SUPPLEMENTARY METHODS

Algorithm to Determine Non-overlapping Clusters

The purity of a set of genes, *G*, is defined as the fraction of genes that have expression change in the same direction as the majority:

$$purity(G) = \frac{\max(\,|\{g: g \in G \ \wedge \ exprdiff(g) \geq 2\}|,\ |\{g: g \in G \ \wedge \ exprdiff(g) \leq 2\}|\,)}{|G|}$$

where *exprdiff*(*g*) denotes the log-fold expression change for gene *g*. Next, all significant clusters meeting a minimum purity threshold are considered. Unless otherwise stated, we used a minimum purity of 0.85. Clusters are processed and added to a final list in decreasing order of purity. Each cluster c' is considered as follows. If c' does not overlap any other clusters already on the list, it is added. If c' is the descendent of another cluster that has been added to the list, it is not added. If c' is the parent of one or more clusters from the list, the purity of c' is compared to $purity(\,\bigcup_{c_i}\{g: g \in c_i\}\,)$, where $\{c_i\}$ are the children of c' already on the list. If the children's purity is within 0.3 of the purity of c', then the children are removed from the list and c' is added. Otherwise, the children are retained and c' is not added.

Methyl-MAPS Experiments

MCF7 and T47D cell lines were maintained in RPMI 1640 medium supplemented with 5% fetal bovine serum, 10 mmol/L HEPES, 4.5 g/L glucose, 2 mmol/L L-glutamine, 1 mmol/L sodium pyruvate, and 50 µg/ml gentamicin in a humidified 37°C incubator containing 5% carbon dioxide. Unmethylated and methylated compartments were obtained by limit DNA digestions as described in Rollins et al. (29) and Edwards et al. (10). Paired-end libraries were prepared from the methylated and unmethylated DNA compartments using an adaptation of Applied Biosystems' SOLiD System Mate-paired Library Preparation Protocol (10). Data filtering, normalization and methylation score calculation was performed as in Edwards et al. (10). A summary of the sequencing statistics is in Supplementary Table S4. Methyl-MAPS data has been deposited in GEO (GSE45337).

RNA-Seq Experiments

RNA-Seq libraries were constructed using the NEBNext Kit from New England Biolabs. Briefly, 10 ug of total RNA was isolated from MCF7 and T47D cells, DNase treated, and twice oligo(dT) selected using the Dynabeads mRNA purification kit (Invitrogen). Isolated mRNA was subsequently fragmented using the supplied RNA fragmentation buffer. The fragmented mRNA was randomly primed and reverse-transcribed. After second-strand synthesis, the cDNA was end-repaired, ligated to barcoded adaptors, size selected on agarose gel (200-400 bp) and PCR amplified. The libraries were sequenced using the

Illumina Genome Analyzer IIx according to the manufacturer's instruction. Two replicates were performed for each sample. RNA-Seq data has been deposited in GEO (GSE45337).

Expression Analysis

RNA-Seq data from MCF7 and T47D cells were experimentally obtained as stated above. A summary of the sequencing statistics is in Supplementary Table S4. RNA-Seq data for H1 and IMR90 cells was downloaded from GEO Accession Number GSE16256, and for HMEC and HCC1954 cells from GEO Accession Number GSE29119. RNA-Seq reads were mapped to the human genome (hg18) using Tophat v. 1.3.3 (30). Aligned reads were filtered to eliminate reads that mapped to rRNA and RNA repeats (snRNA, scRNA, srpRNA, tRNA and RNA). Htseq-count (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html) was used to obtain raw read counts based on Ensembl gene annotations (hg18 v54) using the union method. Genes mapping to ribosomal and mitochondrial proteins were filtered prior to differential testing. Differentially expressed genes (FDR = 0.05 and 2 fold change) were identified using edgeR (31) with TMM normalization (32) and tagwise dispersion. Genes that had no samples with expression of at least 5 cpm were removed prior to differential comparison. A floor on the quantile normalized pseudocounts was applied before the fold-change is calculated. This avoids artificially high fold changes resulting from one sample having a very small number of counts and omits genes that were not expressed in either sample. The floor for the HMEC, HCC1954, IMR90, and H1 samples is 5, and the floor for the MCF7 and T47D samples is 10. Genes with an expression change less than two-fold are discarded. Ensembl gene annotations were converted to RefSeq IDs using the corresponding tables from the UCSC Genome Browser.

DMR and Promoter-Based Approach Optimization

DMRs were found using a two-pass sliding window approach (9,14). The following algorithm is repeated for positively- and negatively-differentially methylated sites independently. In the first pass, a fixed-width window is slid across the genome until it identifies an area with at least a minimum number of differentially methylated sites. The window is then slid, and the DMR extended, until it no longer contains the minimum number sites. After all preliminary DMRs are identified, a second pass removes all DMRs shorter than a certain length or lacking a minimum number of CpGs. Only CpG sites with a minimum difference in methylation between the samples are considered. A site is considered differentially methylated according to Fisher's Exact Test, controlling the FDR with the Benjamini-Hochberg procedure, at the desired significance level. Optimizations are considered both with and without the second pass window. Genes were assigned methylation state of the closest DMR within the distance threshold set. A full list of parameters is in Supplementary Table S2.

Differentially methylated promoters were identified for the promoter-based approach by calculating the average methylation change of all CpGs in the window between the two samples or by calculating the average methylation change at only significantly differentially methylated CpGs. Differentially methylated

CpGs were identified as for the DMR approach above. Calculations were performed on various windows around the TSS. Windows were created by varying the window's size upstream and downstream of the TSS. A full list of parameters is in Supplementary Table S2. For each approach the best points (see Supplementary Figure S19) with an inverse correlation between methylation and expression were identified and plotted in Figure 5.

