Supplementary methods

Data collection, mapping and annotation

The general features of the DNA methylomes used for different cell lines of FB, iPS and ES are described in Table 1. The raw data (fastq files) of each methylome were downloaded from the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena/home) and processed using the following procedure:

1. The bisulfite reads were trimmed with BRAT-BW (Harris et al., 2012), fastqc and cutadapt (Martin, 2011) by:
    - Adapters.
    - Base call quality (keeping only nucleotide with base call quality ≥20).
    - Length (keeping only the reads with length ≥ 20bp).
    - Number N_s of internal of bisulfite reads (keeping only the reads with $N_s \leq 2$).
2. The trimmed reads were aligned with BRAT-BW to the human genome version 19 (hg19) obtained from the UCSC Human Genome Browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/).
3. The aligned reads with duplicates which are PCR byproducts were removed. The PCR duplicates were detected with BRAT-BW by searching for reads with the same length that map to the same *locus* in the hg19 reference genome.
4. For the deduplicated reads we calculated using BRAT-BW the methylation calls as the number of C/T or G/A at each CpG position in the reference genome.
5. From the methylation calls we calculated the methylation ratios for each CpG *locus* ($MR(CpG)$) like in (Laurent et al., 2010), as the number of reads with a CpG in such *locus* divided by the number of reads covering the *locus*.
6. To obtain high quality sequenced CpGs we choose only those CpGs whose cytosine coverage is ≥ 10 in both DNA strands simultaneously.

We used the RefSeq "Genes track" to annotate the hg19 genes. We used an extended concept of promoter, defining the promoter of each gene as the region within 29 Kb upstream and 1 Kb downstream from the Transcription Start Site (TSS).

Generation of "conserved" DNA methylome for each cell type

The methylomes of different cell lines from the same cell type are grouped. Then all the CpGs are aligned based on the cytosine genomic positions. Thus, we define a CpG site as conserved when the methylation ratios across all the cell lines are low fluctuating. We assess the methylation ratios fluctuation using the Kolmogorov–Smirnov test for goodness of fit with the null hypothesis that the CpG methylation ratios follow a uniform distribution in the range [0,1] at a significance level α=0.001. The alternative hypothesis is that the CpG methylation ratios are constant. E.g., when the alternative hypothesis is true, the CpG is conserved. The Kolmogorov–Smirnov test tests were performed with the function KStest of the stats package of scipy.

Analysis of sequence and methylation similarity between the two DNA strands

We define the methylation dissimilarity $MetDis(CpG_i)$ of the methylation in each CpG *locus i* (*i* marks the cytosine position) of the two DNA strands as the absolute difference between the methylation ratios of the cytosines at the positive $MR(CpG^+)$ and negative $MR(CpG^-)$ strands

$$MetDis(CpG_i) = \left| MR(CpG_i^+) - MR(CpG_i^-) \right| . \quad (1)$$

Since the methylation ratios are determined in the range [0 1], we can define the methylation similarity *MetSim* as:

$$MetSim(CpG_i) = 1 - MetDis(CpG_i) . \quad (2)$$

We define the sequence similarity $SeqSim(CpG_i^w)$ of length *w* sequences at a *locus i*, as the matching between the sequences of length *w* in the positive CpG_i^{w+} and negative CpG_i^{w-} strands centered in the CpG of each *loci i*, normalized by the sequence length *w*,

$$SeqSim(CpG_i^w) = \frac{\sum_j^w match(CpG_i^{w+}(j), CpG_i^{w-}(j))}{w} , \quad (3)$$

where $CpG_i^{w+}(j)$ and $CpG_i^{w-}(j)$ are the nucleotides at the position *j* of the CpG sequences at the genomic positions *i* of the positive and negative strands, respectively. The sequence match is the Hamming distance between each nucleotide in the positive strand, and the paired nucleotide in the negative strand. This similarity is a measurement of the degree of palindromy of the sequence. We applied Eqs. (1-3) for $12 \leq w \leq 52$, with a step of 2 (one nucleotide in both directions outwards the central CpG, whose length is also accounted for the sequence length) and depicted the methylation similarity *versus*

the sequence similarity using heatmaps that represent the dot density on a color scale.

Figure 1b shows that the methylation similarity between both DNA strands is low. This justifies the practice in NGS-based methylomics studies (Hackenberg et al., 2011) to use the average methylation in the pairing *loci*. Such approach is valid for the search of differentially methylation regions (DMRs), but it can lead to loss of information determinant for the discrimination of CpGs whose methylation depends on their sequence context. Therefore, in the present study we have kept track of the methylation of the two DNA strands separately, which allowed us to analyze the regulation of the DNA methylation of both strands disjointly.

Discriminative CpGMM discovery algorithm

The algorithm is based on the compilation of two CpG-centered DNA-word dictionaries, one collected from low methylated, and another from high methylated regions. Each CpG word can have a different length w. The shorter CpG words are fused into longer ones to avoid the appearance of submotifs inside motifs. We applied the following pipeline to both DNA strands. In order to avoid cumbersome notations, we describe the procedure for the positive strand. The same steps are applied for the negative strand. The method workflow is depicted in Fig. S5.

1. Compilation of CpG word dictionaries

The minimal and maximal CpG word lengths were based on the results depicted in Figure S2a. The minimal sequence length w_{min}=12 is selected, since a typical methylome has around 10 million high quality sequenced CpGs. The number of all possible CpG-centered words of length w is $N_{CpG}(w)=4^{w-2}$. Then, the number of possible different CpG-centered sequences of length 12, is $N_{CpG}(12)=1048576$. This number is ten times smaller than the number of CpGs. Thus, the probability for all possible CpG words of length $w\leq12$ to be compiled in the dictionary is very high, hence, the dictionary of CpG words of length $w\leq12$ is a comprehensive dictionary. $N_{CpG}(w)$ is an increasing function of w, and w=12 is the longest length, for which all the CpG words are found in the methylome. This is reflected in the high peak of frequencies at length w=12. A quick frequency decay is observed for longer sequence lengths, since they correspond to the less unique CpG words appearing in the genome. Based on such decay and the insufficient number of unique sequences for a next clustering step, we stop to compile dictionaries

for lengths $w>44$. For all cell lines, we found frequency distributions with similar behavior to those previously described. Therefore we use minimal and maximal sequence lengths $w_{min}=12$ and $w_{max}=44$, respectively.

