Intermediate DNA methylation is a conserved signature of genome regulation

GiNell Elliott  
*Washington University School of Medicine in St. Louis*

Xiaoyun Xing  
*Washington University School of Medicine in St. Louis*

Xin Zhou  
*Washington University School of Medicine in St. Louis*

Daofeng Li  
*Washington University School of Medicine in St. Louis*

Gary D. Stormo  
*Washington University School of Medicine in St. Louis*

*See next page for additional authors*

Follow this and additional works at: [https://digitalcommons.wustl.edu/open_access_pubs](https://digitalcommons.wustl.edu/open_access_pubs)

Please let us know how this document benefits you.

**Recommended Citation**

[https://digitalcommons.wustl.edu/open_access_pubs/4344](https://digitalcommons.wustl.edu/open_access_pubs/4344)

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact vanam@wustl.edu.
Authors
GiNell Elliott, Xiaoyun Xing, Xin Zhou, Daofeng Li, Gary D. Stormo, Ting Wang, and et al
Intermediate DNA methylation is a conserved signature of genome regulation


The role of intermediate methylation states in DNA is unclear. Here, to comprehensively identify regions of intermediate methylation and their quantitative relationship with gene activity, we apply integrative and comparative epigenomics to 25 human primary cell and tissue samples. We report 18,452 intermediate methylation regions located near 36% of genes and enriched at enhancers, exons and DNase I hypersensitivity sites. Intermediate methylation regions average 57% methylation, are predominantly allele-independent and are conserved across individuals and between mouse and human, suggesting a conserved function. These regions have an intermediate level of active chromatin marks and their associated genes have intermediate transcriptional activity. Exonic intermediate methylation correlates with exon inclusion at a level between that of fully methylated and unmethylated exons, highlighting gene context-dependent functions. We conclude that intermediate DNA methylation is a conserved signature of gene regulation and exon usage.
DNA methylation plays a central role in human development and cellular identity. Dynamic CpG methylation throughout cell differentiation correlates with cell type-specific gene regulation and expression levels, with loss of methylation reflecting enhancer or gene activation. The majority of CpGs in differentiated cells are uniformly methylated or unmethylated between homologous chromosomes and within cell populations composed of a single cell type. The bimodality of DNA methylation implies a binary on-off control over gene expression, yet a significant number of loci throughout the genome do not fit within this model.

Regions of intermediate methylation (IM) may exhibit allelic methylation, intercellular variability or clusters of interpersed methylated and unmethylated CpGs within each cell. A small number of genes are marked by allele-specific methylation (ASM), including imprinting control regions (ICRs), which are essential for mammalian development. Other IM states, including non-allelic IM, may also be functional and potentially dynamically regulated during development. The level of methylation reported in IM regions in prior studies is variable, but typically much less than the 50% that typifies ICRs. The prevalence of IM states and their function within genetically and phenotypically homogenous cell populations has been enigmatic. Here, we use genome-wide DNA methylation profiling to identify and compare IM regions across human tissues and primary cell types, and among individuals. We define boundaries of IM loci by combining two independent and complementary enrichment-based methods that, respectively, identify methylated and unmethylated regions of DNA. We then validate and quantify methylation levels at baspair resolution within IM regions using whole-genome bisulfite sequencing (WGBS) and methylation array data. We analyse multiple cell types from the same individual to distinguish cell type epigenetic differences from genetic effects on the epigenome. We then examine the relationship between IM states, associated histone marks of gene regulation, mRNA levels and exon usage from the same samples used to discover the IM regions. We further use these data, along with in vivo enhancer assay data and evolutionary conservation, to discover potential functions of IM states.

