2019

ME-Class2 reveals context dependent regulatory roles for 5-hydroxymethylcytosine

Christopher E. Schlosberg
Washington University School of Medicine in St. Louis
Dennis Y. Wu
Washington University School of Medicine in St. Louis
Harrison W. Gabel
Washington University School of Medicine in St. Louis
John R. Edwards
Washington University School of Medicine in St. Louis

Published online 16 January 2019

Nucleic Acids Research, 2019, Vol. 47, No. 5 ***e28***
doi: 10.1093/nar/gkz001

ME-Class2 reveals context dependent regulatory roles for 5-hydroxymethylcytosine

Christopher E. Schlosberg[1], Dennis Y. Wu[2], Harrison W. Gabel[2] and John R. Edwards[1,*]

[1]Center for Pharmacogenomics, Department of Medicine, Washington University in St. Louis School of Medicine, St. Louis, MO 63110, USA and [2]Department of Neuroscience, Washington University in St. Louis School of Medicine, St. Louis, MO 63110, USA

Received May 09, 2018; Revised December 04, 2018; Editorial Decision January 02, 2019; Accepted January 03, 2019

ABSTRACT

Since the discovery of 5-hydroxymethylcytosine (5hmC) as a prominent DNA modification found in mammalian genomes, an emergent question has been what role this mark plays in gene regulation. 5hmC is hypothesized to function as an intermediate in the demethylation of 5-methylcytosine (5mC) and in the reactivation of silenced promoters and enhancers. Further, weak positive correlations are observed between gene body 5hmC and gene expression. We previously demonstrated that ME-Class is an effective tool to understand relationships between whole-genome bisulfite sequencing data and expression. In this work, we present ME-Class2, a machine-learning based tool to perform integrative 5mCG, 5hmCG and expression analysis. Using ME-Class2 we analyze whole-genome single-base resolution 5mCG and 5hmCG datasets from 20 primary tissue and cell samples to reveal relationships between 5hmCG and expression. Our analysis indicates that conversion of 5mCG to 5hmCG within 2 kb of the transcription start site associates with distinct functions depending on the summed level of 5mCG + 5hmCG. Unchanged levels of 5mCG + 5hmCG (conversion from 5mCG to stable 5hmCG) associate with repression. Meanwhile, decreases in 5mCG + 5hmCG (5hmCG-mediated demethylation) associate with gene activation. Our results demonstrate that ME-Class2 will prove invaluable to interpret genome-wide 5mC and 5hmC datasets and guide mechanistic studies into the function of 5hmCG.

INTRODUCTION

In mammalian genomes, cytosines are frequently covalently modified at the 5-position with methyl-, hydroxymethyl-, formyl-, and carboxy- groups (1). The initial modification occurs by addition of a methyl- group to the 5-position of the cytosine (5-methylcytosine, 5mC) by a DNA methyltransferase (Dnmt) (2). The subsequent modifications are then formed through successive oxidation of 5mC by the Ten-eleven translocation (Tet) family of enzymes (Tet1, Tet2, Tet3, reviewed in (1)). While 5mC occurs at nearly 70% of all CG dinucleotides (CpG) in the genome in all tissues, 5-hydroxymethylcytosine (5hmC) appears to be primarily limited to embryonic stem cells, neurons, liver, breast, testis, and placenta tissues, occurring at 2–17% of CpGs depending on the tissue type (1–3). Meanwhile 5hmC's oxidized derivatives, 5-formylcytosine (5fC) and 5-carboxycytosine (5caC), are only found at very low levels, 10–1000-fold less than 5hmC (3). It is still unclear whether 5fC and 5caC are short-lived intermediates (4) or whether they have an independent biological function *in vivo* (5,6). Of these marks, 5mC has been the most studied and is a well-established player in maintaining inactivation of the silenced X chromosome, mono-allelic gene expression at imprinted loci, and silencing retrotransposons (2). Abnormal patterns of 5mC are also linked to transcriptional dysregulation in cancer (7).

One limitation to our prior understanding of 5hmC is that technologies such as whole-genome bisulfite sequencing (WGBS) which have been used to map 5mC in the genome, cannot distinguish between 5mC and 5hmC. Recently Tet-assisted bisulfite sequencing (TAB-seq) and oxidative bisulfite sequencing (oxBS-seq) have been developed to complement WGBS analysis to determine the levels of 5mC and 5hmC throughout the genome (8,9). In TAB-seq, 5hmCs are first glucosylated. Subsequently, 5mC and 5fC bases are converted to 5caC using recombinant Tet proteins. Bisulfite treatment and sequencing are then performed. In TAB-seq, 5hmC is sequenced as a C and all other cytosine forms are measured as a T. In oxBS-seq, 5hmC is first chemically oxidized to 5fC. Bisulfite treatment and sequencing is then performed. In oxBS-seq, 5hmC, 5fC and 5caC are all measured as T. By combining WGBS data with either TAB-seq or oxBS-seq data, levels for 5mC and 5hmC can then be estimated using programs such as MLML, which provides

*To whom correspondence should be addressed. Tel: +1 314 362 6935; Fax: +1 314 362 8826; Email: jredwards@wustl.edu

a simultaneous maximum likelihood based on binomial estimates of 5hmC and 5mC (10).

At the biochemical level, 5hmC likely plays a role in demethylation through both passive and active mechanisms (1). While Dnmt1 is responsible for copying and propagating 5mC during cell division (2), no similar mechanism has yet been discovered for 5hmC. Notably, 5hmC, 5fC and 5caC are found at their highest levels in post-mitotic cells, such as neurons, and are passively diluted during cell division (1). For example, 5hmC starts at low levels in the developing brain, but accumulates in the adult brain (11). Active demethylation occurs through conversion of 5mC to 5caC via 5hmC and 5fC intermediates. 5caC is then converted to unmethylated cytosine through base excision repair or decarboxylation (12). In support of the role of Tet enzymes and 5hmC in demethylation (13), Tet2–/– mouse brains exhibit low level gains in methylation (11) and Tet1-, Tet2-, Tet3-triple KO mice display significant promoter hypermethylation (14).

Genomic analyses show that 5mC and 5hmC have distinct targets throughout the genome. 5mC marks the majority of the genome except for CpG islands (CGIs), gene promoters, and enhancers (15). High levels of 5mC at gene promoters, CGI shores, and enhancers is associated with expression repression (2). Like 5mC, 5hmC is depleted from CGIs in ES cells and neurons, however 5hmC is depleted from intergenic regions in ES cells (4–6). Meanwhile, 5hmC is enriched at enhancers, gene bodies, and CGI shores (11,16).

Based on these data, 5hmC has been hypothesized to serve as an intermediate in promoter demethylation (i.e. removal of 5mCG) that could reactivate gene expression (16,17), however additional evidence suggests 5hmC can also play a regulatory role independent of 5mC. For example, MeCP2 displays reduced affinity for hmCG compared to mCG, and therefore conversion of mCG to stable hmCG in the neuronal genome may lead to loss of functional binding sites for MeCP2 (18). The mechanism for how 5hmC may play an independent gene regulatory role has remained elusive as screens for 5hmC interacting factors have uncovered few 5hmC-specific interactors, although many 5mC binding proteins have reduced affinity to 5hmC (6). Gene body 5hmC, as both a stable mark (as in neurons) and as an intermediate for demethylation, is frequently associated with gene expression (11). 5hmC marked promoters have also been associated with gene repression (17,19). More recent studies in human liver and lung tissues observed 5hmC as a marker of active transcription associated with H3K4me1 at CpG island shores (16). 5hmC may also play a role in enhancer regulation, as Tet2 deletion causes an increase in enhancer 5mC levels and reduced enhancer activity (20). One limitation has been that current analysis methods label promoters as either marked or unmarked by 5hmC, they do not tease apart when 5hmC may be a result of demethylation versus when it may exist as an independent regulatory mark. Further, whether and how 5mC and 5hmC signals in promoters and gene bodies act in concert to affect gene silencing has not been studied.

