Genomic landscape of human allele-specific DNA methylation

Fang Fang[a], Emily Hodges[b], Antoine Molaro[b], Matthew Dean[a], Gregory J. Hannon[b], and Andrew D. Smith[a,1]

[a]Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, California; and [b]Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724

Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved March 8, 2012 (received for review January 24, 2012)

DNA methylation mediates imprinted gene expression by passing an epigenomic state across generations and differentially marking specific regulatory regions on maternal and paternal alleles. Imprinting has been tied to the evolution of the placenta in mammals and defects of imprinting have been associated with human diseases. Although recent advances in genome sequencing have revolutionized the study of DNA methylation, existing methylome data remain largely untapped in the study of imprinting. We present a statistical model to describe allele-specific methylation (ASM) in data from high-throughput short-read bisulfite sequencing. Simulation results indicate technical specifications of existing methylome data, such as read length and coverage, are sufficient for full-genome ASM profiling based on our model. We used our model to analyze methylomes for a diverse set of human cell types, including cultured and uncultured differentiated cells, embryonic stem cells and induced pluripotent stem cells. Regions of ASM identified most consistently across methylomes are tightly connected with known imprinted genes and precisely delineate the boundaries of several known imprinting control regions. Predicted regions of ASM common to multiple cell types frequently mark noncoding RNA promoters and represent promising starting points for targeted validation. More generally, our model provides the analytical complement to cutting-edge experimental technologies for surveying ASM in specific cell types and across species.

Genomic imprinting refers to genes that are preferentially expressed from either the maternal or paternal allele without genotype dependence (1). In mammals, such parent-of-origin gene expression is believed to have evolved along with the placenta, serving to mediate resource distribution between a mother and her offspring (2, 3), though other theories have been proposed (4–6).

The connection between imprinting and DNA methylation was uncovered shortly after the first identification of imprinted genes in mammals (7). Imprinted gene expression, in all known cases, is regulated by allele-specific methylation (ASM) of some *cis*-acting regulatory regions. We use the term allelically methylated region (AMR) in reference to any genomic interval of ASM, whether or not it is associated with imprinted regulation. Typically, an entire imprinted locus is organized as a cluster and regulated by an imprinting control region (ICR) and several other AMRs. The allelic methylation patterns of ICRs are set during gametogenesis and stably maintained throughout somatic development in the offspring (8), irrespective of gene expression levels. The remaining AMRs may be established after fertilization (9), possibly under the control of nearby ICRs or other epigenetic signals.

The identification of imprinted genes and a detailed understanding of their regulation has become increasingly important, along with the realization that aberrant genomic imprinting contributes to several complex diseases (10). Much effort has been directed toward locating imprinted genes using expression screen-based approaches (11, 12). One limitation of such approaches is that many imprinted genes may only show allele-specific expressions in particular tissues at appropriate developmental stages (13). ASM screen-based approaches might overcome the effect of temporal and spatial expression patterns because the ICRs are expected to exist through developmental stages preceding the context in which they become active. Such methods have been successfully applied to identify unique imprinted genes (14–17).

Advances in DNA sequencing technology have been leveraged for high-throughput identification of imprinted genes. The "BS-seq" technology couples bisulfite treatment with high-throughput short-read sequencing, and has enabled genome-wide profiling of DNA methylation in mammalian genomes at single-CpG (cytosine guanine dinucleotide) resolution (18). Li et al. (19) produced a methylome from peripheral blood of a single individual and recognized the potential of using such data to profile ASM. They employed a method based on associating heterozygous SNPs with differential methylation, and identified hundreds of ASM regions. Methods such as this, however, must be applied to data from a single individual and for which matching genotypic data are available. There are two shortcomings of approaches that depend on genotype. First, they can be confounded by ASM that is associated with genotype, but which may not have any regulatory effect. The amount of ASM typically associated with genotype is not well understood, but recent reports suggest it is significant (20). More importantly, because imprinted methylation is not necessarily associated with genotypic variation, these methods will be inherently blind to some portion of ASM.

We present a probabilistic model to describe ASM based on data from BS-seq experiments. Our model is independent of genotype, and therefore has broad applicability to identify ASM in the context of imprinting. In essence, our model describes the degree to which methylation states in reads appear to reflect two distinct patterns, each pattern representing roughly half the data. We validated our method using semisimulated data in which methylation states were simulated within actual reads from BS-seq experiments. Our results indicate that technical characteristics of existing public methylomes (i.e., read length and coverage) are sufficient to accurately identify AMRs. By applying our model to 22 human methylomes, emphasizing those from uncultured cells, we identified a set of candidate AMRs involved in imprinted gene regulation. Candidates consistently identified across methylomes display remarkable concordance with known imprinted genes and allow boundaries of known AMRs to be precisely defined. Many candidates not associated with known imprinted genes mark the promoters of long noncoding RNAs (lncRNAs) and are also supported by similar analyses at orthologous regions in chimp; these provide a starting point for identifying additional imprinted genes, ICRs, and possibly imprinted clusters. Our model, therefore, is an essential analytical complement to recently emerged experimental methods for understanding the role of DNA methylation in genomic imprinting.

Author contributions: F.F., E.H., A.M., M.D., G.J.H., and A.D.S. designed research; F.F., E.H., A.M., and A.D.S. performed research; F.F. and A.D.S. analyzed data; and F.F. and A.D.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: andrewds@usc.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1201310109/-/DCSupplemental.

Modeling Allele-Specific Methylation in BS-Seq Data

We begin this section with a verbal description of our question and the main issues that are addressed by our model. We assume any read has been sequenced after bisulfite treatment and mapped uniquely to the reference genome. Because we are interested in mammalian methylation, we restrict our attention to CpG sites both in the genome and in the reads. Reads not mapping over a CpG are ignored. Our goal is to identify intervals of the genome where it appears that the two alleles have different methylation patterns—typically, in such a case, one allele will be highly methylated and the other not. There are two kinds of important information our model must capture: (*i*) The set of reads mapping into the interval should appear to represent two distinct methylation patterns, and (*ii*) the subsets of reads corresponding to those two patterns should be in roughly equal proportions because the alleles themselves are present in equal proportions. One can consider a methylation pattern as analogous to a haplotype, but with a strong stochastic component. Therefore, reads that contain only a single CpG will provide us with relatively little information, and we would like reads to cover as many CpGs as possible. We can then ask whether neighboring CpG sites on the same read tend to share methylation states, and whether other reads cover the same CpG sites but with the alternative shared methylation state. Our approach is to apply a single-allele model to the data, then apply an allele-specific model, and to compare the fit for these models to determine if the data support ASM.

Modeling Site-Specific DNA Methylation in a Single Allele. We associate each CpG with a single parameter indicating the probability that the CpG is methylated in the cells of interest. For a genomic interval containing n CpGs, the single-allele model is $\Theta = (\theta_1, \ldots, \theta_n)$. Given a set of reads R, the likelihood in the single-allele model within the interval is

$$L_1(\Theta|R) = \Pr(R|\Theta) \propto \prod_{i=1}^{n} \theta_i^{m(R,i)}(1-\theta_i)^{u(R,i)}, \qquad [1]$$

where $m(R,i)$ and $u(R,i)$ give the numbers of methylated and unmethylated observations from reads mapping over the ith CpG. Estimates for each θ_i are obtained assuming a binomial distribution for methylation states $m(R,i)$.

Modeling Regions of Allele-Specific Methylation. Within regions of ASM, we use a two-allele model that associates two distinct methylation probabilities with each CpG. Assuming there are n CpGs in the genomic interval, the two-allele model has the structure $\Theta = \{(\theta_{11}, \theta_{12}), \ldots, (\theta_{n1}, \theta_{n2})\}$, with θ_{i1} and θ_{i2} representing the methylation probabilities at the ith CpG on allele one and allele two, respectively. Under this model, reads mapping over the same genomic CpG may have different probabilities of methylation for their CpGs depending on the allele from which they originate. The allele of origin for any read is missing data, and for a given set R of reads we express this missing data as the partition $\gamma = \{\gamma_1, \gamma_2\}$ defined by $R = \gamma_1 \cup \gamma_2$, where $|R| = |\gamma_1| + |\gamma_2|$. For any $r \in R$, if $r \in \gamma_j$, we say that r originates from allele j. Because we are modeling alleles in the context of data from a diploid cell population, the probability that any read originates from a given allele is 0.5. Thus the likelihood is

$$L_2(\Theta|R,\gamma) = \Pr(R|\gamma,\Theta)\Pr(\gamma) \qquad [2]$$

because the partition γ is independent of Θ. The probability $\Pr(\gamma)$ is effectively a prior on the read partition assuming

$$|\gamma_1| \sim \text{binomial}(|\gamma_1| + |\gamma_2|, 0.5)$$

because each allele is present with equal frequency. Therefore,

$$L2(\Theta|R,\gamma) = \binom{|R|}{|\gamma_1|} 0.5^{|R|} \prod_{i=1}^{n}\prod_{j=1}^{2} \theta_{ij}^{m(\gamma_j,i)}(1-\theta_{ij})^{u(\gamma_j,i)}, \qquad [3]$$

where the m and u are as defined for Eq. **1**. Because the allele of origin for each read is missing data, we fit the two-allele model using expectation maximization (21), obtaining expectations on membership in γ_1 and γ_2. Details are provided in *SI Text*.