For all the genomic CpGs, and for lengths $w_{min}\leq w\leq w_{max}$, with a step of 2, we collected the sequence centered in each CpG (the central CpG is included in the w calculation). This procedure generates a CpG-centered word genomic dictionary harboring all sequences of length w centered in each CpG. The repeated sequences are grouped into unique ones. We call their repetition numbers F_{CpGw} frequency of the sequence of length w. Hence, we generate a set of unique CpG^{w} sequences of different lengths w. We assume that similar sequences have similar methylation ratios MR_{CpGw}. Therefore, we assigned to each unique sequence the average of the methylation ratios of the CpGs from the same unique group. We denote this step as unique sequence search step. A final scanning analysis selects the CpGMMs that discriminate between methylation-prone and -resistant regions. Next, to reduce the noise we implemented a filter step. The unique sequences are filtered by their frequencies. Only unique patterns with frequencies ≥ 3 are retained. This threshold is based on the fact that in our data compilation, a methylome has around 10 million high-quality sequenced CpGs, and the probability to find a sequence of length 12, centered in a CpG, is $1/4^{10}=1/1048576$. The probability to find (using the binomial distribution) 3 repetitions of the same sequence is 0.01093773. This value is in the null hypothesis acceptance twilight. Therefore, we reckon this threshold as the minimal number of repetitions of a sequence of length $w \geq 12$ to consider its frequency statistically significant. The resultant set of sequences is classified into two subsets based on their methylation ratios. If the central CpG methylation ratio of a sequence is ≥ 0.85, we assign such sequence to the methylation-prone subset. Conversely, if it is ≤ 0.5, we assign it to the methylation-resistant subset.

To check whether the assumption that similar sequences have similar methylation ratios MR_{CpGw}, for each CpG word of length w we discretized the CpG methylation ratios into three categories as in Stadler et al (2011): $0.0\leq MR_{CpGw}<0.1$ for unmethylated sites (UMSs), $0.1\leq MR_{CpGw}<0.5$ for low methylated sites (LMSs) and $0.5\leq MR_{CpGw}\leq 1.0$ for high methylated sites (HMSs). Then, for each CpG word we count its membership percentage to one of the three categories. If the percentage is higher than 90%, we assign a 1 to that word, and a 0, if it is less than 90%. Finally calculate the percentage of ones in relation to the total number of CpG words.

2. *Fusion of CpG word dictionaries*

During the dictionary compilations, the sequences are extended in both directions outwards the central CpG. Therefore, some sequences of length w could appear inside sequences of length w+2. Hence, we designed a sequence fusion method that avoids shorter submotifs centered inside longer ones. This method is implemented in an iterative way, starting from the shortest length w_{min}. Thus, for each length w, if a sequence CpG^{w} is included in the center of a sequence CpG^{w+2} of length w+2, the shorter sequence is fused inside the longer one. The methylation ratio of the new sequence is updated as the weight averaged methylated ratios of the fused sequences

$$MR^{update}_{CpG^{w+2}} = \frac{F_{CpG^{w}} MR_{CpG^{w}} + F_{CpG^{w+2}} MR_{CpG^{w+2}}}{F_{CpG^{w}} + F_{CpG^{w+2}}} .$$

Before the sequence fusion step, each unique sequence CpG^{w+2} has an associated scalar frequency F_{CpGw+2} that imputes the same frequency to all sequence nucleotides. After the fusion, to keep track of the individual frequency position in the fused sequence, the scalar frequency is converted into a frequency vector F_{CpGw+2} of length w+2, that stores for each nucleotide position j its respective frequency. Thus, for the central common positions in the sequence CpG^{w+2}, the vectorial frequencies are $F_{CpGw+2}(j)=F_{CpGw+2}(j)+F_{CpGw}(j\text{-}1)$. The peripheral positions preserve the original frequencies of the longer sequence. The scalar frequency of the new sequence is updated as the sum of the scalar frequencies of the fused sequences $F^{update}_{CpGw+2}=F_{CpGw}+F_{CpGw+2}$. After fusion, the shorter sequence CpG^{w} is eliminated from the dictionary. The fusion method iterates till reaching the longest sequence length w_{max}, fusing whenever possible shorter sequences centered in the central CpG into the longer sequences. An example of the distribution of the number of CpG words before and after fusion with respect to their length for the highly methylated CpGs of the negative strand is depicted in Fig. S2a.

3. *Motif discovery through hierarchical clustering*

After the fusion step, the information associated to each CpG word i of length w of each dictionary is integrated into a Position Occurrence Matrix POM_i^{w} of dimension $(4\times w)$ that collects in every column j the frequency $F_{CpGw}(j)$ and stores it in the row indexed by the nucleotide in the position j of the sequence $CpG_i^{w}(j)$. Initially, the POM_i^{w} column has only a non-null value. For each length w, all the

POMs are grouped with a hierarchical clustering algorithm, using the cosine metric calculated with Eq. (4) and the complete linkage method. Before calculating the distances, the (4×*w*) bi-dimensional matrices are vectorized into 4*w* length vectors.

$$dist\left(POM_i^w, POM_j^w\right)=1-\frac{POM_i^w\cdot\left(POM_j^w\right)^T}{\|POM_i^w\|_2\|POM_j^w\|_2}\,. \quad (4)$$

Since the number of sequences to be clustered is very high (in the order of 40000), to accelerate the process, we performed the hierarchical clustering with the Python library fastcluster (*fastcluster:* Fast hierarchical clustering routines for R and Python). The cut-off parameter for the cluster distance was set to 0.75. This parameter is learned from the ADS-iPSC promoter methylome dataset, using silhouettes algorithm (Crooks et al., 2004), in-house implemented in Python. We performed 1001 cut-offs of the hierarchical clustering from 0 ≤ cut-off ≤ 1, with a step of 0.001. As a final cut-off, we chose the one (0.75) that maximizes the average silhouette width. For each cluster *c*, all the sequences are merged into a new averaged POM_c^w that represents the motif cluster.