In addition to the CG context, adult neurons also contain high amounts of non-CpG methylation, which comes close to or surpasses the total amount of mCG in neurons (11,21,22). 5mCH is deposited by DNMT3A across the neuronal genome and is associated with gene repression. Removal of DNMT3A in the mouse brain leads to increased expression of genes that are marked with high levels of 5mCH in the gene body (23,24). Associations have further been observed between genes with high levels of gene body mCA and those repressed by MeCP2, suggesting a potential mechanism by which non-CpG methylation may regulate expression in neurons (21,23,25,26). Further, while 5hmCG in gene bodies is thought to result in reduced MeCP2 binding and increased expression, 5hmCA accumulates in regions flanking enhancers without altering MeCP2 binding (27). Together these data demonstrate that both 5mC and 5hmC can have different effects depending on both the local sequence context (e.g. CG versus CA) and the local genomic context (promoter, gene body, enhancer). However, integrated analysis tools that can separate the effects of mC and 5hmC in both different sequence and genomic contexts on expression do not exist, and we still have a poor understanding of how different marks act independently and/or in combination to affect gene expression in different contexts.

Methods for integrated analysis of 5mC and 5hmC generally consist of the application of meta-gene analysis plots and their variants (11,17,19,23) or DMR (differentially methylated region) analysis (11,16). We previously developed ME-Class to model methylation both at gene promoters and in gene bodies to identify genes with a high probability of association between 5mC and gene expression (28). Our model was more effective at predicting differentially expression changes than models based on DMR analyses and demonstrated the limitations of considering DMRs removed from their local genomic context. Relative to meta-gene type analyses, ME-Class moves away from comparing methylation patterns at high and low expressing genes within a single sample, and instead looks at how changes in methylation between samples at individual genes associate with expression.

Here, we extend ME-Class' functionality to include multiple epigenetic marks as inputs and add a post-classification clustering tool to facilitate understanding the underlying potential regulatory mechanisms. We use these new functionalities to systematically interrogate how changes in 5mC and 5hmC associate with gene expression. Our results indicate that models that include both 5mC and 5hmC out-perform 5mC only models, but only in tissues or cells (such as neuronal tissues) that have high levels of 5hmC. Further, our results indicate that 5hmC associates with gene activation when it is involved in demethylation and with gene repression when it is stably present at and around the promoter of a gene.

MATERIALS AND METHODS

WGBS, TAB-seq, oxBS-seq and RNA-seq data

Mapped sequence reads for whole genome bisulfite sequencing (WGBS), Tet-assisted bisulfite sequencing (TAB-seq), and RNA-seq in liver and lung tumor and matched normal samples were obtained from Li *et al.* (16) and from dendritic cells from Pacis *et al.* (29). WGBS, oxidative bisulfite sequencing (oxBS-seq), and RNA-seq from fetal and 6-

week mouse brain samples were obtained from Lister *et al.* (11) and granule cells from Mellen *et al.* (27). WGBS and TAB-seq for human cortex are from Wen *et al.* (30). Corresponding RNA-seq data are from Brawand *et al.* (31) as used by Wen *et al.* (30).

Estimation of 5mC and 5hmC levels

5mC and 5hmC levels were estimated using maximum likelihood methylation levels (MLML) from either TAB-seq or oxBS-seq (10). We used MLML with a significance level of $\alpha = 0.05$ for the binomial test at each CpG (and CpH for non-CpG analyses) site and an expectation maximization convergence threshold of 1e–10. Counts of individual CpGs with estimated 5hmC and 5mC in all samples can be found in Supplementary Table S1.

Differential expression from RNA-seq

RNA-seq data from human liver, lung, and cortex samples were mapped to hg19 using HISAT2 (32). We used featureCounts to estimate feature counts over RefSeq reads (33). Differentially expressed genes were defined as abs[fold change] ≥ 2 after applying a floor of cpm $= 1$. To create a standardized gene set with high quality methylation data, we excluded genes with ambiguous or incomplete transcription start site (TSS) annotations, genes shorter than 5 kb, genes with <40 CpGs assayed within ±5 kb of the TSS, genes where, for all CpGs within ±5 kb of the TSS, the change in methylation (mCG/CG) was <0.2, and genes with alternative promoters. These filters were used to exclude non-coding genes, pseudogenes, genes shorter than the interpolation boundary (see HRPS model description below), genes with low numbers of CpGs (to reduce bias caused by error in individual CpG measurements), and genes with no methylation changes at their respective promoters. We only included RefSeq genes with cdsStartStat and cdsEndStat flags marked as 'cmpl' according to the UCSC Table Browser. For any RefSeq genes with multiple RefSeq IDs corresponding to the same TSS location, we used a single RefSeq ID with the lowest accession number and excluded the remainder. This is a conservative method to simplify the annotations of genes with alternative promoter annotations. A full summary of differentially expressed filtered gene counts can be found in Supplementary Table S2.

5hmC incorporation in ME-Class

MLML produces an estimate of 5mC and 5hmC for each CpG site. ME-Class high-resolution promoter signature (HRPS), region of interest (ROI), and whole-scale gene (WSG) models described in Schlosberg *et al.* (28) were extended to add 5hmCG features (Figure 1A and B). For the HRPS model, 5mCG and 5hmCG data were independently interpolated using PCHIP interpolation and Gaussian smoothing (50bp bandwidth) across the window ±5 kb relative to each gene's TSS. Interpolated curves for Δ5hmCG/CG and Δ5mCG/CG (i.e. the difference in 5hmCG and mCG levels between samples) were discretized to create feature vectors for classification using the average methylation in each 20 bp segment. Bins for the ROI model were inspired by Lou *et al.* (34). Differential 5mCG and 5hmCG levels were computed for each bin, which was then used in the resultant feature vector. For the WSG model, 5mCG and 5hmCG data were scaled to a constant length between the TSS and RefSeq annotated transcription end site (TES). Feature vectors were created from 125 bins upstream of the gene, 125 bins downstream of the gene, and 500 bins from the area between the TSS and TES. Differential 5mCG and 5hmCG levels were both computed for the entire set of bins and then combined to form the final feature vector. Δ5mCG/CG corresponds to the 5mC feature vector and Δ5hmCG/CG corresponds to the 5hmC feature vector. Δ5mCG/CG and Δ5hmCG/CG corresponds to concatenating 5mCG and 5hmCG feature vectors together for the classification. Δ5mCG/CG + Δ5hmCG/CG corresponds to summing 5mCG and 5hmCG values before creating the feature vector for classification.

Evaluation framework

ME-Class2 uses a random forest classifier (implemented using scikit learn (35)) which uses feature vectors from 5mCG, 5hmCG or both data to predict the direction of expression change. Random forests were built using 5001 trees. The number of trees was optimized on a tuning dataset. For the fetal to 6-week mouse brain comparison and dendritic cell analysis we used an intra-sample 10-fold cross validation. For the normal liver-lung and normal-tumor comparisons we performed cross-fold validation similar to that in Schlosberg *et al.* (28). In brief, we hold out each sample one by one for evaluation and then train on the remaining samples. To further minimize over-fitting, all genes from the validation sample are excluded from the samples used for training. For cortex-liver and cortex-lung comparisons, models were trained using 10-fold cross-validation, and all genes used for validation are excluded from the samples used for training.

Unsupervised clustering of 5hmCG and 5mCG

Unsupervised hierarchical agglomerative clustering (complete linkage) was performed on Δ5mCG/CG and Δ5hmCG/CG in the region [0, +2 kb] for the TSS for correctly predicted genes from ME-Class2 (implemented using scikit learn (35)). Sub-setting our predictions required setting a working threshold for the probability of prediction. Therefore, we set the following range of probabilities of prediction for each experiment based on 90% accuracy at: [0.68, 1.0] fetal-6-week mouse, [0.8, 1.0] normal liver-tumor, [0.7, 1.0] normal-tumor liver and lung. Ranges were set at [0.8–1.0] for cortex-liver and cortex-lung based on 95% accuracy due to the large number of genes accurately predicted for these samples. In the metagene plots of unsupervised results, Δ5mCG/CG + Δ5hmCG/CG corresponds to the summation of 5mCG and 5hmCG.

