomly generated genomic regions. The program runs 100,000 iterations for each set of UMRs. During iteration it randomly generates a set of genomic regions with the same number of UMRs and the same size of each UMR. This simulation allows us to assign p value to each set of UMRs according to the distribution of the fold enrichments from 100,000 sets of randomly generated genomic regions.

References to Methods

1. Eaves HL, Gao Y (2009) MOM: maximum oligonucleotide mapping. *Bioinformatics* 25:969-970.
2. Widnell CC, Tata JR (1964) A procedure for the isolation of enzymically active rat-liver nuclei. *Biochem J* 92:313-317.
3. Olek A., Oswald J, Walter J (1996) A modified and improved method for bisulphite based cytosine methylation analysis. *Nucleic Acids Res*. 24:5064-5066.
4. Fujita PA, *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* 39:D876-882.
5. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. *J Mol Biol* 196:261-282.
6. Takai D, Jones PA (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A* 99:3740-3745.
7. Hackenberg M, *et al.* (2006) CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics* 7:446.
8. Bock C, Walter J, Paulsen M, Lengauer T (2007) CpG island mapping by epigenome prediction. *PLoS Comput Biol* 3:e110.

9. Liu J, *et al.* (2000) Identification of a methylation imprint mark within the mouse Gnas locus. *Mol Cell Biol* 20:5808-5817.

Supplementary Tables

Table S1. A- CpG composition of the mouse and human genomes. B- Distribution of CpGs among methylation sensitive restriction enzymes sites (HpaII, HinP1I, HpyCH4IV and AciI).

A. Abundance of Cs and CpGs

	Mouse (mm9)	Found	Human (hg18)	Found	Expected
Size (bp)	2,725,765,481	100%	3,096,521,113	100%	100%
Cs	564,289,400	20%	587,700,260	19%	25%
CpGs	21,722,925	0.8%*	28,341,026	0.9%*	6.25%

*, 12.8% of expected if A+T = C+G.

B. Restriction sites in genomes

Enzyme		Mouse Genome (mm9)		Human Genome (hg18)	
HpaII	CCGG	1,620,442		2,306,167	
HinP1I	GCGC	1,114,955		1,662,835	
HpyCH4IV	ACGT	1,789,709		2,158,019	
AciI	CCGC/GCGG	2,535,019		4,126,405	
4 Enzymes		7,060,125	32.5%*	10,253,426	36.4%*

*, refers to the number of CpGs shown in **A.**

Table S2. General distribution of CpGs in mouse (mm9) genomic DNA and Its addressability by 4 methylation sensitive restriction enzymes*.

A. ***Repetitive DNA (Repeat Masker)***

	Size in Genome		CpGs in Region		CpG Density**	Addressable CpGs	Survey Pot.***
	bp	%^	number	%^^	bp	number	%
Repetitive	1,119,921,152	41.0	8,961,361	41.2	125	2,969,829	48.6
Non-Repetitive	1,534,990,365	56.3	12,341,418	56.8	124	3,985,280	47.2
Random	70,853,964	2.60	380,178	1.75	--	105,016	47.7
Total	2,725,765,481	--	21,722,957	--	125	7,060,125	32.5

B. Genomic Regions

Genes (RefSeq)	Size in Genome		CpGs in Region		CpG Density**	Addressable CpGs	Survey Pot.***
	bp	%^	number	%^^	bp	number	%
TSS (-3kb-+2kb)#	93,465,135	3.4	1,926,965	8.9	49	859,655	44.6
Exons	42,983,899	1.6	829,613	3.8	52	310,601	37.4
Introns	809,496,067	29.7	6,501,976	30.0	125	2,142,463	33.0
3'R (3kb post)	48,870,486	6.9	487,656	2.2	100	162,612	33.3
Total	957,651,129	36.2	9,745,923	44.9	100	3,324,468	31.4
Intergenic	1,697,243,926	62.3	11,596,547	53.4	146	3,630,725	31.3
Unmapped	70,853,964	2.26	380,168	1.8	--	104,951	27.6
CpG Islands							
GG&F	59,730,409	2.19	3,249,501	5.0	18	1,534,895	48.6
T&J	21,531,848	0.79	1,515,858	7.0	14	844,408	47.2
CpGCluster	37,894,803	1.49	2,782,597	12.8	14	1,327,481	47.7

*****, HpaII (CCGG), HinP1I (GCGC), HpyCH4IV (ACGT), AciI (CCGC/GCGG)
******, average interCpG distance; ***, survey potential: addressables in region
^, 100%=2,725,765,481; ^^,100%=21,722,957; ^^^,100%=7,060,125
#, from UCSC RefSeq track (total 26,431); if two -3kb to + 2kb regions overlaped, they were fused and treated as one

Table S3. Distribution of Addressable CpGs in Mouse Repetitive DNA*

Class of DNA Repeat (from RepeatMasker)#	Enzyme used to probe methylation status of CpGs									
	HpaII (CCGG)		HinP1I (GCGC)		HpyCH4IV (ACGT)		AciI (CCGC/GCGG)		4Enz	
	Total CpGs	%Unique	Total CpGs	%Unique	Total CpGs	%Unique	Total CpGs	%Unique	Total CpGs	%Unique
LINE	327,231	38.4	92,207	59.7	234,312	74.8	312,662	56.1	966,412	55.0
SINE	175,124	93.8	143,082	95.1	177,749	97.2	227,363	80.2	723,318	90.6
LTR	191,423	62.3	137,658	50.5	203,183	72.2	311,518	64.0	843,782	63.4
Low complexity (LC)	20,236	97.6	22,237	98.1	10,052	97.3	59,640	97.6	112,165	97.7
Simple repeats (SR)	18,287	94.6	82,858	92.0	57,370	95.0	66,777	95.5	225,292	94.0
DNA	11,383	99.0	8,008	98.9	26,082	98.8	23,002	98.5	68,475	98.7
Other	6,779	91.3	4,819	81.6	8,287	80.4	10,500	91.6	30,385	86.9
RNA	84	94.0	75	82.7	104	100.0	131	93.9	394	93.4
rRNA	450	72.9	425	62.8	189	86.2	739	68.9	1,803	70.3
scRNA	669	89.1	888	71.6	443	100.0	596	94.1	2,596	86.1
snRNA	129	77.5	383	54.6	256	93.4	267	85.4	1,035	75.0
srpRNA	45	80.0	41	82.9	28	100.0	62	90.3	176	87.5
tRNA	620	81.0	475	82.5	474	96.0	637	81.0	2,206	84.5
RC Transp.	22	100.0	24	100.0	96	100.0	59	100.0	201	100.0
Satellite	603	83.3	438	70.8	1,349	43.8	891	86.2	3,281	66.2
Other	3,742	98.1	1,841	97.1	3,760	96.6	6,369	97.9	15,712	97.6
Unknown	415	84.8	229	93.4	1,588	58.5	749	76.0	2,981	69.2
Addressable in Repeats	750,463	61.8	490,869	75.5	717,035	82.5	1,011,462	70.3	2,969,829	72.0
Addressable in Non-Repeats	843,674	97.5	611,717	97.2	49,441	52.1	1,490,584	97.6	3,985,280	97.5
Random DNA									104,951	
*, CpGs flanked by at least one 25-nt tag that maps to a unique site of the genome							Total Addressable CpGs		7,060,125	
							Addressable CpGs reporting on methylation status of DNA repeats		42.1 %	

#, URL: http://repeatmasker.org

Table S4. Partitioning of 25-nt CpG-tag sequences (reads) returned by Illumina Genome Analyzer (Solexa technology).