4. Selection of the motifs with discrimination capability based on binding energy scanning method

With the given potential discriminative methylation motif sets, one for methylation-prone and other for -resistant, we searched for motifs that are specific to discriminate between high and low methylated CpGs. For such purpose, we took advantage of the analogy of the TF-DNA binding energy, using the Berg-von Hippel method (Berg and Von Hippel, 1987). Based on such analogy, we treat the methylation motif as TFBM, and we consider that a methylation motif has a good match with a genomic region center in a CpG, if the "virtual" binding energy estimated by the Berg-von Hippel method is high. To perform such energy calculation, first, we normalize the POMs, creating the so called Position Weight Matrices (PWMs)

$$PWM^w(i,j)=\frac{POM^w(i,j)}{\sum_k^4 POM^w(k,j)}\,.$$

Thus, all the motifs PWM^w are scanned as in a typical TFBS search (Sarkar et al., 2008) against each CpG of the methylation-prone and -resistant sets, using the binding energy equation of the Berg von Hippel method

$$matchingScore(PWM^w, CpG^w) = \sum_i^w \ln\left(\frac{PWM^w(CpG^w(i), i) + \beta}{max(PWM^w(:, i)) + \beta}\right), \quad (5)$$

where β = 0.00001. The addition of β is necessary to avoid division by zero. The specific value of β was chosen after empirical study to maximize the score dynamic range. CpG^w is the CpG word of length w centered in the genomic CpG *locus*. For better computational performance, we used a different but equivalent implementation of Eq. (5). Higher matching score corresponds to more specific similarity of the motif with the target sequence. We split the matching scores for each motif into two distributions, one for high and another for low methylation regions. Next, we checked whether the motif can discriminate between the two distributions using the Kolmogorov–Smirnov test (KStest) for two samples (with the stats package of scipy) with a significance level α = 0.00001. The motifs passing this test are retained and subjected to a second filter with a double objective. On one hand, the filter estimates the minimal matching score (threshold of the right tail) that has to have a potential target DNA sequence to be "bound" by the motif. The thresholds T_r of the right tails are computed with the equation

$$T_r = \min(\mu + \sigma\lambda, \theta), \quad (6)$$

where θ is the threshold of the right tail of the matching score distribution (methylation-prone distribution, if the underlying motif is a potential methylation-prone CpGMM), μ is the matching score distribution mean, σ is the matching score distribution standard deviation, and λ is set to 2, based on an empirical study. On the other hand, considering as true targets the sequences that pass the filter (6), to strengthen the discriminating capabilities of the motif, we select only those with a False Discovery Rate (FDR) ≤ 0.05

$$FDR = \frac{FalsePositive}{FalsePositive + TruePositive}, \quad (7)$$

where *FalsePositive* is the number of scores ≥ θ in the methylation-resistant distribution, if the underlying motif is a potential methylation-prone CpGMM, *TruePositive* is the number of scores ≥ θ in the methylation-prone distribution, if the underlying motif is a potential methylation-prone CpGMM. We use the right tail (high matching score) of the two distributions. The selected motifs are represented as motif logos, using WebLogo 3.0 (Crooks et al., 2004). All the found motifs were annotated with the gene ontology of their corresponding targets using the R-bioconductor package GOstat (Falcon and Gentleman, 2007). To assess the stability of the motif discovery we performed a bootstrapping with

100 time samplings with replacement over the ADS-iPSC methylomics data. We used as an estimation of the CpGMMs the percentage of CpGMMs that are recovered in at least half of the bootstrapping samples.

Search of cell-type specific and somatic memory CpGMMs

For all the methylation motifs in each pair of cell types (ES/FB/iPS), we computed the Pearson correlation.

$$\rho(PWM_i, PWM_j) = \frac{(PWM_i - \overline{PWM_i}) \cdot (PWM_j - \overline{PWM_j})^T}{\|PWM_i - \overline{PWM_i}\|_2 \|PWM_j - \overline{PWM_j}\|_2}, \quad (8)$$

where PWM_i and PWM_j are PWMs of the cell type i and j, respectively. When the two PWMs have different lengths, e.g. if length(PWM_i)>length(PWM_j), we substitute in the Pearson correlation (8) the longest matrix PWM_i by the overlapped central block of the PWM_i matrix with the same length of the shorter matrix PWM_j. Before applying Eq. (8), both PWMs should be converted from ($4\times w$) bidimensional matrices to $4w$-length vectors. Following the standards, defined in the field of DNA motifs to estimate the pairwise similarities (Stormo, 2000), we consider that two motifs are similar, if their Pearson correlation $\rho > 0.85$. We define as cell-type specific motifs those whit low correlation ($\rho \leq 0.5$) with any other motif in any cell type. The persistent somatic memory CpGMMs are those that have high correlation ($\rho \geq 0.85$) between iPS cells and FBs and simultaneously have low-correlation ($\rho \leq 0.5$) with ES cells. The absent somatic CpGMMs are the ES specific motifs, thus, those ES motifs with low-correlation ($\rho \leq 0.5$) with any other motif in iPS and FB. We searched separately for methylation-prone and -resistant CpGMMs.

Nucleotide enrichment analysis of CpGMMs

The four types of nucleotides for each methylation-resistant and -prone CpGMM are counted and normalized with the motif length. The Wilcoxon-Mann-Whitney-test is applied for each pair distribution of methylation-prone -resistant CpGMMs of the same nucleotide to find the significantly different enrichments with a significance level $\alpha = 0.01$.

Analysis of conservation of CpGMM targets

All targets of methylation-resistant and -prone CpGMMs targets are mapped to phastCons46way conservation track in primates (Siepel et al., 2005) downloaded from:

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/phastCons46wayPrimates.txt.gz. From that file we get the conservation score for each nucleotide position. The Kolmogorov–Smirnov test is applied for two distributions of conservation scores of methylation-prone/resistant CpGMMs with a significance level $\alpha = 0.01$.

Analysis of co-localization of CpGMM targets with genetic loci

1. Analysis of co-localization of CpGMM targets with TSS

The distance from all targets of methylation-resistant and -prone CpGMMs to the corresponding TSS are computed. All target genes and TSSs annotation are taken from the UCSC genome browser Refseq hg19 (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/knownToRefSeq.txt.gz). The Kolmogorov–Smirnov test is applied for two distance distributions of methylation-prone/resistant CpGMMs with a significance level $\alpha = 0.01$.

2. Analysis of co-localization of CpGMM targets with CpG islands

We downloaded the conserved the CpG islands annotation from the UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/cpgIslandExt.txt.gz). We counted for each cell type the number of targets of methylation-resistant and -prone CpGMMs occurring inside or outside CpG islands.

3. Analysis of co-localization of CpGMM targets with TFBSs

We downloaded the conserved TFBS from the UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/tfbsConsSites.txt.gz and http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/tfbsConsFactors.txt.gz). The search of TFs that share binding sites with the CpGMMs is not straightforward, since the CpGMMs have a central CpG anchor but the TFBMs do not necessarily have it. Even focusing on TFBMs with a strong CpG signal, such a signal is not always in the TFBM center. If the targets of a CpGMM co-occur with a TFBS, we consider that the CpG methytlation motif and the TFBMs are associated. Then, to compare CpGMMs and TFBMs, we designed a technique based on detecting co-occurrences between the targets of the two motif types. Thus, for the conserved TFBS co-localization analysis, we counted for each cell

type the number of targets of methylation-resistant and -prone CpGMMs occurring inside or outside conserved TFBS.