Identifying Intervals of Allele-Specific Methylation. We use Bayesian information criterion (BIC) (22) as a model selection criterion in determining whether a fixed interval is best described using a single-allele [Eq. **1**] or two-allele [Eq. **2**] model. A single-allele model has one parameter for each of the n CpGs, and the number of observations is equal to $|R|$:

$$\text{BIC(single)} = n\ln|R| - 2\ln L_1(\Theta|R). \qquad [4]$$

For the two-allele model, there are two parameters for each CpG:

$$\text{BIC(pair)} = 2n\ln|R| - 2\ln L_2(\Theta|R,\gamma). \qquad [5]$$

An interval is identified as having allele-specific methylation if and only if BIC(pair) < BIC(single).

We identify regions of ASM genome-wide by using a fixed-width sliding window (i.e., fixed number of CpG sites) and determining for each whether the single- or two-allele model better describes the data. Results we present are based on a sliding window of 10 CpGs, and issues related to selecting a window size are discussed in the *SI Text*. Intervals in close proximity are merged, and we also excluded intervals overlapping large subunit ribosomal RNA (LSU rRNA) genes from our final analyses because we suspected problems with their assembly in the reference genome (see *SI Text*).

Semisimulated Allele-Specific Methylation Data

We conducted simulations to evaluate how the performance of our model relates to several critical parameters of the underlying dataset. To reflect performance characteristics on real datasets, we used a strategy called "semisimulated" data. The locations of mapped reads were taken from real data, as were the locations of CpGs within reads and the underlying reference genome. The methylation states inside those reads were determined according to randomly generated allele-specific or single-allele methylation profiles. Briefly, within a region designated as an AMR, we randomly generated two methylation profiles by sampling individual CpG methylation levels as βeta variants skewed toward 0 or 1. Then we assigned each read with equal probability to one of the two alleles, and the methylation states of the CpGs within the read were sampled according to probabilities given by the methylation profile corresponding to that allele. A full description of this procedure is provided in the *SI Text*.

With current methylomes from BS-seq, we expected the variation in coverage along chromosomes to be a critical factor for the performance of our model. In addition, the variation in inter-CpG distance may prevent our method from capturing ASM in regions of low CpG density for a fixed read length. We examined how well our method could identify ASM in a given genomic interval by manipulating three independent variables:

- Mean coverages were {5×, 10×, 15×}, corresponding to current methylomes from BS-seq.
- Read lengths were {50, 100, 150} bases corresponding roughly with current short-read sequencing technologies.
- CpG density distributions took three different settings: CpG islands (CGIs) defined as in ref. 23, non-CGI promoters defined as 1 kb upstream of transcription start site (TSS) in National Center for Biotechnology Information reference sequences but not CGIs, and randomly sampled genomic background with CpG density (observed/expected) between 0.2 and 0.4.

Details concerning the number of simulated datasets for each parameter combination can be found in the *SI Text*.

Specificity was generally very high (approximately 99%) for all simulation parameter combinations, reflecting our conservative model selection criterion (Eqs. **4** and **5**). In contrast, sensitivity showed greater dependence on properties of the datasets. Sensitivity was higher for regions of higher CpG density, as expected because our model depends on the relationships between CpG states inside a read. As shown in Fig. 1, inside CGIs sensitivity reached above 95% for all read lengths when the mean coverage was above 10×. Sensitivity reached approximately 70% for intergenic regions but required both 10× coverage and read length 100, which compensates for the decrease in CpG density. As expected, greater coverage and read length improved accuracy, and the effect of read length is equivalent to that of CpG density. These results indicate that methylomes with read lengths around 100 bp and mean coverage above 10× appear sufficient for our model to accurately identify ASM. These criteria are met by most existing methylomes from BS-seq experiments.

Properties of the Methylomes Analyzed

We analyzed 22 publicly available methylomes, including five uncultured primary cell types, eight cultured differentiated cell lines, four embryonic stem cells (ESCs), and five induced pluripotent stem cells (iPSCs) from following studies. Additional details about each of the methylomes can be found in Dataset S1.

Hodges et al. (24) produced four uncultured methylomes from blood cells: hematopoietic stem and progenitor cells (HSPC), B cells, neutrophils, and CD133+ cord blood cells. Because the first three samples were pooled from six unrelated individuals, ASM caused by genetic variants should not be apparent due to the effect of pooling. The last sample was generated from one individual. Li et al. (19) produced the other uncultured methylome from peripheral blood mononuclear cells (PBMC) of one individual.

The study of Laurent et al. (25) produced three methylomes: foreskin fibroblasts, H9 ESCs, and fibroblasts derived from H9 ESCs. Lister et al. (18) produced methylomes for IMR90 cells and H1 ESCs, two replicates each which we treat as distinct methylomes. In a separate study, Lister et al. (26) produced methylomes for 10 cell types. Included among these were H9 ESCs, adipose-derived stem cells (ADS), adipocytes derived from ADS cells (ADS Adipose) and foreskin fibroblasts (FF). Induced pluripotent stem cells derived from ADS, IMR90, and FF cells were also profiled, with FF iPSCs taken at three different times, the last of which were also profiled after being differentiated in the presence of bone morphogenic protein 4 (BMP4).

Allele-Specific Methylation on the X Chromosome

Though not a form of imprinting, dosage compensation is associated with ASM differentially marking one chromosome X (chrX) in female somatic cells (27). In contrast, only a single allele from chrX is represented in male methylome data. Comparing the results of our analyses between male and female X chromosomes therefore provides a measure of specificity: AMRs identified on chrX in males are likely false-positives. In total, 12 of the analyzed methylomes are female. Although coverage on chrX in males is reduced by half, three male methylomes approached 10× coverage on chrX (H1 ESC rep 2, FF, and FF iPSC BMP4; refs. 18 and 26).

The locations of identified AMRs on chrX are presented in Fig. 2. The fraction of AMRs from chrX in female methylomes ranges from 15% to 36% with a mean of 24%. For the three male methylomes tested, the fraction is in the range of 1% to 2%. These results further support our conclusion from simulations that specificity is high in our AMR prediction. X chromosome inactivation is regulated via the X-inactive specific transcript (XIST) gene, a lncRNA with random allele-specific expression in female somatic cells. Our analyses identified an AMR at the XIST promoter in each female differentiated methylome (Fig. S1), but not in any of the ESCs, iPSCs, or male methylomes as expected (28).

Genome-Wide AMR Identification Predicts Imprinted Genes

The full set of identified AMRs for each of the 22 methylomes is presented in Dataset S1. We emphasized the uncultured blood methylomes in compiling sets of high-confidence AMRs, and generally use the remaining methylomes to provide additional supporting evidence. We found 579 autosomal AMRs that are common to at least three of the five uncultured methylomes (HSPCs, B cells, neutrophils, CD133+ cord blood, and PBMCs), 247 common to at least four out of five, and 81 shared across all five. Table 1 presents the 39 AMRs common to all five uncultured methylomes and that are proximal to promoters (±4 kb of a University of California Santa Cruz KnownGene TSS). Among these, 18 overlap a known imprinted gene and 20 mark a lncRNA promoter. The high concordance between our prediction and known imprinted genes further validates our model and provides strong support for the remaining predictions as candidate imprinted genes. The regulatory activity of lncRNAs has been observed for most imprinted clusters (29), and the frequent overlap of identified AMRs and lncRNA promoters suggests these might have similar activity (30). We also identified AMRs in low-coverage chimp blood cell data from the study of Hodges et al. (24), adding additional evidence to several of our predictions.