	Experiment					
	Slxa1 (4 channels)	%	Slxa2 (9 channels)	%	Slxa3 (5 channels)	%
MSREs used*	4Enz		HpaII & BstUI		4Enz	
Total number of reads examined	39,332,981	100	6,560,021^		50,501,350	100
channel A			1,370,155^		9,855,482	
channel B			1,283,578^		9,935,076	
channel C			1,303,148^		9,949,433	
channel D			1,276,474^		9,922,838	
channel E			1,326,666^		10,838,521	
After curation & and trimming to 25nt	39,119,608	99	36,457,931	100	46,772,599	92.6
Did not map (i.e. >1 mismatch)	3,395,798	8.8	4,179,281	11.4	11,986,180	23.7
Mapped to more than one site	3,711,435	9.5	2,571,580	7.1	4,387,951	8.6
Mapped to unique chromosomal sites**	32,031,094	81.4	29,707,270	81.5	30,398,468	60.2

*,methyl-sentitive restriction enzymes, 4Enz=HpaII+HinP1I+HpyCH4IV+AciI; **, up to one mismatch;
^,HpaII only

Table S5. Distribution Identifiable CpGs* in Mouse Genomic DNA as Seen by Read Recovery

A. CpGs in Genome

	In Non-Repetitive DNA	In Repetitive DNA	In Random DNA	Total
All Addressable	3,985,280	2,969,829	105,016	7,060,125
With Unique Tags	3,887,085	2,137,147	na	6,024,232
With Non-Unique Tags	98,195	832,682	na	930,877
% Unique	98	72	na	85

B. CpGs in Non-Repetitive DNA

	Number	%	Reads	Ratio
Fully Methylated - not found	1,697,877	43.7	-	0
Found in 1-3 channels	1,290,445	33.2	3,441,746	2.7
Found in - 4 channels	280,804	7.2	2,848,883	10.1
Found in - 5 channels	617,959	15.9	17,787,095	28.9
	3,887,085		24,164,724	

C. CpGs in Repetitive DNA

	Number	%	Reads	Ratio
Fully Methylated - not found	1,269,804	60.0	-	0
Found in 1-3 channels	636,073	30.1	1,479,028	2.3
Found in - 4 channels	88,603	4.2	802,618	9.1
Found in - 5 channels	120,448	5.7	3,001,981	24.9
	2,114,928		5,283,627	

D. All CpGs identified

	Number	%	Reads	Ratio
Repetitive & Non-Repetitive DNA	3,034,532	50.4	29,448,351	9.7

*, CpGs recognized by HpaII, HinP1I, HpyCH4IV or AciI and flanked by at least one 25-nt tag that maps to a unique genomic site

Table S6. all-CpG-all-hit frequency table (available as tab delimited file)

Key to headers of the All CpGs all Hits tables.

A CRHM chromosome

B LOCATION position of the second base (always a C) of the 26-mer CpG tag sequence

C SENSE indicates that the CpG Tag is located upstream(rev) or downstream(forward) from the coordinate mentioned in column B (location)

D Uniqueness indicates the number of times that the sequence of 26 bp (26-mer CpG tag sequence) is found in all these CpG all Hits tables

E CUT the recognition sequences of the 4 enzymes used in this study

F to J Channel # the number of reads recovered in each of the five channels used by Illumina to produce the complete set of data

K to M REP_* "name, class and family of repetitive region containing the particular CpG Tags"

O to S 1 = is IN named gene region; 0 = is NOT IN named gene region.

The information about genes and genomic regions in these tables is derived from UCSC Genome Browser table refGene, for track RefSeq Genes, and table rmsk, for track RepeatMasker. For each gene we defined a 5Kb TSS region as starting 3Kb before the transcription start coordinate (txStart for + genes and txEnd for -) and continuing +2Kb into the gene. We carefully compensated for UCSC's zero-based counting for start coordinates. We also defined a 3Kb Three Prime Region (3'R) starting just after the transcription end coordinate (txEnd for + genes and txStart for -). Sequence between a gene's TSS and 3'R regions from another gene is considered INTER Genomic, denoted in column S as INTERGEN.

When a tags coordinate could be considered to be in more than one region, the region was assigned according to the hierarchy TSS > 3'R > GEN > INT.

Columns O to S better define the gene region of each coordinate. The number 1 appears in a column if the coordinate can be considered to be in its defined region at any time according to the RefSeq Genes table. Coordinates can be in more than one region. This is most often true because the TSS region extending into the gene overlaps exons and introns In other analyses, when assigning tag coordinates in the GEN region to exons or introns, exon always had priority over intron if there were an overlap.

Table S7. Genomic prameters of CpG Islands (CGIs) as predicted by different algorithms

1. CGI - Gardiner-Garden & Frommer	Number		Size (bp)		CpGs Total		CpG Density*	CpGs Addressable**	
Genomewide	177,704		59,730,409		3,249,501		18.3	1,534,895	
In Genes									
TSS (-3kb-+2kb)#	26,623	15%	15,468,806	26%	1,138,425	35%	13.6	632,545	41%
Exons & Introns	61,057		16,837,000		781,823		21.5	313,083	
3'R (3kb)	4,385		1,317,052		66,854		19.7	29,211	
Total	92,065	52%	33,622,858	52%	1,987,102	61%	17.1	974,839	64%
Intergenic	85,638		26,107,551		1,262,399		20.6	560,056	

2. CGI - Takai & Jones	Number		Size (bp)		CpGs Total		CpG Density*	CpGs Addressable**	
Genomewide	21 ,246		21,531,848		1,515,858		14.2	844,408	
In Genes									
TSS (-3kb-+2kb)	11,803	55%	13,876,459	64%	1,017,136	67%	13.6	577,441	68%
Exons & Introns	2,904		2,389,952		154,350		15.4	80,725	
3'R (3kb)	415		378,171		25,322		1.49	13,466	
Total	15,122	71%	16,644,582	77%	1,196,818	79%	13.9	671,632	79%
Intergenic	6,125		4,908,266		319,050		15.4	172,776	

3. CpGClusters (p<0.00001)	Number		Size (bp)		CpGs Total		CpG Density*	CpGs Addressable**	
Genomewide	120,156		37,894,803		2,782,597		13.6	1,327,481	
In Genes									
TSS (-3kb-+2kb)	23,553	20%	12,640,956	33%	1,099,781	40%	11.5	605,959	46%
Exons & Introns	35,629		8,547,887		568,019		15.0	227,629	
3'R (3kb)	2,862		762,323		53,807		14.1	23,558	
Total	62,045	51%	21951,166	58%	1,721,607	62%	12.7	857,146	65%
Intergenic	58,114		15,943,637		1,060,990		15.1	470,335	

4. Epi-CGIs	Number		Size (bp)		CpGs Total		CpG Density*	CpGs Addressable**	
Genomewide	27,458		56,471,409		2,337,737		24.1	1,172,185	
In Genes									
TSS (-3kb-+2kb)	11,070	40%	30,672,661	54%	1,279,978	55%	23.9	663,239	57%
Exons & Introns	4,350		7,750,319		318,974		24.3	154,809	
3'R (3kb)	702		1,738,098		74,069		23.5	36,484	
Total	16,122	59%	40,161,078	71%	1,673,021	72%	24.0	854,532	73%
Intergenic	11,336		16,310,331		664,716		24.5	317,653	

*, average interCpG distance in bp; **, HpaII (CCGG), HinP1I (GCGC), HpyCH4IV (ACGT), AciI (CCGC/GCGG); #, from UCSC RefSeq track (total 26,431); if two -3kb to + 2kb regions overlaped, they were fused and treated as one.

