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

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

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,

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

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

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

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

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

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.

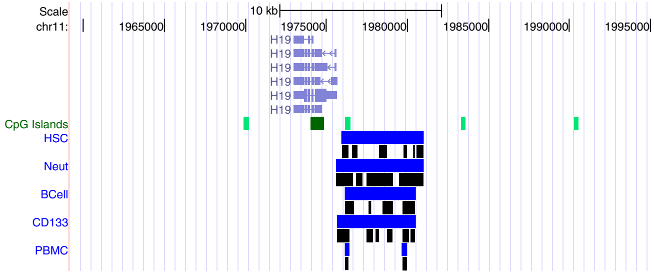
Preprocessing High-Throughput Short-Read Bisulfite Sequencing. For efficient computation, we took the following preprocessing steps before AMR identification.

- 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

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

and

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

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.

	AMRs identified		
Cell type	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.

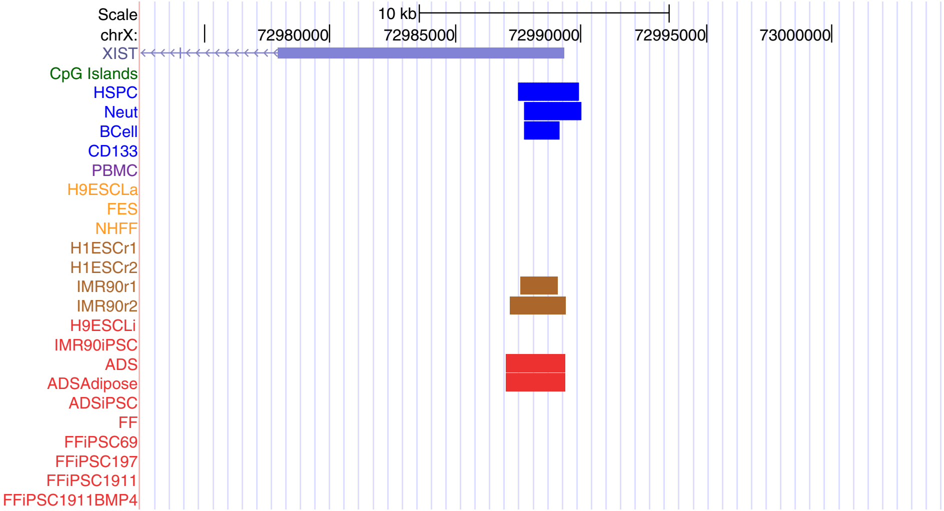
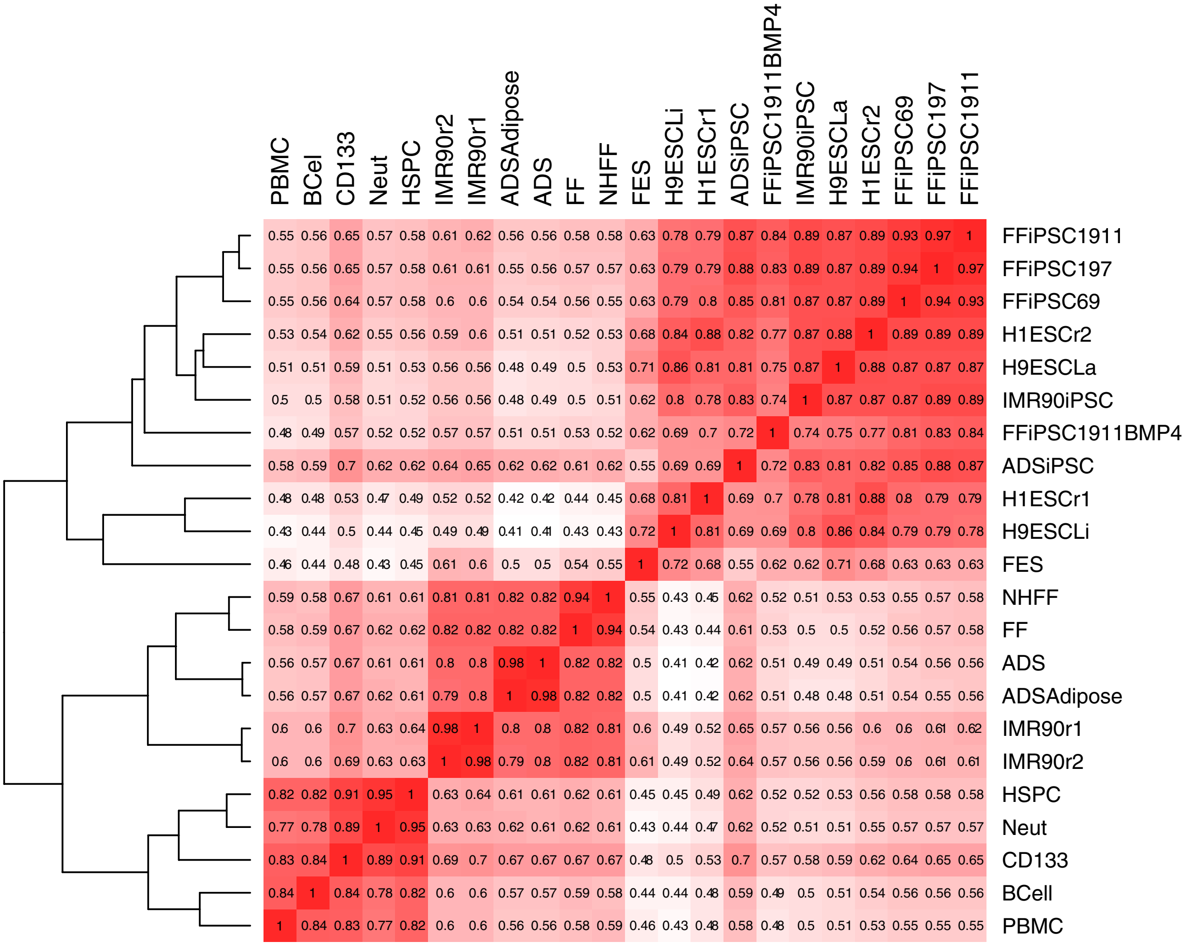