**Results**

**IM is closely associated with genes.** To identify regions of IM, we combined two complementary whole-methylome profiling methods: MeDIP-Seq, which targets methylated DNA using an anti-methylcytosine antibody; and MRE-Seq, which identifies unmethylated DNA by methylation-sensitive restriction enzyme digestions. This experimental approach was chosen because nearly all mapped reads are informative, contrasting to the inefficiency of WGBS in which 70–80% of sequence reads are uninformative because they lack CpG sites. Furthermore, 5-methylcytosine and 5-hydroxymethylcytosine are conflated in WGBS, whereas MeDIP/MRE-Seq detect 5-methylcytosine exclusively. MeDIP/MRE-Seq may also provide greater accuracy in a minority of loci. Most importantly, identification of IM with MeDIP/MRE-Seq does not assume a binary state, as IM are identified by strong and overlapping signals from each assay. We applied MeDIP/MRE-Seq to 23 human primary cell and tissue samples isolated without culturing from three tissue types and seven donors, and two biological replicates of the H1 embryonic stem cell line (H1ES) (see Methods). The diverse set of samples we profiled allowed controlled comparisons of different cell types from the same individual, and identical cell types from different age-matched donors.

A small fraction of CpGs in each sample showed signal enrichment in both MeDIP-Seq and MRE-Seq, indicating simultaneous presence of methylated and unmethylated DNA (Fig. 1a and Supplementary Fig. 1a). We developed a maximum scoring segment algorithm to define boundaries of IM regions representing clusters of neighbouring CpGs with co-occurring enrichment in MeDIP/MRE-Seq (Methods; Supplementary Fig. 1b). Altogether, 18,452 unique autosomal IM regions were detected in one or more samples, containing 2% of the 26.9 million CpGs interrogated. Within IM regions, we estimated the methylation levels at single CpGs using WGBS from H1ES cells and Illumina 450K Infinium Methylation Array data from eight of our samples. WGBS and methylation arrays confirmed the primarily bimodal genome-wide distribution of fully methylated and unmethylated CpGs. Individual CpGs within regions classified as IM, however, had a mean methylation of 57% (WGBS) and 58% (arrays; Fig. 1a). In the majority of regions, the IM state was restricted to one or a subset of tissue or cell types, although 1,754, or 9%, were present across all tissues studied (constitutive IM; Fig. 1b–e). Hierarchical clustering based on the presence or absence of the IM status or based on MeDIP/MRE-Seq read density at the union set of IM regions strongly separated cell types isolated from different tissues (Fig. 1c and Supplementary Fig. 2).

We detected an average of 7,531 autosomal IM regions per sample, with a mean length of 271 bp. We selected the union of IM regions across all samples (n = 18,452) as a reference set for further analysis. Over half of IM regions were intragenic or were within 3 kb of 6,798 autosomal genes (36% of genes; odds ratio = 1.27; P-value < 0.001, χ²). IM regions were enriched at promoters, exons and DNase I hypersensitivity sites (DHSs), and confirmed 17 of 19 known imprinted loci (Fig. 2a,b and Supplementary Figs 2a and 3). Thus, IM regions were associated with a significant number of genes and potential regulatory regions.

Approximately 50% of IM regions overlapped loci identified as differentially methylated across cell and tissue types, consistent with our observation that IM is often tissue-specific. However, most studies comparing methylation between cell types examine only the binary direction of the methylation change (hypo- or hypermethylated) rather than the methylation level maintained within each cell type. Thus, differential methylation as previously defined does not identify specific regions as IM. The majority (69%) of IM regions reported here have not been previously identified as having allele-specific or stable IM states within a population of cells of a single type (17,18,25,26) (Supplementary Fig. 5a,b).

**Level of histone modification and DHS at IM.** To determine whether IM regions could have regulatory function, we next performed chromatin immunoprecipitation sequencing (ChIP-Seq) for selected histone modifications that demarcate active regulatory elements, using the same cell samples in which IM regions were mapped. Within a given sample, we explored the relationship between DNA methylation level and histone modification levels by comparing histone ChIP-Seq signal at IM regions to signal at regions that were methylated and unmethylated in the same sample used for ChIP-Seq, but which had IM status in other samples. Consistent with prior studies, the normalized read densities for the active marks of H3K4me1 and H3K4me3 were anti-correlated with DNA methylation. Interestingly, signal strength for these modifications at regions with IM status consistently fell between signals at methylated and unmethylated sites, providing a novel association of IM with intermediately active chromatin states (Fig. 2c). Furthermore, DHS in H1ES cells or in fetal brain had a similar intermediate state at their IM regions, respectively (Fig. 2d).
Enhancer regions predicted by chromHMM27 using the cell-type-matched ChiP-Seq data were more enriched for IM CpGs than the genomic background of methylated (MeDIP-Seq only) or unmethylated (MRE-Seq only) CpGs (Supplementary Methods and Supplementary Fig. 3). We therefore asked if IM regions coincide with functional regulatory elements defined by in vivo transgenic reporter assays. Of 70 candidate enhancers defined by the VISTA project that we also identified as IM regions, 50 drove transgene expression in mouse embryos28—a higher validation rate than VISTA candidates tested from non-IM regions. Methylation levels of enhancers in the reporter construct are unknown, however, this result demonstrates that many IM regions coincide with functional enhancers (Supplementary Data 1). A small portion of IM regions (9.5%) in fetal brain also shows increased 5-hmC in fetal brain relative to adult brain29 (Supplementary Fig. 5c).

