Supplementary Figures

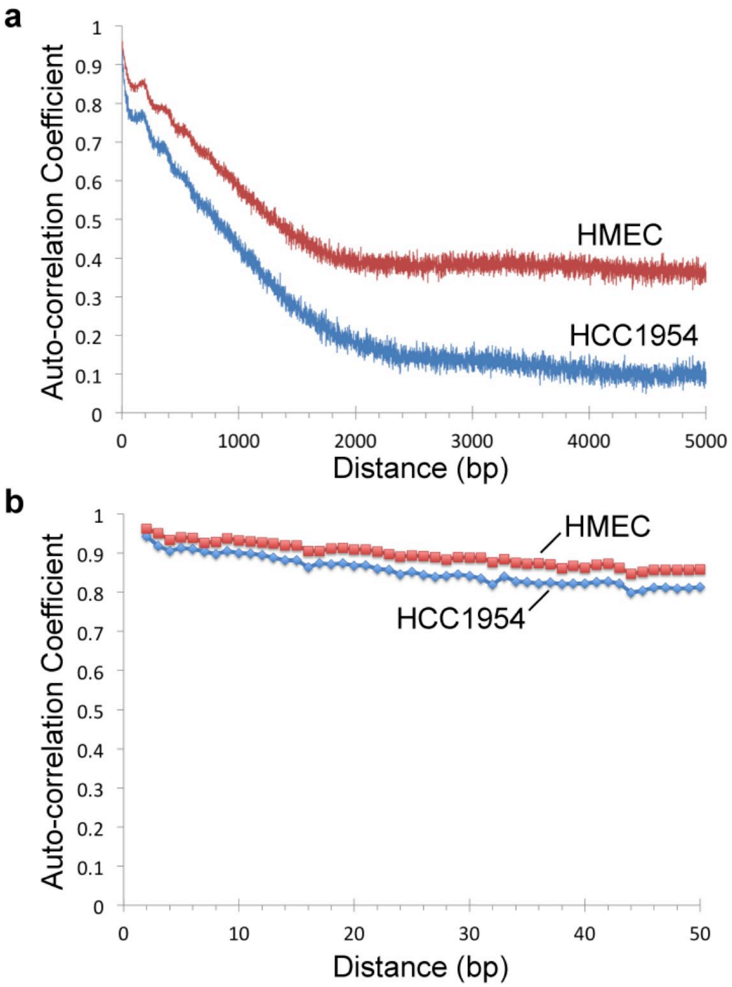


Figure S1 Autocorrelation of methylation score for HMEC and HCC1954 WGBS samples. (**a**) Distances ranging from 0 to 5000 bp. (**b**) Distances ranging from 0 to 50 bp.

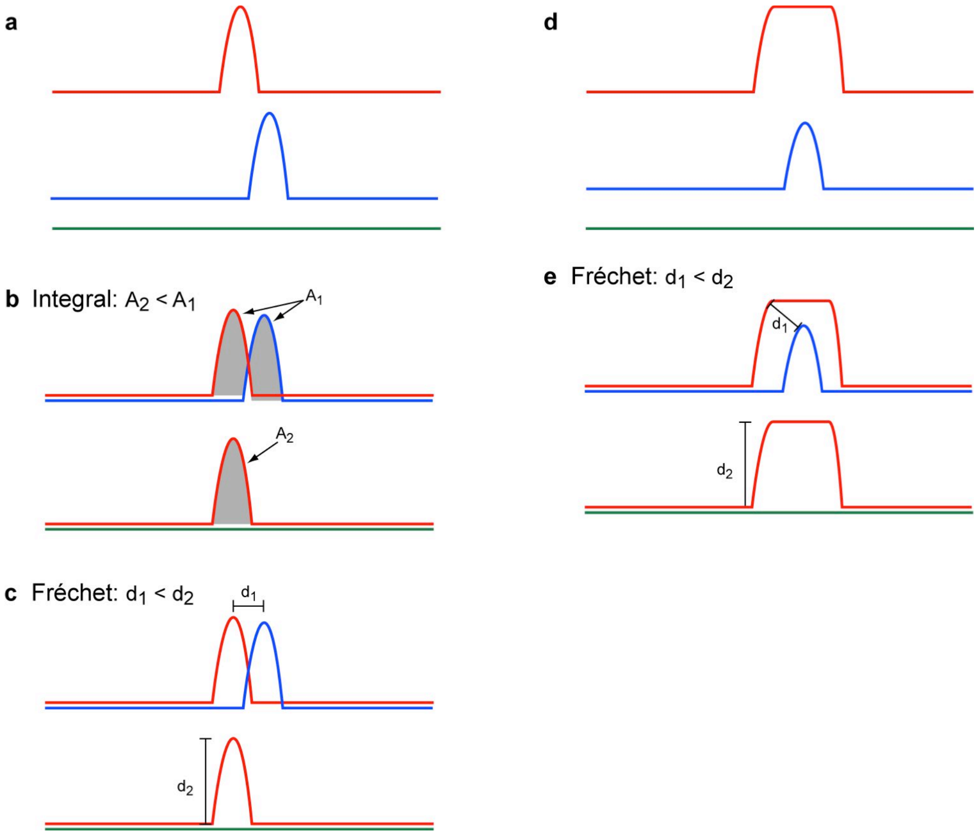


Figure S2 When comparing methylation signatures, CpG positions vary between different promoters. Therefore, if two signatures have methylation peaks with similar shape, but shifted or stretched slightly, then they should still be considered similar. (**a-c**) Pairwise comparison of (**a**) three curves (red vs. blue and red vs. green) by either (**b**) integrating to find the area between the curves (A_1, A_2) or (**c**) the Fréchet distance (d_1, d_2). The integral approach finds greater similarity between the red and green curves ($A_2<A_1$), while the Fréchet distance finds greater similarity between red and blue curves (d1<d2). Note that for both approaches, the distance between identical curves is 0. (**d-e**) Another example comparing three curves with the Fréchet distance. In this case, the red and blue curves are shown to be more similar than the red and green ($d_1<d_2$).

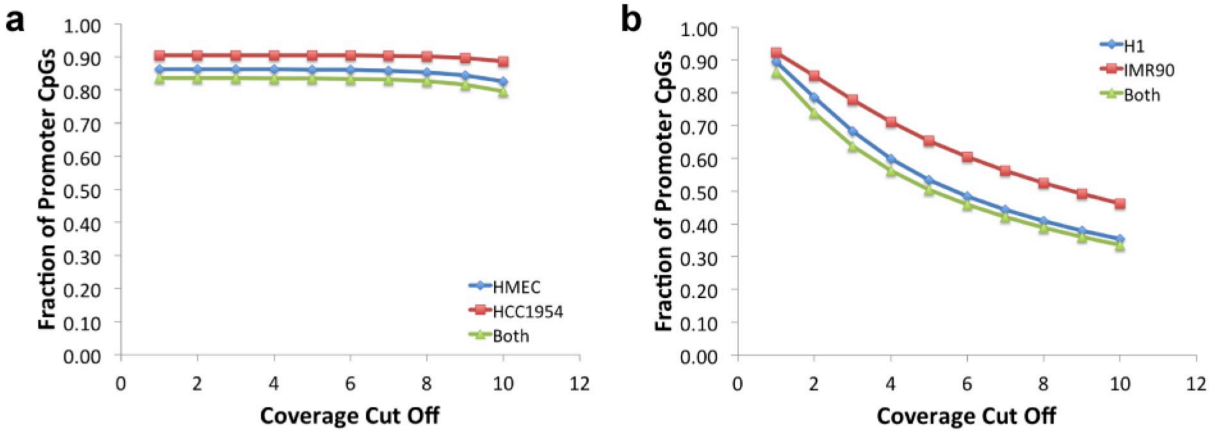


Figure S3 Coverage at promoters is high for HMEC-HCC1954 WGBS data, but low for H1-IMR90 data. (**a**) Fraction of promoter CpGs covered at various coverage cutoffs for HMEC, HCC1954, or both samples. (**b**) Fraction of promoter CpGs covered for IMR90, H1 or both samples. Promoter regions are defined as ± 500 bp from the TSS.