Fig. S1. Allele-specific methylation identified at the XIST promoter. Consistent with earlier findings, allele-specific methylation is found in exactly those methylomes that are (1) female, and (2) not from ESCs or iPSCs.



PBMC	BCel	CD133	Neut	HSPC	IMR90r2	IMR90r1	ADSAdipose	ADS	FF	NHFF	FES	H9ESCLi	H1ESCr1	ADSiPSC	FFiPSC1911BMP4	IMR90iPSC	H9ESCLa	H1ESCr2	FFiPSC69	FFiPSC197	FFiPSC1911	
0.55	0.56	0.65	0.57	0.58	0.61	0.62	0.56	0.56	0.58	0.58	0.63	0.78	0.79	0.87	0.84	0.89	0.87	0.89	0.93	0.97	1	FFiPSC1911
0.55	0.56	0.65	0.57	0.58	0.61	0.61	0.55	0.56	0.57	0.57	0.63	0.79	0.79	0.88	0.83	0.89	0.87	0.89	0.94	1	0.97	FFiPSC197
0.55	0.56	0.64	0.57	0.58	0.6	0.6	0.54	0.54	0.56	0.55	0.63	0.79	0.8	0.85	0.81	0.87	0.87	0.89	1	0.94	0.93	FFiPSC69
0.53	0.54	0.62	0.55	0.56	0.59	0.6	0.51	0.51	0.52	0.53	0.68	0.84	0.88	0.82	0.77	0.87	0.88	1	0.89	0.89	0.89	H1ESCr2
0.51	0.51	0.59	0.51	0.53	0.56	0.56	0.48	0.49	0.5	0.53	0.71	0.86	0.81	0.81	0.75	0.87	1	0.88	0.87	0.87	0.87	H9ESCLa
0.5	0.5	0.58	0.51	0.52	0.56	0.56	0.48	0.49	0.5	0.51	0.62	0.8	0.78	0.83	0.74	1	0.87	0.87	0.87	0.89	0.89	IMR90iPSC
0.48	0.49	0.57	0.52	0.52	0.57	0.57	0.51	0.51	0.53	0.52	0.62	0.69	0.7	0.72	1	0.74	0.75	0.77	0.81	0.83	0.84	FFiPSC1911BMP4
0.58	0.59	0.7	0.62	0.62	0.64	0.65	0.62	0.62	0.61	0.62	0.55	0.69	0.69	1	0.72	0.83	0.81	0.82	0.85	0.88	0.87	ADSiPSC
0.48	0.48	0.53	0.47	0.49	0.52	0.52	0.42	0.42	0.44	0.45	0.68	0.81	1	0.69	0.7	0.78	0.81	0.88	0.8	0.79	0.79	H1ESCr1
0.43	0.44	0.5	0.44	0.45	0.49	0.49	0.41	0.41	0.43	0.43	0.72	1	0.81	0.69	0.69	0.8	0.86	0.84	0.79	0.79	0.78	H9ESCLi
0.46	0.44	0.48	0.43	0.45	0.61	0.6	0.5	0.5	0.54	0.55	1	0.72	0.68	0.55	0.62	0.62	0.71	0.68	0.63	0.63	0.63	FES
0.59	0.58	0.67	0.61	0.61	0.81	0.81	0.82	0.82	0.94	1	0.55	0.43	0.45	0.62	0.52	0.51	0.53	0.53	0.55	0.57	0.58	NHFF
0.58	0.59	0.67	0.62	0.62	0.82	0.82	0.82	0.82	1	0.94	0.54	0.43	0.44	0.61	0.53	0.5	0.5	0.52	0.56	0.57	0.58	FF
0.56	0.57	0.67	0.61	0.61	0.8	0.8	0.98	1	0.82	0.82	0.5	0.41	0.42	0.62	0.51	0.49	0.49	0.51	0.54	0.56	0.56	ADS
0.56	0.57	0.67	0.62	0.61	0.79	0.8	1	0.98	0.82	0.82	0.5	0.41	0.42	0.62	0.51	0.48	0.48	0.51	0.54	0.55	0.56	ADSAdipose
0.6	0.6	0.7	0.63	0.64	0.98	1	0.8	0.8	0.82	0.81	0.6	0.49	0.52	0.65	0.57	0.56	0.56	0.6	0.6	0.61	0.62	IMR90r1
0.6	0.6	0.69	0.63	0.63	1	0.98	0.79	0.8	0.82	0.81	0.61	0.49	0.52	0.64	0.57	0.56	0.56	0.59	0.6	0.61	0.61	IMR90r2
0.82	0.82	0.91	0.95	1	0.63	0.64	0.61	0.61	0.62	0.61	0.45	0.45	0.49	0.62	0.52	0.52	0.53	0.56	0.58	0.58	0.58	HSPC
0.77	0.78	0.89	1	0.95	0.63	0.63	0.62	0.61	0.62	0.61	0.43	0.44	0.47	0.62	0.52	0.51	0.51	0.55	0.57	0.57	0.57	Neut
0.83	0.84	1	0.89	0.91	0.69	0.7	0.67	0.67	0.67	0.67	0.48	0.5	0.53	0.7	0.57	0.58	0.59	0.62	0.64	0.65	0.65	CD133
0.84	1	0.84	0.78	0.82	0.6	0.6	0.57	0.57	0.59	0.58	0.44	0.44	0.48	0.59	0.49	0.5	0.51	0.54	0.56	0.56	0.56	BCell
1	0.84	0.83	0.77	0.82	0.6	0.6	0.56	0.56	0.58	0.59	0.46	0.43	0.48	0.58	0.48	0.5	0.51	0.53	0.55	0.55	0.55	PBMC

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.

