5100–5111 *Nucleic Acids Research, 2017, Vol. 45, No. 9* *Published online 7 February 2017*
doi: 10.1093/nar/gkx078

Modeling complex patterns of differential DNA methylation that associate with gene expression changes

Christopher E. Schlosberg, Nathan D. VanderKraats and John R. Edwards*

Center for Pharmacogenomics, Department of Medicine, Washington University in St. Louis School of Medicine, St. Louis, MO, USA

Received January 05, 2017; Editorial Decision January 24, 2017; Accepted January 26, 2017

ABSTRACT

Numerous genomic studies are underway to determine which genes are abnormally regulated by DNA methylation in disease. However, we have a poor understanding of how disease-specific methylation changes affect expression. We thus developed an integrative analysis tool, Methylation-based Gene Expression Classification (ME-Class), to explain specific variation in methylation that associates with expression change. This model captures the complexity of methylation changes around a gene promoter. Using 17 whole-genome bisulfite sequencing and RNA-seq datasets from different tissues from the Roadmap Epigenomics Project, ME-Class significantly outperforms standard methods using methylation to predict differential gene expression change. To demonstrate its utility, we used ME-Class to analyze 32 datasets from different hematopoietic cell types from the Blueprint Epigenome project. Expression-associated methylation changes were predominantly found when comparing cells from distantly related lineages, implying that changes in the cell's transcriptional program precede associated methylation changes. Training ME-Class on normal-tumor pairs from The Cancer Genome Atlas indicated that cancer-specific expression-associated methylation changes differ from tissue-specific changes. We further show that ME-Class can detect functionally relevant cancer-specific, expression-associated methylation changes that are reversed upon the removal of methylation. ME-Class is thus a powerful tool to identify genes that are dysregulated by DNA methylation in disease.

INTRODUCTION

Establishment of specific patterns of DNA methylation at CG dinucleotides (CpGs) is necessary for normal development (1,2), and aberrant methylation is frequently observed in cancer (3,4). CpG rich-regions, often called CpG islands (CGIs), are typically unmethylated and associated with ∼70% of mammalian gene promoters (5). Hypermethylation of CpG islands overlapping the transcription start site (TSS) is hypothesized to downregulate tumor suppressor genes, thus promoting tumorigenesis (6,7). Typically, promoters are labeled as either methylated and silenced or unmethylated and potentially active based on the methylation levels near the transcription start site (TSS) (8,9). However, studies that rely upon this simple binary characterization (10) to correlate methylation with expression find only modest negative correlations with expression levels (11–13).

The most common approach to associate DNA methylation and expression change is to first identify differentially methylated regions (DMRs) and then associate them with nearby genes. Numerous statistical tools have been developed to identify DMRs (10). Generally, DMRs are found by segmenting the genome into equally spaced regions and identifying which regions have statistically significant differences in methylation. DMRs are then associated with genes or other genomic regulatory elements within a certain distance to gain biological insight into their potential function. While DMR-based methods have been critically important in identifying imprinted loci (14), studies often find only weak correlations between DMRs near gene promoters and differential gene expression (11,12,15). One drawback of DMR methods is that they rely on a set of arbitrarily defined thresholds for the size and number of CpGs to include in the DMR. It is often recommended to adjust these parameters for each individual dataset since the choice of these parameters has substantial implications in the numbers of DMRs identified and putatively associated genes.

One possible reason DMR methods fail to find a strong association between differential methylation and expression is they reduce DNA methylation to a single differential

*To whom correspondence should be addressed. Tel: +1 314 362 6935; Fax: +1 314 362 8844; Email: jredwards@wustl.edu

value removed from its local context. Recent work, however, has indicated that a large number of methylation patterns associate with differential gene expression (16). For example, methylation at CpG island-shores, regions of decreased CpG density flanking CpG islands, correlate with differential gene expression in colon cancer (17). Further, long hypomethylated domains in cancer often contain down-regulated genes (17). Positive correlations between gene body methylation and gene expression have also been frequently observed (18,19).

Here, we present a new approach to predict gene expression changes that accounts for all methylation changes around the TSS. We have previously shown the importance of capturing methylation changes around the TSS to find patterns of methylation change that associate with expression changes using an unsupervised approach (16,20,21). We now build upon these results to develop a supervised method called ME-Class (Methylation-based Expression Classification), which classifies differential expression using signatures of differential methylation.

We use ME-Class to investigate alternate representations of DNA methylation and CpG density to identify methylation features that are most important in predicting expression change using data from the Roadmap Epigenomics Project. We then use ME-Class to examine the role methylation associated expression changes play in hematopoiesis using data from the Blueprint Epigenome project. Lastly, we demonstrate that ME-Class can identify a set of genes with cancer-specific expression-associated DNA methylation changes that are silenced in tumor cells, but that are re-expressed when methylation is removed.

MATERIALS AND METHODS

Roadmap Epigenomics Project (REP) WGBS and mRNA-seq

Samples from 17 primary tissues with matched whole genome bisulfite sequencing (WGBS) and RNA-seq were obtained from the Roadmap Epigenomics Project (REP, Supplementary Table S1) (22). Fractional methylation (M) is defined as mCG/CG. Differential methylation (ΔM) is defined as: $\Delta M = M_{S_2} - M_{S_1}$, where S_1 and S_2 correspond to the first and second sample in the differential comparison, respectively. We obtained consolidated methylation data, which was previously cross-assay standardized and uniformly processed. All CpG sites were filtered for 4× coverage or greater and analysis was performed using the hg19 genome assembly according to analysis standards established in REP. We used uniformly processed protein-coding gene level annotations from Genecode V10 to obtain standardized FPKM values. Each Genecode V10 annotation was converted to RefSeq annotations using the mygene python package (23). To create a standardized gene set with high quality methylation data, we excluded genes with ambiguous or incomplete TSS annotations, genes shorter than 5 kb, genes with <40 CpGs assayed within ±5 kb of the TSS, genes where all CpGs within ±5 kb of the TSS had <0.2 methylation change, and alternative promoters. These filters were used to exclude non-coding and pseudogenes, genes shorter than the interpolation boundary, genes with low numbers of CpGs to reduce bias caused by individual CpGs, and genes with no methylation changes near their promoter, respectively. We only included RefSeq genes with cdsStartStat ($n = 47\,637$ genes) and cdsEndStat ($n = 47\,621$ genes) with 'cmpl' according to the UCSC Table Browser. To eliminate redundant annotations, for any RefSeq genes with multiple RefSeq IDs corresponding to the same TSS location, we only used a single RefSeq ID with the lowest accession number. In analyses with the ROI classifier (see below), all genes with less than four exons were removed from analysis. Differentially expressed genes were defined as genes with ≥2-fold difference between samples after an applied floor of five FPKM to provide a conservative estimate of expression change. These filtering criteria have minimal effect on the fraction of CpG Island (CGI) -associated promoters, which went from 65.9% of genes before filtering to 67.3% after. CpG Islands were defined based on the CGI track from the UCSC Genome Browser (24). A full summary of filtered gene counts is in Supplementary Table S2.

Blueprint Epigenome project WGBS and mRNA-seq

WGBS and RNA-seq from 32 venous and cord blood samples were obtained from the Blueprint Epigenome project (25). Genome coordinates from hg38 were converted to hg19 using liftOver (24). All other analysis steps were performed identically to the REP data above.

Cancer WGBS and mRNA-seq

Breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), rectum adenocarcinoma (READ) and uterine corpus endometrial carcinoma (UCEC) matched normal-tumor WGBS and mRNA-seq samples were obtained from The Cancer Genome Atlas (TCGA) to train a model of tumorigenesis. Normal sigmoid colon (E106) WGBS and mRNA-seq data were obtained from REP (22). HCT116 DKO1 (bi-allelic knockout of DNMT1 and DNMT3b) WGBS and mRNA-seq data were obtained from Blattler *et al.* (26). All other analysis steps were performed identically to the REP data above.