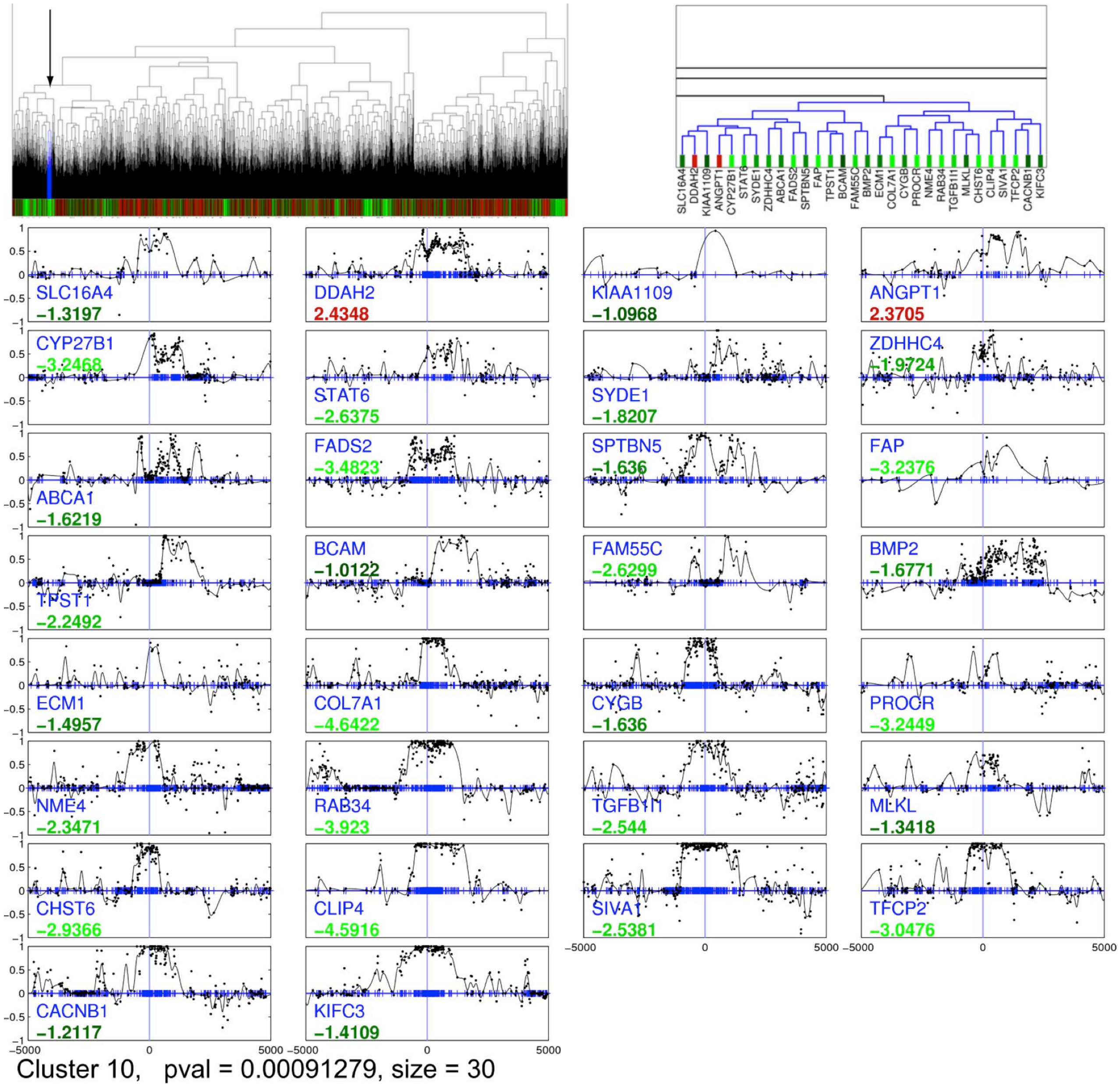


Figure S4 Full cluster for the genes from Figure 2B, comparing HMEC and HCC1954 WGBS data and representing the TSS2 pattern. Clustering was performed on a 10 kb region. Upper left panel shows the full dendrogram; this cluster is marked in blue. Expression values are plotted on a heat scale below the dendrogram (red is upregulated, green is downregulated). Upper right panel shows the section of the dendrogram with the identified cluster. Lower panel portrays the individual methylation signatures for each gene in the cluster. For each signature, the y-axis represents the methylation difference score (HCC1954 minus HMEC), ranging from -1 to 1. The x-axis denotes distance from the TSS, which is indicated by the vertical blue line at the center of each box. Blue tick marks indicate all CpG locations. Black dots represent all experimentally measured methylation difference scores. The black curve is the final signature used for clustering. Values in green denote the $\log_2$ fold expression change.

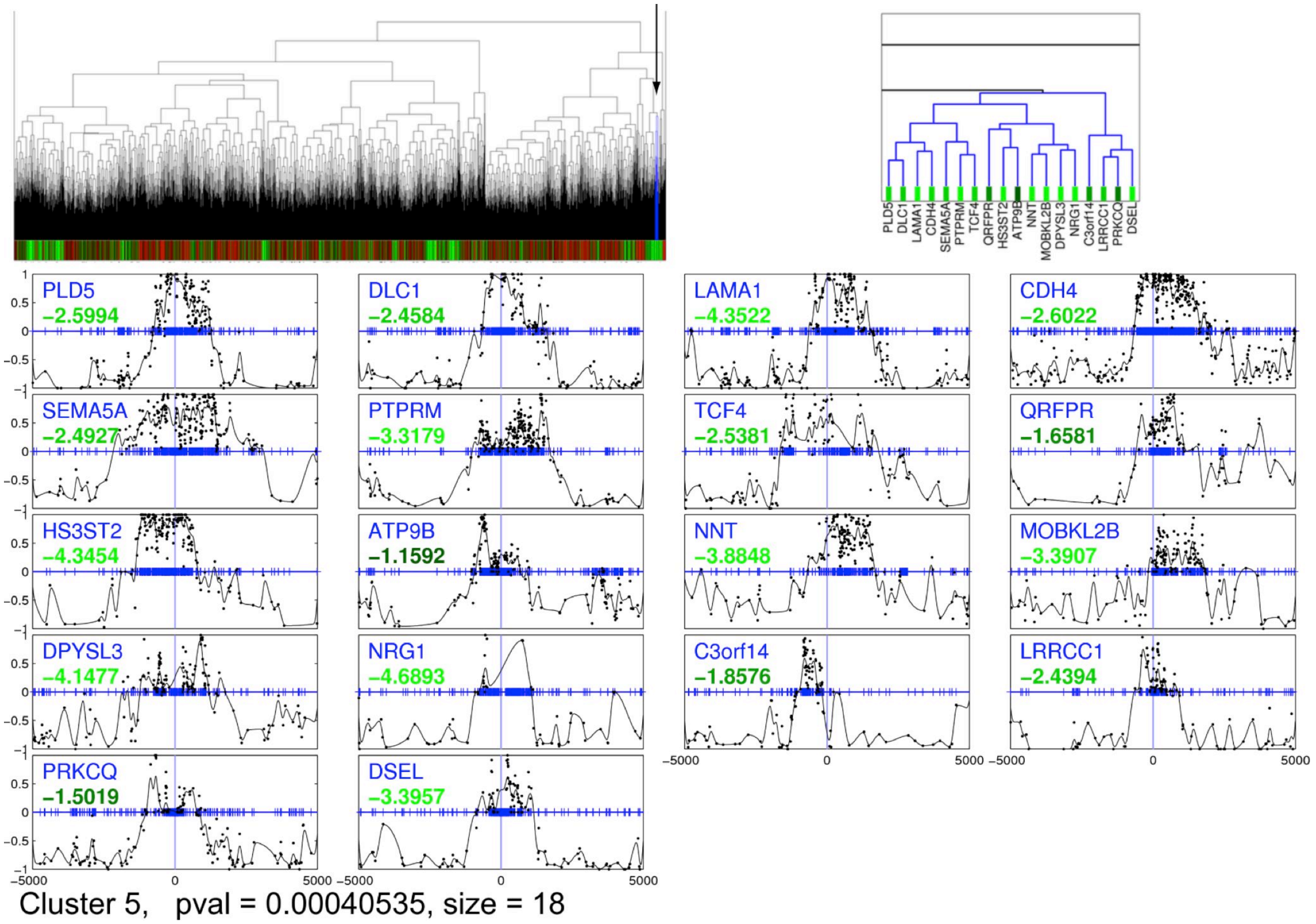


Figure S5 Cluster from HMEC-HCC1954 data containing a differential methylation pattern consisting of hypermethylation at the TSS set in long hypomethylated domains (TSS1 pattern). Clustering was performed on a 10kb region. We further characterized this signature by examining signatures in a 30kb region centered at the TSS, shown in Figure S6. For a full description of the parts of this figure, see Figure S4.

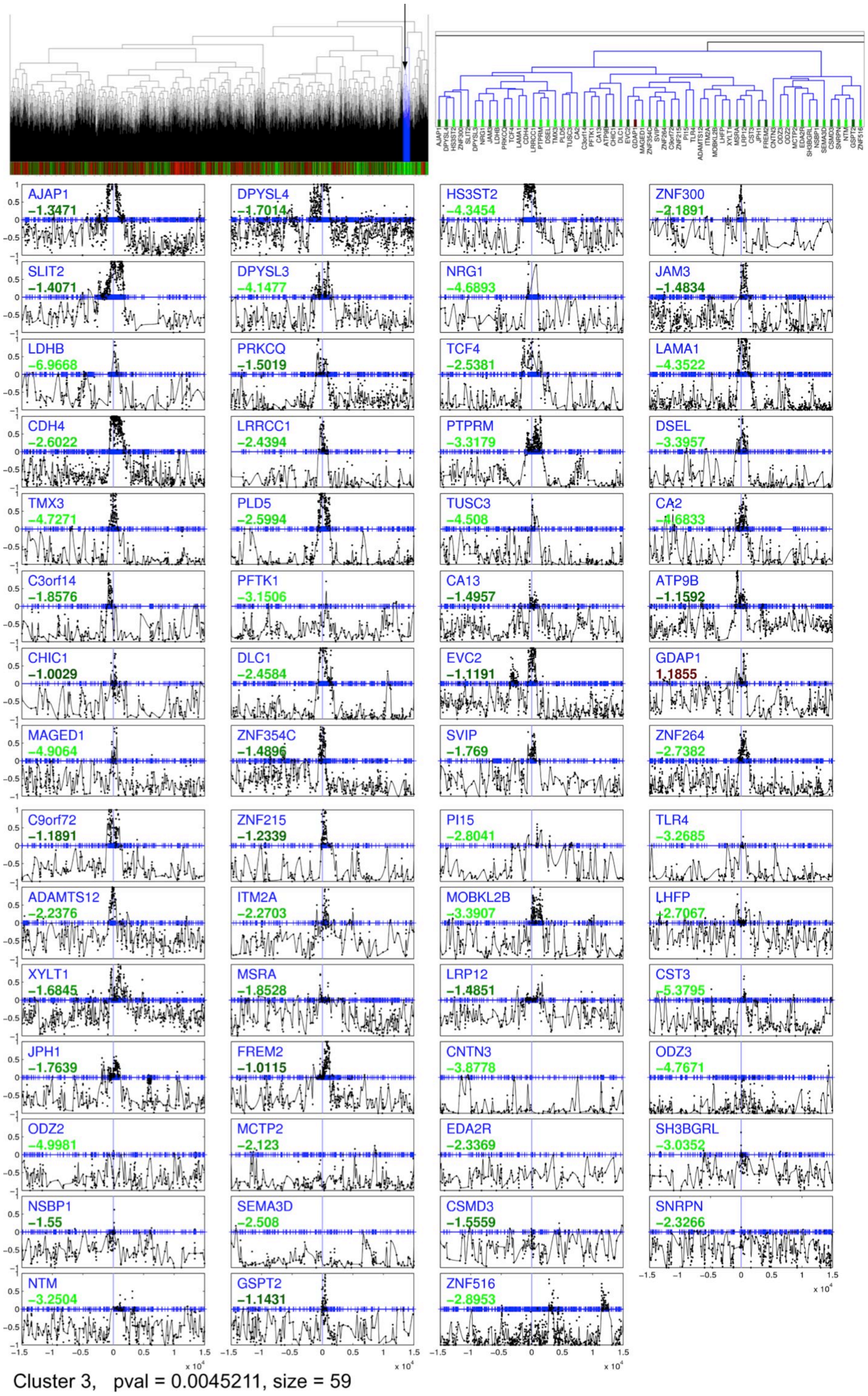


Figure S6 Full cluster for the genes shown in Figure 4A, containing a methylation pattern with hypermethylation at TSS set in long hypomethylated domains from HMEC-HCC1954 data (TSS1 pattern). Clustering was performed on 30kb regions. The last 13 genes of the cluster show the LONG0 pattern. The methylation signatures in this example are defined for the region ± 15kb of the TSS. For a full description of the parts of this figure, see Figure S4.