Table S8. Methylation status of different sets of CGIs

CGI	AvgMSCC score	GG&F Number	%	T&J Number	%	CpGClusters Number	%	Epi-CGIs Number	%
In Genome*		177,704		21 ,246		120,156		27,458	
Analyzed**		92,997	100	16,731	100	67,448	100	16,592	100
Heavily Methylated	<11	60,768	65	1,738	10	39,022	58	2,549	15
Un-methylated	>17	23,318	25	12,460	74	21,386	32	11,191	67

*, from Table S7; **, CGIs with at least 33% of their CpGs addressable by 4Enz

Table S9. Hypomethylated CpG Clusters and UMRs (available as tab delimited file)

Table S10. Distribution of identified sites in hypomethyated CpG clusters.

	N^{o} of CpGs	N^{o} of reads	avg MSCC score
Total identified CpGs	3,034,332	29,448,351	3.3
Total 5-channel CpGs	738,407	20,876,076	28.3
Total clustered CpGs	847,357	19,226,625	22.7
Clustered 5-channel CpGs	559,901	17,246,090	30.8
Clustered 1- to 4-channel CpGs	287,456	1,980,535	6.9
Stand alone* 5-channel CpGs	178,506	3,629,986	20.3
Stand alone 1- to 4-channel CpGs	2,008,469	6,591,740	3.3

*, CpGs that were not included in any of the 64,266 clusters are called stand alones.

Table S11. Genomic Parameters of 4Enz 5-ch w300 UMRs with avg MSCC score ≧17

		Completely Included in					
	UMRs	TSSs	Exons	Introns	3'R	Intergenic	Other*
Number of UMRs	46,804	15,390	931	10,693	1,394	15,646	2,750
Addressable CpGs	801,692	550,529	5,459	47,125	9,515	112,303	76,761
Identified CpGs (HIts)	708,256	469,246	5,198	44,714	8,976	103,637	69,661
Coverage (%)	88	86	95.	95	91.	92	91
Reads	17,276,422	10,894,995	132,088	1,226,065	226,124	3,166,348	1,630,802
avgMSCC Score	24.4	22.9	25.4	27.4	25.2	30.6	23.4
Total CpGs	1,372,296	943,969	9,315	75,803	16,412	189,975	136,822
Survey Potential (%)	58	58	59	62	56	59	56
Average UMR (bp)	490	859	197	215	285	286	857
Total Size (bp)	22,927,212	13,225,454	182,950	2,293,760	396,949	4,471,664	2,356,435
CpG Density (bp)**	16.7	14.0	19.6	30.3	24.2	23.5	17.2

*, UMRs that partially overlap TSSs, Intros and Exons; **, average distance between CpGs

Table S12. Enrichment of UMRs in ORegAnno & Mammalian Most Conserved (MMC) Elements

			Avg	ORegAnno Elements			MMC Elements		
UMRs	Number	Size (bp)	Size (bp)	Number	Enrich-ment**	p-pvalue***	Number	Enrich-ment**	p-value
All	46,804	22,927,212	490	4,325	29.1	<0.00011 (9K)	98,658	5.7	NA
Non-CGI	24,399	4,417,330	197	2,427	84.8	<0.00006 (18K)	14,353	4.3	NA
at TSSs	2,979	579,653	195	304	81.0	<0.00001 (100K)	2,423	5.5	<0.0008 (1.4K)
in Gene body	10,032	1,725,003	172	1,134	101.5	<0.00002 (44K)	5,063	3.9	<0.0025 (0.4K)
in Intergenic	10,433	1,842,653	177	858	71.9	<0.00008 (13K)	5,291	3.7	NA
CGI-like	22,405	18,509,882	759	1,898	15.8	<0.0009 (11K)	84,305	6.0	NA
at TSSs	12,411	12,645,801	1,019	1,053	12.9	<0.00006 (18K)	59,972	6.2	<0.005 (0.2K)
in Gene body	2,986	1,148,656	385	289	38.8	<0.00002 (83K)	4,978	5.7	<0.0025 (0.4K)
in Intergenic	5,213	2,629,011	504	405	23.8	<0.00008 (13K	9,483	4.7	NA
Genome (mm9)		2,620,346,127*		16,976			1,990,870		

TSS, -3kb to +2kb of transcription start site; Gene body, exons plus introns plus 3kb of 3' not transcribed.
*, corrected for 105,419,354 Ns; **, see supplemenatry Methods; ***, number in parenthesis is the number of iterations in simulations; NA, not avalibale due to limitations in cpu time