Single window (SW)

We computed the average methylation across a single, fixed ±1 kb window around the TSS of each gene (17,27). We performed logistic regression to predict differential expression from the average methylation change around the TSS (Figure 1A and B). Logistic regression cross-validation was run with 1000 maximum iterations of the optimization algorithm.

Differentially methylated regions (DMRs)

We used DSS-single to compare DMRs between individual samples (28). We identified DMRs ($P < 0.01$) and used their size (bp), average differential methylation, and stranded distance (bp) to the closest TSS (0 if overlapping by ≥1 bp) as features for gene expression change classification with a Random Forest (RF) classifier with 1001 estimators (Figure 1A and B).

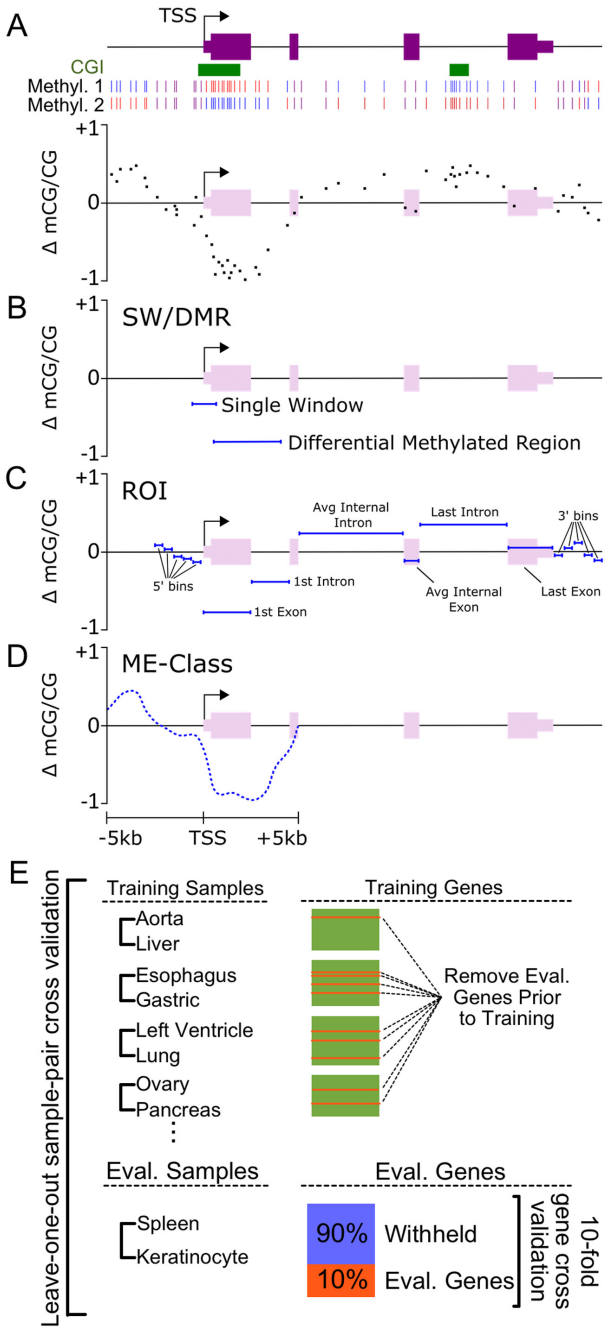

Figure 1. Models of DNA methylation and validation framework for predicting differential gene expression change from differential DNA methylation. (**A**) Heat map indicates methylation status at individual CpG sites—red is fully methylated, blue is fully unmethylated—for an example gene in two samples (Methyl. 1 and Methyl. 2). Individual points below indicate differential DNA methylation (Methyl. 2—Methyl. 1) across the example gene at individual CpG sites. (**B**) Example regions that would be used to calculate the single window (SW) and differentially methylated region (DMR) using the data in (A). (**C**) Regions used to calculate methylation features for the Region of Interest (ROI) representation of the gene in (A). (**D**) ME-Class representation of the gene in (A). Each individual point is the differential methylation value used as a feature in a Random Forest after interpolation and smoothing. (**E**) Cross-validation comparison framework. Evaluation is performed sample-wise across the 17 sample comparisons. In the evaluation comparison, genes are split into 10-folds. Prior to training, each evaluation gene is removed for all other tissues. Further model details are in Table 1. CGI = CpG island.

Regions of interest (ROI)

The Regions of Interest (ROI) classifier reduces DNA methylation to multiple averaged values across annotated gene elements (upstream, exon, intron and downstream bins) as features for a Random Forest (RF) to predict expression class (Figure 1C). ROI classifier features were implemented as described in (29) to predict differential expression class rather than single sample binned expression values. We used a RF classifier with 100 trees as originally described. Increasing the number of trees to 1001 increased run time substantially without an appreciable increase in performance (Supplementary Figure S1).

ME-Class

Gene signatures were constructed as in VanderKraats *et al.* with minor modification (16). This signature allows the model to incorporate the entire profile of methylation changes across the gene's promoter region including any CGI and CGI-shore regions (Figure 1A and D). In addition, these signatures allow comparison of methylation differences between genes, which have CpGs in different locations. We applied a localized z-score normalization of each differential methylation value in a 10 kb window surrounding the TSS based upon the distribution of methylation values in a 100 kb surrounding anchor window. We created methylation signatures using a piecewise cubic hermite interpolating polynomial (PCHIP) to interpolate a curve of z-score normalized differential methylation values in the 10 kb window around the TSS for each differentially expressed gene. The interpolated curve was then subjected to Gaussian smoothing with a bandwidth of 50 bp. Since CpG methylation values are highly autocorrelated (30), interpolation and smoothing of the data decrease the influence of sequencing error at individual CpGs (16). Similar smoothing approaches have shown a marked improvement in the ability to determine DMRs (31). To obtain discrete features, we subsampled our interpolated methylation signature at 20 bp resolution. We then used these features with a RF classifier with 1001 estimators. We initially chose a RF classifier since it provides a nonparametric model, has a low number of hyperparameters, generates an internal unbiased estimate of testing error, identifies feature importance, and typically performs near-optimally with minimal tuning (32). Logistic Regression (LR) (max_iter = 1001), Gradient Boosted Classification Trees (GBCT) (n_estimators = 1001), Gaussian Naïve Bayes (NB), L2 distance-based k-Nearest Neighbors (kNN) ($k = 21$), and Dynamic Time Warping (DTW) kNN ($k = 21$) were implemented with default parameters other than stated modifications. All machine learning methods were implemented with scikit-learn and mlpy (DTW only) python packages (33,34).

Whole gene methylation models

We also implemented three alternative representations of methylation data to incorporate the full profile of methylation changes across the entire gene (Supplementary Figure S2A): Whole Scaled Gene (WSG), Whole Gene (WG), and Uniform Gene Features (UGF). For each representation, we created 125 bins in the regions 5 kb upstream of the

TSS and downstream of the TES (Transcription End Site) (20 bp resolution). These regions/features were then added to specific features for each representation as follows. The WSG representation is an emerging representation in the literature to describe methylation changes by linearly scaling the methylation profile across the entire gene (22,26,35). To obtain discrete features, we used 500 bins across the gene (Supplementary Figure S2B). For the WG representation, we modeled methylation data as a curve (subsampled to 20 bp resolution) across the entire length of the gene (Supplementary Figure S2C). For the UGF approach, we represented each exon with 10 scaled bins and each intron with 30 scaled bins. Multiple exons or introns were not averaged together (Supplementary Figure S2D). For WSG, we used a RF classifier. For both WG and UGF, we used curve similarity as defined with DTW (36), and classified expression changes using kNN ($k = 21$).