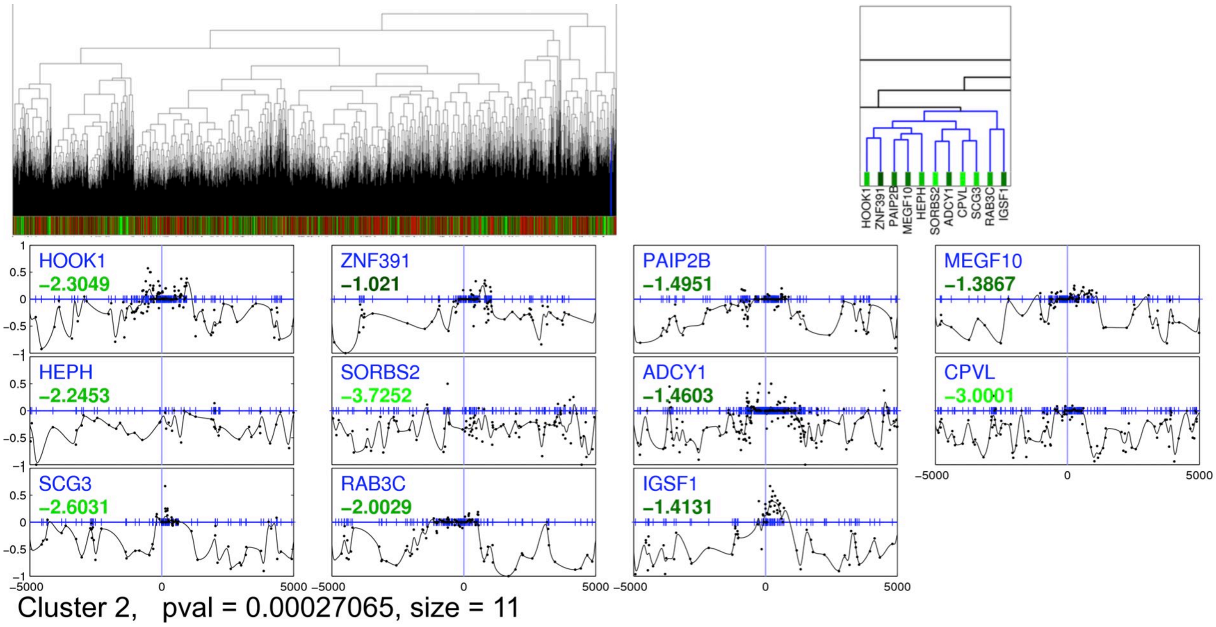


Figure S7 Example of the LONG0 pattern containing no methylation change at the TSS set in long hypomethylated domains from H9-IMR90 data. The methylation signatures in this example are defined for the region ± 5kb of the TSS. For a full description of the parts of this figure, see Figure S4.

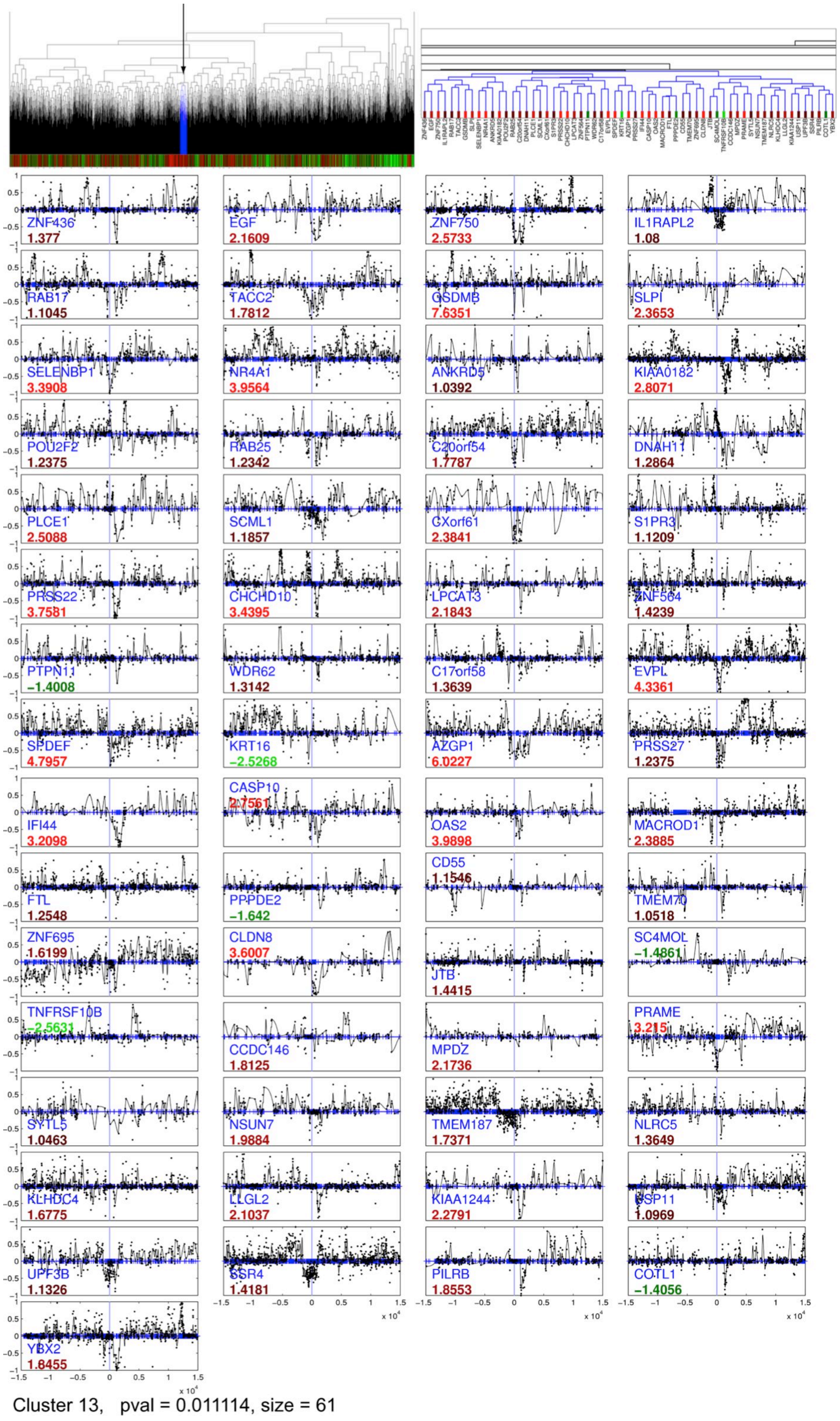


Figure S8 Full cluster for the genes shown in Figure 4B, containing a methylation pattern with hypomethylation at the TSS set in long hypermethylated domains (TSS1i pattern). Clustering was performed on 30kb regions. For a full description of the parts of this figure, see Figure S4.

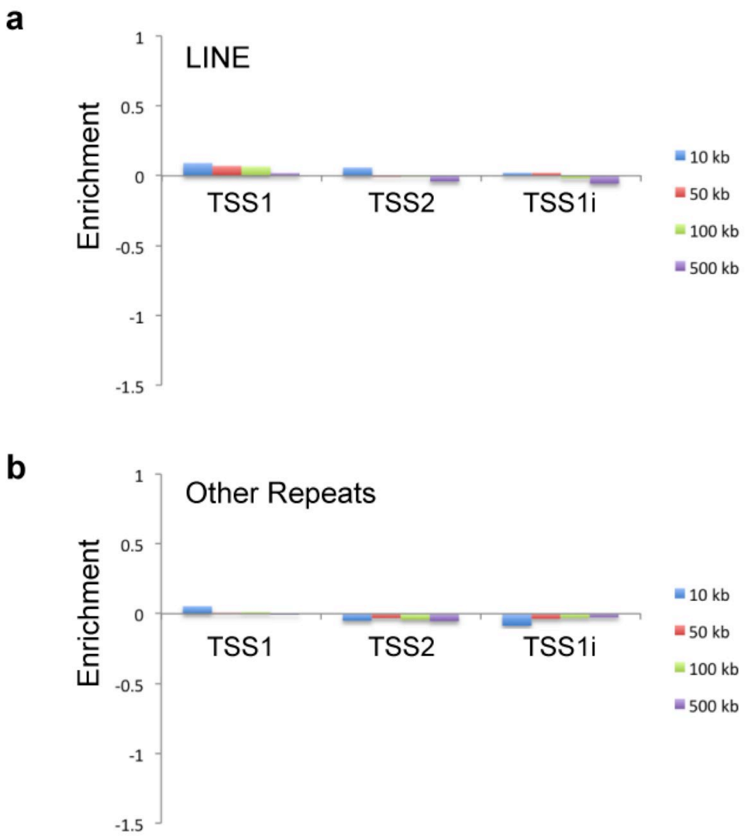


Figure S9 Lack of enrichment of LINE and other repeats identified by Repeatmasker in regions around TSS patterns. Plotted is the $\log_2$ fold enrichment of (**a**) LINE elements and (**b**) all repeats other than Alu and LINE elements.