Supplementary Figures

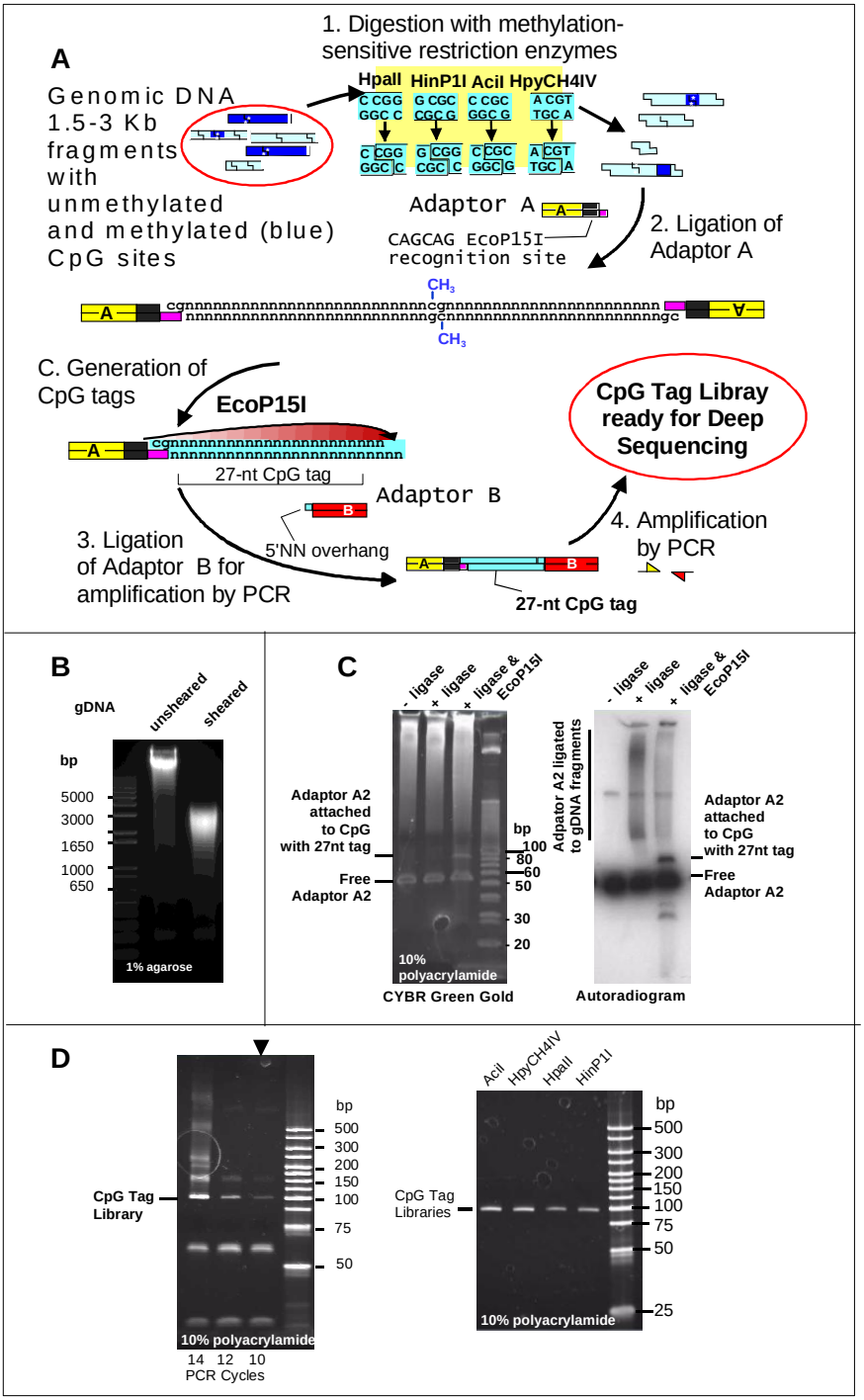


Figure S1. **CpG-tag library preparation. (A)** Strategy for preparation of CpG-tag libraries using four methylation-sensitive restriction enzymes which expose un-methylated CpGs and for retrieval of CpG tags using the type III restriction enzyme

EcoP15I. **(B).** Mouse liver genomic DNA size before and after hydroshearing (Harvard Apparatus), 0.8% agarose gel. **(C).** Ligation of ^{32}P-labeled adaptor A2 to sheared and HpaII-digested genomic DNA: polyacrylamide gel electrophoretic analyses of the samples before ligation, after over-night ligation, and after digestion with EcoP15I-*Left*, visualization with SYBR Green Gold (Invitrogen); *right*, autoradiogram. **(D).** *Left,* PCR amplification of adpt-A-CpG tag-adpt-B template (botton Fig 1A). *Right,* PAGE analysis of four libraries prepared from genomic mouse liver digested with the indicated methyl-sensitive restriction enzymes using a 12-step PCR amplification.

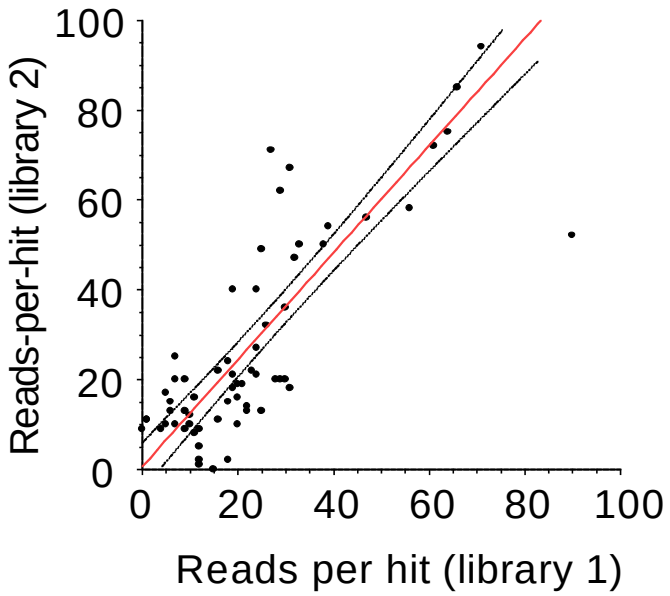


Figure S2 Comparison of read recovery in two experiments shows sequence dependent bias. Two independent mouse CpG-tag libraries to which lambda genomic DNA had been added as internal standard were prepared and sequence. The scatter plot compares the number of reads recovered from HinP1I GCGC sites located at the same position of the lambda genome in the two libraries and shows that the MSCC scores for CpGs at certain positions are affected by systematic bias which introduces variation in the final counts in a methylation-independent manner.

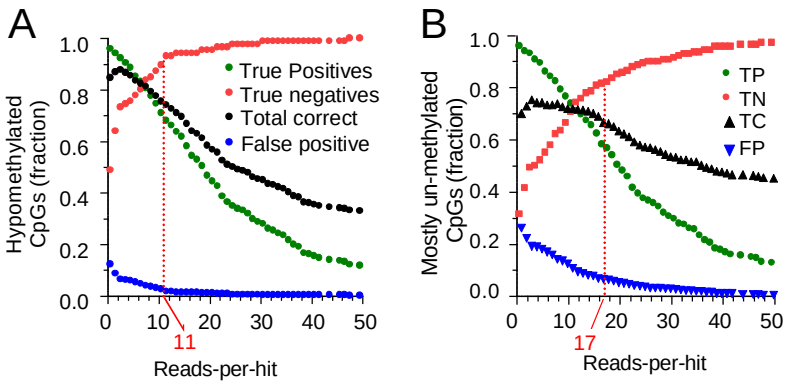


Figure S3. The optimal cutoff values, that were used in the assignment of methylation states in Fig 3C, were selected according to the relation benefits (fraction of correctly categorized) and cost (false positives and low sensitivity). **A.** We found 11 as the optimal cutoff to classify a CpG as hypomethylated (methylation rate < 75%). Once the value of 11 is passed, the fraction of heavily methylated sites properly found remains almost invariant (red curve), but the sensitivity with which we detect hypomethylated sites continues to decline (green curve). Moreover, beyond this value, the fraction of false positives does not decrease significantly (blue curve). Fig. 3C shows that at this read-per-hit cutoff we could correctly classify 75% of the CpGs analyzed, with a false discovered rate of 2.5% (FDR). The high rate of false negatives (22%) might be caused by the lack of coverage of the libraries used for this analysis (16% according to the number of lambda 1-3 channel hits). **B.** A similar analysis was done to evaluate the capacity of the method to detect bonafide unmethylated sites (< 25% methylation). It can be seen that a cutoff of higher than 25 reads-per-hit yields a good classification (low false positives) based on strong evidence (high number of reads-per-hit), but gives the method poor sensitivity. Moreover, as soon as a reads-per-hit value of 17 is passed, the effect of loss of sensitivity from increasing the threshold value becomes more evident. When a 17 reads-per-hit ratio is used as cutoff value we correctly classify 67% of CpGs as mostly un-methylated with a FDR of 7.2% (Fig. 3C).