Classifier performance

To evaluate the amount of data needed to train ME-Class, we divided the 17 REP datasets into eight samples that were held out for evaluation and nine samples that were used for training. For a given number of training samples (n), nine random permutations of n pairwise comparisons were chosen from the training samples and used to train nine ME-Class classifiers. The resulting nine classifiers were then evaluated on a fixed set of eight comparisons from the holdout evaluation samples (Supplementary Figure S10).

To evaluate each classifier, we implemented a conservative two-stage cross-validation framework (Figure 1E) to ensure that the model does not overfit to any given sample or individual gene. For a given evaluation, we performed the following procedure: (i) leave-one-out sample pair cross validation: We divided all differential training samples into a training set and an evaluation set. This ensured that no individual sample from the training set appears in the evaluation set; (ii) 10-fold gene cross validation: we randomly divided the genes from the evaluation sample into 10-folds. To evaluate each fold, we first removed examples of all genes in the evaluation fold from all samples in the training set prior to training. Thus, if gene A is in the evaluation data, no examples of gene A for any tissue are used for training. We then trained on the training samples/genes and evaluated the chosen fold of genes in the evaluation samples. We then repeated this process 10 times for each fold of the evaluation sample and for each differential sample in the dataset. This process helped the classifier generalize across genes and samples by ensuring that each evaluation gene does not observe either an example of itself or any other genes from its individual sample.

Average performance based on the accuracy, reject rate, positive predictive value (PPV), and negative predictive value (NPV) was reported across all genes treated as a single pool from all samples. Testing accuracy was defined as the number of genes with correctly predicted expression divided by total genes returned. Receiver Operator Characteristic (ROC) and Precision Recall (PR) curves were averaged to provide a sample-wise level of reporting. RF feature importance was estimated as Gini importance. Unless otherwise stated, all statistical comparisons were performed using FDR-corrected paired, pairwise Wilcoxon rank sum test in R.

Gene ontology analysis

Gene Ontology analysis was performed by analyzing gene lists with Functional Annotation Clustering with default parameters from DAVID (37). Blueprint analysis gene lists were identified by any genes with $\geq$90% probability of classification in $\geq$2 differential samples. Colon cancer gene lists were identified by any down-regulated genes with a $\geq$90% probability of classification by ME-Class (see discussion of reject rate in Results).

RESULTS

ME-Class predicts gene expression change from differential methylation in tissue samples

Since the goal of most genome-wide methylation studies is to identify how changes in methylation alter expression, we examined the ability of methylation to predict differential expression change. We first sought to understand whether a methylation signature approach (i.e. modeling the entirety of methylation changes around a gene's TSS) could outperform current DMR, single window (SW), and ROI methods in finding genes with associated differential methylation and expression (Figure 1, Table 1). To compare supervised classifiers, we used WGBS DNA methylation and RNA-seq data from the Roadmap Epigenomics Project for 17 tissue samples (Supplementary Table S1) (22). We implemented a conservative sample-wise and 10-fold gene-wise cross-validation framework that ensures the genes in the evaluation step have not been seen in any tissue during training (Figure 1E). Since patterns of differential DNA methylation can be very similar between datasets, this evaluation framework tests the strength of the DNA methylation representation and the universality of DNA methylation patterns rather than the ability to simply recall an observed gene's methylation signature.

Using this framework, ME-Class outperformed all methods by receiver operator characteristic (ROC) curve analysis ($P < 10^{-3}$ for ME-Class compared to each of DMR, SW and ROI, Figure 2A) and precision-recall (PR) analysis ($P < 10^{-3}$ for ME-Class compared to each of DMR, SW and ROI, Supplementary Figure S3C). In addition, ME-Class performed better than or equal to each of the other methods analyzed for any individual comparison based on ROC AUC (area under the curve; Figure 2C). Interestingly, there was a large variability in the classification performance of the differential samples. While in the case of keratinocyte comparisons this could likely be explained by poor sequencing coverage (Figure 2E and F), this per-sample performance difference appeared primarily to be due to fundamental differences in the methylation profiles of the expression classes. The ROC AUC was strongly correlated (Figure 2D, $R^2 = 0.87$) with the average methylation difference between up- and down-regulated genes in the region +0.5 kb to +2.5 kb relative to the TSS (the most important region for classification, Figure 2B). While we cannot rule out there is some other technical artifact in the data causing this

Table 1. DNA methylation features and classification methods for each model in Figure 1

DNA methylation representation	Features	Gene expression change classification method
Single Window (SW)	ΔmCG/CG ±1 kb of TSS	Logistic regression (LR)
Differentially Methylated Regions (DMR)	Distance from TSS to DMR (bp) DMR width (bp) Avg. ΔmCG/CG	Random forest (RF)
Regions of Interest (ROI)	Avg. ΔmCG/CG: Five 400 bp bins 5′ of TSS First Exon First Intron Avg. internal exon Avg. internal intron Last exon Last intron Five 400 bp bins 3′ of txEnd	Random forest (RF)
Methylation-based Expression Classification (ME-Class)	ΔmCG/CG of 500 bins (20 bp) ±5 kb of TSS	Random forest (RF)

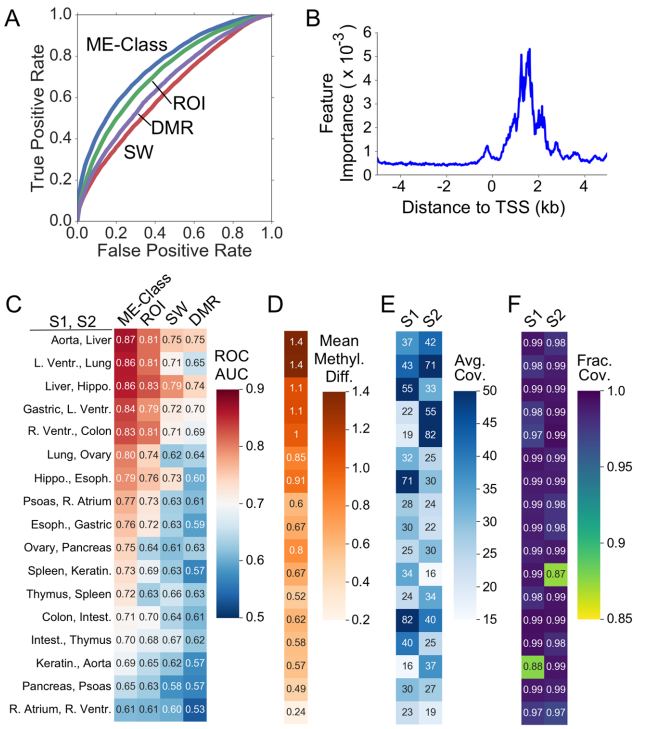

Figure 2. ME-Class outperforms standard methods for tissue-specific expression classification. Methods were evaluated using 17 tissue samples from the REP with the two-stage cross-validation framework in Figure 1E. (**A**) ROC analysis from a combination of all 17 samples (ROC AUC: ME-Class, 0.76; ROI, 0.71; DMR, 0.63; SW, 0.66). (**B**) RF feature importance from ME-Class trained on all 17 differential comparisons from REP. (**C**) ROC AUC for each of the 17 samples comparisons. (**D**) Mean *z*-score normalized methylation difference of the region [+0.5 kb, +2.5 kb] relative to the TSS. (**E**) Average CpG coverage and (**F**) average fraction of CpGs within the 10 kb window around the TSS for each REP sample. S1 and S2 correspond to the first and second sample in the evaluation differential comparison. L. Ventr. = Left ventricle, Hippo. = Hippocampus, R. Ventr. = Right ventricle, Esoph. = Esophagus, R. Atrium = Right Atrium, Keratin. = Keratinocyte, Intest. = Intestine.