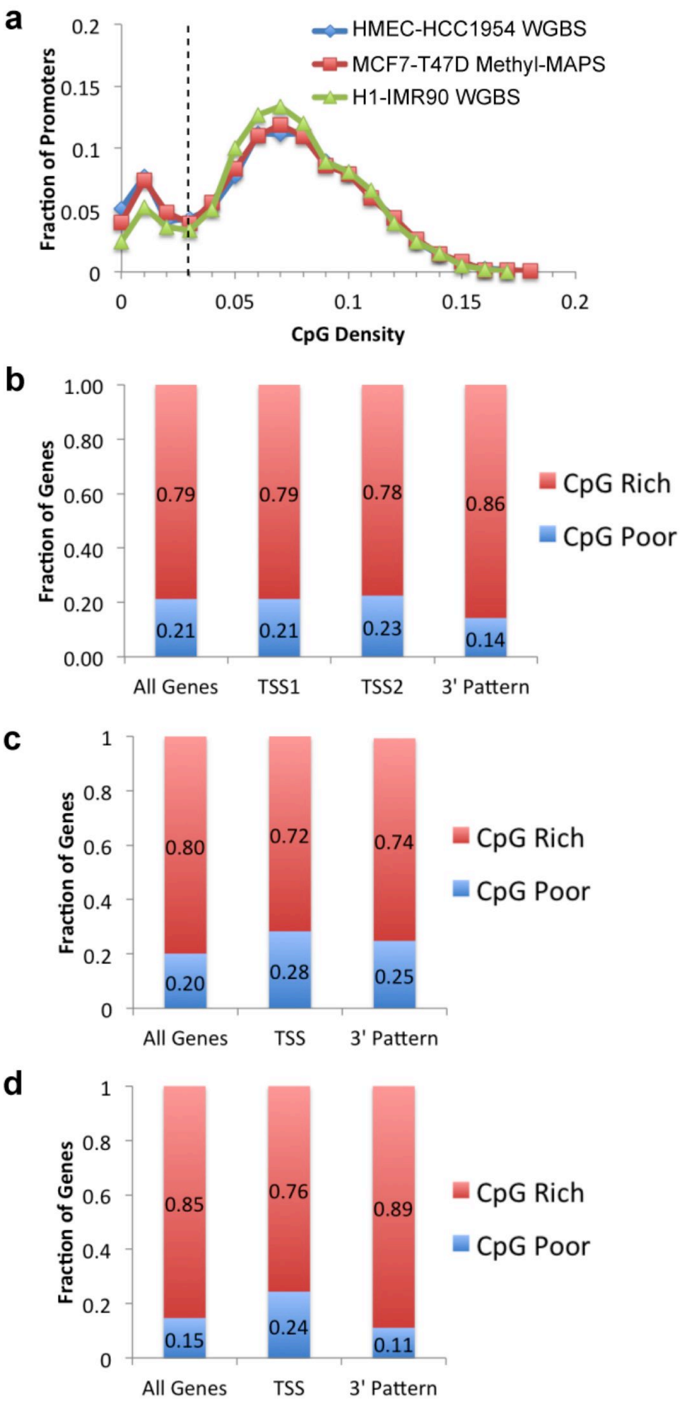
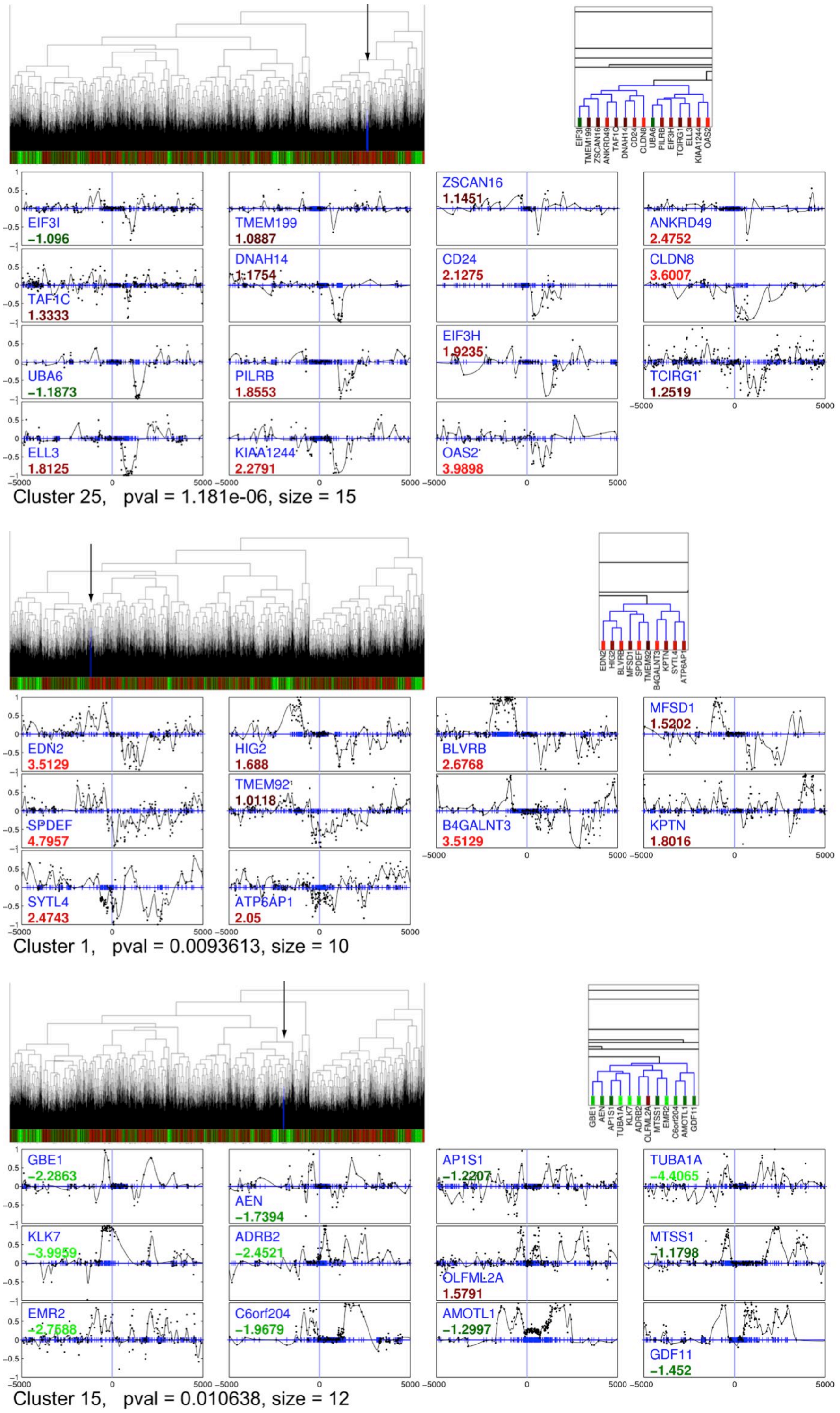


Figure S10 (**a**) CpG density of gene promoter regions in all datasets. The promoter region is defined as ± 500 bp from the TSS. Based on this distribution, we defined CpG-poor promoters as those with density smaller than 0.03 (dashed vertical line), and CpG-rich promoters as the remainder. The fraction of poor versus rich promoters is shown for different methylation patterns for (**b**) HMEC-HCC1954 WGBS, (**c**) MCF7-T47D Methyl-MAPS, and (**d**) IMR90-H1 WGBS data. The coverage skew at promoters in IMR90-H1 data (Figure S3) likely contributes to the increase in CpG-rich genes for this comparison. TSS patterns in IMR90-H1 and MCF7-T47D data are grouped since there were not enough examples to clearly separate TSS1 and TSS2 patterns.



Cluster 25, pval = 1.181e-06, size = 15

Cluster 1, pval = 0.0093613, size = 10

Cluster 15, pval = 0.010638, size = 12

Figure S11 Full clusters from HMEC-HCC1954 WGBS data for patterns characterized by changes in methylation 3' of the TSS. Examples were shown in Figure 2C (top panel, PROX3i pattern), Figure 5A (middle panel, PROX3i pattern), and Figure 5B (bottom panel, PROX2 pattern). Clustering was performed on 10kb regions. For a full description of the parts of this figure, see Figure S4.

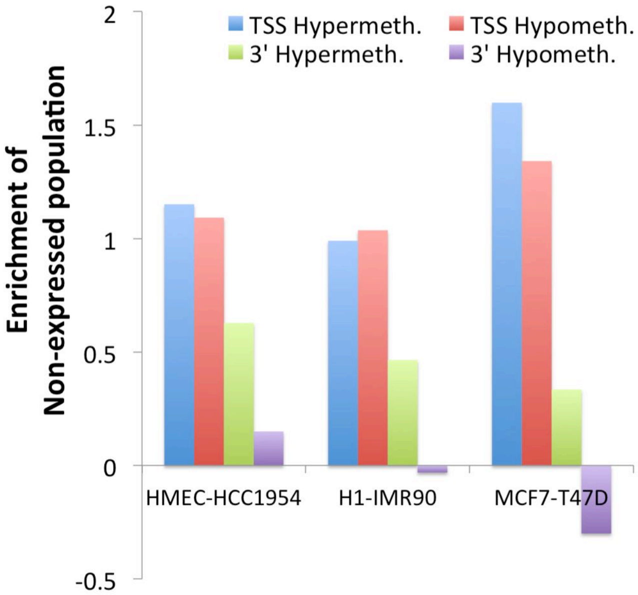


Figure S12 TSS patterns are enriched for genes whose expression is completely silenced, while 3' patterns show no enrichment (its genes tend to be downregulated instead of completely silenced). Shown here is the $\log_2$ fold change of non-expressed genes (cpm < 5) versus expressed genes (cpm >= 5) in the methylated sample. Counts per million (cpm).