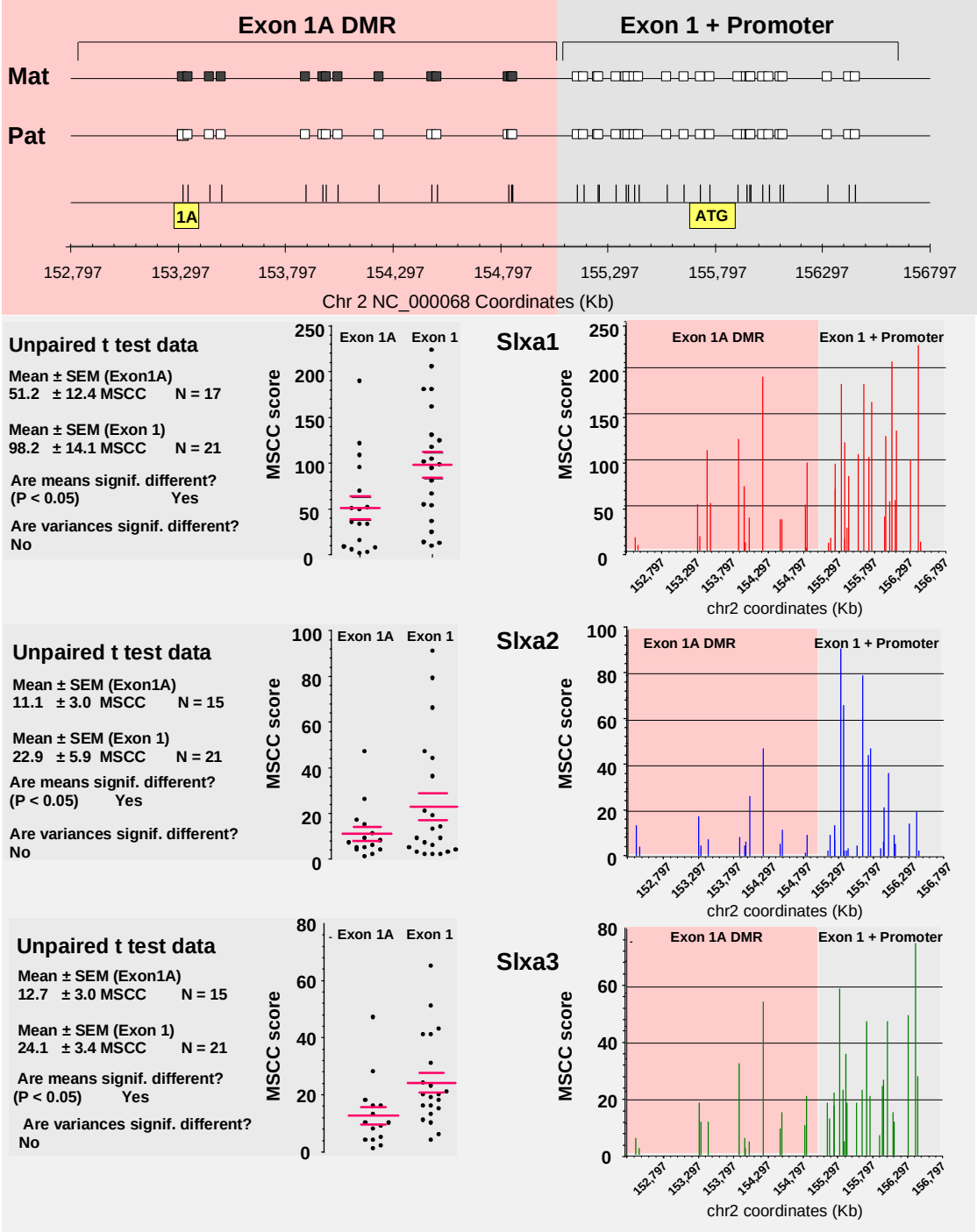
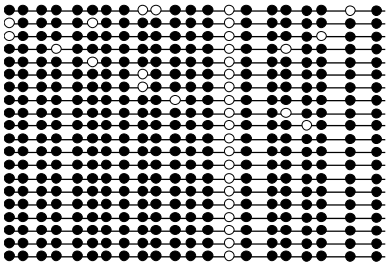


Figure S4

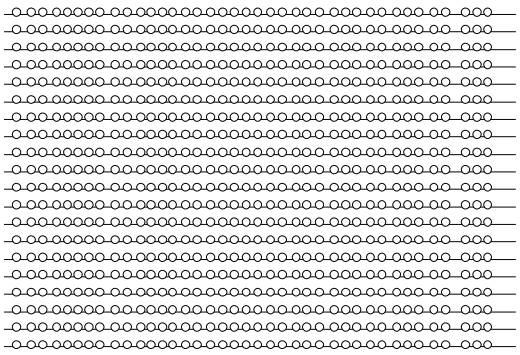
Proportionality between between avgMSCC score and methylation rate. Based on the results obtained with the unmethylated internal standard DNA sample (*lambda* genome) an avgMSCC of 20 or larger is expected for unmethylated regions (Fig. 1A). This result was confirmed by the distribution of avgMSCC scores between CGIs (Fig. 4A), showing that unmethylated CGIs scored 21.5 ± 4.6 reads-per-hit (mean ± SD). The proportionality between avgMSCC score and methylation rate was compared using two genomic regions with a one fold difference in their methylation status. The

methylation pattern of the Gnas Complex Locus on distal mouse chromosome 2 has been extensively studied, both in different organisms and in different tissues. The sequence containing Exons 1A and 1, has the peculiarity of having a differentially methylated region (50%) immediately followed by a region completely unmethylated (9). This feature makes the locus ideal to test whether or not this difference in methylation can be captured in our CpG sequence tag libraries. We counted the number of reads for each HpaII CCGG site located within the previously defined boundaries of the regions corresponding to Exon 1A and 1 of the Gnas Complex Locus. The average MSCC was calculated for each region, and the means and variances were compared using an unpaired t-test and the Levene's test respectively. Fig. S4 shows the result obtained in three independent HpaII CpG tag libraries made from three mice. We concluded that, within each experiment the mean MSCC score at Exon1A is half the mean value of Exon1 ($P < 0.05$). The set labeled "Slxa3" is derived from the data extensively analyzed in this manuscript. The values obtained for this set (24.2 ± 3.4 for Exon 1 and 12.7 ± 3.1 for Exon 1A) show a remarkable correspondence with what is expected for regions with 50% and 0% methylation (Fig. 4A). Given the reproducibility in the triplicates we conclude that averaging the MSCC score of individual CpGs produces an accurate estimation of the average level of methylation for the sites located at the same regions (e.g. CpG islands).