5mCH analysis

Due to the previously observed impact of gene expression on mCH throughout the gene body in neurons (24), and the observed effect of MeCP2-based repression acting on genes

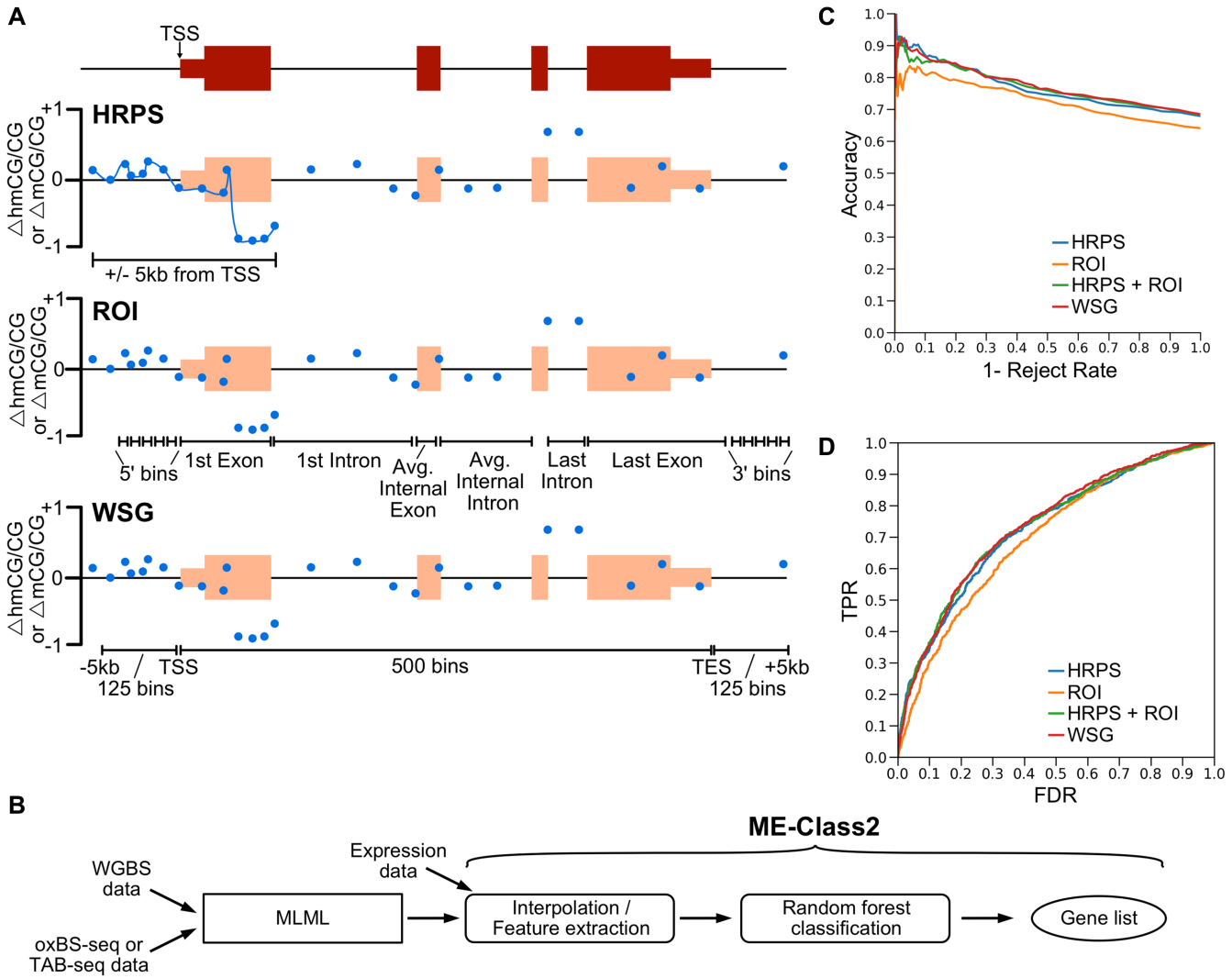

Figure 1. (**A**) Cartoon example showing different models to encode methylation features for a gene for ME-Class2 analysis. Δ5mCG/CG and Δ5hmCG/CG refer to the differences between two samples. HRPS is high-resolution promoter signature; ROI is region of interest; WSG is whole-scale gene. TSS is transcription start site. Blue dots show example differential methylation (5hmCG or 5mCG). (**B**) ME-Class2 workflow. (C, D) Performance of different gene models using ME-Class2 5mCG and 5hmCG data from fetal and 6-week mouse brain as evaluated using accuracy versus 1 – reject rate (**C**) and ROC (receiver operating characteristic) curve analysis (**D**). 1 – Reject rate is the fraction of genes with predicted associations between methylation and expression. ROC AUC are HRPS: 0.727, HRPS + ROI: 0.735, WSG: 0.739, ROI: 0.699.

with high amounts of gene-body methylation (21), a whole-scaled gene (WSG) model was chosen. This model split each gene into 50 windows. After examination of the feature importance of mCH around genes, the 25 kb window around the TSS and TES of genes was also assessed with features defined by 1 kb bins. This approach was chosen based on the analysis from (21), which showed associations between gene body mCH and expression that ranged up to 25 kb around a gene. These windows were used to assess both mCH as a whole, and mCA, mCC and mCT individually.

Ontology analysis

Ontology analysis was conducted with the functional annotation clustering tool from DAVID with the default set of ontologies and parameters.

RESULTS

Differential 5hmCG at promoter and promoter-proximal regions is more important than gene-body 5hmCG in predicting expression changes

We extended ME-Class to simultaneously incorporate 5mCG and 5hmCG information from high resolution genomic data from WGBS, TAB-seq and oxBS-seq. We used the three best performing models for associating 5mCG and gene expression (28) to understand which model performed the best at the new task of using combined 5mCG and 5hmCG data (Figure 1A,B). The first is a high-resolution promoter signature (HRPS) that interpolates a signature around the window ±5kb of the TSS for both 5mCG and 5hmCG signals. We previously identified this model as optimal for associating 5mCG and expression changes (28). The second model, which we call 'regions of interest' (ROI),

bins methylation data upstream of the TSS and across gene features such as first and internal exons and introns (34). We further compared these methods to a whole-scaled gene (WSG) approach, which is based on a scaling method to compare whole gene signals across genes and is commonly used to capture correlations between gene body methylation and expression (11).

We initially benchmarked these models using a set of WGBS and TAB-seq data from fetal and 6-week mouse brains. To evaluate performance, we plot the accuracy versus 1- reject rate for each model. This performance metric allows us to focus on only the genes with the highest quality predictions given some confidence threshold. The underlying premise is that only some genes should have associated DNA methylation and expression changes, not all. We demonstrate good performance to predict gene expression change as measured by both accuracy versus 1 – reject rate (Figure 1C) and ROC analysis (Figure 1D) for all models. In the HRPS model we predict differential expression in 216 genes with greater than 90% accuracy, which outperforms ROI and WSG models for which we detect a similar number of genes, but at only 83% and 88% accuracy respectively. Both methods that capture the area around the TSS at high resolution (HRPS and WSG) out-perform other methods. Interestingly, models that incorporate features from the gene body (ROI and WSG) do not perform better than those that only model the data around the TSS at high resolution (HRPS). Further, direct addition of gene-body features to the HRPS model (HRPS + ROI) does not increase performance. Random forest feature importance analysis indicates that 5mCG and 5hmCG changes within 2kb and primarily downstream of the TSS into the first intron are the most important regions for successful classification (Figure 2).

Addition of 5hmCG data improves ME-Class2 performance

We next sought to determine whether models trained using both 5hmCG and 5mCG data outperformed those trained only using 5mCG data. Figure 3A–C shows that including 5mCG and 5hmCG as independent features boosts ME-Class2 performance in the comparison of mouse fetal and 6-week brains and human cortex versus liver and lung using the HRPS model (corresponding ROC curves are in Supplementary Figure S1). For mouse brain comparisons, the model using 5hmCG and 5mCG data predicted 112 genes at >90% accuracy. Using 5mCG or 5hmCG data alone, the accuracy for a similar number of genes was only 82% and 75% respectively. Similar increases in performance with the inclusion of 5hmCG data were observed for other models including WSG, ROI and HRPS + ROI (Supplementary Figure S2). Using the HRPS model, Δ5mCG + Δ5hmC, which is effectively what is measured by only WGBS data in the absence of a 5hmC-specific assay, performed equivalent to Δ5mCG alone (Supplementary Figure S3).