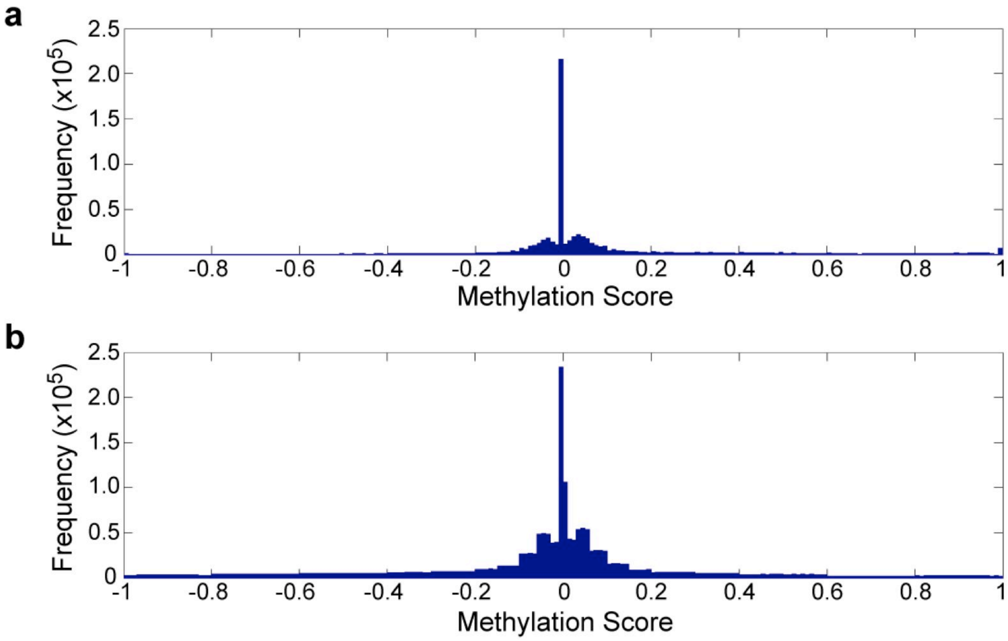


Figure S13 Histograms of methylation difference scores for (**a**) real HMEC-HCC1954 WGBS data and (**b**) simulated HMEC-HCC1954 data.

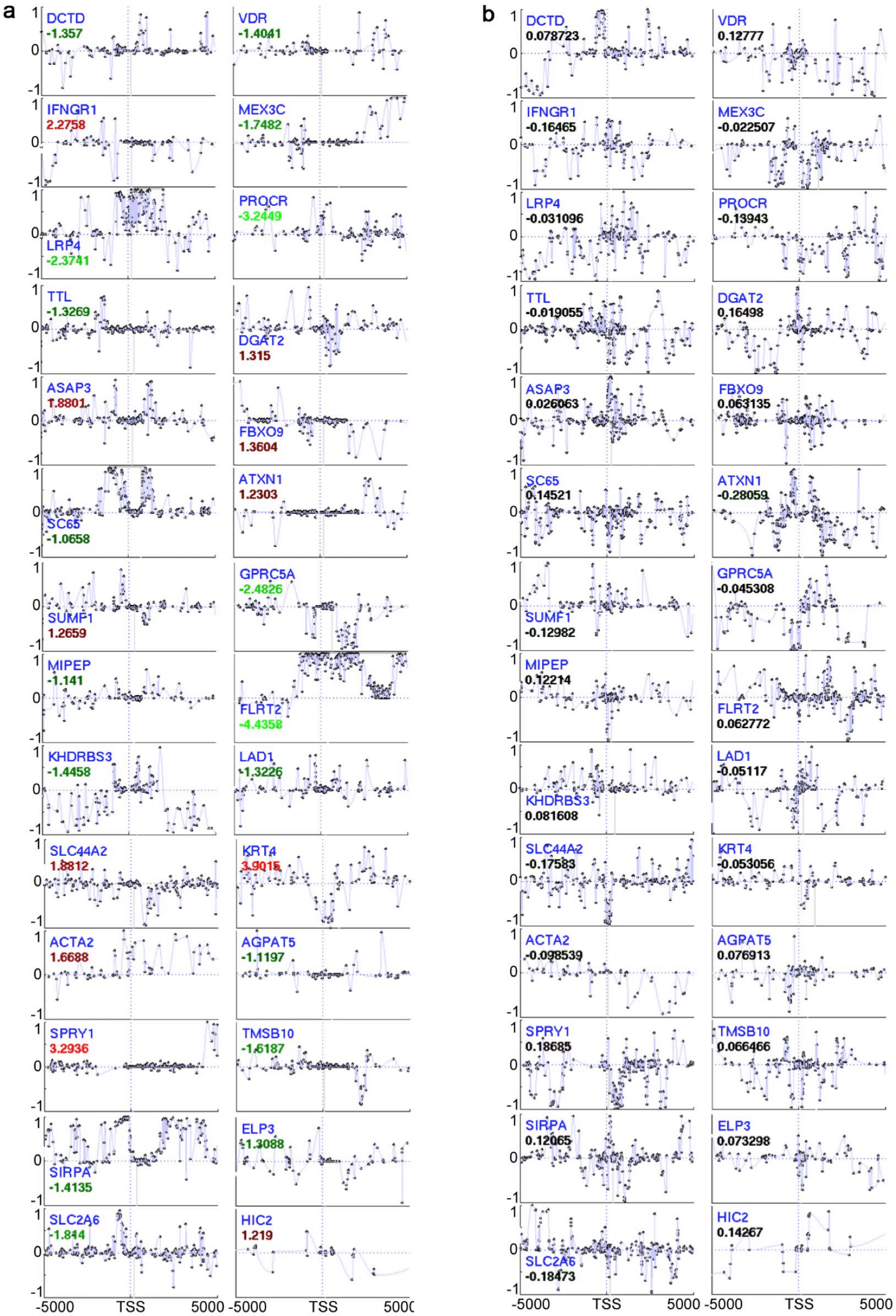


Figure S14 Real methylation signatures from HMEC-HCC1954 WGBS data (**a**) and simulated data for the same genes (**b**). Note that the individual curves should not be the same, but the simulated signatures should represent the background trends found in real data. For each signature, the y-axis represents the methylation difference score, ranging from -1 to 1. The x-axis denotes distance from the TSS, which is indicated by the vertical blue line at the center of each box. Black dots represent all experimentally measured methylation difference scores. The blue curve is the final signature used for clustering.

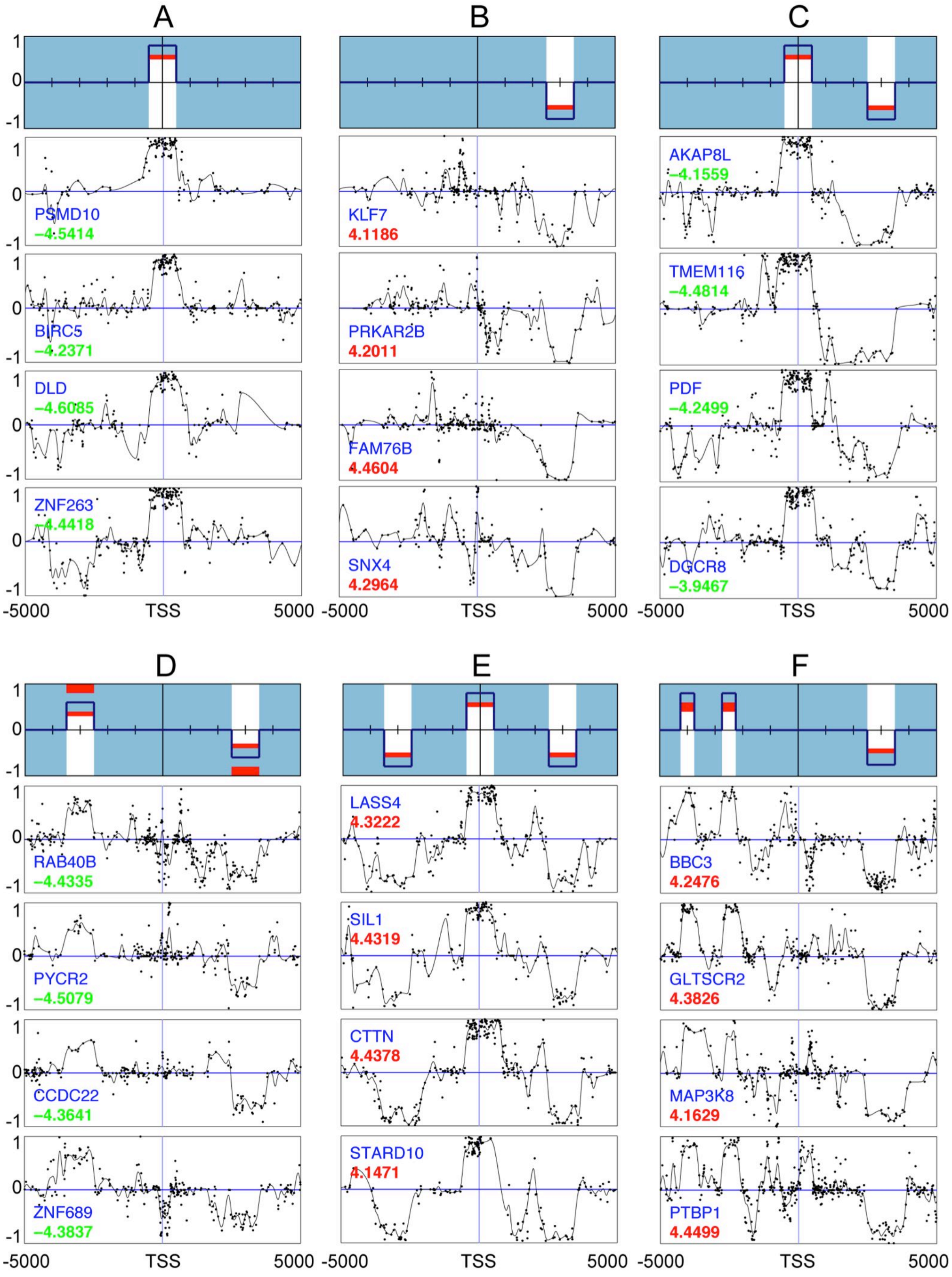


Figure S15 Six simulated patterns used to test the method's ability to detect different types of patterns. For each pattern (A-F), the top panel is a cartoon showing the target regions of the pattern. Blue shading depicts the areas through which a simulated curve with this pattern may be generated. The dark blue line traces the center of the allowable regions. White regions depict areas through which a curve with this pattern (positive class) may not pass, but a curve not designated with this pattern (negative class, background) may pass. Red regions depict buffer areas where neither positive nor negative class curves may pass. The lower panel for each pattern shows examples of positive curves simulated with the pattern. For both the cartoons and signatures, the y-axis represents the methylation difference score, ranging from -1 to 1. The x-axis denotes distance from the TSS, which is indicated by the vertical line at the center of each box. Black dots represent all experimentally measured methylation difference scores. The black curve is the final signature used for clustering.