4. Analysis of co-localization of CpGMM targets with histone marks

We downloaded the twelve histone marks broad peak signals from the ENCODE project (The ENCODE Project Consortium, 2011) for ES cells (H1) and fibroblasts (NHLF). For each histone mark and for each CpGMM type (methylation-prone or -resistant), we developed the following algorithm to calculate the correlation between CpGMM targets and ES histone mark signals:

1. Identify in which *loci* the histone mark signals and the CpGMM targets co-occur. Such *loci* are those for which CpGMM target and the support of the broad peak signal overlap in at least one nucleotide.
2. Collect the co-occurring scores of the histone mark signal over the overlapping region with the CpGMM target (Fig. S6) for all the CpGMM targets of all the CpGMMs.
3. Calculate the histogram of the co-occurring scores (Fig. 5c).
4. Calculate the mean of the co-occurring scores (Fig. 5a).
5. Calculate the difference between the mean of the co-occurring scores of the methylation-resistant CpGMMs and the mean of the co-occurring scores methylation-prone CpGMMs (Fig. 5b).

Correlation analysis between CpGMM targets, CTCF and gene expression

The transcriptomics RNA-seq data and the CTCF binding data of ES cells (H1) and fibroblasts (NHLF) are downloaded from the ENCODE project (The ENCODE Project Consortium, 2011). We focused on the extended promoter regions as defined in the *Data collection and annotation subsection.* We collect all the CpGMMs targets and the broad-peak CTCF signals inside of the extended promoter regions. The target genes are classified into six groups:

1. Genes with only methylation-prone CpGMMs.
2. Genes with only methylation-resistant CpGMMs.
3. Genes with methylation-prone CpGMMs near TSS, and CTCF signal in-between methylation-resistant and -prone CpGMMs.
4. Genes with methylation-resistant CpGMMs near TSS, and CTCF signal in-between

methylation-resistant and -prone CpGMMs.

5. Genes with methylation-resistant CpGMMs, methylation-prone CpGMMs near TSS, and absent CTCF signal in-between.
6. Genes with methylation-prone CpGMMs, methylation-resistant CpGMMs near TSS, and absent CTCF signal in-between.

The expected number of CTCFs inside the extended promoter regions of genes with bivalent composition of CpGMMs is defined as the pool of the lengths covered by the broad-peak CTCF signals that lie inside the extended promoter of the bivalent composition genes divided by the pool of the total lengths of the extended promoters of the bivalent composition genes. The expression values (FPKM) of these classified genes are computed based on the RNA-seq data.

Algorithm implementation

All annotation information was compiled in a relational database, with the database management performed with MySQL (version 5.1.61). The algorithms were implemented in Python (version 2.6.5), with numerical package numpy (version 1.6.1) and scipy (version 0.9.0), in an Ubuntu (version 10.04.1) environment.

Supplementary references

Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK, Schübeler D. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011; 480(7378):490-5.

Harris EY, Ponts N, Le Roch KG, Lonardi S. BRAT-BW: efficient and accurate mapping of bisulfite-treated reads. Bioinformatics. 2012; 28(13):1795-6.

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011.