**Level of gene expression and exon inclusion near IM.** We hypothesized that the DNA methylation level in IM regions might correspond to a difference in the level of transcription of the associated gene. To examine this relationship, we initially focused on IM regions occurring within 10 kb of transcription start sites in breast myoepithelial cells. For comparison, we applied the same selection criteria to methylated and fully unmethylated regions in myoepithelial cells that had IM status in other cell types. Methylated and unmethylated status at these regions distinguished proximal genes with significantly different mean expression values, following the established inverse correlation between DNA methylation at enhancers and gene expression. Remarkably, despite an average DNA methylation level of 63% in myoepithelial cells based on methylation array, regions with IM status corresponded to a set of genes with mean expression distinct from both the methylated and unmethylated sets ($P<0.005$, Wilcoxon), suggesting that enhancers with IM are associated with intermediate levels of gene expression (Fig. 2e).

**IM is predominantly allele independent.** IM CpGs could reflect ASM or allele-independent methylation (AIM). Interestingly, most previously identified ASM has not been connected to allelic gene expression14,17. To distinguish ASM from AIM, we identified heterozygous single-nucleotide polymorphisms (SNPs) in the autosomal IM regions and determined the allelic preference of sequencing reads from MeDIP/MRE-Seq, where ASM segregated heterozygous SNPs between the two assays (Methods). We identified a total of 2,072 ASM SNPs (721 unique across samples) and 5,895 AIM SNPs (3,262 unique; Fig. 3a,b). We then categorized each IM region as ASM or AIM if it contained two or more ASM or AIM SNPs (Methods). We identified 109 ASM regions and 927 AIM regions (Fig. 3b).
From these data, we infer that the methylation pattern underlying a majority of IM regions is allele independent rather than allele specific.

ASM should have approximately 50% methylation at individual CpGs, whereas IM due to intercellular heterogeneity could theoretically fall at any IM value. On the methylation arrays, methylation scores of CpGs within ICRs and our predicted ASM had a relatively narrow distribution centred near 50%. Methylation scores at AIM regions peaked near 50% but had a broader distribution, consistent with the presence of more intercellular variation at individual CpGs, and potentially across neighbouring CpGs (Fig. 3c).

We selected 36 high confidence IM regions (based on MeDIP/MRE-Seq signal strength) for validation by bisulfitie, PCR, cloning and sequencing. A total of 35 loci validated as IM (97%), including 12 of 14 ASM (86%; 2 predicted as ASM were found to be AIM; Fig. 3a) and 12 of 12 AIM (100%) regions. Interestingly, the AIM regions exhibited an interspersed mixture of methylated and unmethylated CpGs on each clone (Supplementary Fig. 6 and Supplementary Data 2). Thus, the intermediate signals in H3K4me1, H3K4me3 and DHSs at IM regions cannot be attributed to fully unmethylated alleles in an ASM configuration. Potential configurations include intercellular heterogeneity, differences in nucleosome spacing and differences between individual subunits within each nucleosome.