We computed the methylation level in sperm at each of the identified AMRs using data from a previous study (31). Among the 579 autosomal AMRs common to three out of five uncultured methylomes, 146 were methylated (>50%) in sperm. As indicated in Table 1, among the 39 predicted AMRs common across all uncultured cell types, only three are methylated in sperm. Among these is the H19 ICR, which is well known to be methylated on the paternal allele (32). If we use the methylation level in sperm as an indicator of methylation on the paternal allele, these results point to an asymmetry in the paternal and maternal mechanisms of imprinting DNA methylation.

One of the central questions related to the use of iPSCs, in research or therapeutically, is the degree to which they resemble true ESCs. The landmark study of Lister et al. demonstrated significant reprogramming variability between iPSCs (26). Evidence from cloning studies suggests that imprinting might be especially difficult to reprogram (33). We assembled the union of all identified AMRs in all methylomes. For each of these AMRs, we computed average methylation in each methylome. We then clustered the methylomes hierarchically according to correlation of

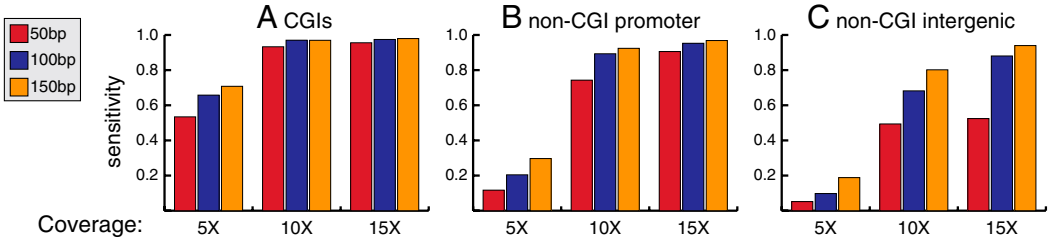


Fig. 1. Sensitivity of AMR identification based on semisimulated data. Coverages of 5, 10, and 15×, and read lengths of 50, 100, and 150 bp were used. CpG densities were controlled by simulating within (*A*) CGIs, (*B*) non-CGI promoter regions, and (*C*) non-CGI intergenic regions.

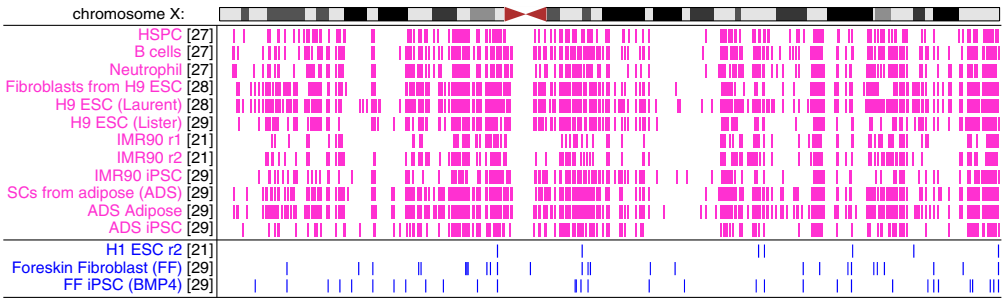


Fig. 2. Locations of AMRs identified on chrX. All female data (pink) was included. Only male methylomes (blue) with sufficient coverage on chrX are shown because these have coverage reduced by 50% compared with autosomes. Numbers in brackets indicate references for data sources. See text and Dataset S1 for information about methylomes.

methylation levels through these intervals (Fig. S2). The iPSCs correlated better with ESCs than with the somatic cells from which they are derived, suggesting that ASM has in general been successfully reprogrammed in these iPSCs.

However, we found several examples where the iPSCs appear to diverge from the ESCs in terms of ASM (Fig. S3). An AMR was identified at the GNAS-1 promoter in 18 of the methylomes, but this interval was hypomethylated in the ADS iPSCs. Similarly, for the AMR identified at the 3′ promoter of ZNF331 gene, the ADS iPSCs are methylated at 50% and resemble differentiated cells more closely than ESCs or other iPSCs, suggesting failed reprogramming of ADS iPSCs at these locations. It has been proposed that a single imprinted cluster might be sufficient to diagnose iPSC reprogramming in mouse (34). The diversity of ASM we observe between iPSCs and even between ESC lines suggests such diagnosis will be more complex in human.

Table 1. AMRs common to all uncultured cells that overlap gene promoters

Gene symbols	NC	CGI	Sp	ESC	iPSC	Tot	Chimp
* GNAS,…	Y	Y		4	5	22	3
* GNAS-AS1,…	Y	Y		4	5	22	3
* MESTIT1/MEST,…	Y	Y		4	5	22	3
* SGCE,PEG10		Y		4	5	22	3
NHP2L1		Y		4	5	22	3
* ZNF597,NAA60		Y		4	5	22	3
* SNRPN,SNURF		Y		4	5	22	2
* AMPD3,…	Y			4	5	22	0
PMF1-BGLAP		Y		4	5	22	0
LOC554226,…	Y	Y		4	5	22	0
UNC45B				4	5	22	0
LINC00273	Y	Y		4	5	22	0
* NAP1L5				4	5	21	3
KCNQ1OT1	Y	Y		3	5	21	0
LOC284801,…	Y	Y		4	4	21	0
* PSIMCT1	Y	Y		2	5	20	0
* H19,…	Y	Y	Y	2	5	20	0
TRAPPC9		Y		3	4	20	0
CR590796	Y			4	3	20	0
* DIRAS3		Y		2	5	19	2
AX748049				2	5	19	2
BC028329	Y			3	4	19	2
ZNF718,ZNF595		Y		3	5	19	0
BC023516	Y	Y		3	5	18	3
LOC100130522	Y	Y		2	4	18	3
* FANK1		Y		2	5	18	0
* GNAS	Y	Y		2	3	18	0
VTRNA2-1	Y	Y		3	4	17	3
MTRNR2L3			Y	4	1	15	0
* BLCAP,NNAT		Y		2	2	14	2
LOC728024	Y			0	3	14	2
RPS2P32	Y	Y		0	3	14	2
* ZIM2,PEG3,…	Y	Y		1	0	12	0
* MEG3		Y		0	0	11	3
LOC100132167	Y			1	0	11	0
* HOXA6,HOXA5,…		Y	Y	0	0	8	2
KIAA0934,DIP2C		Y	Y	0	0	8	2
LOC440570,…		Y		0	0	5	3
LOC100335030		Y		0	0	5	1

Columns indicate whether the gene is noncoding (NC), the AMR overlaps a CGI promoter (CGI), or is hypermethylated in sperm (Sp). Counts indicate the number of ESC, iPSC, total human methylomes, and chimp methylomes in which the AMR is found. Ellipsis indicate additional gene names can be found in Dataset S1. Asterisk * indicates known imprinted gene.

Analysis of Known Imprinting Control Regions

There are approximately 65 human genes currently validated as imprinted and these reside in 32 imprinted clusters (see Dataset S1). We asked for what proportion of these clusters do we identify an AMR shared between cells, and do these shared AMRs coincide with experimentally validated AMRs? As can be seen from Table 2, 24 of the clusters contain validated AMRs, and in 21 of those cases we correctly identify a known AMR common to four out of five uncultured cells. For the IGF2R and INPP5F clusters, we only identified AMRs shared between two and three of the uncultured cells, respectively. The AMPD3 gene has no validated AMR to our knowledge. Our algorithm finds an AMR shared across all 22 methylomes, indicating a likely candidate for validation. To our knowledge, no AMRs have yet been identified for the remaining clusters, and our algorithm fails to predict any AMRs that are shared between methylomes.

Knowledge of the location of true AMRs around several of these imprinted genes allowed us to apply a more intensive analysis to examine them with greater sensitivity and precision. We designed a dynamic programming algorithm to optimize the locations of AMR boundaries by evaluating each possible AMR size rather than joining overlapping sliding windows. This algorithm uses a scoring function based on the likelihoods of Eqs. **1** and **2** but remains too computationally demanding for genome-wide application (details are provided in the *SI Text*). We refer to AMRs identified with this algorithm as "refined" AMRs.

The imprinted cluster on chr14 consists of seven genes controlled by the maternally expressed lncRNA MEG3. The region harbors an AMR at the MEG3 promoter, and another intergenic AMR approximately 15 kb upstream of the MEG3 TSS, both paternally methylated with the upstream AMR shown to act as an ICR (35). Our genome-wide scan found the MEG3 promoter AMR in 11/13 differentiated cells. The boundaries of refined AMRs were identified in each uncultured methylome at nearly the exact same location, covering an interval that is hypomethylated in sperm (Fig. S4). A refined AMR was identified in each uncultured methylome precisely at the known ICR location, which is methylated in sperm. Interestingly, each of the ESC/iPSC methylomes shows full methylation through the ICR, suggesting possible imprinting defects in these cells.