Figures Supplementary material

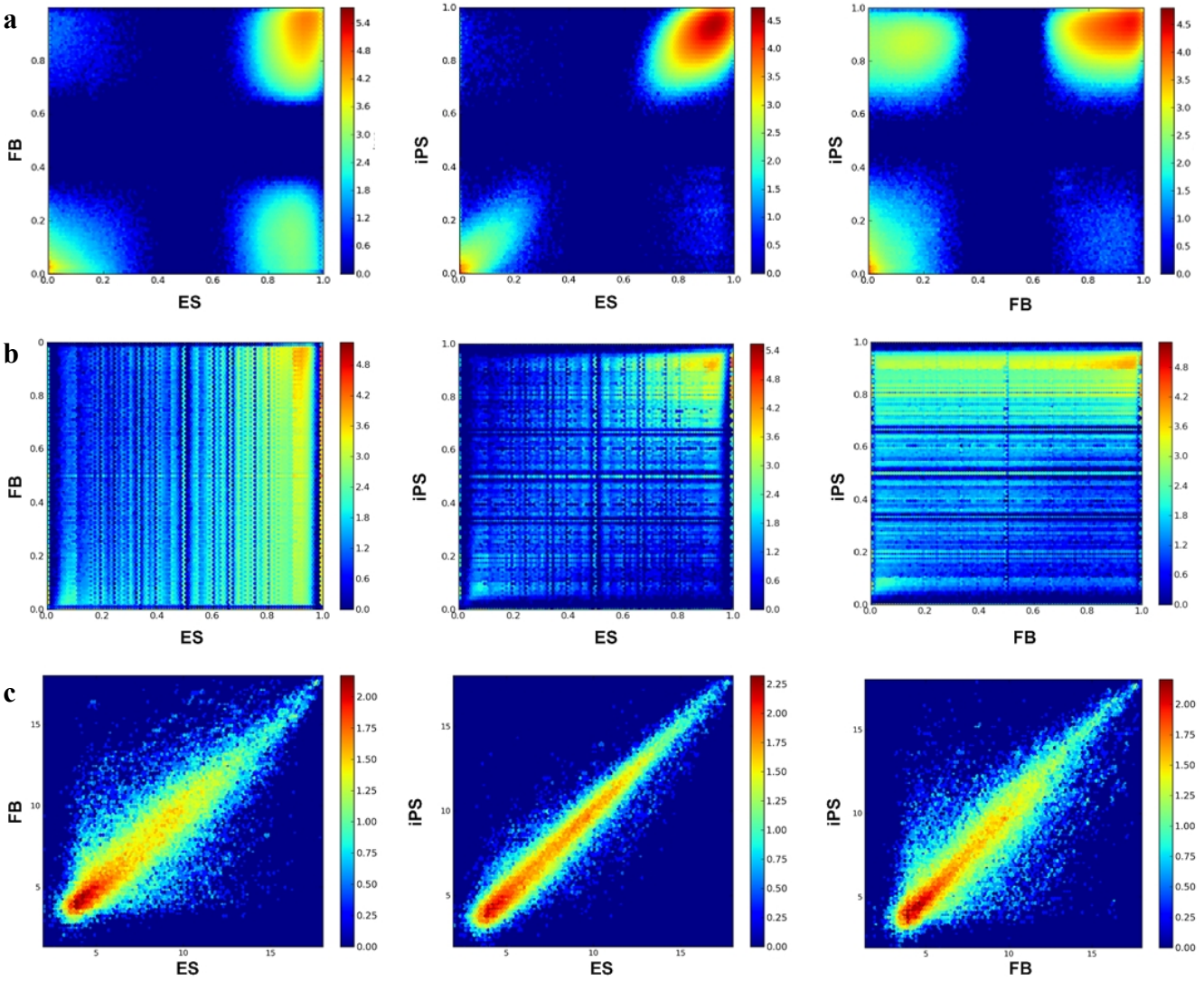

Figure S1. **Comparison between conserved methylomes of the different cell types across the different cell lines, and cell line methylomes and transcriptomes**. (a) Pairwise scatter plots of the positive DNA strand of the conserved methylomes of each of the three cell types used. (b) Pairwise scatter plots of the positive DNA strand of the fibroblast, FF iPS and H9 ESC methylomes. (c) Pairwise scatter plots of the fibroblast, iPS and H9 ESC transcriptomes. The transcriptomics data were taken from Takahashi and Yamanaka (2006) and produced the scatter plots of Fig. S1c. The scatter density is represented in $\log_{10}$ scale by the colorbar to the right of each plot.

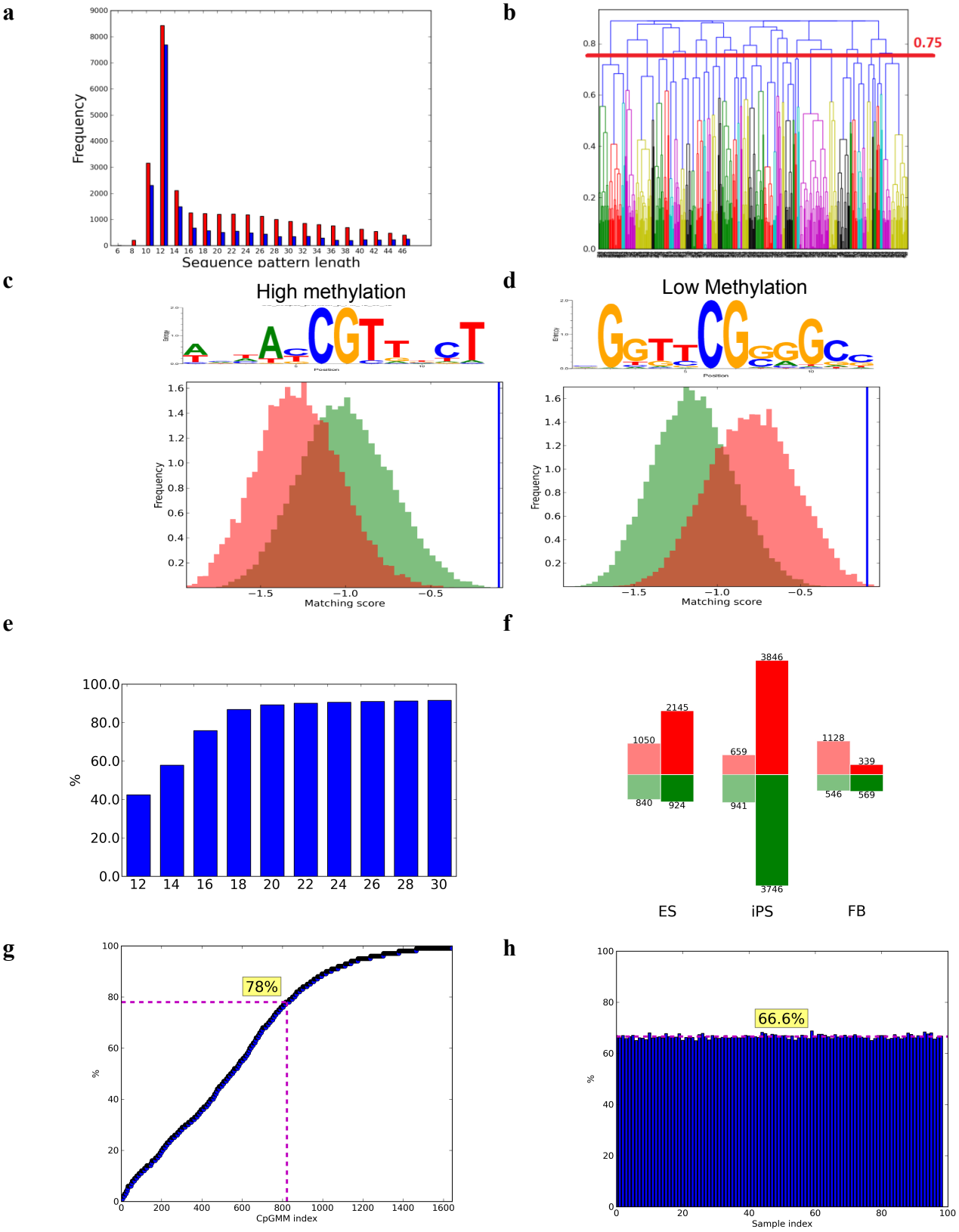
a
Frequency
Sequence pattern length
b
0.75
c
High methylation
Frequency
Matching score
d
Low Methylation
Frequency
Matching score
e
%
f
3846
2145
1050
659
1128
339
840
924
941
546
569
3746
ES
iPS
FB
g
78%
%
CpGMM index
h
66.6%
%
Sample index

Figure S2. **Intermediary results from the CpGMM discovery method pipeline.** (a) Distribution of the frequency of sequences before (red), and after (blue) fusion for different sequence lengths w for high-methylated CpGs of the negative strand of the ADS-iPSC case. (b) Hierarchical clustering of fused sequences for the negative strand of the ADS-iPSC case for sequence length w=16. The cut-off value that decides the final clusters is marked with a horizontal red line. (c) Histogram of the matching score of a typical methylation-prone CpGMM. (d) Histogram of the matching-score of a typical methylation-resistant CpGMM. The upper panels in (c) and (d) show the matching-score distribution, generated when scanning the motifs over high methylated regions, and the lower panels when scanning over low-methylated regions. The vertical blue line indicates the position of the threshold, calculated with Eq. (6) and used to collect the specific motif targets. The corresponding methylation motifs overlay the matching score distributions obtained in the high-methylated regions. (e) Bar plot of the percentages of sequences of length w that preserves the same category of methylation ratios (UMS, LMS or HMS) across the ADS-iPSC methylome. (f) Number of methylation-resistant (red top bars) and methylation-prone (green bottom bars) CpGMMs that appear in one cell line (left bars) and in more than one cell line (right bars) of each cell type. (g) Number of times each ADS-iPSC CpGMM is recovered by the bootstrapping process. (h) Percentage of recoved ADS-iPSC CpGMMs for each of the 100 time samplings of the bootstrapping process.