To examine the relationship between ASM and histone modifications, we measured allelic preference for histone signals at ASM loci and compared this to the allelic preference of MeDIP/MRE-Seq signals at the same locations (Fig. 3d). We focused on heterozygous SNPs from the two fetal brain samples with genotype validated by whole-genome sequencing. At loci classified as ASM, the active chromatin marks H3K4me3 and H3K4me1 showed a clear preference for the unmethylated allele, whereas the repressive mark H3K9me3 had a bias towards the methylated allele. Interestingly, the repressive mark H3K27me3 preferentially occurred on the unmethylated allele. In contrast, histone modifications at AIM and fully unmethylated regions did not exhibit a strong allelic preference (Supplementary Fig. 4).
IM state is evolutionarily conserved. As a complementary comparative epigenomics approach to investigate whether IM regions may have function, we identified IM states using MeDIP/MRE-Seq data from murine embryonic stem (ES) cells and fetal neurons, taking the union of regions in both cell types as the reference IM set for mouse. A total of 13,623 IM regions were detected in the two mouse samples. We then determined how frequently IM states in the mouse genome were conserved as IM in human. After mapping mouse IM regions to their syntenic loci in the human genome, 17.4% directly overlapped human IM regions, representing a 14-fold enrichment over random expectations in the human genome, 17.4% directly overlapped human IM regions. We then determined how often IM regions were also profiled from mouse, the enrichment increased to 41.3%. Averaged phastCons scores indicated conservation of the IM state (Fig. 4a,b). When restricting the reference IM set for mouse to only the two human tissues (ES and fetal brain) A value of 0 is unmethylated, and a value of 1 is fully methylated. (MRE-seq, top panel) and methylated DNA (MeDIP-seq, bottom panel) at ASM SNPs from fetal brain, in which the heterozygous genotype was verified by whole-genome sequencing. A positive correlation indicates that signals are on the same allele, whereas negative correlation indicates that signals are on opposite alleles.

Discussion

We have defined 18,452 discrete regions of the genome that maintain DNA methylation levels near 50% in one or more cell types across multiple individuals, and are associated with intermediately active rather than suppressed gene expression. Although the precise function of intermediate DNA methylation states is challenging to prove using current methods, the IM signature enriches for regions with multiple indicators of regulatory function, particularly those associated with enhancers. The intermediate level of active histone modification and chromatin accessibility at IM regions implies these are regulatory sites distinct from repressive, fully methylated states or permissive, unmethylated states. Intermediate signals cannot be explained trivially as cell-type differences within tissues, because they are equally intermediate in tissue, unsorted peripheral blood and six highly purified cell types. The strong association of exonic IM with an intermediate level of exon inclusion provides independent validation of the general concept that IM is an epigenomic signature of context-dependent function. Significant interspecies conservation, and conservation among different
individuals at IM regions further suggests an important function and potentially a shared mechanism for their establishment and maintenance.

Although an individual CpG is either methylated or unmethylated, our results raise the possibility that local clusters of CpGs provide quantitative instructions for gene regulation or exon usage that exists stably between fully ‘on’ and ‘off’ states within single-cell types. Intriguingly, we approximate that 18% or fewer IM regions are allele-specifically methylated, whereas the majority is allele independent, suggesting that the predominant mode of IM establishment relies on epigenetic heterogeneity at a precise loci within cell populations. Methods to create and abolish IM states in a locus-specific manner will be required to further examine these possibilities31.

Widespread intercellular and allelic stochasticity of gene expression has been observed in single-cell mRNA sequencing data32–34; however, it is not clear why some genes exhibit stochastic expression that exists stably between fully ‘on’ and ‘off’ states within single-cell types. A quantitative model of epigenetic gene regulation that includes stable, tissue-specific intermediate states may help explain variations in gene activity between cells of the same type. Although this remains to be proven, it would have far-reaching implications for the multitude of diseases linked to disruption of DNA methylation, including the epigenetic heterogeneity in cancer cells35,36.

Methods
Sample preparation. All assays were performed as part of the repository of the NIH Roadmap Epigenomics Mapping Centers for the reference Human Epigenome Atlas35.