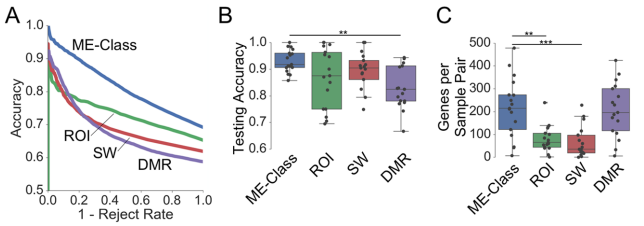

Figure 3. ME-Class identifies more genes at higher accuracy with expression-associated methylation changes in tissue-specific differential comparisons in the REP dataset. (**A**) Classifier accuracy versus 1—reject rate. (**B**) Accuracy of testing sample at 90% operating probability of classification. (**C**) Number of genes identified with expression-associated methylation changes at 90% operating probability of classification. Points in (B) and (C) indicate the performance of individual REP sample comparisons. ** indicates $P < 0.005$ and *** indicates $P < 0.001$. All other comparisons were not significant for $\alpha = 0.05$.

correlation, it appears that most normal tissue methylation-based expression classification derives from TSS 3′ proximal methylation changes.

ME-Class generates a list of genes with associated differential methylation and expression

Transcription can be influenced by multiple factors other than DNA methylation, such as transcription factors or chromatin modifications. Thus, the key issue is whether ME-Class can be used to identify a subset of genes that have high quality associations between differential methylation and expression. For this purpose, we introduced a reject rate into the classifier that allows us to control for external factors other than methylation that indicate gene expression. The reject rate excluded genes that cannot be reliably predicted (i.e. they likely do not have methylation-associated expression changes) using a threshold for the probability of classification output by each classifier. In practical terms, the reject rate allows one to set a parameter based on the cross-validation evaluation error that can control the false positive rate when running ME-Class on unseen samples. In Figure 3A, we observe that ME-Class outperformed ROI,

SW and DMR methods in accuracy and proportion of the genes returned across all rejection rates.

We next set the classification probability at 90% and examined how many genes were returned by each method and the accuracy of this list. ME-Class returned significantly more genes than the SW and ROI methods (ROI: $P = 2.9 \times 10^{-3}$; SW: $P = 9.2 \times 10^{-4}$) and was significantly more accurate at 90% probability of classification than the DMR method (DMR: $P = 3.0 \times 10^{-3}$) (Figure 3B and C). ME-Class returned the largest average number of genes, 217, at the highest level of accuracy (93.1%). The ROI and SW methods achieved lower levels of accuracy and returned a much lower average number of genes (ROI: 81 genes, 86.9% accuracy; SW: 66 genes, 84.5% accuracy). The DMR method returned a similar average number of genes, 207, but at the cost of a much lower level of accuracy (83.3%). This implies that the nearest isolated DMR is often insufficient to predict the expression class even when tuning the DMR parameters to find optimal segmentation parameters.

At 90% probability of classification, ME-Class did not show any bias towards the positive (up-regulated) or negative (down-regulated) class (Supplementary Figure S3A and B). ME-Class matched or exceeded the accuracy given the probability of classification, indicating that this probability can be used as an estimate of the final classification error in cross-sample comparisons (Supplementary Figure S3D). This demonstrates that when running ME-Class on new samples, the probability of classification can be used as an estimate of the false discovery rate. Meanwhile, the ROI, SW and DMR approaches all have a lower accuracy than the probability of classification at high probabilities indicating that they are likely overfit and do not generalize well to other datasets.

To understand the difference in why ME-Class was highly predictive of some genes and not others, we examined metagene plots of the methylation signal at predicted genes subset by the probability of prediction. We observed that the highest predicted genes have the greatest methylation difference between downregulated and upregulated genes in a [+0.5 kb, +2.5 kb] region around the TSS of each gene, and that this signal decays with decreasing probability (Supplementary Figure S4). This dampening of the methylation signal can likely be attributed to either noise from the WGBS assay or biological noise from cell-type heterogeneity.

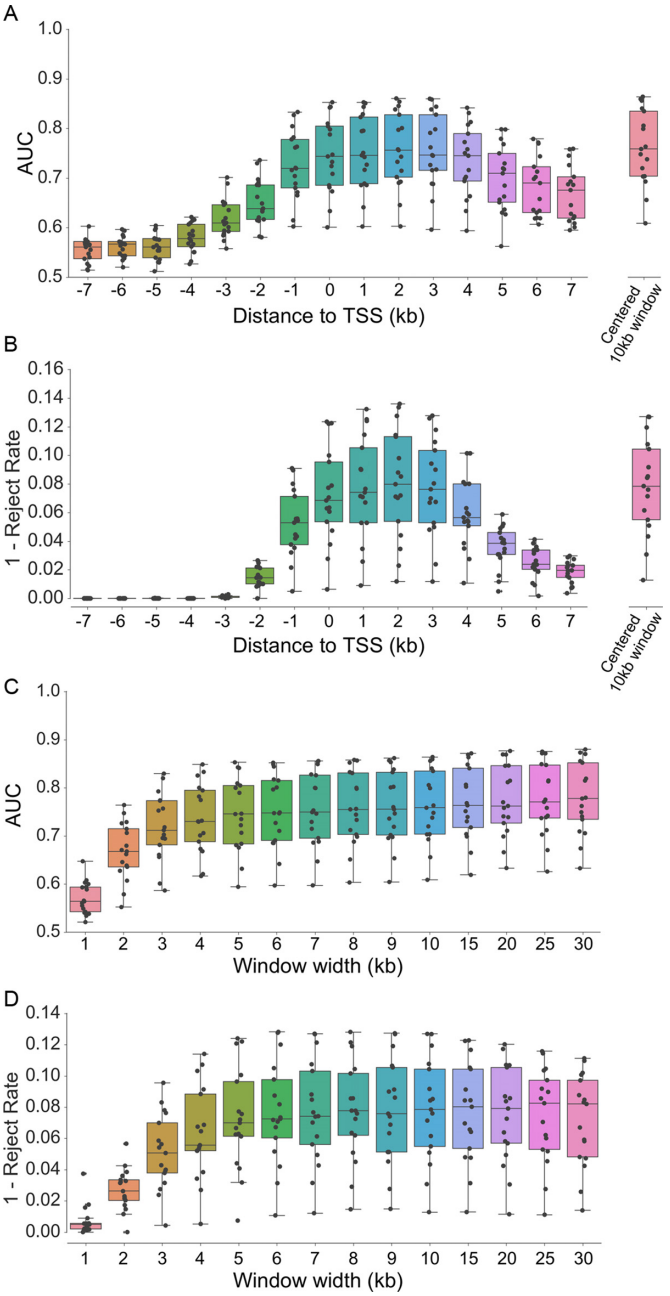

Figure 4. Importance of DNA methylation changes 3′ proximal to TSS for tissue-specific expression classification. Methods were evaluated using 17 REP tissue samples with the two-stage cross-validation framework in Figure 1E. (**A**) ROC AUC and (**B**) 1-reject rate for ME-Class methylation signatures created from fixed 5 kb windows centered at varying distances to the TSS. (**C**) ROC AUC and (**D**) 1-reject rate for ME-Class methylation signatures created using increasing window widths centered at the TSS. Individual points are the performance of individual REP sample comparisons.

3′ proximal and TSS regions are most predictive of differential expression

We next sought to understand why ME-Class performed better by examining which features are most important for classification. We first developed a series of ME-Class classifiers each using signatures from 5 kb windows centered at varying distances away from the TSS (Figure 4A and B). Performance peaks for a window centered 2 kb downstream of the TSS, indicating that the most important features exist downstream of the TSS. There was no substantial difference observed in ROC AUC between the entire 10 kb window compared for the 2 kb downstream centered 5 kb window (Figure 4B). This agrees with our analysis of RF feature importance, which showed that the most important features for gene expression classification occur downstream +0.5 to +2.5 kb of the TSS (Figure 2B).