Imprinted expression in the GNAS locus is highly complex, with maternally, paternally, and biallelically expressed transcripts sharing sets of exons (36). This locus includes four AMRs at alternative promoters (NESP55, GNAS-AS1, XLαs, Exon A/B)

Table 2. Imprinted clusters and associated AMRs

Cluster	Known	ICR	Unc	ESC	iPSC	Tot	Chimp
GNAS	Y	Y	5	4	5	22	3
SGCE/PEG10	Y	Y	5	4	5	22	3
MESTIT1/MEST	Y	Y	5	4	5	22	3
ZNF597/NAA60	Y	Y	5	4	5	22	2
SNRPN/SNURF	Y	Y	5	4	5	22	1
AMPD3			5	4	5	22	0
NAP1L5	Y	Y	5	4	5	21	3
KCNQ1OT1	Y	Y	5	3	5	21	0
PSIMCT-1/HM13	Y	Y	5	2	5	20	3
KCNK9	Y	Y	5	3	4	20	3
INS-IGF2-H19	Y	Y	5	2	5	20	0
DIRAS3	Y	Y	5	2	5	19	3
ZDBF2	Y	Y	5	3	4	19	2
FANK1	Y		5	2	5	18	0
BLCAP/NNAT	Y	Y	5	2	2	14	2
ZIM2/PEG3	Y	Y	5	1	0	12	0
DLK1/MEG3	Y		5	0	0	11	3
RB1	Y	Y	5	0	0	7	3
L3MBTL	Y	Y	4	3	5	19	3
DDC/GRB10	Y	Y	4	3	4	19	3
PLAGL1/HYMAI	Y	Y	4	2	4	18	3
FAM50B	Y	Y	4	0	1	12	3
TCEB3C	Y	Y	4	1	1	11	0
INPP5F	Y	Y	3	3	5	18	2
IGF2R	Y		2	0	4	13	1
DXLGAP2			1	1	0	2	0
TP73			1	0	0	1	0
ANKRD11			1	0	0	1	0
DLX5			0	0	0	2	0
ABCA1			0	0	0	1	0
WT1			0	0	0	0	0
RBP5			0	0	0	0	0

Columns indicate whether the AMR was previously known, an ICR, and the number of uncultured (Unc), ESCs, iPSCs, total (Tot), and chimp methylomes in which each AMR was found. Genomic locations of AMRs can be found in Dataset S1

(37). We identified refined AMRs at these locations in all uncultured methylomes (Fig. 3). Between different methylomes, boundaries of refined AMRs fluctuated by fewer than 10 CpGs and frequently were identified at identical locations. In each case, two separate refined AMRs were identified at the GNAS-AS1 and XLαs promoters. The consistent location of the refined AMR boundary between the GNAS-A/B and GNAS-1 TSS, which coincides with the center of a CGI, suggests a strict partition of regulatory sequence between these two transcripts.

The LTR-derived PEG10 and the adjacent SGCE are part of an imprinted gene cluster on chr7 sharing complete synteny with imprinted orthologs in mouse (38). PEG10 and SGCE are separated by less than 100 bp, are divergently transcribed, and have a single CGI overlapping both TSS. Our analyses revealed an AMR at their shared promoter in all 22 methylomes, with a positional bias in the direction of PEG10. As can be seen from Fig. 3, the refined AMRs in uncultured methylomes have identical boundaries precisely between the PEG10 and SGCE TSS at the center of a CGI similar to the GNAS-A/B case described above. Each refined AMR is fully contained inside the body of PEG10, consistent with the LTR origin of PEG10, which implies that PEG10 carries internal regulatory elements. This internal PEG10 promoter appears responsible for imprinted regulation of both genes, despite the hypomethylation reaching into SGCE in all methylomes. One plausible scenario is that regulatory elements within the AMR interact with those nearby in the hypomethylated portion of the CGI to regulate SGCE.

Discussion

We presented a computational strategy for identifying ASM in methylomes produced by BS-seq technology. Our method does not depend on the existence of genotypic variation and is therefore able to identify ASM associated solely with parent-of-origin. Results on simulated data indicate that our method has generally high specificity, and that sensitivity increases with read length as well as mean coverage throughout the genome. Our results also show that this model is accurate even for current read lengths and depths of coverage, both of which are critical technical parameters in connecting methylation states of individual molecules. We applied our method to 22 publicly available human methylomes and validated its accuracy on real data by comparing ASM identified on female and male X chromosomes. Our most consistent predictions across methylomes showed high concordance with known AMRs associated with imprinted genes. The remaining predictions represent likely candidate ICRs for imprinted loci, with several overlapping lncRNA promoters and supported by similar analysis at orthologous regions based on low-coverage methylomes from chimp.

Our top predicted autosomal AMRs show remarkable concordance with known AMRs controlling imprinted gene expression. Among the 39 common to all uncultured methylomes and prox-

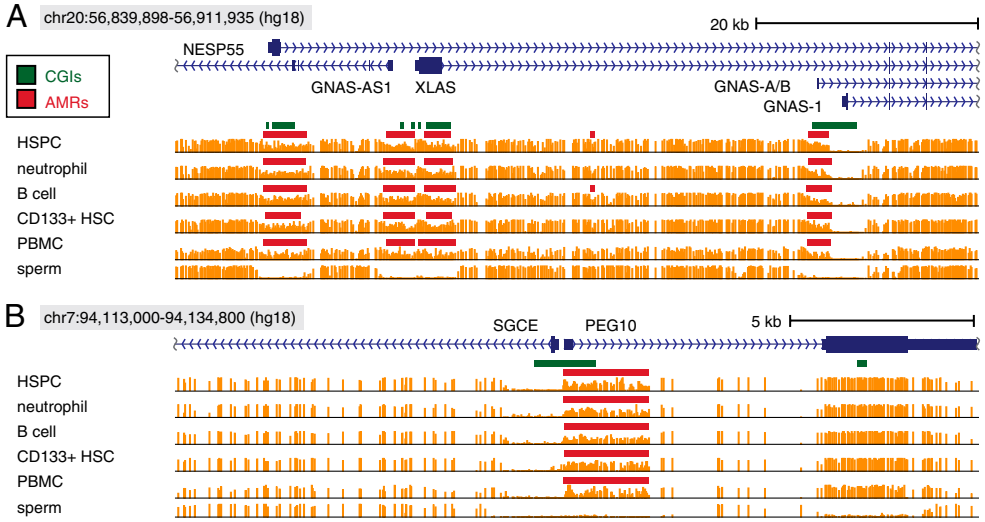


Fig. 3. Regions of allele-specific methylation through (*A*) the GNAS and (*B*) SGCE/PEG10 loci in five uncultured blood cells (HSPC, neutrophils, B cells, and CD 133+ cord blood cells). Vertical orange bars indicate methylation levels of CpGs. In both examples refined AMRs show highly consistent boundaries across methylomes, and each includes an AMR with a precise boundary inside a CGI, distinguishing the regulatory regions of distinct TSS.

imal to annotated promoters, 18 are marking known imprinted genes. It appears as though the AMRs that are already known are also those identified most consistently across methylomes. This finding can be interpreted in several ways. One possibility is that a significant portion of the imprinted genes or clusters, possibly more than half, have already been identified. Estimates of the total number of imprinted genes in human hover around 100–200 (39, 40), and many parent-of-origin disease phenotypes have already been explained through known imprinted genes. Among our remaining predictions (several hundred putative AMRs) many could represent weaker ASM signals possibly without functional relevance, and these are identified with less consistency across datasets because of a lack of sensitivity in our method. Another possibility is that many genes are imprinted with cell-type specificity, and that the known AMRs are biased toward those that can be identified in a greater variety of cell types. Our top predictions were based on consistency across the available methylomes from uncultured cells, but these all happened to be from blood.

Another important finding to emerge from our analyses is the precision with which AMRs are defined across cell types. The GNAS and SGCE/PEG10 examples illustrate this strong consistency of AMR boundaries: In both of these examples, there are pairs of TSS in very close proximity, sharing CGIs, but for which one has ASM methylation and the other does not. Methods like ours that can delineate the boundaries of AMRs will assist future efforts to precisely map the elements inside these regulatory regions.