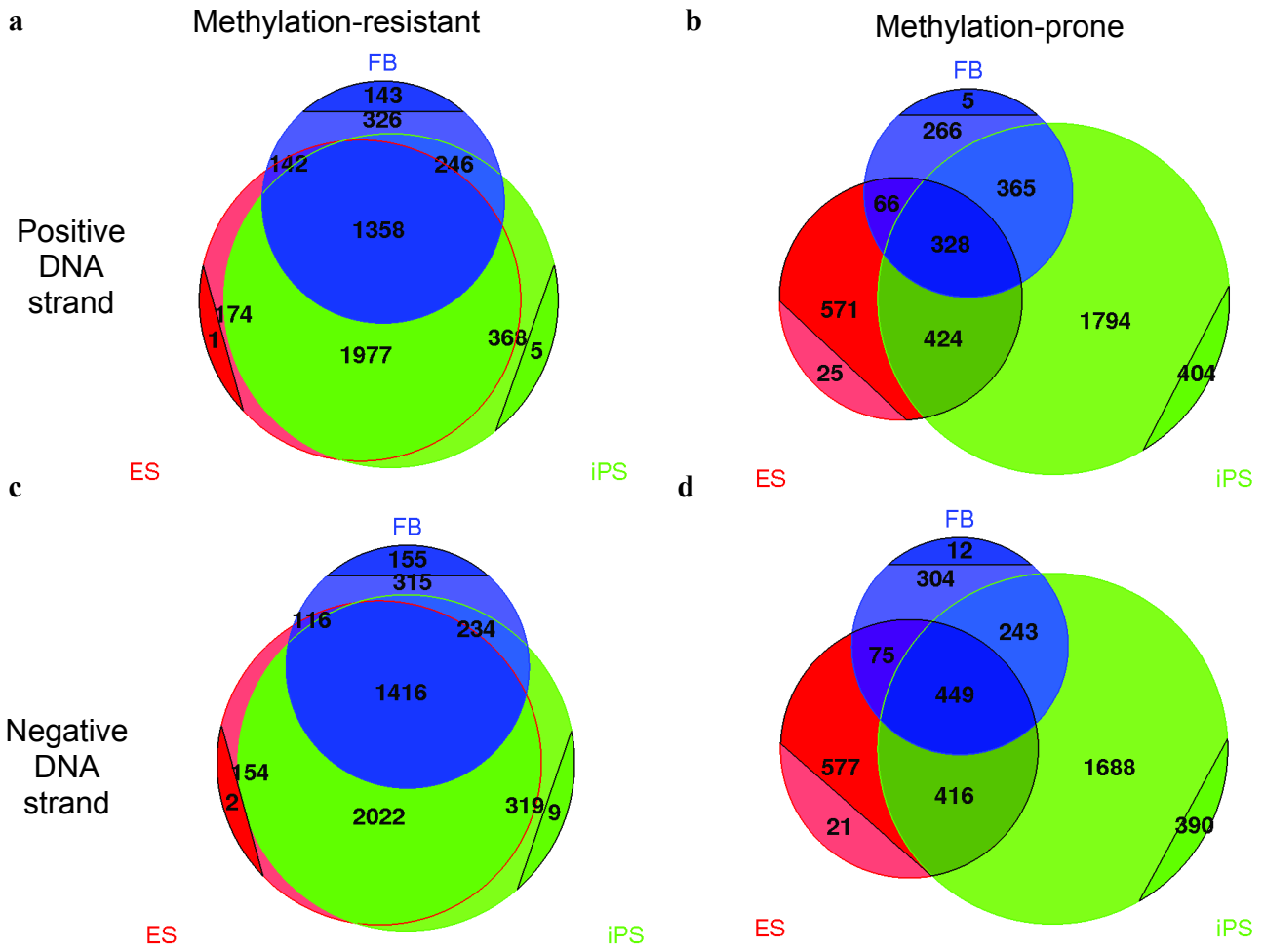

Figure S3. **Venn diagrams of the numbers of CpGMM for each DNA strand and each methylation mode clustered in each cell type.** (a) Methylation-resistant CpGMMs in the positive DNA strand. (b) Methylation-prone CpGMMs in the positive DNA strand. (c) Methylation-resistant CpGMMs in the negative DNA strand. (d) Methylation-prone CpGMMs in the negative DNA strand. In red ES CpGMM numbers, in green iPS and in blue fibroblast. The numbers enclosed by the circular segments are the numbers of cell-type specific motifs (those with a Pearson correlation with any CpGMM of any other cell type population less than 0.5).