We next evaluated the use of a Most Important (MI) window within the [+0.5 kb, +2.5 kb] region around the TSS of each gene in the REP dataset (Supplementary Figure S5). Both a SW classifier trained on only this region, and a

ME-Class classifier using only features from this region performed substantially worse than ME-Class given data from the full [–5 kb, +5 kb] region (Supplementary Figure S5), underscoring the importance of using all the data around the TSS to represent DNA methylation.

We then designed a series of ME-Class classifiers to examine how the size of the TSS-centered window affects performance. Increasing the window size beyond 10 kb did not show substantial improvements in ROC AUC (Figure 4C). However, after the window size increases greater than 10 kb there is a decrease in the number of genes returned at 90% probability of classification ($P = 4.5 \times 10^{-3}$) (Figure 4D). Combined, these results indicate that high-resolution features across a window of at least 5 kb wide and shifted 3′ proximal of the TSS is sufficient to capture the complexity of methylation signal around the promoter required for expression prediction in the REP samples.

Alternative models and features to improve ME-Class

We next sought to determine whether the underlying model used in ME-Class was sufficient for predicting expression. We were inspired from our previous unsupervised analysis (16) to design ME-Class to model the changes in methylation levels around the TSS. However, other features, including CpG density, the density of methylated CpG sites, and gene body methylation, have been described in the literature to have correlations with expression. Thus, we sought to understand whether adding these features would improve classification performance.

CpG density does not improve ME-Class but CpG-poor genes are more predictive

The density of methylated CpGs has been implied as the important feature for why CGI methylation affects gene silencing (38,39). We thus compared gene signatures computed from the normalized methylation density (mCG/bp), CpG density or fractional methylation (mCG/CG) to see which feature performed best. ROC and PR analysis, as well as examining the relationship between accuracy and reject rate, show no substantial increased effect of using mCG/CG rather than mCG/bp. Unsurprisingly, a model based on CpG density alone performs nearly equivalent to random guessing (Supplementary Figure S6a). While the direct addition of CpG density did not improve performance, we found that ME-Class performed worse on CGI-associated and CpG-rich promoters (Supplementary Figure S7). This is in agreement with prior findings that there is a stronger correlation between methylation and expression for genes which had no CpG island as compared to those with CpG Islands (40). However, since more genes contain CGIs, ME-Class identifies a strong association (90% probability of prediction) between differential methylation and expression for more CGI-associated genes (mean = 135 for REP samples) than CGI-poor genes (mean = 108). A similar trend is observed for CpG-rich (mean = 170) versus CpG-poor (mean = 73) promoters. Previously, it has been hypothesized that CGI-associated and CpG-rich genes tend to remain unmethylated in normal cell-types irrespective of their expression levels (41); however, our analysis suggests that while there are better associations between methylation and expression for non-CGI-associated genes, there are more CGI-associated genes that show strong associations.

The addition of gene body methylation changes does not improve ME-Class

Since gene body methylation has been shown to be positively correlated with gene expression (18,19,40), we examined if we could improve ME-Class by adding additional gene features that modeled methylation changes in the gene body. ME-Class performance was not substantially improved by adding features for averaged gene features similar to that of the ROI method such as the average methylation of internal exons, introns, and region downstream of the gene (Supplementary Figure S6B). This was unsurprising, since feature importance analysis of the ROI classifier indicated that the most important features for classification were the methylation levels of the first exon and first intron, which substantially overlap the region from the TSS to +5 kb.

We also investigated whether other gene representations could determine whether methylation information from the gene body could improve classification performance. Therefore, we implemented three alternative approaches to model DNA methylation throughout a gene: Whole Scaled Gene (WSG), Whole Gene (WG), and Uniform Gene Features (UGF) (Supplementary Figure S2B–D). In the WSG approach, the methylation profile is interpolated across the entire gene and then all genes are rescaled to a uniform length. This is a common method to visualize genomic trends in genome-wide methylation data (22,26,35). WG is similar, but the genes are not scaled after interpolation. Lastly, in the UGF approach methylation is interpolated then methylation features are extracted using a uniform number of bins for each exon and intron. All alternative approaches include regions –5 kb of the TSS and +5 kb of the TES. Using ROC AUC analysis, the TSS-centric (default ME-Class representation) model outperforms the WG ($P = 3.7 \times 10^{-5}$) and UGF ($P = 2.3 \times 10^{-5}$) approaches (Supplementary Figure S2E–G). Further, the TSS-centric approach identifies more genes on average (TSS: 178, WSG: 112) at a higher average level of accuracy (TSS: 93%, WSG: 89%) than the WSG approach.

All combined, these results suggest that models that incorporate gene body methylation, whether through average features or whole gene representations, do not substantially outperform models comprising only information from around the TSS. Our results demonstrate that features in the gene body and downstream of the TES are minimally important when using differential methylation to classify expression change. Thus, even though there are correlations between differential gene body methylation and differential expression, there is minimal new information in the gene body relative to the information already found in the ±5 kb region around the TSS.

Optimizing ME-Class

Before using ME-Class to analyze additional samples, we examined whether we could tune ME-class for better per-

formance. Using the REP dataset, we compared the performance of the RF, Logistic Regression, Gradient Boosted Classification Trees (GBCT), Naïve Bayes, and k-NN. We also compared the RF-based approach to a method using DTW as a curve similarity metric for kNN classification. RF, Logistic Regression and GBCT outperformed the remaining machine learning methods by both ROC AUC analysis and examining the relationship between accuracy and reject rate (Supplementary Figure S8). We also found that ME-Class performed similarly well as long as smoothing parameters were maintained below 200 bp (Supplementary Figure S9A and B). Interpolation and smoothing serve to decrease inaccuracy of low coverage methylation calls, as has been observed in DMR callers (31). Changes in the interpolation method also had no substantial effect on performance (Supplementary Figure S9C). To assess how much training data ME-Class requires for accurate classification, we separated the REP dataset into nine differential samples for training held out 8 differential samples for evaluation (see details in Materials and Methods). ME-Class was consistent within 0.02 ROC AUC of the full training set after using three samples (Supplementary Figure S10A) and showed minimal increases in obtaining consistent gene sets as the full training set when using more than five samples (Supplementary Figure S10B) or 20 000 genes (Supplementary Figure S10C).

Myeloid/Lymphoid differential methylation comparisons are most predictive of expression change

We next applied ME-Class to identify methylation-associated expression changes in hematopoiesis. We used WGBS and mRNA-seq datasets provided by the Blueprint Epigenome project in cord and venous blood composed of 32 isolated samples from 10 cell types. We retrained ME-Class using the entire 17 samples from the REP data above and then used this model to find methylation-associated expression changes in 469 hematopoietic lineage-wise differential comparisons. We found a large variation in the number of genes identified based on the cell-types being compared. Comparison of relatively distantly related lymphoid and myeloid lineages resulted in 54–218 genes (mean = 88) returned at 90% probability of classification (Figure 5A and B). Gene ontology analysis suggests that genes identified in myeloid-lymphoid comparisons are enriched for genes involved in T-cell activation, leukocyte differentiation and hematopoiesis. Similar performance results were found if we characterize the classifier based on ROC AUC (Figure 5B). This contrasts with a comparison of closely related myeloid cells such as macrophages, monocytes and dendritic cells, which identified between 0 and 55 genes when comparing any two cell types (mean = 16 genes; $P = 2 \times 10^{-16}$) and demonstrate performance near random guessing (ROC AUC = 0.55 ± 0.04). Neutrophil samples stood out as particularly poor performers in all comparisons suggesting that either there are not methylation-associated expression differences in these cells, or that methylation profiles in these cells are fundamentally different from that of other tissues (Supplementary Figure S11).

Based on ROC analysis there was an inverse relationship between the relatedness of the cells being compared and the ROC AUC from ME-Class (Figure 5C). Similar results were also obtained using an ME-class model trained a combination of closely- and distantly-related hematopoietic cell types (Supplementary Figure S12). While other analyses have examined associations upon myeloid (42) and B-cell differentiation (43), this analysis demonstrates that truly predictive differences in differential DNA methylation primarily reside between the myeloid and lymphoid lineages, the two major lineages derived from hematopoietic stem cells.