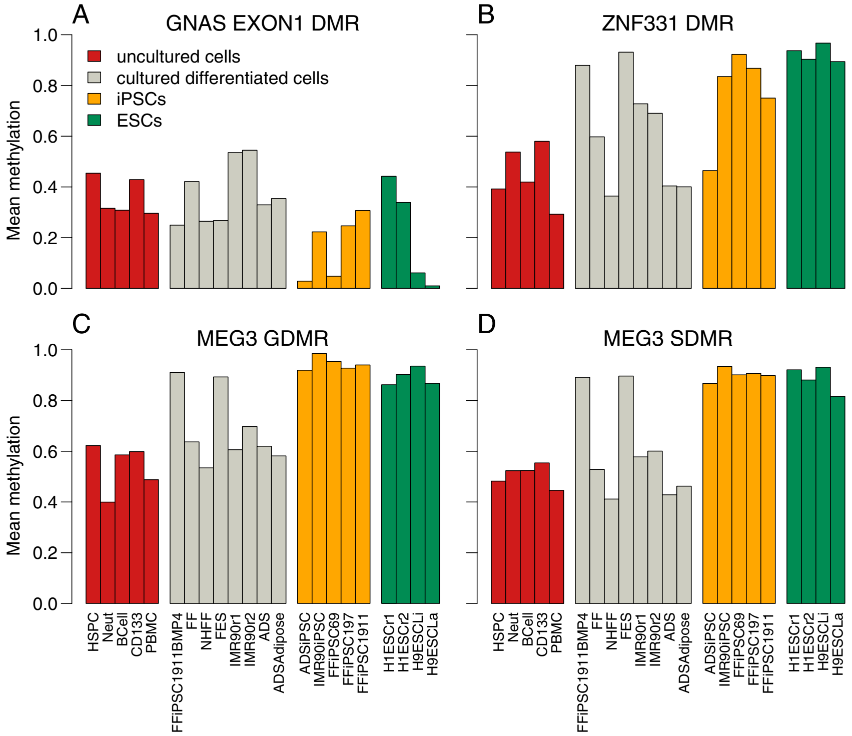


Fig. S3. Examples for iPSC reprogramming of AMRs. Twenty-two methylomes were divided into four groups: uncultured cells, cultured differentiated cells, iPSCs, and ESCs. (*A*) Differentially methylated region (DMR) in GNAS EXON1. Some ESCs and iPSCs lost the allele-specific methylation in this region. (*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.

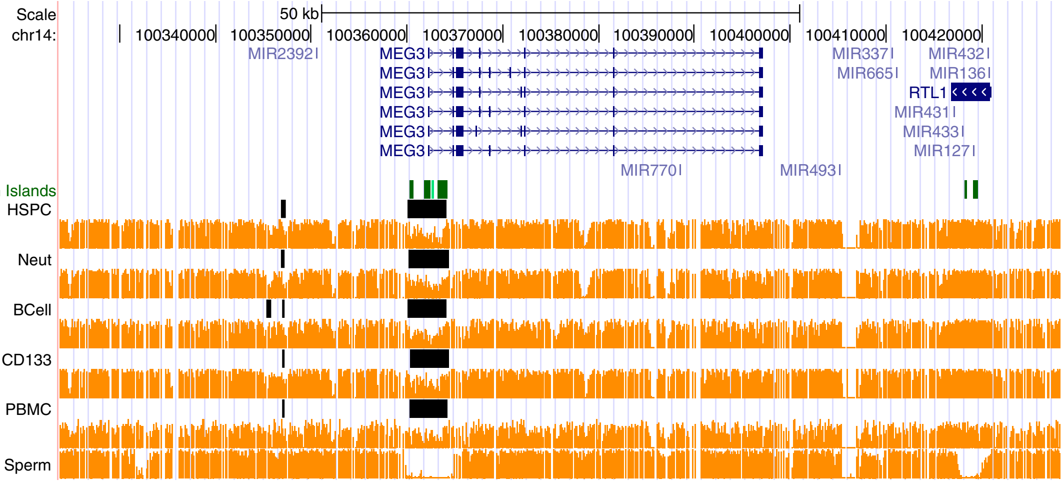


Fig. S4. Refined AMRs near MEG3. One AMR consistently appears in the promoter region of the lncRNA MEG3, and another AMR approximately 15 kb upstream of the MEG3 TSS appears in six differentiated cells. Black bars are AMRs, and green bars are CGIs. Orange bins indicate the methylation levels at CpG sites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

Other Supportiong Information Files

Dataset S1 (XLSX)