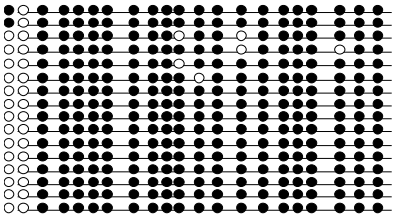
T&J CpG Island at
chr8:3,678,745 - 3,679,621
CpG addresed by 4-Enz: 40%
avg MSCC: 1
Bisulfite analysis:
3,679,140 - 3,679,460



T&J CpG Island at
chr6:128,387,615-128,389,191
CpG addresed by 4-Enz**: 50%**
avg MSCC**: 20.1**
Bisulfite analysis:
128,388,049-128,388,397



CpGCluster at
chr1:37,932,177 - 37,932,772
CpG addresed by 4-Enz**: 40%**
avg MSCC**: 0**
Bisulfite analysis:
37,932,177 - 37,932,526



CpGCluster at
chr11:102,177,975 - 102,179,925
CpG addresed by 4-Enz**: 74%**
avg MSCC**: 19.5**
Bisulfite analysis:
102,179,347 - 102,179,538

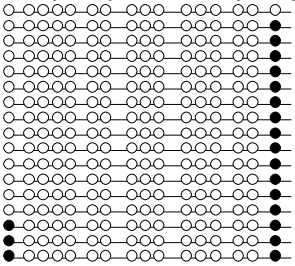


Figure S5. Bisulfite sequencing analysis of CGIs: Confirmation of methylated and unmethylated states of CGIs representing each of the peaks of the bimodal avgMSCC score distributions shown in Fig. 4.

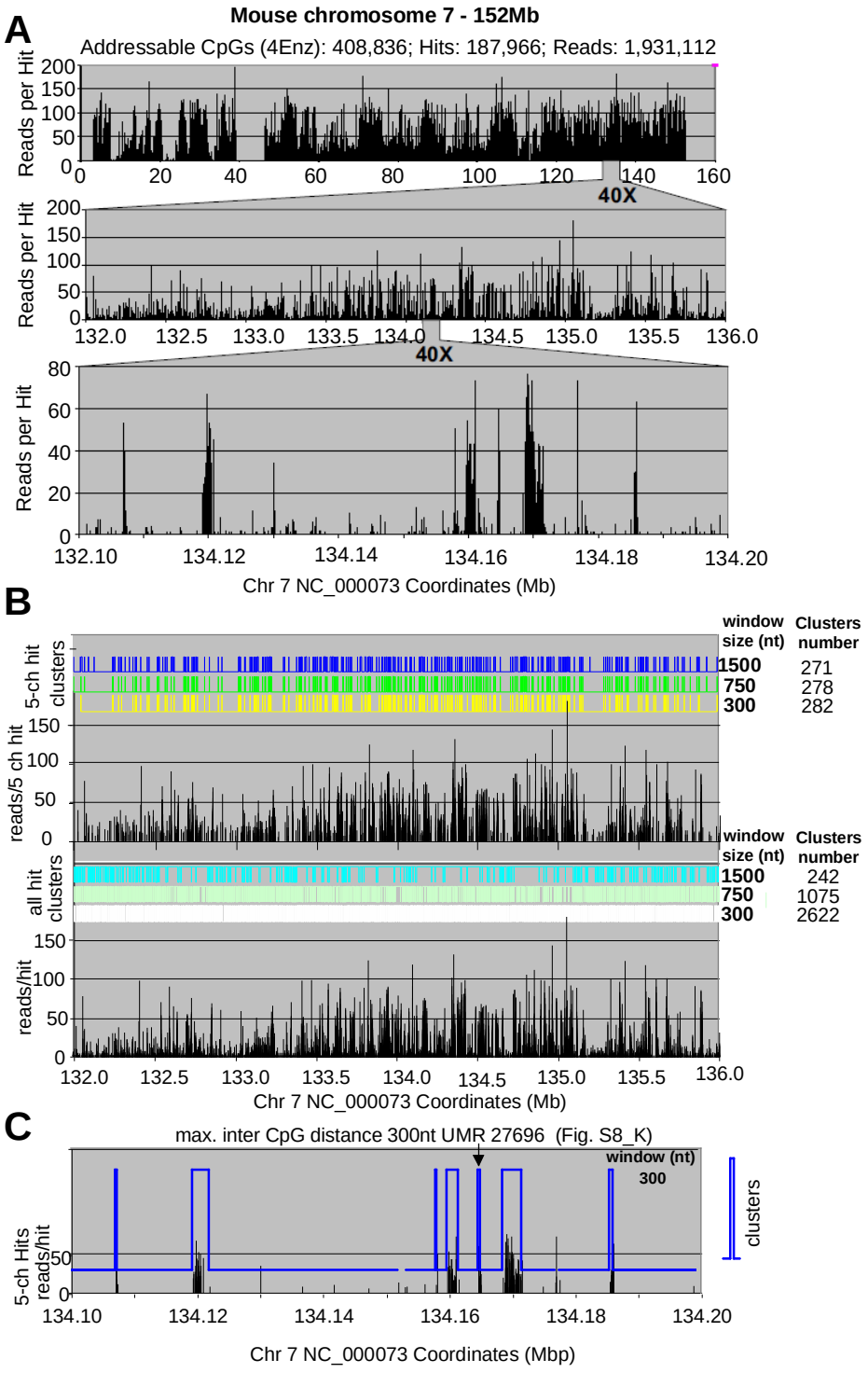


Figure S6. Genome wide map of hypomethylated CpG clusters. A. Mouse chromosome 7 CpGs identified by the indicated read/hit ratios are shown at three levels of resolution. **B.** Two approaches to finding clusters of hypomethylated CpGs. *Bottom panel*, all identified CpGs were used to produce clusters. Three different maximum inter-neighbor distances were considered to produce the clusters. The resulting number of clusters, as shown in the corresponding tracks, differs widely with the size of the sliding window used. *Top panel*, CpGs identified in 5 of 5 channels are shown. Most of these

CpGs are hypomethylated (Fig. 1C). When the map of hypomethylation is built using these CpGs, the number and composition of the clusters is rather insensitive to the selected inter CpG distance. **C.** Distribution of unmethylated and methylated CpGs is shown for the indicated chromosomal interval. *Blue lines*, the corresponding w300 clusters obtained by grouping 5-channel hits located at inter-CpG distances equal or less than 300 bp. *Arrow*, cluster 27,696 (Table S9) with avgMSCC score of 17.9 for which the bisulfite sequencing analysis shows an unmethylation rate of 82.5%, conferring to it the assignment of “unmethylated region” or UMR (Fig. S8-K).