We also observe similar performance gains in human cortex versus liver and lung comparisons (Figure 3). For cortex versus liver, differential expression of an average of 480 genes per sample (493, 493 and 453 for each sample respectively) could be predicted at 90% accuracy using 5hmCG and 5mCG changes, but this accuracy fell to 82% and 77% for 5mCG and 5hmCG alone respectively. Meanwhile for cortex versus lung, differential expression of an average of 278 genes per sample (359, 266 and 209 for each sample respectively) could be predicted at 90% accuracy using 5hmCG and 5mCG changes, but this accuracy fell to 81% and 72% for 5mCG and 5hmCG alone respectively. We also

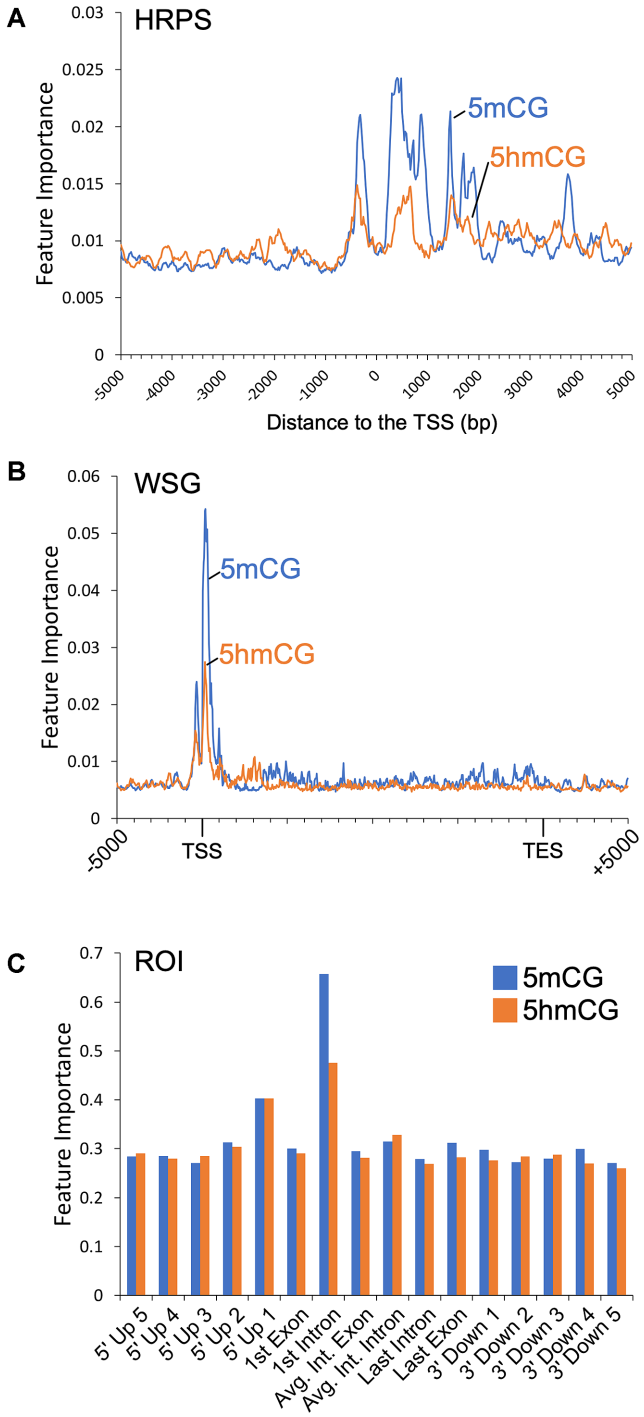

Figure 2. Feature importance for the ME-Class2 random forest classifier for fetal-6wk mouse brain 5mCG and 5hmCG data for (**A**) HRPS, (**B**) WSG and (**C**) ROI data representations. Binning schemes for each model are in Figure 1A.

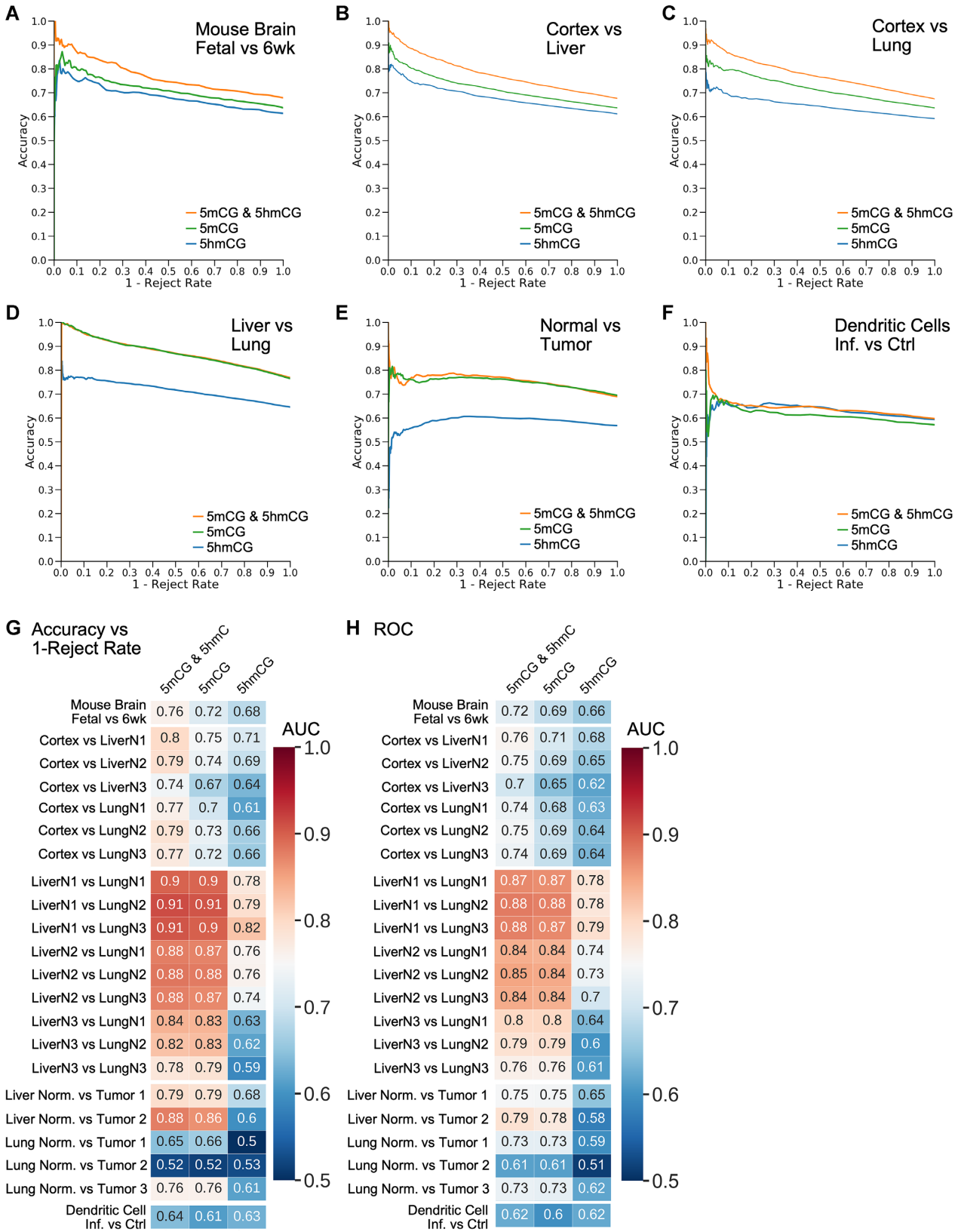

Figure 3. (**A**–**F**) ME-Class2 performance (accuracy versus 1 – reject rate) for different 5mCG and 5hmCG datasets using the HRPS feature model. The 5mCG and 5mCG & 5hmCG curves directly overlap in panel D. Corresponding detailed ROC (receiver operating characteristic) curves are in Supplementary Figure S2. (**G**) Area under the curve (AUC) for the accuracy versus 1 – reject rate curves and (**H**) ROC AUC for each individual sample comparison used in A–F.

observed an increase in performance comparing bacterially infected and non-infected dendritic cells, although the addition of 5hmCG data only allowed the prediction of differential expression for 16 total genes at greater than 93% accuracy (Figure 3F, Supplementary Table S3). However, we do not observe such performance gains for all samples. We did not observe any substantial difference between 5mCG only and 5mCG and 5hmCG models in comparisons involving human lung and liver tissues across three individuals (Figure 3D, G, H), or in normal-tumor comparisons from three lung and two liver tumors (Figure 3E, G, H). Feature importance analysis of cortex vs liver, cortex versus lung, and infected dendritic cells all support the region within 2–3 kb of the TSS as most important for predicting expression change (Supplementary Figure S4).