Metagene Analysis of Individual Patterns

Metagene analyses for **Figure 3C**,**E** were performed by averaging methylation scores across all CpG sites with minimum coverage appropriate for that sample in a sliding window positioned relative to the gene's TSS. For 20 Mb regions, the sliding window was 1000 bp and was shifted 100 bp for each calculation. For 20 kb plots, the sliding window was 100 bp and shifted 1 bp for each calculation. CpG density was calculated in the same fashion, but for all CpG sites independent of sample coverage.

Simulated Background Methylation Data

The most straightforward way to obtain realistic DNA methylation data would be to reuse existing methylation curves and simply generate new expression values as necessary. However, the very presence of the correlations we seek to discover heavily influences the existence and frequency of methylation patterns throughout the data. Any set of existing curves will be highly enriched for the patterns that actually correlate with expression in the real dataset, regardless of the expression labels. Instead, we need an approach that can simulate unbiased background methylation data without any underlying, enriched patterns. We simulated differential methylation across the genome using a modified second order Markov chain, as follows. For each CpG site, we recorded both the differential methylation levels of the two previous sites and the distances from each of the two previous sites to the CpG that follows them. Histograms of differential methylation given this information at two previous sites were created from the HMEC-HCC1954 dataset. To simulate the methylation level at a CpG site, we select a bin from the histogram corresponding to the distances and methylation levels of the two previous sites, according to the relatively probabilities of each bin. For the special case of the first CpG on each chromosome, the methylation value is selected at random from the overall distribution of differential methylation in the sample pair. For the special case of the second CpG site on each chromosome, we randomly select from the sum of the appropriate histograms.

Once a histogram bin is selected, the exact differential methylation value is chosen at random, uniformly, from the range of the bin. The methylation bins were chosen to reflect the distribution of differential methylation values, including one tight bin around 0 to capture the large number of CpGs with no methylation change between samples. The methylation bin boundaries were: -1, -0.8, -0.6, -0.4, -0.3, -0.2, -0.15, -0.1, -0.06, -0.03, -0.0001, 0.0001, 0.03, 0.06, 0.1, 0.15, 0.2, 0.3, 0.4, 0.6, 0.8, and 1. The distance bins were chosen so that all bins contained roughly the same number of sites. The distance bin boundaries, in base pairs, were: 1, 10, 25, 80, 200, and infinity. Too many bins for these histograms

would lead to undersampling when training on a fixed-size dataset; the number of bins used here was near maximal for the HMEC-HCC1954 dataset without experiencing undersampling.

In real datasets, the methylation level often remains constant over a run of many bases. Reproducing this phenomenon would require decreasing the bin sizes to the point of severe undersampling. Instead, we added a special case for whenever the new bin prescribed by the model is the same as the previous. We stored an additional set of histograms representing the differential methylation within the constant bin. These bin boundaries were: -0.2, -0.15, -0.1, -0.07, -0.04, -0.02, -0.01, -0.00001, 0.00001, 0.01, 0.02, 0.04, 0.07, 0.1, 0.15, and 0.2. During the course of simulation, if a new site's methylation bin is chosen to be the same as the methylation bin for the previous site, a bin is chosen from this additional histogram. A value is selected at random from the chosen bin and added to the precise value of the previous site to determine the value at the new site. If the new value does not lie in the differential methylation bin that was originally selected the process is repeated.

The overall distributions of differential methylation in the real data and our simulated genome are compared in Supplementary Figure S22. Examples of simulated differential methylation signatures for randomly chosen genes are in Supplementary Figure S23. The real methylation signatures are included in Supplementary Figure S24 for comparison. Our approach models the local behavior of differential methylation for a given sample pair. It is intended to be a reasonable approximation of what background differential methylation would look like without directly modeling larger, unknown features in a dataset.

Simulated Methylation Patterns

The background simulation strategy is used to create neutral methylation patterns with no non-local structural features. To introduce a pattern into the dataset, we first determine the target regions that the pattern will govern in each fixed window for each simulated signature. The regions are defined in relation to the TSS, which has a constant position in each window. We then describe the desired range of differential methylation for each pattern for all target regions. The pattern is added to a randomly-chosen subset of the data during the simulation of the genome by restricting the random selection from the appropriate histogram across the target regions. Any region not covered by the pattern definition is simulated using the background strategy. Curves for which a pattern is introduced are called positive examples, and curves that do not exhibit the pattern are called negative examples. For all negative examples, the random selection is restricted across the pattern's target regions to ensure that the pattern does not occur by chance.

The fraction of curves containing each pattern is set for each experiment. In addition to setting the boundaries of the ranges for each target region, we include a buffer to allow for optional separation between positive and negative examples. No values can be sampled within the buffer for positive or negative examples. Examples are shown in Supplementary Figure S25.

Simulated Expression Data

Positive examples can have either upregulated or downregulated differential expression values, depending on the pattern. Upregulated expression is simulated by drawing an expression value for the first sample from a normal distribution with $\mu=10$ and $\sigma = 1$, then drawing a value for the second sample from a normal distribution with $\mu=200$ and $\sigma = 10$. Downregulated expression is simulated by drawing from the same distributions in the opposite order. Negative examples' differential expression values are simulated by drawing a value at random for each sample from one of these two normal distributions.