ME-Class identifies subsets of genes sensitive to demethylation in colon cancer

We next examined whether ME-Class can accurately identify genes that are hypermethylated and silenced in a model of colon cancer. For this problem, we first trained a cancer-specific ME-Class model using WGBS and mRNA-seq data from four different normal-tumor differential comparisons from The Cancer Genome Atlas (TCGA) (BRCA, COAD, READ and UCEC). We then used this cancer-specific ME-Class model to identify methylation-associated changes in expression in a colon cancer cell line (HCT116) relative to normal colon tissue (Figure 6A). We observed that the primary peak of feature importance in our trained model shifts from the region [+0.5 kb, +2.5 kb] downstream of the TSS in the REP tissue-specific model to the region [–0.5 kb, +1.5 kb] overlapping the TSS in the TCGA normal-tumor model (Figure 6B). We found a severe class imbalance; 187 genes with methylation-associated expression changes were identified as down-regulated, but no genes were predicted as up-regulated. Functional annotation clustering of gene ontology of the 187 down-regulated, identified genes showed that these genes are enriched for C2H2 zinc fingers, previously shown to be hypermethylated and silenced in carcinogenesis, and are involved in cell adhesion, whose dysregulation is important for tumorigenesis (44). To understand whether the tumor-associated hypermethylation was functional, we examined what happened to these genes after removal of methylation by double knockout of DNMT1 and 3b (HCT116 DKO1). Genes with an ME-Class signature showed a significant upregulation of expression relative to down-regulated genes not identified by ME-Class (Figure 6C, $P = 1 \times 10^{-14}$). Our analysis from this model of colon cancer demonstrates that ME-Class can identify genes likely regulated by DNA methylation in human disease and is consistent with the hypothesis that hypermethylation near the TSS plays a primary role in modulating gene activity cancer (41,45,46).

DISCUSSION

One challenge in the field of DNA methylation analysis has been the difficulty of integrative analysis of genome-wide DNA methylation and expression data. While methods exist to facilitate this task, they have many parameters that must be set, often with no clear way to make intelligent choices for their values. For example, DMR and SW approaches require the user to set several parameters (such

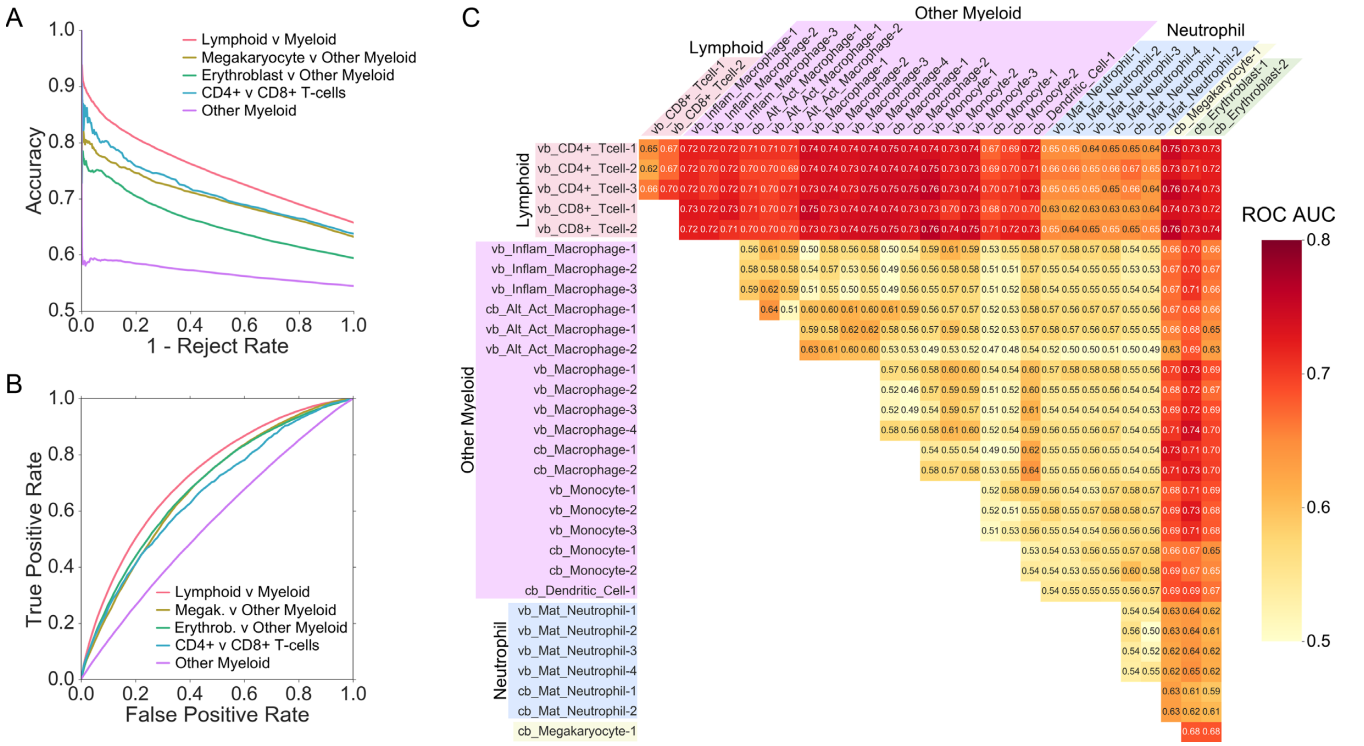

Figure 5. ME-Class performance is higher for cell comparisons between distally related cell lineages as opposed to directly related ones. (**A**) Accuracy versus 1-reject rate and (**B**) ROC of selected cell-types (ROC AUC: Lymphoid versus Myeloid, 0.72 ± 0.02; Megak. versus other Myeloid, 0.68 ± 0.02; Erythrob. versus Other Myeloid, 0.68 ± 0.02; CD4+ versus CD8+ T cells, 0.66 ± 0.02; Other Myeloid, 0.55 ± 0.03). (**C**) ME-Class ROC AUC performance for each sample-wise comparison of hematopoiesis samples. Vb = venous blood, cb = cord blood, Erythrob. = Erythroblast, Alt_Act_Macrophage = Alternating activated Macrophage, Inflam_Macrophage = Inflammatory Macrophage, Mat_Neutrophil = Mature Neutrophil, Megak. = Megakaryocyte. ROC AUC error is the standard deviation.

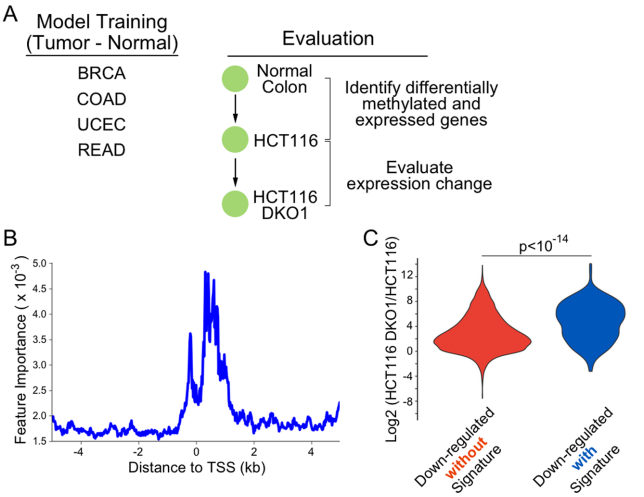

Figure 6. ME-Class identifies genes re-expressed after removal of DNA methylation in a model of colon cancer. (**A**) Experimental scheme. We first built an ME-class model using WGBS data from four tumor-normal pairs from TCGA. We then used this model to identify genes with expression-associated methylation changes upon tumorigenesis (HCT116) from normal colon (REP) and evaluate the demethylation effect by genetic manipulation (DKO1: bi-allelic knockout of DNMT1 and DNMT3b in HCT116). (**B**) Feature importance for the ME-Class training model built from four TCGA tumor-normal samples (COAD, BRCA, READ, UCEC). C) Violin plots showing the expression fold change in HCT116 DKO1 cells using down regulated differentially expressed gene sets identified with ($n = 187$ genes) and without ($n = 5370$ genes) identified methylation signatures.