Predictive methylation signatures in non-brain tissues and tumors are solely dependent on changes in 5mCG

Similar unsupervised clustering of highly predictive liver-lung and cancer-specific genes show why the addition of 5hmCG data did not increase performance in these comparisons (Supplementary Figure S5). The differential methylation signatures produced from these clusters in each case show that there is little difference in 5hmCG, such that the net Δ5mCG + Δ5hmCG levels closely follow the Δ5mCG levels. This suggests that the level 5hmCG does not vary substantially across non-brain tissues for genes where methylation changes are predictive of expression changes. The observed 5mCG patterns in each cluster resemble those we previously found in other tissues (28) and cancer cell lines (36). This implies that promoter 5mCG, rather than 5hmCG, is primarily associated with gene expression change in cancer.

ME-Class2 identifies 5hmCG and 5mCG signatures in brain tissues

To better understand why we observed a boost in performance by including 5hmCG in brain and cortex comparisons, we conducted post-hoc unsupervised clustering analysis of identified signatures of 5hmCG and 5mCG that associate with expression change using the mouse fetal and 6-week brain comparison. We observe three distinct classes of differential 5hmCG and 5mCG signatures that predict expression changes (Figure 4A–C). In Figure 4A, we observe increases in both 5mCG and 5hmCG 3′ proximal to the TSS, which associate with a decrease in expression (Cluster C1, $n = 76$). This contrasts to the signature observed in cluster C2 ($n = 29$, Figure 4B). These genes also decrease in expression; however, while the 5mCG signal decreases 3′ proximal to the TSS, the 5hmCG increases over the same region. There is no substantial change in the net Δ5mCG + Δ5hmCG level, indicating that the primary feature in this cluster is a conversion from 5mCG to stable 5hmCG rather than demethylation. While most of the observed patterns cluster because of changes in 5mC data, observation of the C2 cluster is entirely dependent on the addition of 5hmC data. A third cluster (C3, $n = 70$, Figure 4C) comprises a set of genes that increase in expression and are again characterized by 5mCG decreases and 5hmCG increase 3′ proximal to the TSS. In this case however, the net amount of 5mCG + 5hmCG decreases indicating 5hmCG plays a role as an intermediate toward demethylation. Differential methylation signatures similar to those found in clusters C1 and C3 were also observed in human cortex versus liver and lung comparisons (Supplementary Figure S6).

To better understand whether stably 5hmCG marked promoters were associated with gene repression we examined expression levels of genes in each cluster across mouse development. Cluster C2 genes which are marked by 5mCG alone in the fetal cortex have much lower expression as a whole than genes that gain 5mCG and 5hmCG found in cluster C1 ($P < 0.009$, Wilcoxon test, Figure 4D–F). This agrees with our finding that 5hmCG within 2 kb of the TSS associates with transcriptional repression. To test whether this conclusion would hold true in an alternative dataset, we first used feature importance analysis (Figure 2A) to identify the region from [–800 bp, 2100 bp] around the TSS for both 5mCG and 5hmCG signals that contributes the greatest to classification in the fetal-6-week brain comparison. Next, we calculated the average 5mC, 5hmC and unmodified C content across this region for all genes in granule cells. Agreeing with our hypothesis, genes primarily marked with high levels of either 5mCG ($P < 2e–16$, Bonferroni adjusted Wilcoxon test) or 5hmCG ($P < 2e–7$, Bonferroni adjusted Wilcoxon test) are generally not expressed (Figure 4G).

Lastly, to understand whether mCG conversion or demethylation events are potential causes of transcriptional change, we examined whether genes from each cluster identified in the fetal-6-week comparison above were differentially expressed in the Tet1–/– mouse cortex (Figure 4H) (37). Cluster C1 is characterized by predominantly increased 5mCG levels and thus, as expected, there was no significant difference in the expression of these genes after removal of Tet1. Genes in cluster C2 that were downregulated in 6-week mouse brain, which had undergone a conversion of 5mCG to 5hmCG (with no net decrease of 5mCG + 5hmCG), were found to generally increase in expression in Tet1–/– mouse cortex relative to WT ($P = 0.037$, Bonferroni adjusted Wilcoxon test). Surprisingly, there was also no change in expression for genes undergoing Tet-mediated demethylation (cluster C3). This could be because 5hmC-mediated demethylation occurs as a consequence of transcription. In agreement, transcription factor complexes have been implicated to recruit Tet1 leading to 5hmC mediated demethylation mediated by PPARγ in differentiated ES cells (38). However, our analysis has several limitations that could explain the lack of an observed effect. The promoters of selected genes that are differentially expressed in Tet1−/− cortex and hippocampus were shown to increase in 5mCG levels by only 11–50%, which may be insufficient for many genes to show a change in expression (37). Additionally, we cannot rule out that other Tet members play a compensatory role in the absence of Tet1. For example, Tet1–/– mice do not show impaired differentiation (3,14) and have normal brain morphology (37) in contrast to Tet1-, Tet2-, Tet3-triple KO mice, which display impaired differentiation, impaired embryonic development, and significant promoter hypermethylation (14). In support of the fact that the observed demethylation may activate transcription, a gene found in cluster C3, regulator of G pro-