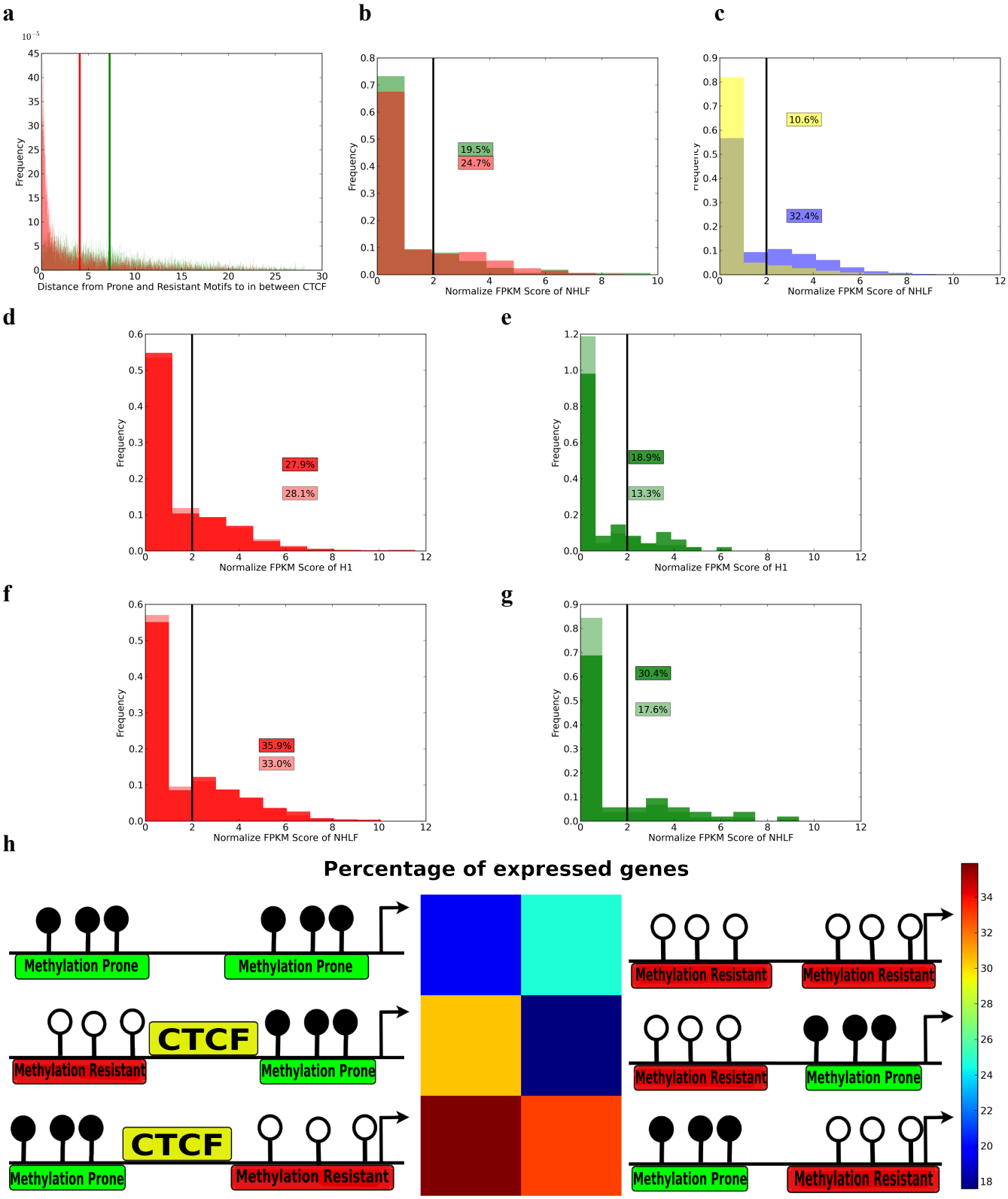
a
b
c
Frequency
Distance from Prone and Resistant Motifs to in between CTCF
Normalize FPKM Score of NHLF
19.5%
24.7%
10.6%
32.4%
d
e
Normalize FPKM Score of H1
27.9%
28.1%
18.9%
13.3%
f
g
35.9%
33.0%
30.4%
17.6%
h
Percentage of expressed genes
Methylation Prone
Methylation Resistant
CTCF

Figure S4. **Discriminative features between mixed promotors contained bivalent- and monovalent-CpGMMs.** (a) Histograms of the distances of methylation-resistant CpGMM targets (red), methylation-prone CpGMM targets (green) to the in-between CTCF for promoters containing simultaneously methylation-resistant and -prone CpGMMs in fribroblasts. The red and green vertical lines show mean values of the distances to CTCF or methylation-resistant and -prone CpGMM *loci*, respectively. (b) Histogram of the fibroblast expression of genes with only methylation-resistant or -prone CpGMMs. (c) Histogram of the fibroblast expression of all the genes (yellow) and genes with promoters containing simultaneously methylation-resistant and -prone CpGMMs (blue). (d,e,f,g) Histograms of the expression of the genes with mixed structure of methylation-resistant and -prone CpGMMs in their 1Kb upstream promoters. (d) ES cell gene-expression histograms for the case of methylation-resistant CpGMM close to the TSS. (e) ES cell gene expression histrograms for the case of methylation-prone CpGMM close to the TSS. (f) Fibroblast gene-expression histrograms for the case of methylation-resistant CpGMM close to the TSS. (g) Fibroblast gene-expression histrograms for the case of methylation-prone CpGMM close to the TSS. In dark red are shown the cases with a CTCF in-between a methylation-prone and -resistant CpGMM region close to the TSS. In light red are shown the cases without a CTCF in-between a methylation-prone and -resistant CpGMM region close to the TSS. In dark green are shown the cases with a CTCF in-between a methylation-prone and -resistant CpGMM region close to the TSS. In light green are shown the cases without a CTCF in-between a methylation-prone and -resistant CpGMM region close to the TSS. The numbers inside boxes are the percentages of expressed genes. (h) Heatmap of the percentage of fibroblast expressed genes with mixed and unmixed structure of methylation-resistant and methylation-prone CpGMMs in their 1Kb upstream promoters. The corresponding genomic structure (with the positions relative to the TSS marked with an arrow) of the methylation-resistant and -prone CpGMMs, and the CTCF is represented on the side of each cell.