Among the 22 methylomes analyzed, the total number of identified AMRs varied substantially across methylomes. Much of this variation is likely due to variation in coverage and number of CpGs/read, and relates to sensitivity as illustrated in our simulations. A substantial part may also be due to epigenotypic variation between the cells: Those from culture will exhibit the associated epigenomic effects. It has been known for some time that permanent cell lines have altered DNA methylation (41, 42), and more recently it was shown that aberrant methylation is correlated with passage number (43). We also observed some examples which might indicate an effect of culture conditions, such as the MEG3 AMRs (Fig. S4) where cells differentiated in culture show different properties from other somatic cells. Moreover, any methylome datasets derived from single individuals might exhibit the effects of genotype (44).

Combining experimental approaches with analytic methods such as ours on appropriate biological samples will reveal the breadth of cell-type specific ASM in humans and further clarify the mechanisms underlying imprinted gene expression in human diseases. Moreover, conducting similar studies in appropriately selected mammalian species will help to elucidate the enigmatic role of sexual conflict influencing mammalian evolution.

ACKNOWLEDGMENTS. The authors thank Mike Waterman, Simon Tavaré, Fengzhu Sun, and members of the Smith, Hannon, and Dean labs for helpful discussions. The work was supported in part by grants from the National Institutes of Health (R01HG005238 and 5P01CA013106-45) and by a kind gift from Kathryn W. Davis.

1. Barlow DP, Stöger R, Herrmann BG, Saito K, Schweifer N (1991) The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. *Nature* 349:84–87.
2. Haig D, Westoby M (1989) Parent-specific gene expression and the triploid endosperm. *Am Nat* 134:147–155.
3. Moore T, Haig D (1991) Genomic imprinting in mammalian development: a parental tug-of-war. *Trends Genet* 7:45–49.
4. Barlow DP (1993) Methylation and imprinting: From host defense to gene regulation? *Science* 260:309–310.
5. Varmuza S, Mann M (1994) Genomic imprinting-defusing the ovarian time bomb. *Trends Genet* 10:118–123.
6. Pardo-Manuel de Villena F, de la Casa-Esperón E, Sapienza C (2000) Natural selection and the function of genome imprinting: Beyond the silenced minority. *Trends Genet* 16:573–579.
7. Zhang Y, et al. (1993) Imprinting of human H19: Allele-specific CpG methylation, loss of the active allele in Wilms tumor, and potential for somatic allele switching. *Am J Hum Genet* 53:113–124.
8. Davis TL, Yang GJ, McCarrey JR, Bartolomei MS (2000) The H19 methylation imprint is erased and re-established differentially on the parental alleles during male germ cell development. *Hum Mol Genet* 9:2885–2894.
9. El-Maarri O, et al. (2001) Maternal methylation imprints on human chromosome 15 are established during or after fertilization. *Nat Genet* 27:341–344.
10. Monk D (2010) Deciphering the cancer imprintome. *Brief Funct Genomics* 9:329–339.
11. Nikaido I, et al. (2003) Discovery of imprinted transcripts in the mouse transcriptome using large-scale expression profiling. *Genome Res* 13:1402–1409.
12. Pollard KS, et al. (2008) A genome-wide approach to identifying novel-imprinted genes. *Hum Genet* 122:625–634.
13. Deltour L, Montagutelli X, Guenet JL, Jami J, Páldi A (1995) Tissue-and developmental stage-specific imprinting of the mouse proinsulin gene, Ins2. *Dev Biol* 168:686–688.
14. Peters J, et al. (1999) A cluster of oppositely imprinted transcripts at the Gnas locus in the distal imprinting region of mouse chromosome 2. *Proc Natl Acad Sci USA* 96:3830–3835.
15. Smith RJ, Dean W, Konfortova G, Kelsey G (2003) Identification of novel imprinted genes in a genome-wide screen for maternal methylation. *Genome Res* 13:558–569.
16. Shoemaker R, Deng J, Wang W, Zhang K (2010) Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res* 20:883–889.
17. Choufani S, et al. (2011) A novel approach identifies new differentially methylated regions (DMRs) associated with imprinted genes. *Genome Res* 21:465–476.
18. Lister R, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462:315–322.
19. Li Y, et al. (2010) The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol* 8:e1000533.
20. Kerkel K, et al. (2008) Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat Genet* 40:904–908.
21. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B Stat Methodol* 39:1–38.
22. Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464.
23. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. *J Mol Biol* 196:261–282.
24. Hodges E, et al. (2011) Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol Cell* 44:17–28.
25. Laurent L, et al. (2010) Dynamic changes in the human methylome during differentiation. *Genome Res* 20:320–331.
26. Lister R, et al. (2011) Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* 471:68–73.
27. Migeon BR (1990) Insights into X chromosome inactivation from studies of species variation, DNA methylation and replication, and vice versa. *Genet Res* 56:91–98.
28. Wutz A (2011) Gene silencing in X-chromosome inactivation: Advances in understanding facultative heterochromatin formation. *Nat Rev Genet* 12:542–553.
29. O'Neill MJ (2005) The influence of non-coding RNAs on allele-specific gene expression in mammals. *Hum Mol Genet* 14(Suppl 1):R113–120.
30. Koerner MV, Pauler FM, Huang R, Barlow DP (2009) The function of non-coding RNAs in genomic imprinting. *Development* 136:1771–1783.
31. Molaro A, et al. (2011) Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell* 146:1029–1041.
32. Tremblay KD, Saam JR, Ingram RS, Tilghman SM, Bartolomei MS (1995) A paternal-specific methylation imprint marks the alleles of the mouse H19 gene. *Nat Genet* 9:407–413.
33. Rideout WM, Eggan K, Jaenisch R (2001) Nuclear cloning and epigenetic reprogramming of the genome. *Science* 293:1093–1098.
34. Stadtfeld M, et al. (2010) Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. *Nature* 465:175–181.
35. Kagami M, et al. (2008) Deletions and epimutations affecting the human 14q322 imprinted region in individuals with paternal and maternal upd(14)-like phenotypes. *Nat Genet* 40:237–242.
36. Williamson CM, et al. (2006) Identification of an imprinting control region affecting the expression of all transcripts in the Gnas cluster. *Nat Genet* 38:350–355.
37. Fröhlich LF, et al. (2010) Targeted deletion of the Nesp55 DMR defines another Gnas imprinting control region and provides a mouse model of autosomal dominant PHP-Ib. *Proc Natl Acad Sci USA* 107:9275–9280.
38. Ono R, et al. (2001) A retrotransposon-derived gene, PEG10, is a novel imprinted gene located on human chromosome 7q21. *Genomics* 73:232–237.
39. Barlow DP (1995) Gametic imprinting in mammals. *Science* 270:1610–1613.
40. Luedi PP, et al. (2007) Computational and experimental identification of novel human imprinted genes. *Genome Res* 17:1723–1730.
41. Doherty AS, Mann MRW, Tremblay KD, Bartolomei MS, Schultz RM (2000) Differential effects of culture on imprinted H19 expression in the preimplantation mouse embryo. *Biol Reprod* 62:1526–1535.
42. Fernández-Gonzalez R, et al. (2004) Long-term effect of in vitro culture of mouse embryos with serum on mRNA expression of imprinting genes, development, and behavior. *Proc Natl Acad Sci USA* 101:5880–5885.
43. Meissner A, et al. (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454:766–770.
44. Gertz J, et al. (2011) Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. *PLoS Genet* 7:e1002228.

Supporting Information

Fang et al. 10.1073/pnas.1201310109

SI Text

Expectation Maximization for the Two-Allele Model. When computing the likelihood according to our two-allele model, we require a partition $R = \gamma_1 \cup \gamma_2$ of the reads assigning each read to one of two alleles. This partition is missing information, and we infer the expected partition by assigning indicator variables for the events that individual reads have membership in γ_2. Assuming two alleles, and therefore two methylation probabilities for each CpG (cytosine guanine dinucleotide), we let θ_{i1} and θ_{i2} be the methylation probabilities at CpG i for allele 1 and 2, respectively. The read set R is partitioned into two subsets γ_1 and γ_2 according to the allele of origin for each read. When calculating the likelihood, the methylation probabilities are the parameters $\Theta = \{(\theta_{11}, \theta_{12}), \ldots, (\theta_{n1}, \theta_{n2})\}$. Let $\mu_i, i \in \{1, 2\}$ denote the probability that a read comes from allele i, so $\mu_1 = \mu_2 = 0.5$. We use the indicator functions $I_1(r_i)$ and $I_2(r_i) = 1\text{-}I_1(r_i)$ for events that r_i originated from allele 1 and allele 2, respectively. The complete data likelihood is

$$L(\Theta|R, \gamma) = \prod_{i=1}^{m}\prod_{j=1}^{2}(\mu_j \prod_{k=1}^{n} \theta_{kj}^{m(r_i,k)}(1\text{-}\theta_{kj})^{u(r_i,k)})^{I_j(r_i)} \quad [S1]$$

where $m(r_i, k)$ and $u(r_i, k)$ are indicators for the methylation state of the read r_i at the kth CpG, and we let $m(r_i, k) = u(r_i, k) = 0$ when the kth CpG is not covered by r_i.