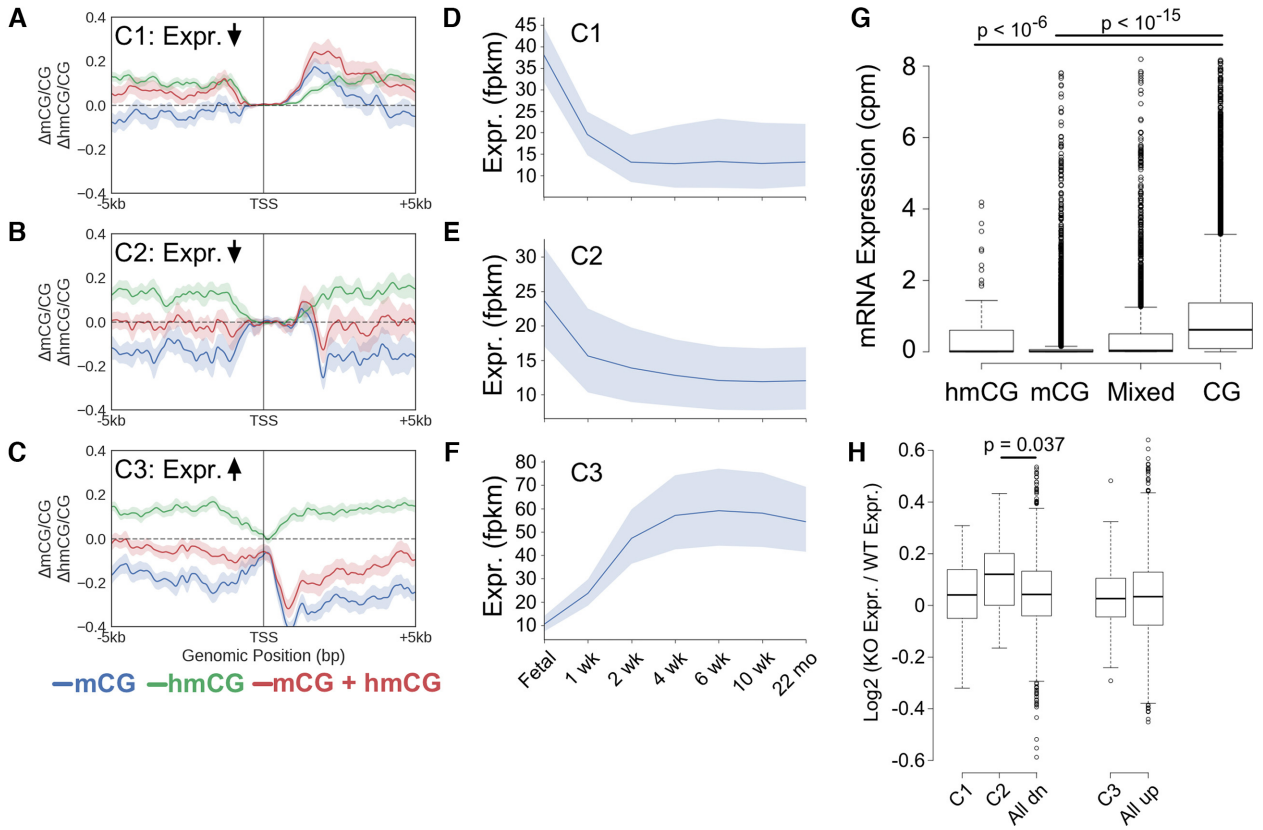

Figure 4. (**A**–**C**) Metagene plots for clusters of similar differential methylation signatures (6-week-fetal) that are predictive of expression in the fetal-6-week mouse brain comparison. Shading indicates the 68% bootstrapped confidence interval. Cluster C1: $n = 76$, C2: $n = 29$ and C3: $n = 70$. (**D**–**F**) Average expression of all genes found in C1, C2 and C3 clusters across mouse brain development. Shading indicates the 95% confidence interval. (**G**) mRNA expression in granule cells of genes whose promoters (defined as [–800 bp, +2 kb] around the TSS) are >50% marked by mCG, hmCG, a combination of mCG and hmCG, or CG (unmethylated). Outliers have been cropped for clarity. The original plot can be found in Supplementary Figure S7. (**H**) Log_2 expression changes in cortex from Tet1–/– mice versus cortex from WT mouse. All dn and all up correspond to all down- and up-regulated genes, respectively, in 6-week compared to fetal mouse brain. *P*-values computed using a Bonferroni adjusted Wilcoxon test.

tein signaling RGS14, was shown previously to up-regulate after demethylation of neural progenitors using Dnmt inhibitors (39). In summary, while these data support a role for 5hmCG as a functional repressor, whether demethylation is a cause or consequence of transcriptional silencing or whether there is a context-dependent component is unclear.

ME-Class2 identifies 5mCH and 5mCG signatures in brain tissues

In addition to mCG, adult neurons also contain high amounts of non-CpG methylation (11). To determine whether this non-CpG methylation has a significant association with gene expression we compared ME-Class2 models trained on only CpH methylation or on both CpG and CpH methylation (Figure 5). Models only using 5hmCH showed very poor performance likely indicating 5hmCH plays a regulatory role in very few if any genes (Supplementary Figure S8). In contrast, using mCH alone resulted in fair performance, and the addition of mCH to mCG models resulted in a substantial improvement in ME-Class2 performance (Figure 5A). Of the three mCH marks, mCA, the mark with the highest levels of methylation (11), returned more genes at higher accuracy than mCC and mCT (265 versus 5 and 40 respectively, Figure 5B). The addition of gene body mCH using the WSG model improved performance as compared to promoter-only models, in agreement with prior data suggesting that mCH at gene bodies may regulate MeCP2 function, and that transcription is associated with levels of gene body 5mCH (23). While feature analysis suggested mCH near the TSS were the most important features (Figure 5C, D), gene body features also were important. Since mCH is severely depleted from the region adjacent to the TSS, TSS mCH methylation may be relevant for only a small number of genes. As a whole, this data suggests that mCH serves a regulatory role in the brain, alongside mCG.

ME-Class2 identifies genes associated with neurodevelopmental disorders and neuronal development

Gene ontology analysis using DAVID (40) revealed genes associated with neurodevelopmental disorders and basic neuronal development in all clusters (Supplementary Table S4). Several of these have been implicated to have disorder-associated differential methylation including Shank2 in cluster C2 and Nrxn1, Pacsin1, and Grin1 in cluster C3. Shank2, a synaptic protein, has previously been shown to

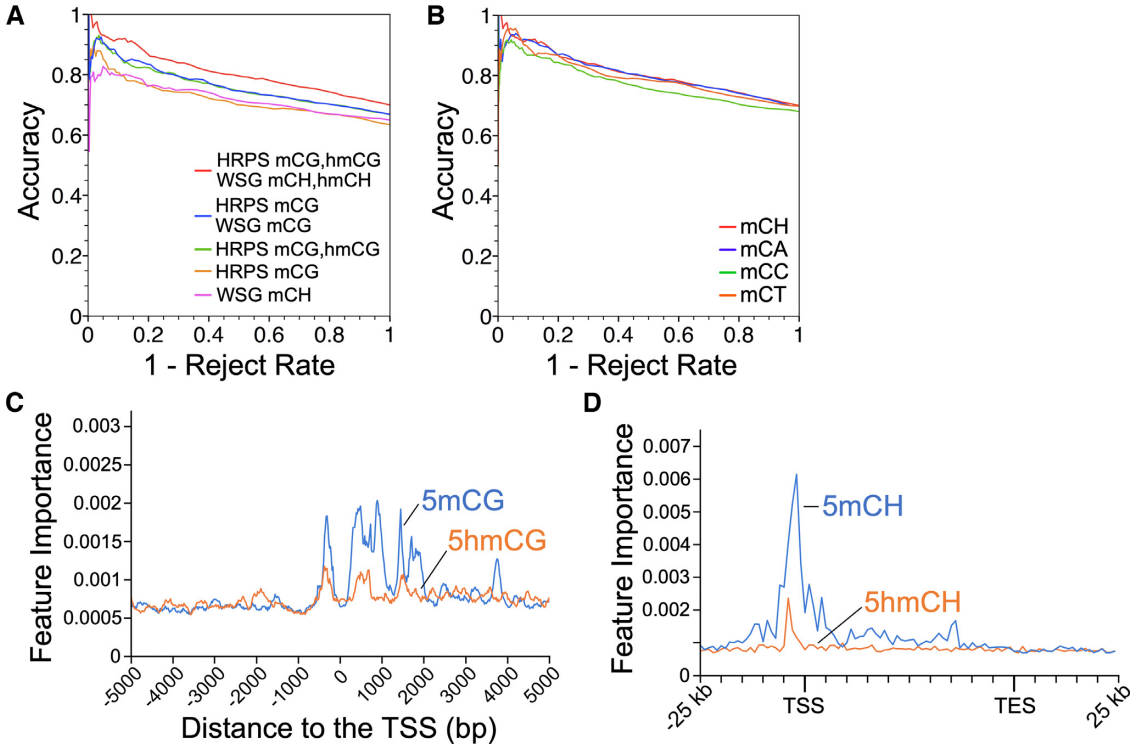

Figure 5. (**A**) ME-Class2 performance as assessed by accuracy versus 1 – reject rate (fraction of genes) for different ME-Class2 models incorporating mCH. (**B**) ME-Class2 performance of the HRPS mCG, hmCG with WSG mCH model. (**C**, **D**) Feature importance analysis of the best performing model that incorporates all features (HRPS mCG, hmCG with WSG mCH, 5hmC).

change methylation in the developing human brain and is associated with neurodevelopmental disorders (41). Methylation of Grin1, a component of NMDA receptor complexes, is associated with depression in children (42). Nrxn1 has previously been discovered as having a high ranking meQTL in 110 human hippocampus samples (43). Age-related DNA methylation changes have been found in Nrxn1, which has been implicated in schizophrenia and autism (44). Methylation of PACSIN1 is associated with substance-use risk (45). Importantly, our analysis suggests that 5hmCG may regulate disease-risk genes differently depending on whether it plays a role in repression or demethylation.