Blood. Buffy coats were obtained from the Stanford Blood Center (Palo Alto, California, USA). Blood was drawn and processed on the same day. Peripheral blood mononuclear cells (PBMCs) were isolated by Histopaque-1077 (Sigma-Aldrich) density-gradient centrifugation according to the manufacturer’s protocol. Further purification of memory CD4+ T lymphocytes was performed using a RosoSep instrument and the isolation kit for each subpopulation (EasySep Human Memory CD4+ T Cell Enrichment kit and Custom Human Naive CD8+ T Cell Enrichment kit; STEMCELL Technologies). Total PBMCs were karyotyped (Molecular Diagnostic Services) and analysed to determine the percentage of cells in each cell cycle stage (G0/G1, S and G2/M). PBMC and T-cell subpopulations were stained with antibodies (anti-CD3 TRI-COLOR (Invitrogen), anti-CD4 PE (BD Biosciences), anti-CD8 FITC (BD Biosciences), anti-CD4 TRI-COLOR (Invitrogen), anti-CD45RO PE (Invitrogen), anti-CD45RA FITC (BD Biosciences) and anti-CD8 TRI-COLOR (Invitrogen)) and analysed by fluorescence-activated cell sorting for purity. Cells were aliquoted for isolation of either DNA or RNA and were washed in PBS. Cell pellets for RNA purification were resuspended in 1 ml of TRIzol reagent (Invitrogen) and frozen at −80 °C. Cell pellets for DNA purification were flash frozen in liquid nitrogen and stored at −80 °C. Antibodies used included TRI-COLOR-conjugated antibody to CD3 (Invitrogen, MHC0306), Phycoerythrin-

Figure 4 | The IM state is conserved in syntenic loci in mouse. (a) A novel, tissue-specific human IM region in an internal exon of DCHS1 shows conserved IM state at the orthologous exon in mouse. Height for all tracks shows a signal range of 0–50 reads. (b) The pie chart indicates the distance to the nearest human IM region from each aligned mouse IM region. The bar graph shows the fold-enrichment of overlap between human and mouse IM regions at the CpG level using the complete set of human IM regions and a set restricted to cell types analogous to those in the mouse IM analysis. (c) Average phastCons conservation scores over all IM regions and regions that do not overlap coding exons. Scores are based on alignment of 46 vertebrate species.
conjugated antibody to CD4 (BD Biosciences, 340419), fluorescein isothiocyanate (FITC)—conjugated antibody to CD8 (BD Biosciences, 561947), TRI-COLOR-conjugated antibody to CD4 (Invitrogen, MHCD4006), PE-conjugated antibody to CD54 (BD Biosciences, MHCD45RA01) and TRI-COLOR-conjugated antibody to CD8 (Invitrogen, MHCD0806). All antibodies were used according to the manufacturer’s instructions.

Breast. Breast tissues were obtained from disease-free premenopausal women undergoing reduction mammoplasty in accordance with institutional review board instructions. Tissues were obtained with appropriate consent according to the School of Medicine approved the experiments. Gestational day 12.5–13 embryos were removed and cells were dissected out and dissociated using 0.5 ml trypsin/ EDTA (0.05%/0.02%; Tissue Culture Support Center, Washington University, St Louis, MO, USA) for no longer than 15 min at 37 °C with two or three times of trypsinization by hand. The resulting cell suspension at 295 for 10 min, and the pellet was washed with RPMI 1640 supplemented with 10% FBS. Clusters enriched in epithelial cells (referred to as organoids) were recovered after serial filtration through 150-μm nylon mesh (Fishér) and 40-μm nylon mesh (Fishér). The final filtrate contained primarily mammary stromal cells (fibroblasts, immune cells and endothelial cells) and some epithelial cells. After centrifugation at 290 for 5 min at 4 °C, epithelial organoids and filtrate were frozen for long-term storage. The day of cell sorting, epithelial organoids were thawed and further digested with 0.5 g/l trypsin, 0.5 mM EDTA, pH 8.0, and dispase DNase 1 (STEMCELL Technologies). Generation of single-cell suspensions was monitored visually. Single-cell suspensions were filtered through a 40-μm cell strainer (Fisher), spun down and allowed to regenerate in mammary epithelial cell growth medium (Lonza) supplemented with 2% FCS for 60–90 min at 37 °C. This regeneration step enables quenching of trypsin and re-expression of the cell surface markers before staining, which was needed because their extracellular domains had been cleaved by trypsin.