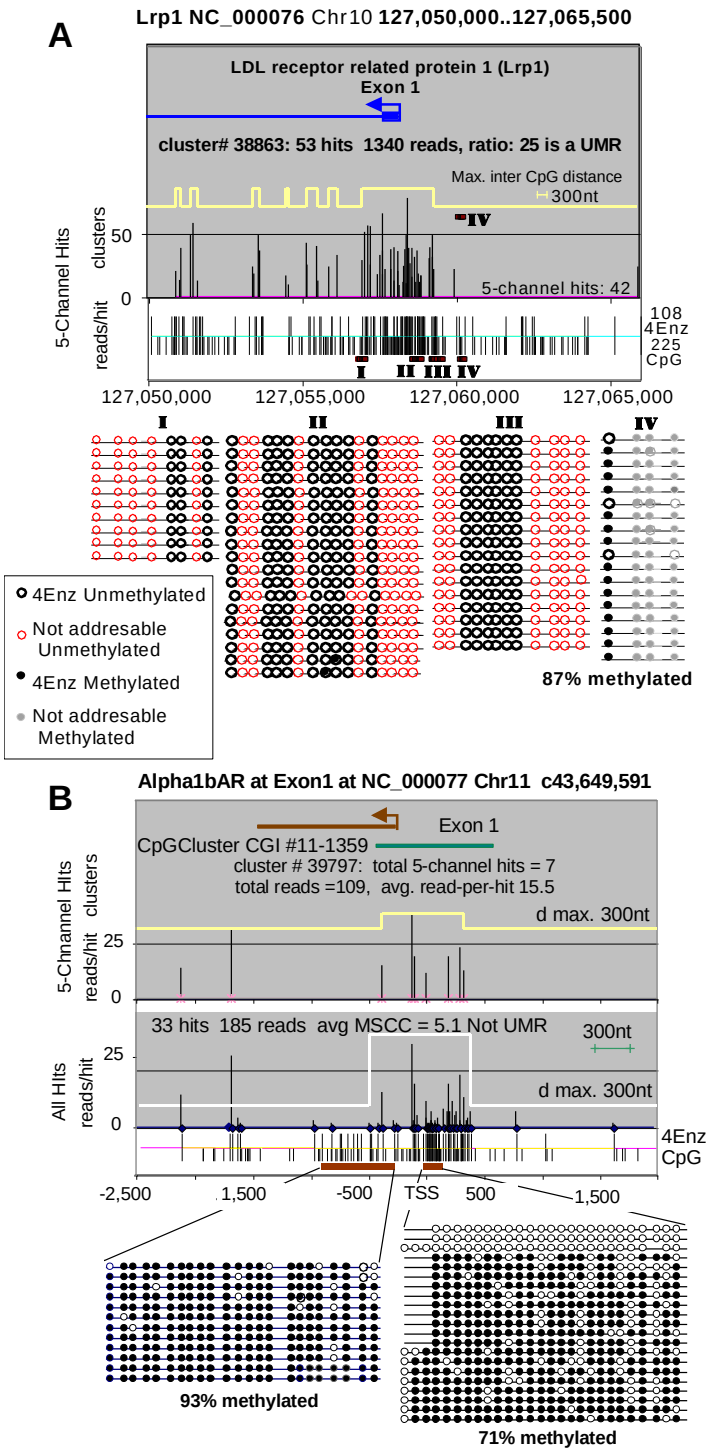
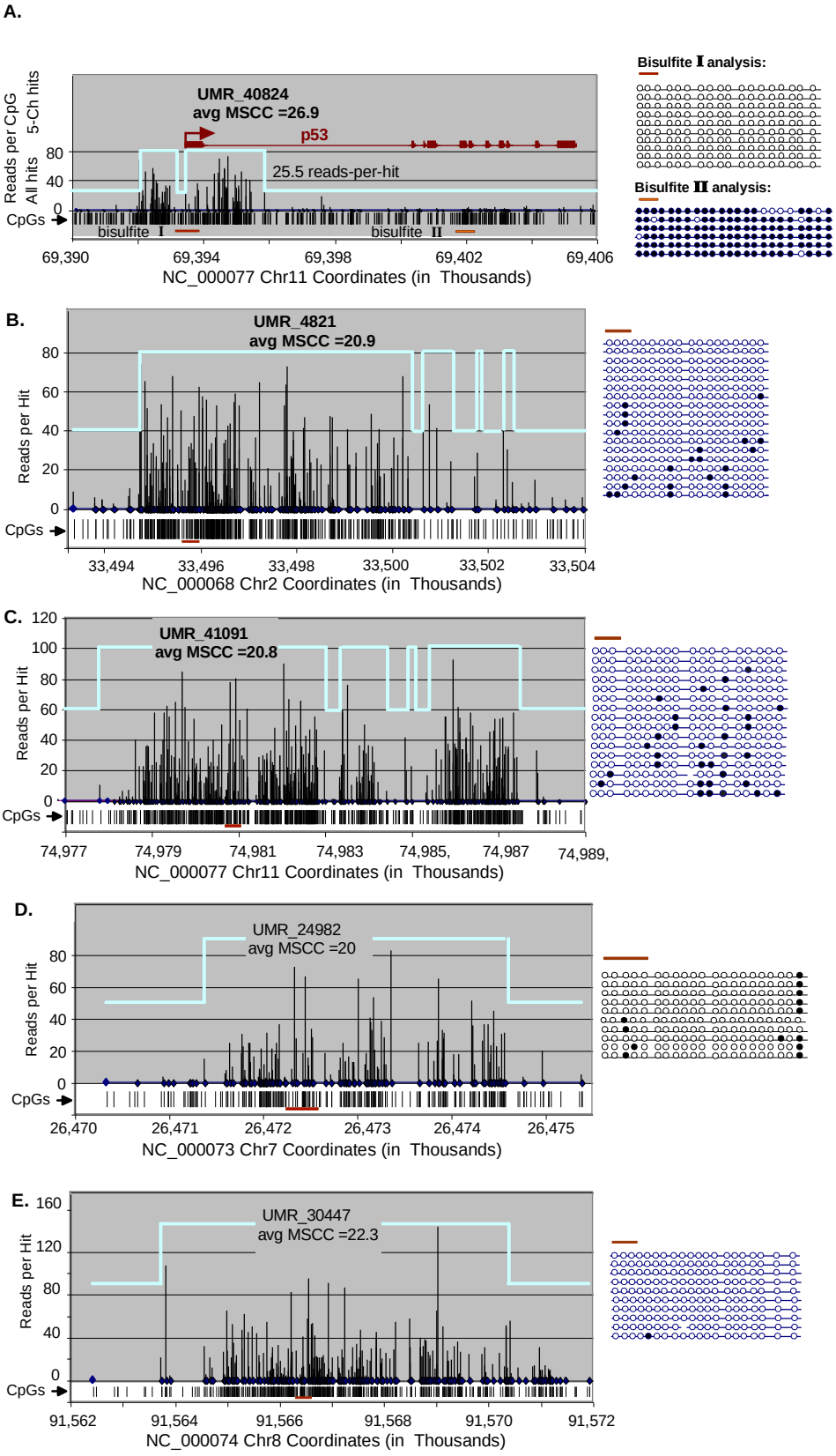


Figure S7. Strategies followed to define unmethylated regions (UMRs). A. Hypomethylated CpG clusters were mapped in the genome based in the inter-CpG-distance between 5-channel hits. Inside each cluster, contiguous identified CpGs are at distances of 300 nt or less. One hypomethylated cluster was found overlapping exon 1 of the Lrp1 gene in chr 10. This cluster was spotted because it contains 40 5-channel hits that meet the criteria of interCpG distance (300 nt or less). These hits produced 1,225 reads and give an avg MSCC score of 30.6. However these CpGs were not the

only ones identified in the spotted region. There are an additional 13 4Enz CpGs that were identified by a relatively small number of reads. In order to produce a final score that measures the unmethylated status of the region we included these hits. The final avg MSCC score including 53 hits that produced a total of 1,340 reads was 25.3. Since the avg MSCC was larger than 17 we clasify the region as an UMR. Below, are shown the results of four bisulfite sequencing experiments. Numbers I, II and III analyzed regions (maroon bars) from within the UMR and show all of the CpGs to be unmethylated. This includes not only CpGs addressable by the 4Enz cocktail, but also CpGs not addressed by the enzymes. Bisulfite sequencing analysis of number IV is from a region just outside the UMR and includes four CpGs, one of which is addressable and was identified by two reads. In this example the border of the UMR shows a sharp change in the level of methylation. **B.** The upper panel shows a hypomethylated CpG cluster that was found overlapping exon 1 of Adra1b (adrenergic receptor α1b) in chr11. Only the CpGs identified in 5 of 5 channels (5-Channel Hits, upper panel) are plotted. However, the the cluster spotted with these 7 CpGs included an additional 26 CpGs that were identified in a few of the sequencing channels and have a few number of reads (bottom panel). When the avg MSCC is scored using all 33 hits, the value droped to 5.1 indicating that the region, although hypomethylated (<75% methylation) it is not an UMR. Bisulfite sequencing analysis confirmed that the region is methylated. Also here not-addressable CpGs share methylation status with their near neighbors.