DISCUSSION

We successfully extended ME-Class to predict gene expression classification from both 5hmCG and 5mCG. Feature importance analysis shows that even in tissues with substantial 5hmCG, 5mCG is still the most useful mark for predicting expression changes (Figure 2, Supplementary Figure S4). 5hmCG models alone perform very poorly, which demonstrates the importance of considering 5hmC in the context of 5mC to understand potential associations and effects on transcription. Unsupervised analysis revealed a set of down-regulated genes with no net change in 5mCG + 5hmCG levels, but for which 5mCG levels decrease and 5hmCG levels rise. Models using only WGBS data would miss these genes since WGBS only observes the net change in 5mCG + 5hmCG. For other tissues with minimal amounts of 5hmCG it is unlikely that obtaining TAB- or oxBS-seq data will provide more information over what is already found using WGBS (5mCG + 5hmCG).

Our results suggest that the incorporation of 5mCG and 5hmCG marks at gene promoters and proximal promoters are the most important features for predicting expression changes. Features in the gene body or outside a 2–3 kb window from the TSS have little impact on the ability to associate methylation and transcription changes. This contrasts with 5mCH which showed importance both in this region and also through the gene body. In the CpG context, the direct addition of gene body features based on differential methylation of internal exons and introns using the ROI approach led to no boost in performance. Feature importance analysis (Figure 2, Supplementary Figure S4) clearly indicates that for all models, the features within 2–3 kb of the TSS are most essential for prediction and that gene body features greater than 2–3 kb from the TSS are of limited utility. Taken together, this implies either that average gene body 5hmCG plays little functional role in the regulation of transcription, or that gene body information is redundant with that found within 2–3 kb of the promoter. This finding is similar to as previously reported for 5mCG (28). Another alternative is that gene body 5hmCG plays a subtle effect on gene regulation that can only be uncovered with additional training data. For example, these models do not effectively incorporate individual regulatory elements such as enhancers, which are known to be regulated by 5hmC (20), and that may have context-dependent contributions on nearby genes.

We cannot rule out that the small amount of 5fC (and possibly 5caC) in the genome may play some as of yet undefined role in this process since these marks are indistinguishable from 5hmC in TAB- and oxBS-seq assays. Recent biochemical evidence has suggested that the majority of detectable 5fC in the genome is stable (46) and 5fC can inhibit the rate of transcript elongation (47).

Our results further show that the addition of 5hmCG data has the greatest effect on performance in samples with substantial amounts of 5hmCG, such as the brain. 5hmCG accumulates in the adult brain as can be observed in Figure 4A, where 5hmCG increases in most regions around the TSS in 6wk relative to fetal mouse brain (11). Post-mitotic neurons have high 5hmCG levels (11) and thus these samples benefit the most from inclusion of 5hmCG for predictions. 5hmCG exists at relatively low levels in liver (2.27–5.68%) and lung (1.94–3.04%) (16) in comparison to mouse brain (17.2%) (11) and human cortex (13%) (30) tissue. 5hmCG is an intermediate cytosine modification which is not replicated during mitosis (1). Lack of gene expression correlating 5hmCG patterns in normal lung and liver may be because dividing cells in these tissues passively dilute 5hmCG from their genomes. Thus, the scarcity of 5hmCG might explain its lack of predictive ability for expression class change in non-neuronal tissues. In agreement, clustering analysis of 5mCG and 5hmCG signals of predictive genes did not reveal a cluster of 5mCG to 5hmCG conversion as we observed in the model of mouse brain development. Instead, the patterns of differential 5hmCG and 5mCG closely follows that of 5mCG alone across all predictive signatures. However, we cannot rule out that 5hmCG inclusion in tissues with low amounts of 5hmCG might facilitate the identification of a few rare genes regulated by 5hmCG, which cannot be assessed with the amount of training data currently available.

Lastly our work points to a potential mechanism of 5hmCG mediated repression of gene proximal promoters independent of that observed by 5mCG. We identify that conversion of 5mCG to 5hmCG primarily is associated with the downregulation of gene expression in brain and many of these genes are up-regulated upon the removal of Tet1. Since there have been very few proteins identified that specifically bind 5hmC relative to 5mC (6), it is possible that 5hmCG-associated gene silencing could instead be caused by Tet1-recruitment of interacting partners, such as Sin3A and OGT, which have been shown to be involved in Tet1-dependent silencing of LINE-1 (48) or PRC2, which forms repressive chromatin (17,19). It is further unclear at this point why 5hmCG would stabilize in some genes versus others, and complicating matters is that how much active versus passive demethylation occurs via 5hmCG is still a point of contention (1). TET enzymes have substantially higher activity on 5mC relative to 5hmC and 5fC substrates, which may facilitate 5hmCG stability (3,49). It may be that demethylation is the dominant mechanism prior to neurons exiting the cell cycle, while stable 5hmCG occurs after.

Here we have demonstrated the power of integrated epigenetic analysis using ME-Class2 to examine the interplay of 5mC and 5hmC. However, ME-Class2 can be easily used with any epigenetic marks. Our application of ME-Class2 to the study of 5mC and 5hmC demonstrates that incorporating 5hmCG information is critical for prediction of gene expression changes in samples with high levels of 5hmC such as the brain and neurons. ME-Class2 has identified a class 5mCG/5hmCG patterns that show the conversion from 5mCG to 5hmCG in the 3′ proximal region of the promoter in a model of mouse brain development. We speculate that these patterns of 5mCG and 5hmCG coordinate with additional silencing factors potentially recruited either directly by 5hmCG or by the Tet enzymes in a context-specific manner. As the field continues to collect genome-wide, differential DNA methylation, tools such as ME-Class2 will prove invaluable for the interpretation of this epigenomic data and will guide mechanistic studies into the integrated function of 5mC and 5hmC in human disease.

DATA AVAILABILITY

ME-Class2 is an open source collaborative initiative available in the GitHub repository (https://github.com/jredwards417/me-class2).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would also like to thank Jerry Fong and Scot Matkovich for critical reading of the manuscript, and Kilian Weinberger for helpful discussions.

FUNDING

National Institutes of Health [R01 GM108811 to J.R.E.]; National Institutes of Health Genome Analysis Training Program [T32 HG000045 to C.E.S.]; The Mathers Foundation [to H.W.G.]. Funding for open access charge: U.S. Department of Health and Human Services; National Institutes of Health, National Institute of General Medical Sciences [R01 GM108811].
Conflict of interest statement. None declared.

REFERENCES

1. Rasmussen,K.D. and Helin,K. (2016) Role of TET enzymes in DNA methylation, development, and cancer. *Genes Dev.*, **30**, 733–750.
2. Edwards,J.R., Yarychkivska,O., Boulard,M. and Bestor,T.H. (2017) DNA methylation and DNA methyltransferases. *Epigenet. Chromatin*, **10**, 23.
3. Ito,S., Shen,L., Dai,Q., Wu,S.C., Collins,L.B., Swenberg,J.A., He,C. and Zhang,Y. (2011) Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, **333**, 1300–1303.
4. Branco,M.R., Ficz,G. and Reik,W. (2012) Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nat. Rev. Genet.*, **13**, 7–13.
5. Raiber,E.-A., Murat,P., Chirgadze,D.Y., Beraldi,D., Luisi,Ben F and Balasubramanian,S. (2015) 5-Formylcytosine alters the structure of the DNA double helix. *Nat. Struct. Mol. Biol.*, **22**, 44–49.
6. Iurlaro,M., Ficz,G., Oxley,D., Raiber,E.-A., Bachman,M., Booth,M.J., Andrews,S., Balasubramanian,S. and Reik,W. (2013) A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol.*, **14**, R119.