Fetal brain. Post-mortem human fetal neural tissues were obtained from a case of twin non-syndrome fetuses whose death was attributed to environmental/placental aetiology. Tissues were obtained with appropriate consent according to the Partner’s Healthcare/Brigham and Women’s Hospital Institutional Review Board guidelines (protocol 2010001144). All samples and tissues were de-identified and linked only with a minimal data set (age, ancestry and, in some cases, parity/gravidity). All study subjects provided written informed consent. Tissue was dissociated mechanically and enzymatically. Briefly, tissue was minced and dissociated in RPMI 1640 with 1 glutamine and 25 mm HEPS (Fishér) supplemented with 10% fetal bovine serum (FBS; JR Scientific), 100 U ml⁻¹ penicillin, 100 μg ml⁻¹ streptomycin sulfate, 0.25 μg ml⁻¹ fungizone, 50 μg ml⁻¹ gentamicin sulfate (Lonza), 200 U ml⁻¹ collagenase 2 (Worthington) and 100 U ml⁻¹ hyaluronidase (Sigma-Aldrich) at 37 °C for 16 h. The cell suspension was then centrifuged at 295 for 10 min, and the pellet was washed with RPMI 1640 supplemented with 10% FBS. Cells were re-suspended in either DMEM/10% FBS or growth medium, plated and cultured in neurobasal medium (Invitrogen #21103-049) with 2% serum-free supplement (Invitrogen #17504-04), glutamine (5 mM) and penicillin/streptomycin. Cells were typically seeded at a density of one forebrain per 24-well plate. Neurons were stimulated with KCl (55 mM) at 5 days in vitro for 4 h (ref. 40).

Chromatin immunoprecipitation sequencing. Standard operating procedures for ChIP-seq library construction are available at http://www.roadmapepigenomics.org/protocols/type/experimental/. ChIP-seq library construction involves the following protocols in order: (i) cross-linking of frozen cell pellet, (ii) DNA sonication using Sonic Dismembrator 550 and (iii) SLX-PET protocol for Illumina sample preparation. Antibodies used in this study were subjected to sensitivity assessment to meet Reference Epigenome Mapping Quality Standards (http://www.roadmapepigenomics.org/protocols) including western blotting of whole-cell extracts, 384 peptide dot blot (Active Motif MODified Histone Peptide Array) and ChIP-seq using control cell pellets (HL60). Antibody vendor, catalogue number and lot number are provided alongside all ChIP-seq data sets and are available through GEO and the NCBI epigenomics portals (for example, http://www.ncbi.nlm.nih.gov. beckerpro.wustl.edu/go/query/acc.cgi?acc=GSM613811). Final library distributions were calculated using an Agilent Bioanalyzer and quantified by fluorometric quantification (Qubit, Life Technologies). Libraries were standardized using single-end 76 nt sequencing chemistry on an Illumina GAiiX or HiSeq2000 following the manufacturer’s protocols (Illumina) as either single or multiplexed libraries using custom index adapters added during library construction.

Sequence reads were aligned to NCBI GRCm38 lite reference using Burrows-Wheeler Alignment tool (BWA) 0.6.2-r126 with default parameters. MethyQa (an unpublished C programme; available at http://methyqafa.sourceforge.net/) was used to directionally extend aligned reads to the average insert size of DNA fragments (150 bp) and to generate a Wig file for downstream visualization. Reads with BWA mapping quality scores < 10 were discarded and reads that aligned to the same genomic coordinate were counted only once.