A.
UMR_40824
avg MSCC =26.9
p53
25.5 reads-per-hit
Reads per CpG
5-Ch hits
All hits
CpGs
bisulfite I
bisulfite II
69,390
69,394
69,398
69,402
69,406
NC_000077 Chr11 Coordinates (in Thousands)
Bisulfite I analysis:
Bisulfite II analysis:
B.
UMR_4821
avg MSCC =20.9
Reads per Hit
CpGs
33,494
33,496
33,498
33,500
33,502
33,504
NC_000068 Chr2 Coordinates (in Thousands)
C.
UMR_41091
avg MSCC =20.8
Reads per Hit
CpGs
74,977
74,979
74,981
74,983
74,985,
74,987
74,989,
NC_000077 Chr11 Coordinates (in Thousands)
D.
UMR_24982
avg MSCC =20
Reads per Hit
CpGs
26,470
26,471
26,472
26,473
26,474
26,475
NC_000073 Chr7 Coordinates (in Thousands)
E.
UMR_30447
avg MSCC =22.3
Reads per Hit
CpGs
91,562
91,564
91,566
91,568
91,570
91,572
NC_000074 Chr8 Coordinates (in Thousands)

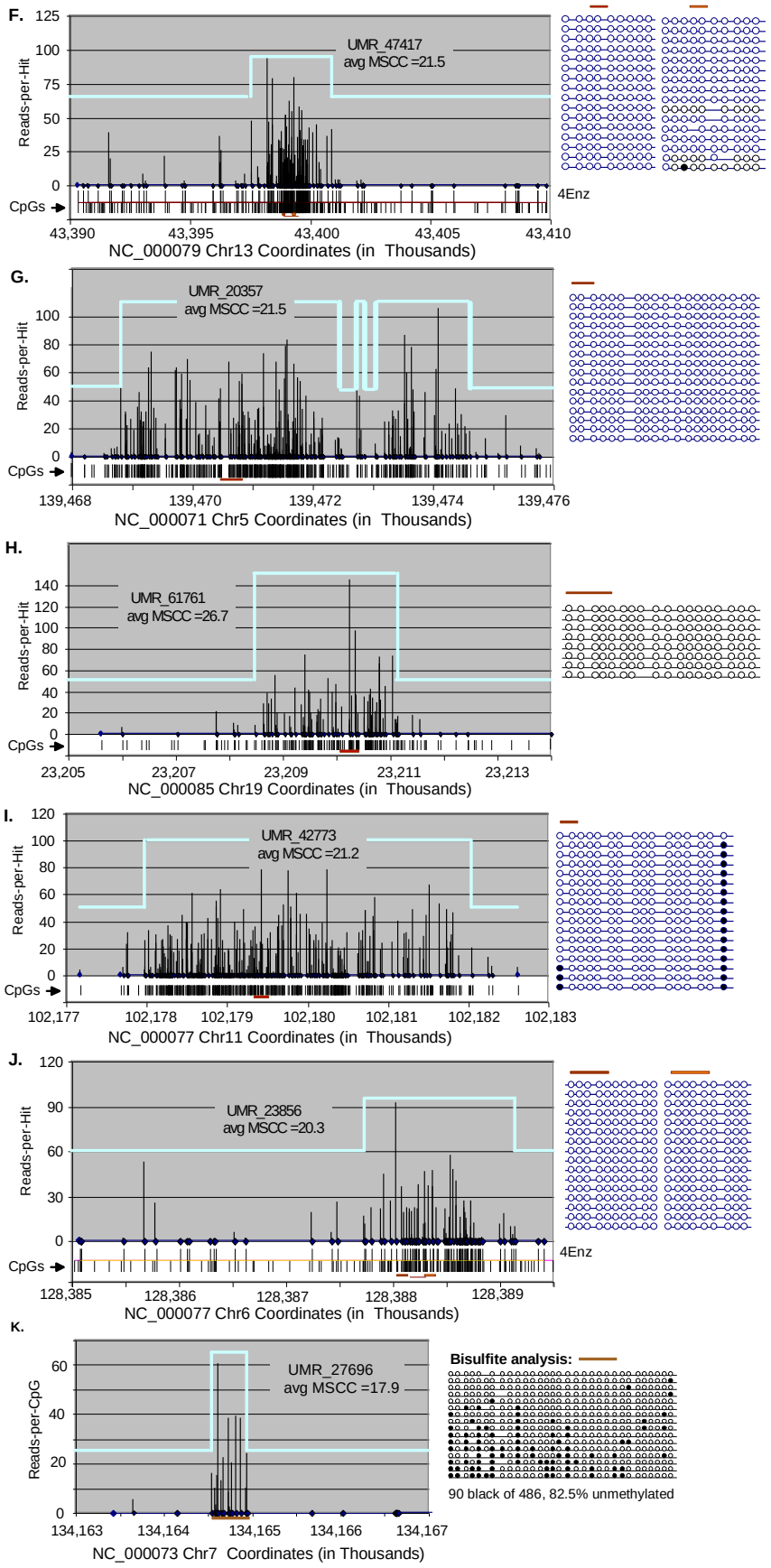

F.
UMR_47417
avg MSCC =21.5
Reads-per-Hit
CpGs
4Enz
NC_000079 Chr13 Coordinates (in Thousands)
G.
UMR_20357
avg MSCC =21.5
NC_000071 Chr5 Coordinates (in Thousands)
H.
UMR_61761
avg MSCC =26.7
NC_000085 Chr19 Coordinates (in Thousands)
I.
UMR_42773
avg MSCC =21.2
NC_000077 Chr11 Coordinates (in Thousands)
J.
UMR_23856
avg MSCC =20.3
NC_000077 Chr6 Coordinates (in Thousands)
K.
UMR_27696
avg MSCC =17.9
Reads-per-CpG
Bisulfite analysis:
90 black of 486, 82.5% unmethylated
NC_000073 Chr7 Coordinates (in Thousands)

Figure S8. Bisulfite sequencing analysis of eleven hypomethylated clusters identified as UMRs (avgMSCC score >17). *Left panels,* reads frequency maps. Brown bars indicate the regions analyzed by bisulfite sequencing. *Right panels*, outcome of bisulfite sequencing experiments. *Open circles,* unmethylated; *closed circles*, methylated. We counted the number of addressable CpGs that were located in the analyzed regions of panels A,E,F,G,H and J. Inspite of these regions being completely unmethylated, 22% of addressable CpGs have an MSCC score below 5, in agreement with the FDR predicted by ROC analysis.