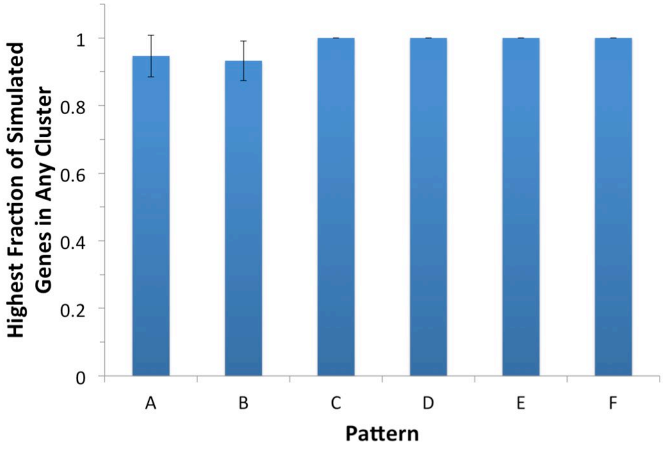


Figure S16 Performance of the pattern discovery method for the six patterns (A-F) described in Figure S15. For an experiment, if any cluster contains a high fraction of positive class genes, then the pattern is said to be discoverable. We record the highest fraction of positive class genes of all clusters and plot the average and standard error for three simulations. Every simulation for every pattern had at least one cluster containing at least 85% positive class genes, indicating a high likelihood that all patterns were discoverable in all cases.

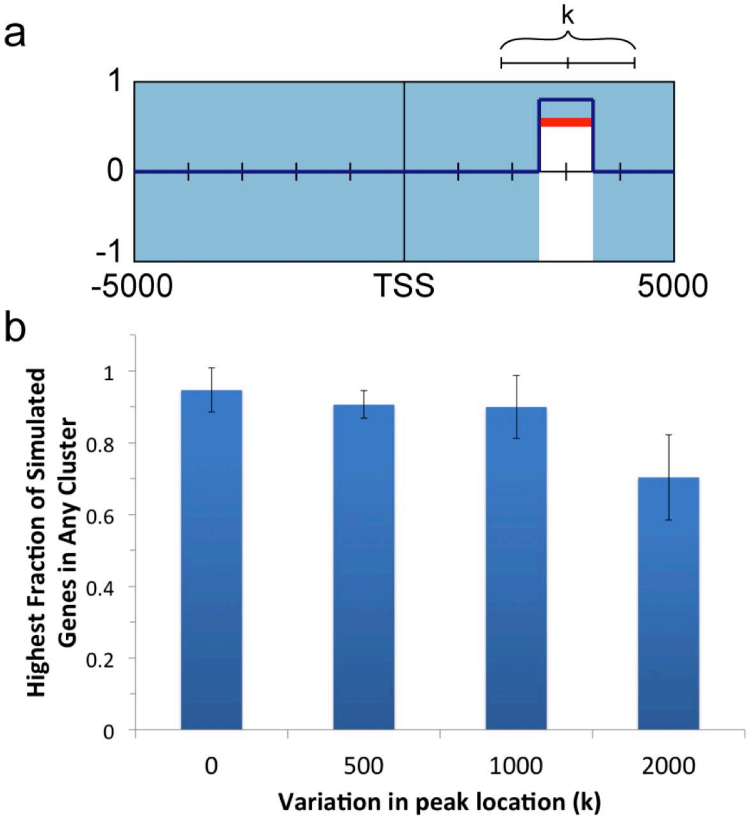


Figure S17 Performance of the pattern discovery method for a peak of fixed width, 1000 bp, whose location varies relative to the TSS. (**a**) A cartoon of the basic peak. Cartoon makeup is described in full in Figure S15. (**b**) Performance, as highest fraction of positive class genes in any identified cluster, for a set of widths over which the peak may range. *k* denotes the width of the region in which the peak may range, which is centered at 3000 bp downstream of the TSS. For example, for *k*=1000, some peaks will be centered as near as 2500 bp downstream of the TSS, and some will be centered as far as 3500 bp downstream of the TSS. The average fraction and standard error are plotted for three simulations for each width. For each *k* up to 1000 bp, at least one cluster contained a high fraction of positive class genes, indicating the pattern would be easily discoverable. Although some degradation occurred when the peak was allowed to vary by 2000 bp, some clusters were always found where the majority of genes were of the positive class. Since negative class (background) genes were allowed to have peaks within the region where positive class patterns could exist, it is likely that the performance drop is due to collisions with negative class genes that are very similar to positive ones.

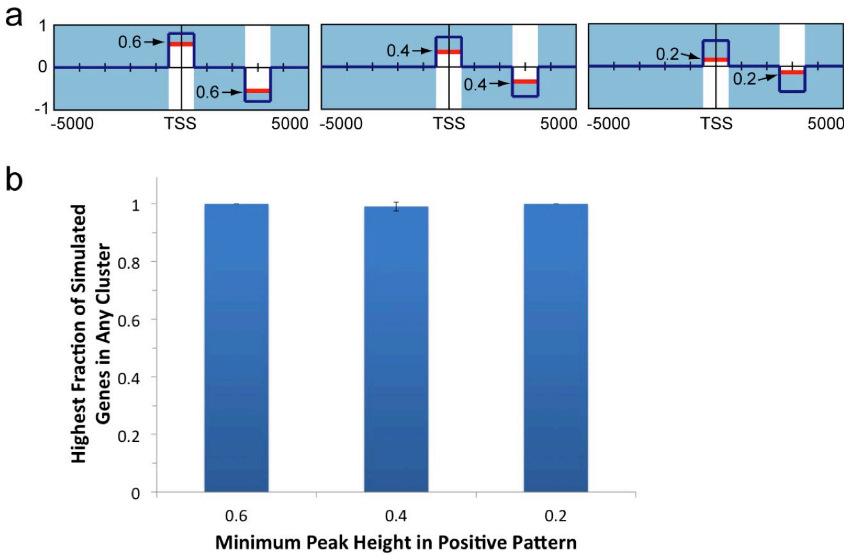


Figure S18 Performance of the pattern discovery method for a pattern with two peaks whose allowable regions vary in height. Three height ranges are used, and they are always symmetric between the two peaks. The ranges for the first peak are [0.6, 1], [0.4, 1], and [0.2, 1], respectively. (**a**) Cartoons depicting the three patterns. Cartoon makeup is described in full in Figure S25. (**b**) Performance, as the highest fraction of positive class genes in any identified cluster, for the three patterns. The average fraction and standard error are plotted for three simulations with each pattern. Every simulation for every pattern had at least one cluster containing at least 95% positive class genes, indicating a high likelihood that all patterns were discoverable in all cases.