The expectation (E) step updates the missing data γ with the observed data R and parameters Θ. We define p_{ji} as the probability that a read r_i comes from allele j. These p_{ji} are essentially the expected values of membership in the subsets γ_1 and γ_2 of the partition. Therefore, p_{ji} can be calculated as the ratio of the probability that the read r_i comes from the allele j and the sum of probabilities that the read comes from either allele. At the nth iteration,

$$p_{ji}^{(n)} = \Pr(I_j(r_i) = 1|R, \Theta) = \frac{\mu_j \prod_{k=1}^{n} \theta_{kj}^{m(r_i,k)}(1\text{-}\theta_{kj})^{u(r_i,k)}}{\sum_{j=1}^{2} \mu_j \prod_{k=1}^{n} \theta_{kj}^{m(r_i,k)}(1\text{-}\theta_{kj})^{u(r_i,k)}}$$

$$= \frac{\prod_{k=1}^{n} \theta_{kj}^{m(r_i,k)}(1\text{-}\theta_{kj})^{u(r_i,k)}}{\sum_{j=1}^{2} \prod_{k=1}^{n} \theta_{kj}^{m(r_i,k)}(1\text{-}\theta_{kj})^{u(r_i,k)}}, \quad [S2]$$

where the parameters on the right-hand side are as estimated in iteration $n - 1$. The maximization (M) step updates the parameters Θ to maximize the likelihood

$$\theta_{k1}^{(n+1)} = \frac{\sum_{i=1}^{m} p_{1i} m(r_i, k)}{\sum_{i=1}^{m} p_{1i}}, \quad [S3]$$

$$\theta_{k2}^{(n+1)} = \frac{\sum_{i=1}^{m} p_{2i} m(r_i, k)}{\sum_{i=1}^{m} p_{2i}}. \quad [S4]$$

With these expectation maximization (EM) steps, we can estimate values for all parameters $\Theta = \{(\theta_{11}, \theta_{12}), \ldots, (\theta_{n1}, \theta_{n2})\}$ and the probabilities for each read originating from either allele.

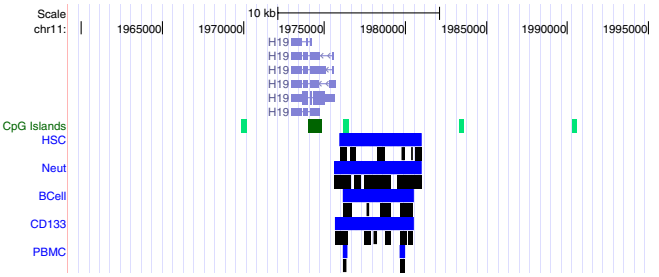
Preprocessing High-Throughput Short-Read Bisulfite Sequencing. For efficient computation, we took the following preprocessing steps before AMR identification.

- Bisulfite sequencing data was mapped with the RMAPBS software (1) after removing adaptor sequences.
- Only one read per mapping location was retained to eliminate bias from PCR duplicates.
- All paired-end reads having both ends map within 1,000 bp were merged as a single read, possibly including a spacer consisting of N characters.
- All reads were converted from genomic coordinates to CpG coordinates, and all non-CpG positions were removed form each read. The characters in the converted reads were C, T, and N, to indicate methylated, unmethylated, and unknown.
- Only reads with at least one non-N character were retained after the conversion.
- When processing reads, positions with N were ignored completely.

Issues Related to Selecting a Sliding Window Size. In selecting a window size, the two main considerations (other than computational speed) are (*i*) the window size must be small enough that allelically methylated region (AMR) boundaries are accurately identified with the desired resolution; (*ii*) to be large enough that we can leverage as much information as possible from the overlapping reads. In general, there is no single window size that will optimally identify AMRs through the entire genome, and different datasets likely will benefit most from using different window sizes (e.g., based on average CpGs per read, and total amount of data).

To select the window size of 10 CpGs, we tested windows of size 5, 10, 15, and 20 using the blood cell methylomes. We examined how these window sizes identify known AMRs in the H19, GNAS, SGCE, SNRPN, KCNQ1, ZIM2, and MEG3 loci.

When experimental technology produce longer reads, it is likely that a larger window size will capture a much greater amount of information about how the reads corresponding to the same allele overlap. However, using a larger window size will still blur boundaries of AMRs, and potentially will cause smaller AMRs to be missed. When a better gold standard training set exists, we will be in a better position to optimize parameters such as the window size.



Rationale for Merging Nearby AMR Fragments. As described, we identified AMRs by applying our model in a sliding window along the chromosomes and any identified AMR "fragments" that were adjacent were merged if they were within 1 kbp of each other. Some motivation for this procedure can be found in the figure above, which shows the difference between the AMRs before (black blocks) and after (blue blocks) merging for the blood methylomes at the H19 imprinting control region (ICR). In this case, due to fluctuations in coverage through the 5 kbp known

H19 ICR, several fragments of AMRs were identified initially, and after merging the intervals covered by the hematopoietic stem/progenitor, B-cell neutrophil and CD133+ cord blood were very similar. Using such a method will always fail to join nearby fragments if they are more distant than the cutoff, as illustrated for the peripheral blood mononuclear cells methylome (also in the figure above).

Removal of LSU-rRNA Genes from Predictions. We observed that several of our top identified AMRs (i.e., those most consistent across methylomes) overlapped LSU-rRNA genes. Such a finding would be consistent with reports of dosage compensation, analogous to X chromosome inactivation, for rRNA genes (2). However, we also noticed that the number of reads mapping over these regions was generally much more than in other top identified AMRs. BLASTing several of these in the National Center for Biotechnology Information nonredundant (NCBI) database revealed that most of them matched only one location in hg18, but matched additional locations in newer assemblies of chromosomes, frequently even newer than are included in hg19. We therefore decided to remove these from our predictions because we believe they are likely artifacts representing methylation states from multiple genomic intervals superimposed on a single interval.

Optimizing of Regions of Allele-Specific Methylation. Our genome-wide AMR identification was based on testing for allele-specific methylation in sliding windows along each chromosome. We also designed an algorithm that did not require a sliding window, allowing us to optimize the boundaries of the identified AMRs so that we might more precisely locate these boundaries. This algorithm is much more computationally expensive, and so it is not appropriate for genome-wide application. This method uses scores that are based on the likelihoods (described in the paper) for either one or two alleles, and is equivalent to testing all ways to partition of a genomic interval into alternating subintervals of allele-specific and single-allele methylation. We did not use Bayesian information criterion (BIC) in this method, but instead used a heuristic penalty term equal to a linear function of the number of reads inside the AMR to offset the difference in model complexity between the allele-specific and single-allele models. This criterion is similar to Akaike information criterion. Because of the logarithmic function in the BIC, it could not be computed incrementally in the dynamic programming recurrence presented below.

The effect of this different penalty term is increased sensitivity, but also decreased specificity. This method is only suitable for applying in regions where we have prior information telling us we should find an AMR, and our goal is to locate the boundaries of that AMR. The importance of this task is evident from examples, such as PEG10 (3, 4) and GNAS (5, 6) promoters (Fig. 3), where precise boundaries seem to distinguish allelic states of nearby promoters.