Methylation-sensitive restriction enzyme-seq. Methylation-sensitive restriction enzyme-seq (MRE-seq) was performed as in Maunakea et al., with modifications as detailed below. Five parallel restriction enzyme digestions (HpaII, BstHII36, Ssal/ACI and Hinfi (Fermentas), and HpyCH4IV (NEB)) were performed, each using 1 μg of DNA per digest for each of the skin cell type samples. Five units of enzyme were initially incubated with DNA for 3 h and then an additional five units of enzyme was added to the digestion for a total of 6 h of digestion time. DNA was purified by phenol/chloroform/isooamyl alcohol extraction, followed by chloroform extraction using phase lock gels. Digested DNA from the different reactions was purified and precipitated with one- tenth volume of 3 μM sodium acetate (pH 3.2) and 2.5 volumes of ethanol. The purified DNA was size selected and purified using Sonic Dismembrator 550 and gel electrophoresis. Library construction was performed as per the Illumina Genomic DNA Sample Prep Kit protocol with the following modifications. During the end-repair reaction, T4 DNA polymerase and T4 PNK were excluded and 1 μl of 1.5 diluted Klenow DNA polymerase was used. For the adapter ligation reaction, 1 μl of 1.1 diluted PE aceohyme and 2 μl of 2x AceO mix was used. Two ChIP-seq reads were added to the sequencing mix. After sequencing, 0.5 μl of adapter ligated DNA was used for the PCR enrichment reaction with PCR PE Primers 1.0 and 2.0. PCR products were size selected and purified (170–420 bp) by gel electrophoresis and Qiagen Quickspin extraction. DNA libraries were checked for quality by Nanodrop (Thermo Scientific) and Agilent DNA Bioanalyzer (Agilent). Reads were aligned to hg19 using the BWA and pre-processed using methyQa (an unpublished C programme; available at http://methyqafa.sourceforge.net/). MRE reads were normalized to account for differing enzyme efficiencies and methylation values were determined by counting reads with Cpgs at fragment ends. To enable comparison between MRE-seq data from blood, brain and breast samples the raw reads and restriction enzyme-seq libraries were re-processed and re-normalized. MRE-seq, Cpg-seq and restriction enzyme-seq, cell type specific MRE-seq reads that resulted from the use of additional restriction enzymes (BstHII36 and HpyCH4IV) were removed. Detailed library

**ARTICLE**

**NATURE COMMUNICATIONS** | DOI: 10.1038/ncomms7363
---

© 2015 Macmillan Publishers Limited. All rights reserved.
Methylated DNA immunoprecipitation-seq. MeDIP-seq was performed as in Maunakea et al.22 Five micrograms of genomic DNA were sonicated to a fragment size of ~100–400 bp using a Bioruptor sonicator (Diagenode). End-repair, addition of 3′A tails and PE adapter ligation with 2 μg of sonicated DNA was performed using the SureSelect Human Genomic DNA Sample Prep Kit (Agilent). Oligonucleotide primers used for the PCR-enrichment reaction were synthesized by Life Technologies and purified by gel electrophoresis. PCR was performed with Phusion DNA Polymerase (New England Biolabs) to yield 220–420 bp PCR fragments. Fragments were purified by agarose gel (Invitrogen) and ligated to the Illumina PE adapters (Illumina). The ligation products were subjected to six to ten cycles of PCR purifying free adaptors. The ligated DNA was amplified and hybridized to an Infinium HumanMethylation450 beadchip (Illumina) following the Infinium HD methylation assay protocol at the UCSF Genomics Core facility. Methylation levels (beta values) were determined using the MethylKit Module of the Illumina GenomeStudio software.

**Bisulfite treatment and library construction for WGBS.** One to five micrograms of genomic DNA were bisulfite converted using the EZ DNA Methylation Kit (Zymo Research) as per the manufacturer’s alternative incubation conditions protocol. The bisulfite-converted DNA was amplified and hybridized to an Infinium HumanMethylation450 beadchip (Illumina) following the Infinium HD methylation assay protocol at the UCSF Genomics Core Facility. Methylation levels (beta values) were determined using the MethylKit Module of the Illumina GenomeStudio software.

**Targeted bisulfite PCR and sequencing.** Further validation of genome-wide data, particularly sites with apparent intermediate DNA methylation, was performed by bisulfite sequencing. Total genomic DNA underwent bisulfite conversion following established protocol41 with a modified conversion conditions of: 95 °C for 1 min, 50 °C for 59 min for a total of 16 cycles. The regions of interest were amplified with bisulfite PCR primers and were subsequently cloned using pCR2.1/TOPO (Invitrogen). Single colony PCR and sequencing (QuintaraBio) were undertaken for analysis.

**Genomic features.** CpG islands, untranslated regions (UTRs), gene bodies, exons and introns were RefSeq gene annotations downloaded from the UCSC Genome Browser (hg19). Promoters were defined as the 3′-kb upstream of the transcription start site (TSS) for all RefSeq genes. Intergenic regions were defined as all regions outside RefSeq gene bodies and promoters.