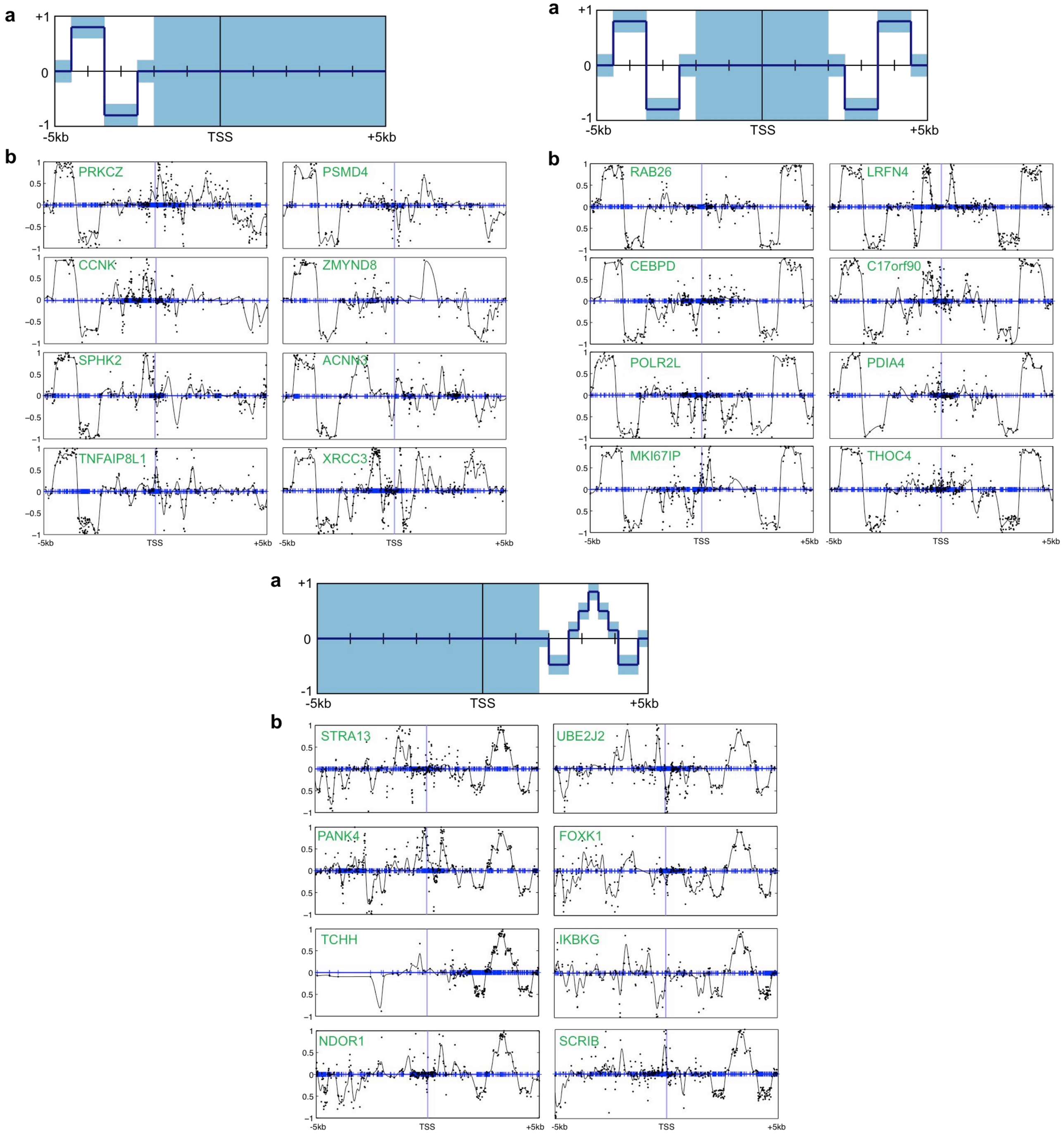


Figure S19 Simulation of the two_peak pattern (top left) four-peak pattern (top right) and Mexican Hat pattern (bottom). (**a**) Graphical depiction of the pattern. The blue box shows the allowable regions for curves generated using this pattern. The blue line traces the center of the allowable regions. No constraints are placed on the shaded blue region of the pattern. (**b**) Eight example curves simulated from each pattern. For each signature, the y-axis represents the methylation difference score, ranging from -1 to 1. The x-axis denotes distance from the TSS, which is indicated by the vertical blue line at the center of each box. Blue tick marks indicate all CpG locations. Black dots represent all experimentally measured methylation difference scores. The black curve is the final signature used for clustering.

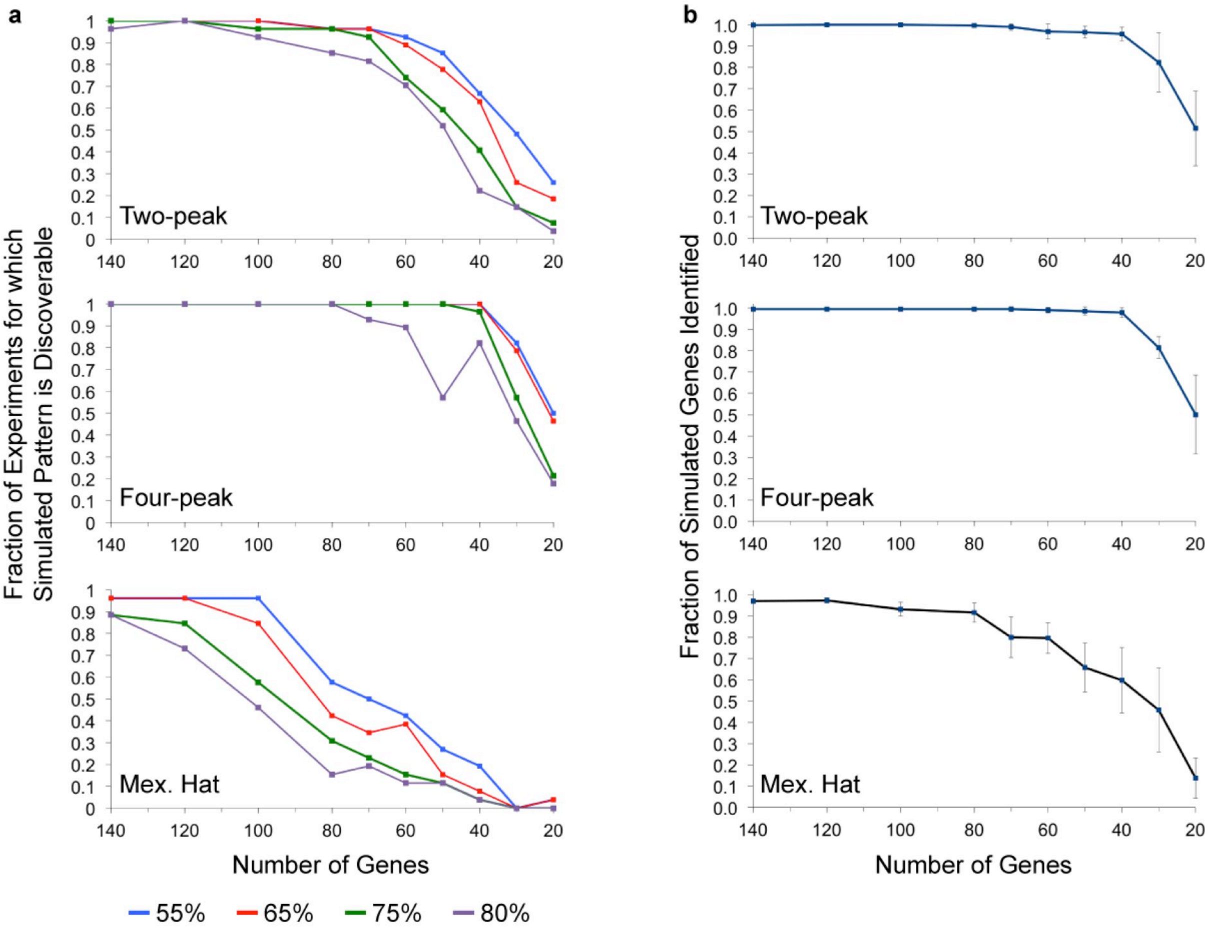


Figure S20 Performance of the discovery and gene list generation methods as varying numbers of simulated genes are added to the original set of 3560 real data signatures from the HMEC-HCC1954 dataset. The x-axes represent the number of simulated genes added using the appropriate pattern. (**a**) For a minimum of 25 experiments per data point, plots the fraction of experiments for which the simulated pattern is discoverable. The pattern is said to be discoverable if at least one cluster is identified such that a large percentage of its members come from the simulated set. The exact percentage is varied from 55% to 80%, and shown using four curves of different colors. (**b**) Fraction of simulated genes identified by the gene list generation method for each pattern. Error bars indicate standard error from the mean. A possible explanation for the worse performance of the Mexican Hat pattern is discussed Figure S21.

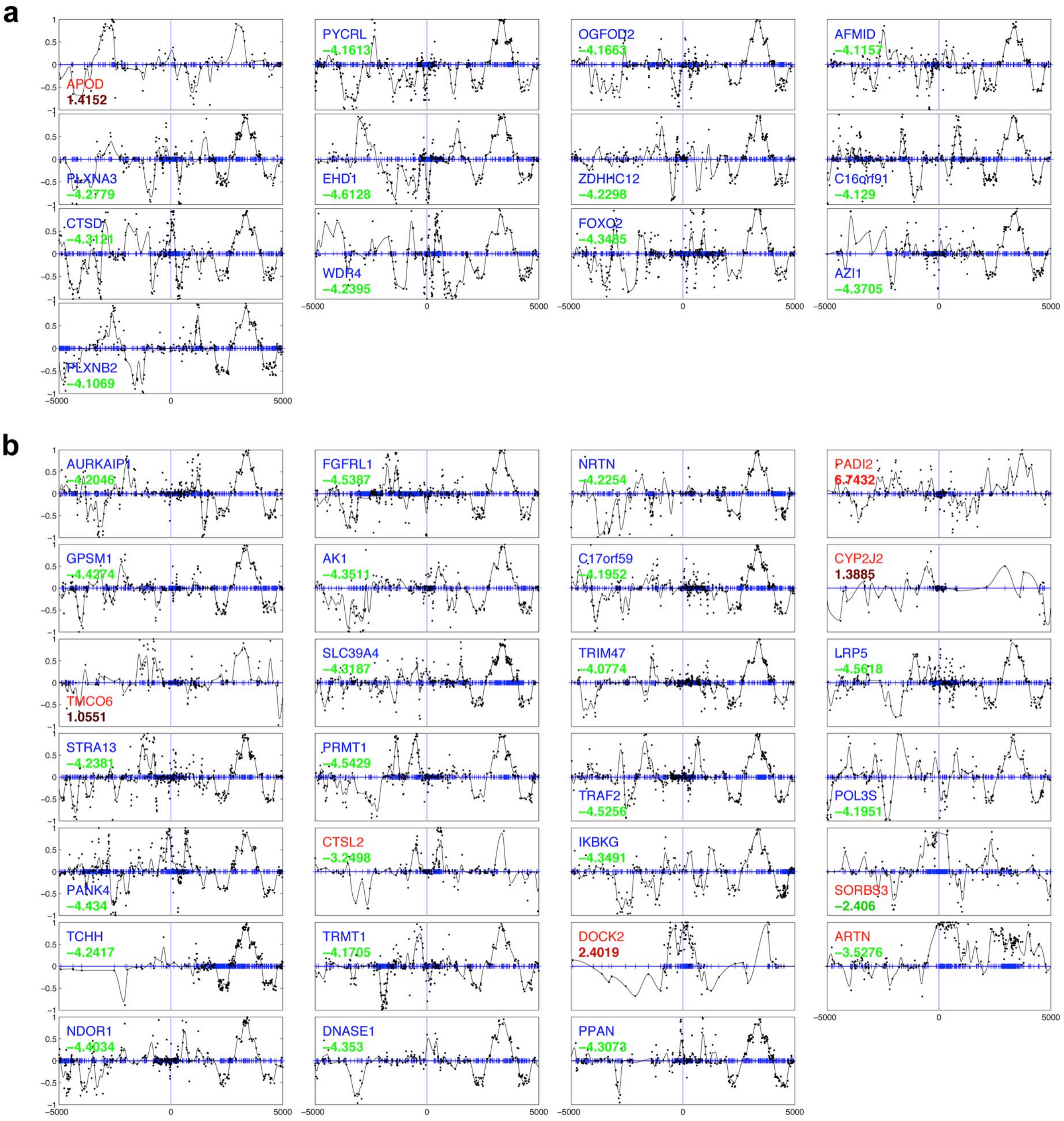


Figure S21 (**a,b**) Two example clusters from the experiments described in Figure S19, with Mexican Hat pattern simulation data added to real HMEC-HCC1954 data. The Mexican Hat pattern often collides with genes in the real data that have similar peaks out in the 3' side. Full descriptions of the methylation signature boxes are given in Figure S4. Blue gene symbols indicate simulated genes with the Mexican Hat pattern. Red gene symbols indicate genes that co-cluster with the Mexican Hat pattern but are background genes.