Let $L_2(i, j)$ denote the maximum likelihood of the two-allele model using only CpGs i through j as estimated by EM and let $L_1(i)$ denote the likelihood of the single-allele model computed only for the ith CpG. For CpG i, we use $\text{score}_1(i)$ to indicate the maximum likelihood of the interval $[1, i]$ assuming the ith CpG has single-allele methylation, and $\text{score}_2(i)$ is the maximum likelihood of the interval $[1, i]$ with the ith CpG as the end of an AMR. Assume the size distribution of non-AMR is a geometric distribution with parameter τ, and the size distribution of AMR (f_2) is arbitrary. Then we use the recurrences

$$\text{score}_2(i) = \max_{1 \le i' < i} \{\log L_2(i', i) + \log f_2(i - i') + \text{score}_1(i')\}, \quad \textbf{[S5]}$$

and

$$\text{score}_1(i) = \log L_1(i) + \max \begin{cases} \text{score}_2(i-1) + \log \tau, \\ \text{score}_1(i-1) + \log(1-\tau). \end{cases} \quad \textbf{[S6]}$$

to compute the maximal values of likelihoods for partial segmentations of the data up to each i. We record such score_1 and score_2 for each CpG. The estimated optimal value is found at the nth CpG and a traceback provides the precise locations of AMRs.

In practice we impose a minimum size (10 CpGs) on the AMRs and spaces between AMRs. The reason why the function f_2 is described as "arbitrary" above is because the value of score_2 cannot be built up incrementally, and each individual value of score_2 must be computed using EM. In this context no one duration distribution will lead to faster computation; because of this there is no speed benefit to using a geometric distribution for the sizes of AMRs in our scoring function. However, we did not evaluate other distributions and simply used a geometric distribution for f_2. The value of τ and the corresponding parameter for f_2 were set by assuming that the mean AMR size was 100 CpGs, and that the mean inter-AMR distance was 10,000 CpGs.

Semisimulated Data. We used a strategy that we call "semisimulated" data to reflect the coverage variance of the real sequencing data. All simulated reads took the locations of real data, and their methylation states were generated according to the simulated methylation probabilities of CpGs in the genome. For each CpG, we randomly generated two methylation profiles by sampling individual CpG methylation levels as Beta variants skewed toward 0 or 1 (e.g., Beta distribution with mean 0.75 for one allele and 0.25 for the other with variance also controlled). For CpGs designated within non-AMR, both alleles' methylation probability was set as one of the two profiles randomly. In this way, the average methylation level through a region was always roughly 0.5, even for single-allele simulations. Then we assigned each read with equal probability to one of the two alleles and the methylation states of the CpGs within the read were sampled according to probabilities given by the methylation profile corresponding to that allele. Mimicking the bisulfite conversion, all unmethylated read cytosines were converted to thymines.

In the simulation, we manipulated three independent variables: mean coverages, read lengths, and CpG density distributions. The mean coverages were {5×, 10×, 15×}, and the read lengths were {50, 100, 150} bases. All reads were taken from the human B-cell and neutrophil methylomes. Different CpG densities were taken from three sets of regions:

1. All CGIs defined in ref. 7, with mean size of 760 bp;
2. non-CGI promoters defined as 1 kb regions upstream of refSeq TSS but not CGIs;
3. non-CGI intergenic regions that were intergenic regions with CpG density (observed/expected) between 0.2 and 0.4 and mean size of 1,457 bp.

For each combination of variables, 100 regions were randomly selected from one of the three sets. Then each region was simulated as AMR and non-AMR 10 times, respectively. In total, there were $2,000 = 2 \times 10 \times 100$ data points in one simulation. To calculate the variances of specificity and sensitivity, we repeated the simulations 100 times for each variable combination.

Estimating False-Discovery Rate (FDR) Using Semisimulated Data. We used the idea of semisimulated data to obtain bounds on false-discovery rate for the five blood methylomes analyzed. Our procedure was as follows. Using the real data from reads, we randomly shuffled methylation states corresponding to each CpG site. In other words, the methylation states were collected from all reads mapping over a specific CpG site, and then randomly permuted before being assigned back to those reads. This simulation preserves exactly the likelihood for any interval under our

single-allele model. We used chr10, and we did 1,000 such random experiments for each of the five blood methylomes, which provided a false-positive rate (type I error rate) that can be used to bound the FDR.

Cell type	AMRs identified		
	In real data	In random data	Estimated type I error rate
Neutrophil	133	0.009	6.8e−05
B cell	160	0.008	5.0e − 05
Hematopoietic stem/progenitor	132	0.038	0.00029
CD133+ cord blood	138	0.008	5.8e−05
Peripheral blood mononuclear cell (PBMC)	58	0.035	0.0006

Because in each case above the number of AMRs identified under our null hypothesis is less than 0.1, we may estimate an upper bound on the FDR as $0.1/x$, where x is the number of AMRs identified. In all cases, this simulation would result in an FDR of less than 0.01.

Caveat: The major caveat associated with estimating an FDR in the way we have above has to do with the underlying biology. Cell populations grow as mixtures of clones. DNA methylation has a stochastic component that remains poorly understood. At the same time, any stochastic changes in methylation will be preserved due to the mitotic inheritance of the methylation. Therefore, any real methylome will likely by chance contain intervals that truly represent a mixture of two different methylation profiles, yet these may be associated with absolutely no biological function (according to our current understanding).

The best way to ensure that identified AMRs are not spurious, therefore, is to analyze replicate experiments where the cells are grown or purified separately. In the case of the methylomes we have analyzed, each comes from a very different population of cells, and therefore AMRs that overlap between cell types should be absent from the intersection of the AMR sets.

1. Smith AD, et al. (2009) Updates to the rmap short-read mapping software. *Bioinformatics* 25:2841–2842.
2. Schlesinger S, Selig S, Bergman Y, Cedar H (2009) Allelic inactivation of rDNA loci. *Genes Dev* 23:2437–2447.
3. Ono R, et al. (2001) A retrotransposon-derived gene, PEG10, is a novel imprinted gene located on human chromosome 7q21. *Genomics* 73:232–237.
4. Ono R, et al. (2005) Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. *Genomics* 38:101–106.
5. Williamson CM, et al. (2006) Identification of an imprinting control region affecting the expression of all transcripts in the Gnas cluster. *Nat Genet* 28:350–355.
6. Fröhlich LF, et al. (2010) Targeted deletion of the Nesp55 DMR defines another Gnas imprinting control region and provides a mouse model of autosomal dominant PHP-Ib. *Proc Natl Acad Sci USA* 107:9275–9280.
7. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. *J Mol Biol* 196:261–282.

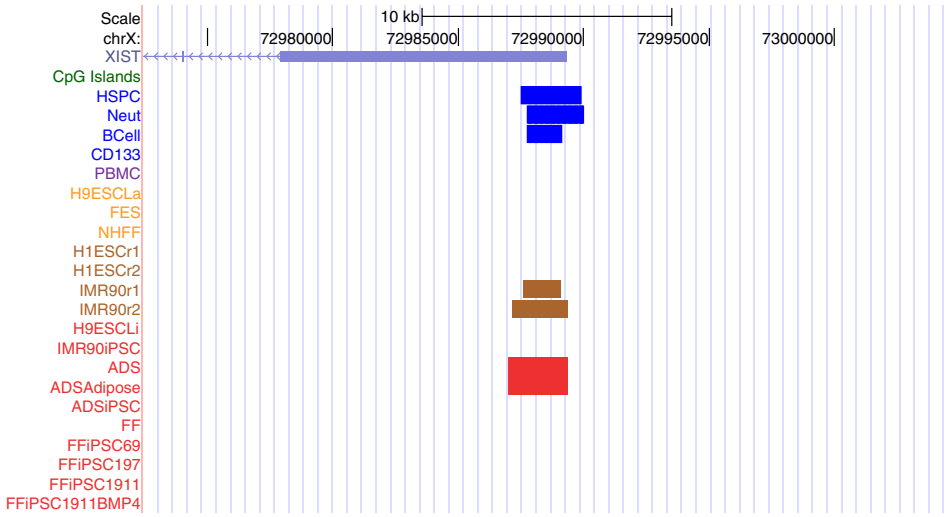


Fig. S1. Allele-specific methylation identified at the XIST promoter. Consistent with earlier findings, allele-specific methylation is found in exactly those methylomes that are (1) female, and (2) not from ESCs or iPSCs.