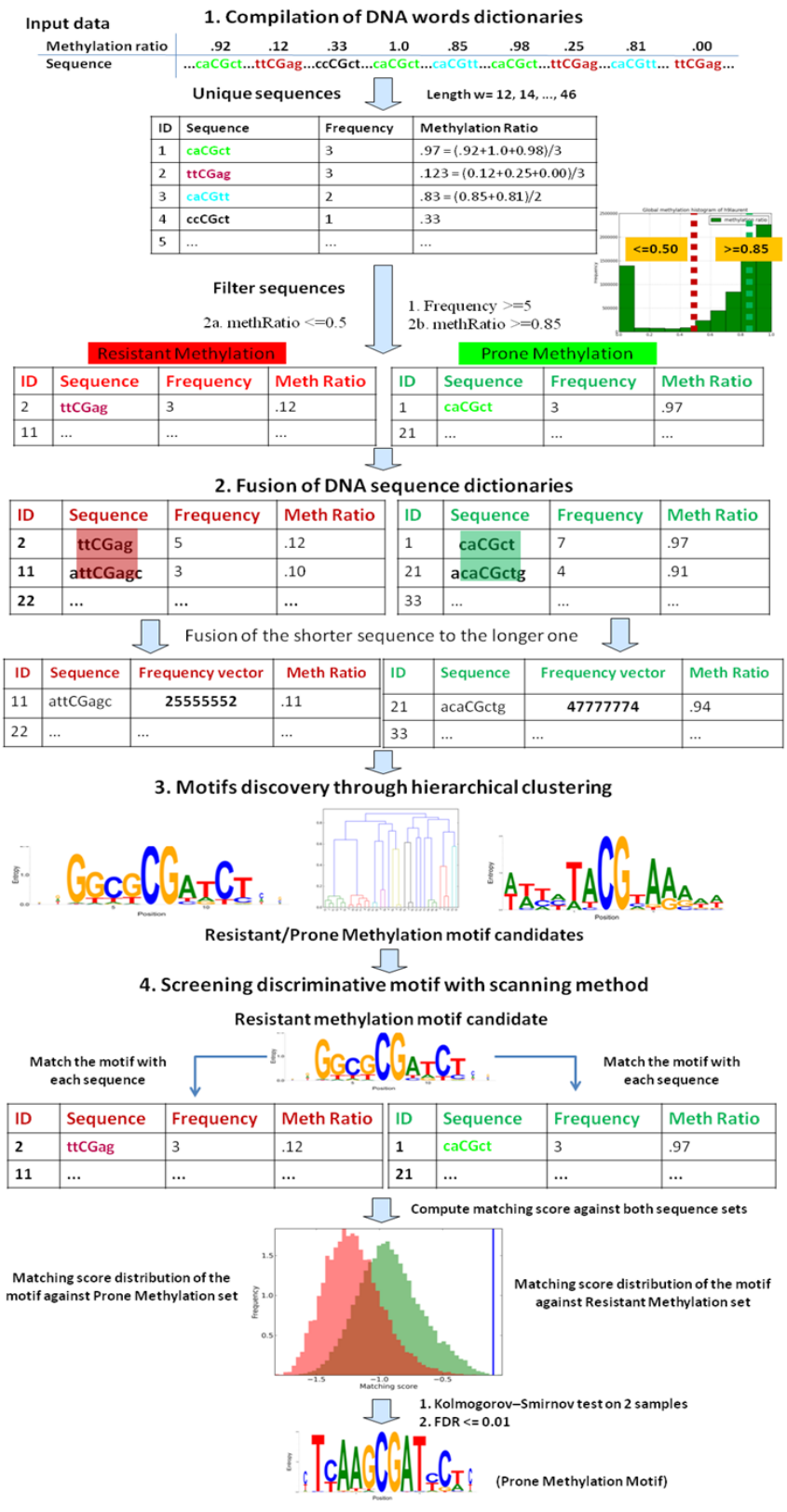
Input data
1. Compilation of DNA words dictionaries
Methylation ratio .92 .12 .33 1.0 .85 .98 .25 .81 .00
Sequence ...caCGct...ttCGag...ccCGct...caCGct...caCGtt...caCGct...ttCGag...caCGtt... ttCGag...
Unique sequences
Length w= 12, 14, ..., 46
ID Sequence Frequency Methylation Ratio
1 caCGct 3 .97 = (.92+1.0+0.98)/3
2 ttCGag 3 .123 = (0.12+0.25+0.00)/3
3 caCGtt 2 .83 = (0.85+0.81)/2
4 ccCGct 1 .33
5 ... ...
<=0.50
>=0.85
Filter sequences
2a. methRatio <=0.5
1. Frequency >=5
2b. methRatio >=0.85
Resistant Methylation
Prone Methylation
ID Sequence Frequency Meth Ratio
2 ttCGag 3 .12
11 ... ...
ID Sequence Frequency Meth Ratio
1 caCGct 3 .97
21 ... ...
2. Fusion of DNA sequence dictionaries
ID Sequence Frequency Meth Ratio
2 ttCGag 5 .12
11 attCGagc 3 .10
22 ... ...
ID Sequence Frequency Meth Ratio
1 caCGct 7 .97
21 acaCGctg 4 .91
33 ... ...
Fusion of the shorter sequence to the longer one
ID Sequence Frequency vector Meth Ratio
11 attCGagc 25555552 .11
22 ... ...
ID Sequence Frequency vector Meth Ratio
21 acaCGctg 47777774 .94
33 ... ...
3. Motifs discovery through hierarchical clustering
Resistant/Prone Methylation motif candidates
4. Screening discriminative motif with scanning method
Resistant methylation motif candidate
Match the motif with each sequence
Match the motif with each sequence
ID Sequence Frequency Meth Ratio
2 ttCGag 3 .12
11 ... ...
ID Sequence Frequency Meth Ratio
1 caCGct 3 .97
21 ... ...
Compute matching score against both sequence sets
Matching score distribution of the motif against Prone Methylation set
Matching score distribution of the motif against Resistant Methylation set
Frequency
Matching score
1. Kolmogorov–Smirnov test on 2 samples
2. FDR <= 0.01
(Prone Methylation Motif)

Figure S5. **CpGMM discovery method pipeline.** Example of processing of a genomic sequence fragment annotated with the methylation level of each CpG. The data collection is illustrated for a simple case of sequence length w=3. The identical sequences are of the same color. We focused on the low (red) and high (green) methylated sequences. The processes of methylation-resistant and -prone CpGMMs are highlighted in red and green, respectively.

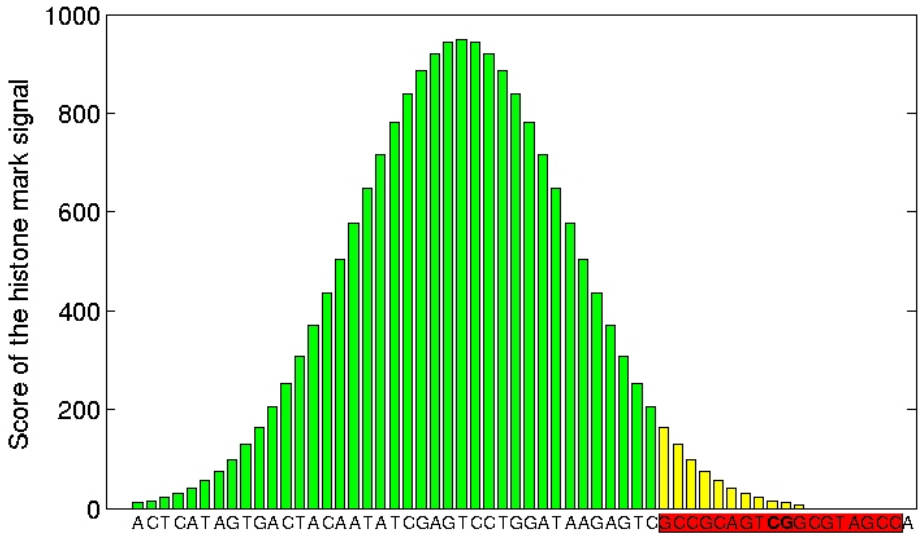

Figure S6. **Example of calculation of the co-occurrence scores between a broad-peak signal histone mark and a CpGMM target.** The position of the CpGMM target is framed by a red box, the scores of the histone mark region not-overlapping with the CpGMM target are drawn with green bars, and those corresponding with the region overlapping the CpGMM target are drawn with yellow bars.