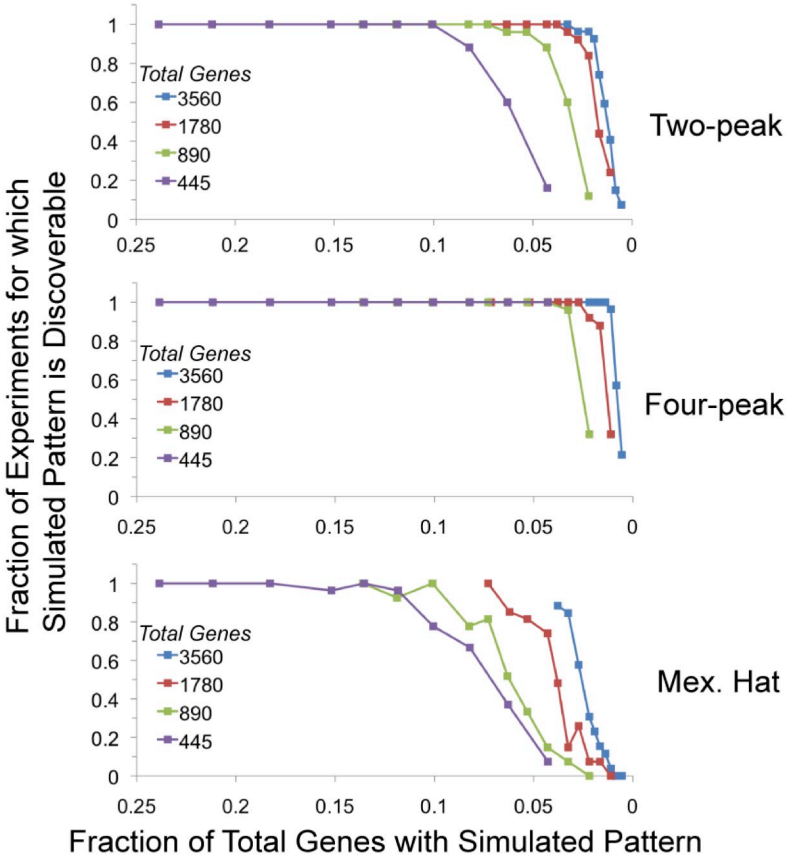


Figure S22 Performance of the discovery method varies with the size of the dataset. For each curve, a varying number of simulated genes are added to a set of real data signatures from the HMEC-HCC1954 dataset. Performance curves are plotted as functions of the fraction of genes in the experiment that are simulated. We see an improvement in the ability to discover patterns as the size of the dataset increases, even if the fraction of genes representing the pattern stays constant. For these plots, the pattern is said to be discoverable if 75% of the genes in some identified cluster are simulated genes. A minimum of 25 experiments are performed per data point. As an example, consider a simulated two_peak pattern introduced at a frequency of 5% of the dataset. If the dataset includes 445 total genes, the pattern can be discovered ~10% of the time. However, if the dataset includes 1780 total genes, it can be discovered almost 100% of the time.

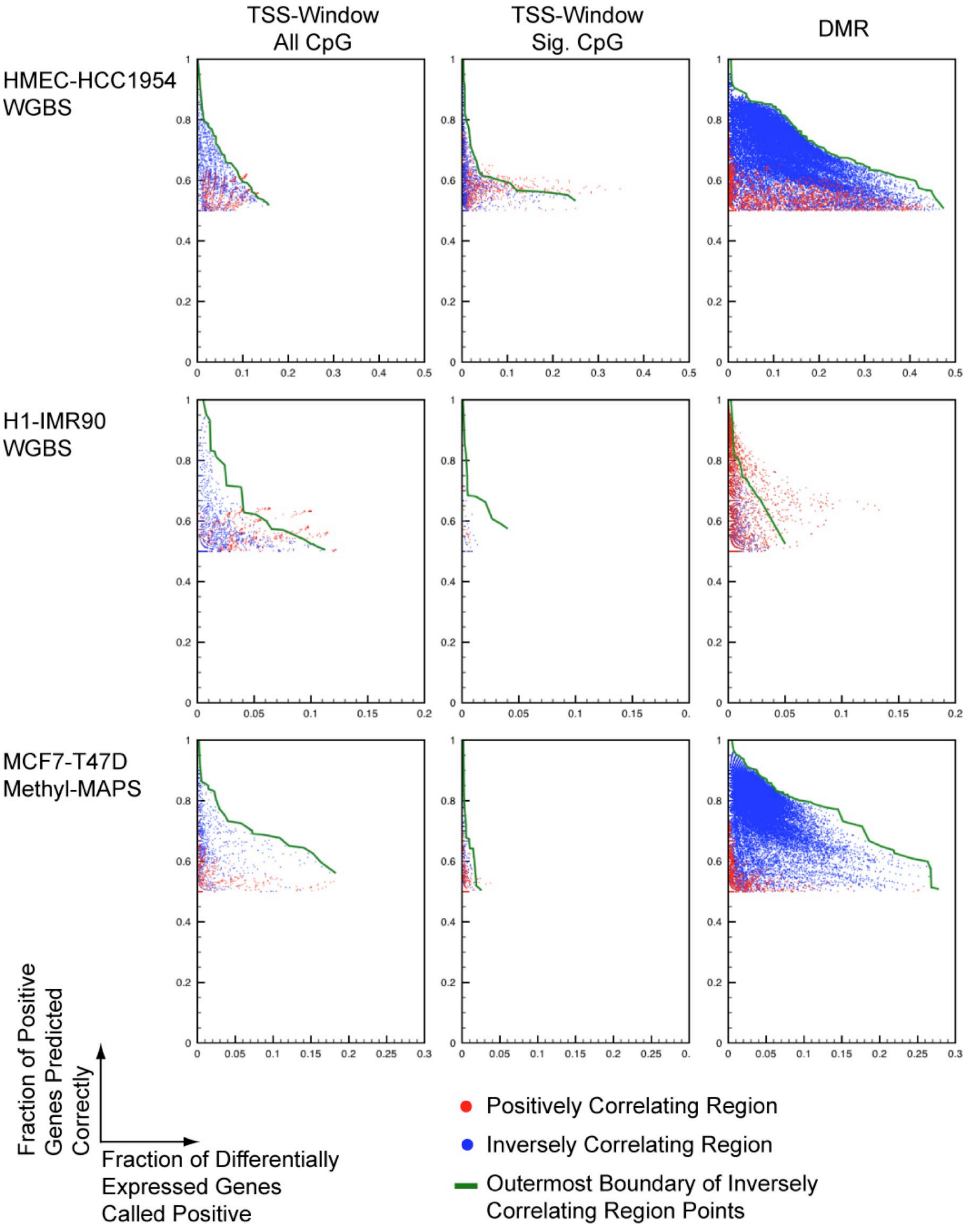


Figure S23 Optimization of DMR- and promoter-based approaches. Up and to the right indicates a better set of parameters. A complete list of parameters is in Table S3. Blue dots indicate performance assuming a negative correlation between methylation and expression; red dots assume positive correlation. The outer boundary points that make up the green line are plotted in Figure 6 and consist of the most optimal negative correlation parameter sets.

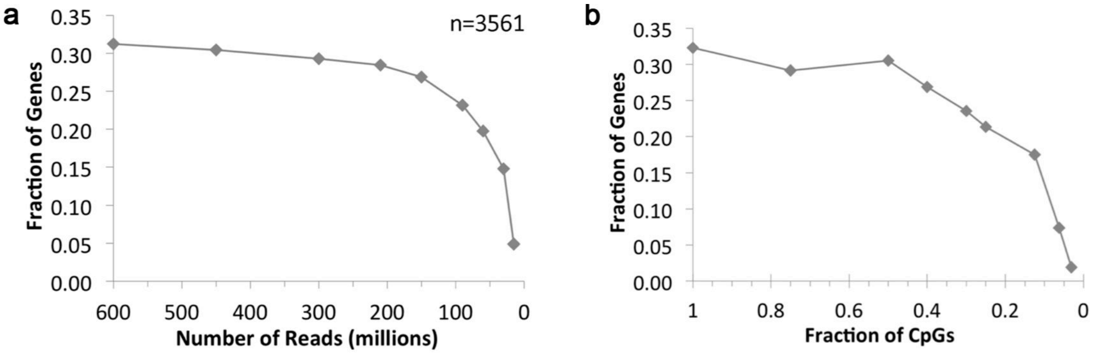


Figure S24 Performance as real data quality is degraded. (**a**) Performance curve as the number of reads decreases. Each data point represents a specific number of reads selected at random from the full set of reads comprising the HMEC-HCC1954 dataset. Performance is the number of genes identified by our gene list method using a minimum purity of 0.85. 600M reads corresponds to ~20x coverage, 300M to ~15x, 210M to 7x, etc. The y-axis is the fraction of genes identified out of the total number of differentially expressed genes (3,561) in the HMEC-HCC1954 dataset after filtering. (**b**) Performance curve for an experiment where data was removed at random from selected fractions of CpG sites, using the HMEC-HCC1954 dataset. Performance is the number of genes identified by our gene list method using a minimum purity of 0.85. The y-axis is the fraction of genes identified out of the total number of differentially expressed genes (3,561) in the HMEC-HCC1954 dataset after filtering.