**IM detection algorithm.** We developed a maximum scoring segment algorithm to identify regions of overlapping MeDIP-Seq and MRE-Seq signals (Supplementary Fig. 1). For each sample, total MeDIP-Seq read counts were normalized to 50 M, and MRE-Seq read counts were normalized to 35 M (ref. 20). Given normalized MeDIP-Seq and MRE-Seq read densities across all CpGs, the algorithm traces through each CpG sequentially, comparing read counts from both assays. An arbitrary score proportional to the read density was increased when the signals overlapped and decreased when they did not, and an additional penalty proportional to the distance between CpGs was assigned. When the score returned to zero at some distance following the initialization of an IM region, the end point of the region was defined as the position with the highest score following the start site. Additional IM detection data sets can be downloaded at http://epigenome.wustl.edu/Intermediate_Methylation/.

**IM data filters.** To limit the false detection of IM, we determined a length and score threshold by comparing IM calls derived from our data to calls from randomly shuffled data. Paired MeDIP-Seq and MRE-Seq read counts were randomly reassigned to CpGs, maintaining their paired relationship. The length and score distributions of IM calls from 1,000 random simulations were then compared with actual distributions, and cutoffs were selected to achieve an estimated false-positive rate of less than 1%. The minimum score was therefore set to 8.0, and the minimum length set to 100 bp (length cutoff was applied to the reference IM set).

**Reference IM list assembly.** The reference list of IM regions is intended as a summary of the IM landscape across all available samples, and represents the union of all IM regions that passed the data filters. IM regions within 100 bp were merged and assigned the age, the data sets required to be present in at least 2 of the 25 samples analysed to further reduce false positives. These tasks were performed using applications from the BEDTools suite42.
Gene expression and relative exon expression. Measured, unmethylated and IM regions were associated with NCBI Reference Sequence (RefSeq) genes if they fell within 10 kb of the transcription start site of that gene (Total gene-associated regions: IM = 6,677; methylated = 3,270; unmethylated = 5,605). Gene expression levels were measured as RPKM for each annotated gene transcript associated with a region. Regions were associated with exons if they occurred within 1 kb of the exon (IM exons = 14,336; methylated exons = 6,642; unmethylated exons = 9,331). Relative exon expression was measured using the following formula:

\[
\text{Relative Exon Expression} = \frac{R_I - R_M}{R_I}
\]

where \( R_I \) is the RPKM value for a given exon and \( R_M \) is the RPKM value for the transcript containing that exon.

**Comparison of human and mouse IM.** Mouse coordinates (mm9) were mapped to orthologous human regions using the UCSC LiftOver tool with minMatch parameter set to 0.8 (80% sequence identity). Additional data tables can be downloaded at http://epigenome.wustl.edu/Intermediate_Methylation/.

**References**


**Acknowledgements**

We thank collaborators in the Reference Epigenome Mapping Centers (REMC), the Epigenome Data Analysis and Coordination Center and NCBI who have generated and processed data used in this project. We thank Jufang Chang and Jeff Milbrandt for providing mouse cortical neuron samples. We acknowledge support from the NIH Roadmap Epigenomics Program, sponsored by the National Institute on Drug Abuse (NIDA) and the National Institute of Environmental Health Sciences (NIEHS). We thank Theresa Kadlec and Arthur Weiss for processing blood samples. J.F.C. and T.W. are supported by NIH grant U01ES017154. G.E. is partly supported by the Chancellor’s Graduate Fellowship Program at the Washington University in St Louis. T.W. and G.E. are also supported by NIH Grant R01HG007354, R01HG007175, R01ES024992 and American Cancer Society grant RSG-14-049-01-DMC. J.F.C. is also supported by NIH R01CA169316.

**Author contributions**


**Additional information**

**Accession codes**: All high-throughput sequencing data have been deposited in the NCBI Gene Expression Omnibus under project accession code GSE16368.

**Supplementary Information** accompanies this paper at http://www.nature.com/naturecommunications

**Competing financial interests**: The authors declare no competing financial interests.

**Reprints and permission** information is available online at http://npg.nature.com/reprintsandpermissions/


---

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/