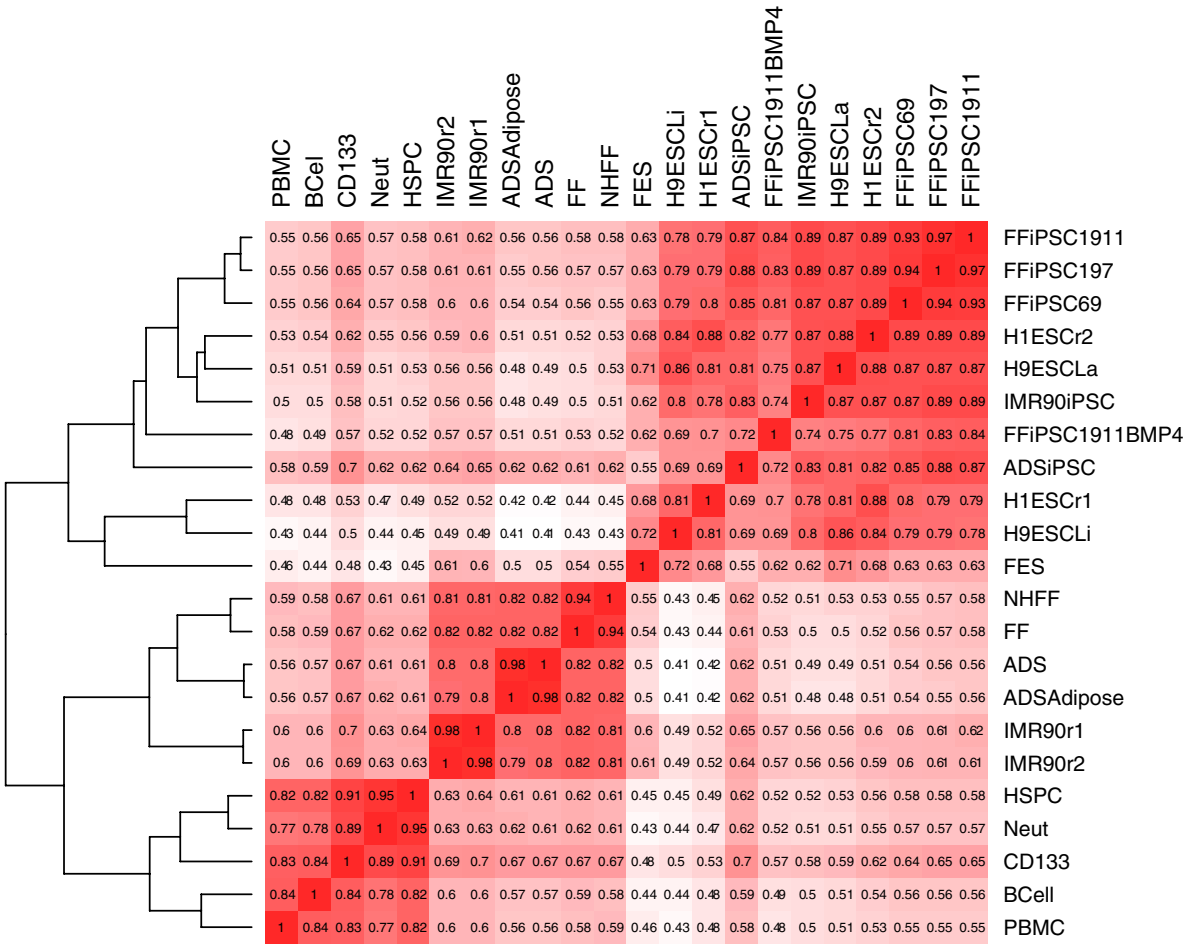


Fig. S2. Clustering of all 22 methylomes according to their methylation patterns in all identified AMRs. The numbers in cells indicate the correlation of methylation patterns between two cell types, and a higher number corresponds to a darker color. Basically, three clusters are formed: (*i*) ESCs/iPSCs; (*ii*) cultured differentiated cells; (*iii*) uncultured cells.

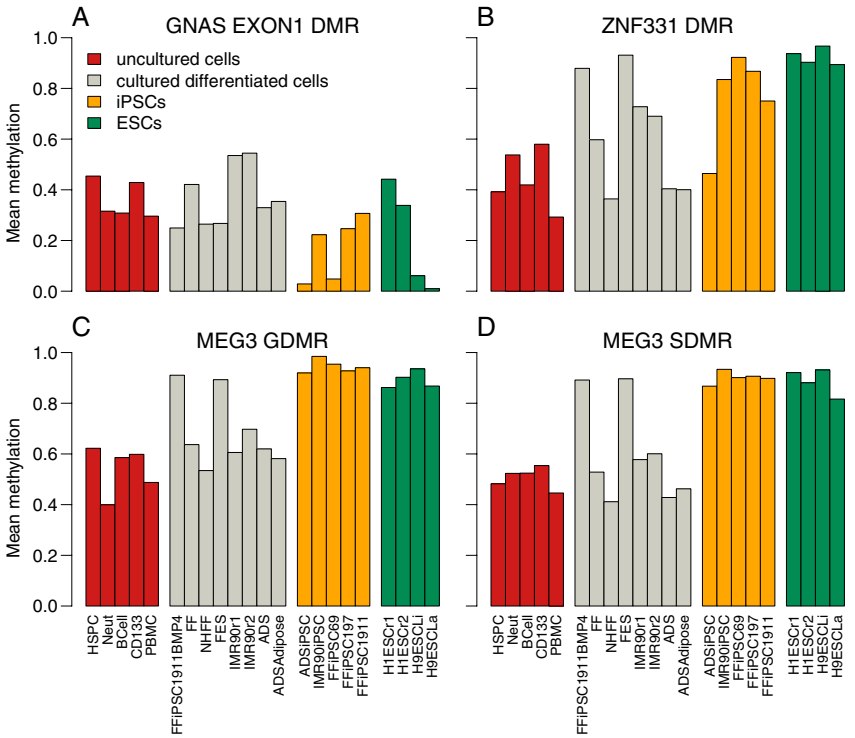


Fig. S3. Examples for iPSC reprogramming of AMRs. Twenty-two methylomes were divided into four groups: uncultured cells, cultured differentiated cells, iPSCs, and ESCs. (*A*) Differentially methylated region (DMR) in GNAS EXON1. Some ESCs and iPSCs lost the allele-specific methylation in this region. (*B*) DMR in ZNF331. All ESCs were hypermethylated, four out of five iPSCs reprogrammed such hypermethylation from allele-specific methylation (ASM) except ADSiPSC. Some cultured differentiated cells also had both alleles methylated (*C and D*) GDMR and SDMR for MEG3. All ESCs and iPSCs showed hypermethylation in these regions. Some cultured differentiated cells lost the ASM marks as well.

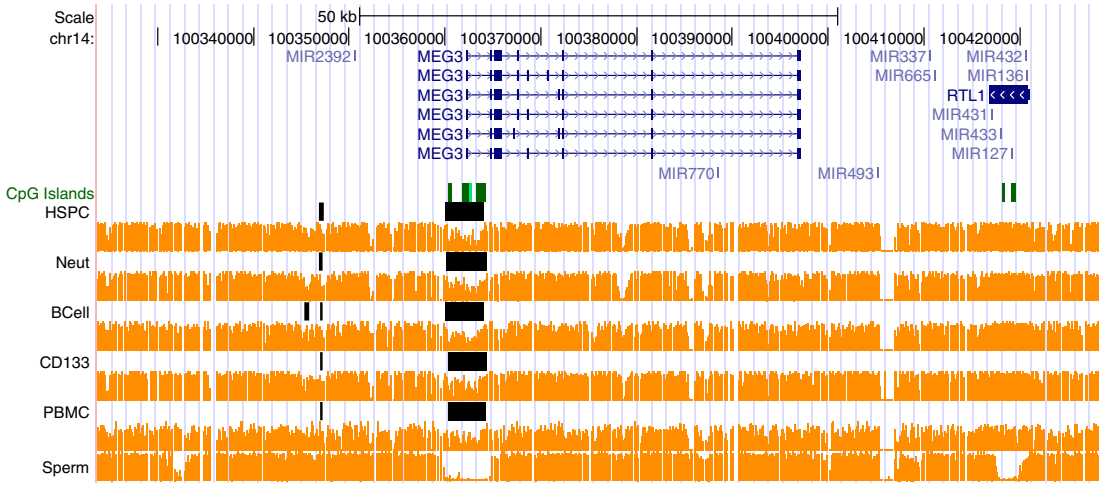


Fig. S4. Refined AMRs near MEG3. One AMR consistently appears in the promoter region of the lncRNA MEG3, and another AMR approximately 15 kb upstream of the MEG3 TSS appears in six differentiated cells. Black bars are AMRs, and green bars are CGIs. Orange bins indicate the methylation levels at CpG sites in each uncultured cell, and red bins are methylation levels in sperm.

Other Supportiong Information Files

Dataset S1 (XLSX)