as minimum window size or a minimum number of CpGs) that can drastically change the list of genes with predicted methylation changes. Our results from REP analysis show that these methods cannot be used to accurately predict expression from DNA methylation, likely because they cannot model the signal complexity necessary to associate methylation and expression change (Figure 2). Incorporating the complexity of patterns, rather than reducing methylation to a single or even multiple averaged values, is critical for the success of ME-Class. Alternative gene representations that incorporate gene body methylation perform no better than representations that focus on the region around a gene's TSS (Supplementary Figure S6B). Thus, the information obtained from correlations between gene body methylation and expression is either too noisy, or it is redundant with information within ±5 kb of the TSS. In the future, it may be possible to improve ME-Class by adding features specific to enhancers, but first we need better computational tools and experimental data across multiple cell types to connect regulatory units with specific genes.

In this study, we asked whether we could build models of DNA methylation that would generalize across different genes and across samples. For this purpose, we established a strict evaluation framework to test changes in methylation to identify these most likely affected genes. Using a training and evaluation paradigm that consists of both cross-sample and cross-gene evaluation, we have shown that ME-Class predictions are not overfit to any given dataset. ME-Class

performs well across high quality datasets without the need for tuning for individual datasets. However, there are limits to this generalization. For example, tissue-specific and cancer-specific datasets require different models to achieve high performance. Further, even though we excluded alternative promoters from this analysis to compare methods across a set of high quality reference genes, ME-Class can use isoform-specific promoter locations and expression data as input to provide an isoform-level analysis.

The role of gene-specific DNA methylation changes in development has been debated for many years (47). If promoter DNA methylation played a primary role in regulating cell-type specific expression changes then one would expect to observe a large fraction of differentially expressed genes with methylation-associated expression changes even among closely related cell types. However, we do not observe this. On the other hand, if methylation were a consequence of expression change, then one would expect to see a large number of methylation-associated expression differences in distantly related cell types or tissues, but not in closely related ones. Our results from ME-class are consistent with the latter statement. ME-Class identifies on average 7.5% (217 genes/sample) of differentially expressed genes from different tissues in REP are associated with promoter-proximal methylation changes at 90% accuracy, and 2.5% (88 genes/sample) of differentially expressed genes for distantly related hematopoietic lineages (myeloid versus lymphoid). However, ME-class performs poorly on sample comparisons of closely related hematopoietic lineages, often identifying few genes (myeloid versus myeloid; mean = 0.6% or 16 genes/sample). Our data is thus consistent with a model where transcriptional changes precede promoter-proximal methylation changes during differentiation, and thus, these tissue- and cell- specific DNA methylation changes are likely a consequence of transcriptional changes. While DNA methylation is unlikely to initiate tissue-specific expression, we cannot rule out the possibility that these promoter-proximal methylation changes play a later role in maintaining these transcriptional programs.

In contrast, the reactivation of genes identified by ME-Class as methylated and silenced upon removal of DNA methylation is consistent with the hypothesis that methylation changes at the promoter in cancer play a direct role in gene regulation. This observation is consistent with our recent work showing that an unsupervised analysis of differential methylation data in AML could identify a set of genes that were likely to be up-regulated upon treatment with demethylating agents (21). We also observe a shift in the most informative region for expression associated methylation changes from [+0.5 kb, +2.5 kb] in a tissue-specific model to [–0.5 kb, +1.5 kb] in a cancer-specific model. While early studies suggested that tissue-specific and cancer-specific expression-associated methylation changes were similar, our results are in agreement with more recent studies that use higher resolution methods and larger numbers of samples (12,40,48,49). Further, the observation that both windows are shifted downstream of the TSS is in agreement with recent studies that have suggested that the transcriptional activator p300 can bind downstream of the TSS at unmethylated CpG Islands to increase gene expression (40), and that decreases in methylation downstream of the TSS co-occur with increases in active H3K4me3 that also shift downstream of the TSS (49,50). This difference in expression-associated methylation changes may explain why methylation appears to play a role in gene silencing in cancer, but not in development. In addition, the context dependent nature of these models has a profound effect on downstream applications indicating that different methylation models may need to be trained for different contexts (i.e. cancer-specific models need to be trained to understand expression-associated methylation changes in cancer). Once these models are trained, they are applicable across other similar datasets.

As more large-scale, genome-wide DNA methylation studies of the differences between matched normal and tumor samples become available, tools such as ME-class will prove invaluable to understand how specific methylation changes affect transcription. In addition, our results show that ME-Class is a powerful tool to identify genes that are silenced by methylation in disease and could be used to facilitate the identification of patients who may benefit from clinically-approved demethylating therapeutics (51).

AVAILABILITY

ME-Class is publicly available on Github at http://github.com/cschlosberg/me-class.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to acknowledge Tao Ju and Kilian Weinberger for helpful discussion in the development of ME-Class. We further thank Jerry Fong, Lisa Rois and Manoj Singh for assistance in testing the ME-Class code and critical feedback on this manuscript.

FUNDING

Siteman Cancer Center, U.S. Department of Defense Congressionally Directed Medical Research Program for Breast Cancer [W81XWH-11-1-0401]; National Institutes of Health [NIGMS 5R01GM108811, NLM R21LM011199] (to J.R.E.); National Institutes of Health T32 Genome Analysis Training Program [2T32HG000045-16] for pre-doctoral support to C.E.S. Funding for open access charge: NIH [R01 GM108811].
Conflict of interest statement. None declared.

REFERENCES

1. Smith,Z.D. and Meissner,A. (2013) DNA methylation: roles in mammalian development. *Nat. Rev. Genet.*, **14**, 204–220.
2. Laurent,L., Wong,E., Li,G., Huynh,T., Tsirigos,A., Ong,C.T., Low,H.M., Kin Sung,K.W., Rigoutsos,I., Loring,J. *et al.* (2010) Dynamic changes in the human methylome during differentiation. *Genome Res.*, **20**, 320–331.
3. Laird,P.W. and Jaenisch,R. (1994) DNA methylation and cancer. *Hum. Mol. Genet.*, **3**(suppl. 1), 1487–1495.

4. Baylin,S.B., Esteller,M., Rountree,M.R., Bachman,K.E., Schuebel,K. and Herman,J.G. (2001) Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Hum. Mol. Genet.*, **10**, 687–692.
5. Lister,R., Pelizzola,M., Dowen,R.H., Hawkins,R.D., Hon,G., Tonti-Filippini,J., Nery,J.R., Lee,L., Ye,Z., Ngo,Q.-M. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.
6. Ehrlich,M. (2002) DNA methylation in cancer: too much, but also too little. *Oncogene*, **21**, 5400–5413.
7. Keshet,I., Schlesinger,Y., Farkash,S., Rand,E., Hecht,M., Segal,E., Pikarski,E., Young,R.A., Niveleau,A., Cedar,H. *et al.* (2006) Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nat. Genet.*, **38**, 149–153.
8. Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.
9. Suzuki,M.M. and Bird,A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.*, **9**, 465–476.
10. Bock,C. (2012) Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.*, **13**, 705–719.
11. Wagner,J.R., Busche,S., Ge,B., Kwan,T., Pastinen,T. and Blanchette,M. (2014) The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol.*, **15**, R37.
12. Schultz,M.D., He,Y., Whitaker,J.W., Hariharan,M., Mukamel,E.A., Leung,D., Rajagopal,N., Nery,J.R., Urich,M.A., Chen,H. *et al.* (2015) Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*, **523**, 212–216.
13. Kulis,M., Heath,S., Bibikova,M., Queiros,A.C., Navarro,A., Clot,G., Martinez-Trillos,A., Castellano,G., Brun-Heath,I., Pinyol,M. *et al.* (2012) Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat. Genet.*, **44**, 1236–1242.
14. Proudhon,C., Duffié,R., Ajjan,S., Cowley,M., Iranzo,J., Carbajosa,G., Saadeh,H., Holland,M.L., Oakey,R.J., Rakyan,V.K. *et al.* (2012) Protection against de novo methylation is instrumental in maintaining parent-of-origin methylation inherited from the gametes. *Mol. Cell*, **47**, 909–920.
15. van Eijk,K.R., de Jong,S., Boks,M.P., Langeveld,T., Colas,F., Veldink,J.H., de Kovel,C.G., Janson,E., Strengman,E., Langfelder,P. *et al.* (2012) Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics*, **13**, 1–13.
16. Vanderkraats,N.D., Hiken,J.F., Decker,K.F. and Edwards,J.R. (2013) Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes. *Nucleic Acids Res.*, **41**, 6816–6827.
17. Irizarry,R.A., Ladd-Acosta,C., Wen,B., Wu,Z., Montano,C., Onyango,P., Cui,H., Gabo,K., Rongione,M., Webster,M. *et al.* (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.*, **41**, 178–186.
18. Yang,X., Han,H., De Carvalho,D.D., Lay,F.D., Jones,P.A. and Liang,G. (2014) Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell*, **26**, 577–590.
19. Ball,M.P., Li,J.B., Gao,Y., Lee,J.-H., LeProust,E.M., Park,I.-H., Xie,B., Daley,G.Q. and Church,G.M. (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotech.*, **27**, 361–368.
20. Cruickshanks,H.A., McBryan,T., Nelson,D.M., VanderKraats,N.D., Shah,P.P., van Tuyn,J., Rai,T.S., Brock,C., Donahue,G., Dunican,D.S. *et al.* (2013) Senescent cells harbour features of the cancer epigenome. *Nat. Cell Biol.*, **15**, 1495–1506.
21. Lund,K., Cole,J.J., VanderKraats,N.D., McBryan,T., Pchelintsev,N.A., Clark,W., Copland,M., Edwards,J.R. and Adams,P.D. (2014) DNMT inhibitors reverse a specific signature of aberrant promoter DNA methylation and associated gene silencing in AML. *Genome Biol.*, **15**, 1906–1920.
22. Consortium,R.E., Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
23. Wu,C., MacLeod,I. and Su,A.I. (2013) BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Res.*, **41**, D561–D565.
24. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,A.D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
25. Adams,D., Altucci,L., Antonarakis,S.E., Ballesteros,J., Beck,S., Bird,A., Bock,C., Boehm,B., Campo,E., Caricasole,A. *et al.* (2012) BLUEPRINT to decode the epigenetic signature written in blood. *Nat. Biotech.*, **30**, 224–226.
26. Blattler,A., Yao,L., Witt,H., Guo,Y., Nicolet,C.M., Berman,B.P. and Farnham,P.J. (2014) Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes. *Genome Biol.*, **15**, 327–316.
27. Jjingo,D., Conley,A.B., Yi,S.V., Lunyak,V.V. and Jordan,I.K. (2012) On the presence and role of human gene-body DNA methylation. *Oncotarget*, **3**, 462–474.
28. Wu,H., Xu,T., Feng,H., Chen,L., Li,B., Yao,B., Qin,Z., Jin,P. and Conneely,K.N. (2015) Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Res.*, **43**, e141.
29. Lou,S., Lee,H.-M., Qin,H., Li,J.-W., Gao,Z., Liu,X., Chan,L.L., Lam,V.K.L., So,W.-Y., Wang,Y. *et al.* (2014) Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. *Genome Biol.*, **15**, 6–21.
30. Eckhardt,F., Lewin,J., Cortese,R., Rakyan,V.K., Attwood,J., Burger,M., Burton,J., Cox,T.V., Davies,R., Down,T.A. *et al.* (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.*, **38**, 1378–1385.
31. Hansen,K.D., Langmead,B. and Irizarry,R.A. (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.*, **13**, R83.
32. Breiman,L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.
33. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V. *et al.* (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
34. Albanese,D., Visintainer,R., Merler,S., Riccadonna,S., Jurman,G. and Furlanello,C. (2012) mlpy: machine learning python. *arXiv*, 1–4.
35. Kretzmer,H., Bernhart,S.H., Wang,W., Haake,A., Weniger,M.A., Bergmann,A.K., Betts,M.J., Carrillo-de-Santa-Pau,E., Doose,G., Gutwein,J. *et al.* (2015) DNA methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control. *Nat. Genet.*, **47**, 1316–1325.
36. Sakoe,H. and Chiba,S. (1978) Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Trans. Acoust. Speech, Signal Process.*, **26**, 43–49.
37. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
38. Hsieh,C.L. (1994) Dependence of transcriptional repression on CpG methylation density. *Mol. Cell. Biol.*, **14**, 5487–5494.
39. Deaton,A.M. and Bird,A. (2011) CpG islands and the regulation of transcription. *Genes Dev.*, **25**, 1010–1022.
40. Varley,K.E., Gertz,J., Bowling,K.M., Parker,S.L., Reddy,T.E., Pauli-Behn,F., Cross,M.K., Williams,B.A., Stamatoyannopoulos,J.A., Crawford,G.E. *et al.* (2013) Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res.*, **23**, 555–567.
41. Jones,P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, **13**, 484–492.
42. Bocker,M.T., Hellwig,I., Breiling,A., Eckstein,V., Ho,A.D. and Lyko,F. (2011) Genome-wide promoter DNA methylation dynamics of human hematopoietic progenitor cells during differentiation and aging. *Blood*, **117**, e182–e189.
43. Kulis,M., Merkel,A., Heath,S., Queiros,A.C., Schuyler,R.P., Castellano,G., Beekman,R., Raineri,E., Esteve,A., Clot,G. *et al.* (2015) Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nat. Genet.*, **47**, 746–756.
44. Severson,P.L., Tokar,E.J., Vrba,L., Waalkes,M.P. and Futscher,B.W. (2014) Coordinate H3K9 and DNA methylation silencing of ZNFs in

toxicant-induced malignant transformation. *Epigenetics*, **8**, 1080–1088.
45. Herman,J.G. and Baylin,S.B. (2003) Gene silencing in cancer in association with promoter hypermethylation. *N. Engl. J. Med.*, **349**, 2042–2054.
46. Robertson,K.D. (2005) DNA methylation and human disease. *Nat. Rev. Genet.*, **6**, 597–610.
47. Bestor,T.H., Edwards,J.R. and Boulard,M. (2015) Notes on the role of dynamic DNA methylation in mammalian development. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 6796–6799.
48. Moarii,M., Boeva,V., Vert,J.-P. and Reyal,F. (2015) Changes in correlation between promoter methylation and gene expression in cancer. *BMC Genomics*, **16**, 873.
49. Hovestadt,V., Jones,D.T.W., Picelli,S., Wang,W., Kool,M., Northcott,P.A., Sultan,M., Stachurski,K., Ryzhova,M., Warnatz,H.-J. *et al.* (2014) Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature*, **510**, 537–541.
50. Hodges,E., Molaro,A., Dos Santos,C.O., Thekkat,P., Song,Q., Uren,P.J., Park,J., Butler,J., Rafii,S., McCombie,W.R. *et al.* Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol. Cell*, **44**, 17–28.
51. Azad,N., Zahnow,C.A., Rudin,C.M. and Baylin,S.B. (2013) The future of epigenetic therapy in solid tumours—lessons from the past. *Nat. Rev. Clin. Oncol.*, **10**, 256–266.