7. Koch,A., Joosten,S.C., Feng,Z., de Ruijter,T.C., Draht,M.X., Melotte,V., Smits,K.M., Veeck,J., Herman,J.G., Van Neste,L. *et al.* (2018) Analysis of DNA methylation in cancer: location revisited. *Nat. Rev. Clin. Oncol.*, **15**, 459–466.
8. Booth,M.J., Branco,M.R., Ficz,G., Oxley,D., Krueger,F., Reik,W. and Balasubramanian,S. (2012) Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science*, **336**, 934–937.
9. Yu,M., Hon,G.C., Szulwach,K.E., Song,C.-X., Zhang,L., Kim,A., Li,X., Dai,Q., Shen,Y., Park,B. *et al.* (2012) Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell*, **149**, 1368–1380.
10. Qu,J., Zhou,M., Song,Q., Hong,E.E. and Smith,A.D. (2013) MLML: consistent simultaneous estimates of DNA methylation and hydroxymethylation. *Bioinformatics*, **29**, 2645–2646.
11. Lister,R., Mukamel,E.A., Nery,J.R., Urich,M., Puddifoot,C.A., Johnson,N.D., Lucero,J., Huang,Y., Dwork,A.J., Schultz,M.D. *et al.* (2013) Global epigenomic reconfiguration during mammalian brain development. *Science*, **341**, 1237905.
12. Kohli,R.M. and Zhang,Y. (2013) TET enzymes, TDG and the dynamics of DNA demethylation. *Nature*, **502**, 472–479.
13. Wu,H. and Zhang,Y. (2014) Reversing DNA Methylation: Mechanisms, genomics, and biological functions. *Cell*, **156**, 45–68.
14. Dawlaty,M.M., Breiling,A., Le,T., Barrasa,M.I., Raddatz,G., Gao,Q., Powell,B.E., Cheng,A.W., Faull,K.F., Lyko,F. *et al.* (2014) Loss of Tet enzymes compromises proper differentiation of embryonic stem cells. *Dev. Cell*, **29**, 102–111.
15. Edwards,J.R., O'Donnell,A.H., Rollins,R.A., Peckham,H.E., Lee,C., Milekic,M.H., Chanrion,B., Fu,Y., Su,T., Hibshoosh,H. *et al.* (2010) Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Res.*, **20**, 972–980.
16. Li,X., Liu,Y., Salz,T., Hansen,K.D. and Feinberg,A. (2016) Whole-genome analysis of the methylome and hydroxymethylome in normal and malignant lung and liver. *Genome Res.*, **26**, 1730–1741.
17. Wu,H., D'Alessio,A.C., Ito,S., Wang,Z., Cui,K., Zhao,K., Sun,Y.E. and Zhang,Y. (2011) Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Genes Dev.*, **25**, 679–684.
18. Mellén,M., Ayata,P. and Heintz,N. (2017) 5-hydroxymethylcytosine accumulation in postmitotic neurons results in functional demethylation of expressed genes. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, E7812–E7821.
19. Neri,F., Incarnato,D., Krepelova,A., Rapelli,S., Pagnani,A., Zecchina,R., Parlato,C. and Oliviero,S. (2013) Genome-wide analysis identifies a functional association of Tet1 and Polycomb repressive complex 2 in mouse embryonic stem cells. *Genome Biol.*, **14**, R91.
20. Hon,G.C., Song,C.-X., Du,T., Jin,F., Selvaraj,S., Lee,A.Y., Yen,C.-A., Ye,Z., Mao,S.-Q., Wang,B.-A. *et al.* (2014) 5mC oxidation by Tet2 modulates enhancer activity and timing of transcriptome reprogramming during differentiation. *Mol. Cell*, **56**, 286–297.
21. Kinde,B., Wu,D.Y., Greenberg,M.E. and Gabel,H.W. (2016) DNA methylation in the gene body influences MeCP2-mediated gene repression. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 15114–15119.
22. Guo,J.U., Su,Y., Shin,J.H., Shin,J., Li,H., Xie,Bin, Zhong,C., Hu,S., Le,T., Fan,G. *et al.* (2014) Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat. Neurosci.*, **17**, 215–222.
23. Gabel,H.W., Kinde,B., Stroud,H., Gilbert,C.S., Harmin,D.A., Kastan,N.R., Hemberg,M., Ebert,D.H. and Greenberg,M.E. (2015) Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature*, **522**, 89–93.
24. Stroud,H., Su,S.C., Hrvatin,S., Greben,A.W., Renthal,W., Boxer,L.D., Nagy,M.A., Hochbaum,D.R., Kinde,B., Gabel,H.W. *et al.* (2017) Early-Life gene expression in Neurons modulates lasting epigenetic states. *Cell*, **171**, 1151–1164.
25. Lagger,S., Connelly,J.C., Schweikert,G., Webb,S., Selfridge,J., Ramsahoye,B.H., Yu,M., He,C., Sanguinetti,G., Sowers,L.C. *et al.* (2017) MeCP2 recognizes cytosine methylated tri-nucleotide and di-nucleotide sequences to tune transcription in the mammalian brain. *PLoS Genet.*, **13**, e1006793.
26. Chen,L., Chen,K., Lavery,L.A., Baker,S.A., Shaw,C.A., Li,W. and Zoghbi,H.Y. (2015) MeCP2 binds to non-CG methylated DNA as neurons mature, influencing transcription and the timing of onset for Rett syndrome. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 5509–5514.
27. Mellén,M., Ayata,P. and Heintz,N. (2017) 5-hydroxymethylcytosine accumulation in postmitotic neurons results in functional demethylation of expressed genes. *PNAS*, **114**, E7812–E7821.
28. Schlosberg,C.E., Vanderkraats,N.D. and Edwards,J.R. (2017) Modeling complex patterns of differential DNA methylation that associate with gene expression changes. *Nucleic Acids Res.*, **45**, 5100–5111.
29. Pacis,A., Tailleux,L., Morin,A.M., Lambourne,J., MacIsaac,J.L., Yotova,V., Dumaine,A., Danckaert,A., Luca,F., Grenier,J.-C. *et al.* (2015) Bacterial infection remodels the DNA methylation landscape of human dendritic cells. - PubMed - NCBI. *Genome Res.*, **25**, 1801–1811.
30. Wen,L., Li,X., Yan,L., Tan,Y., Li,R., Zhao,Y., Wang,Y., Xie,J., Zhang,Y., Song,C. *et al.* (2014) Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. *Genome Biol.*, **15**, R49.
31. Brawand,D., Soumillon,M., Necsulea,A., Julien,P., Csárdi,G., Harrigan,P., Weier,M., Liechti,A., Aximu-Petri,A., Kircher,M. *et al.* (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343–348.
32. Kim,D., Langmead,B. and Salzberg,S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.
33. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
34. Lou,S., Lee,H.-M., Qin,H., Li,J.-W., Gao,Z., Liu,X., Chan,L.L., Kl Lam,V., 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**, 408.
35. 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.
36. 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.
37. Rudenko,A., Dawlaty,M.M., Seo,J., Cheng,A.W., Meng,J., Le,T., Faull,K.F., Jaenisch,R. and Tsai,L.-H. (2013) Tet1 is critical for neuronal activity-regulated gene expression and memory extinction. *Neuron.*, **79**, 1109–1122.
38. Fujiki,K., Shinoda,A., Kano,F., Sato,R., Shirahige,K. and Murata,M. (2013) PPARγ-induced PARylation promotes local DNA demethylation by production of 5-hydroxymethylcytosine. *Nat. Commun.*, **4**, 2262.
39. Tuggle,K., Ali,M.W., Salazar,H. and Hooks,S.B. (2014) Regulator of G protein signaling transcript expression in human neural progenitor Differentiation: R7 subfamily regulation by DNA methylation. *Neurosignals*, **22**, 43–51.
40. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
41. Spiers,H., Hannon,E., Schalkwyk,L.C., Smith,R., Wong,C.C.Y., O'Donovan,M.C., Bray,N.J. and Mill,J. (2015) Methylomic trajectories across human fetal brain development. *Genome Res.*, **25**, 338–352.
42. Weder,N., Zhang,H., Jensen,K., Yang,B.Z., Simen,A., Jackowski,A., Lipschitz,D., Douglas-Palumberi,H., Ge,M., Perepletchikova,F. *et al.* (2014) Child abuse, depression, and methylation in genes involved with stress, neural plasticity, and brain circuitry. *J.Am.Acad.Child Adolescent Psychiatry*, **53**, 417–424.
43. Schulz,H., Ruppert,A.-K., Herms,S., Wolf,C., Mirza-Schreiber,N., Stegle,O., Czamara,D., Forstner,A.J., Sivalingam,S., Schoch,S. *et al.* (2017) Genome-wide mapping of genetic determinants influencing DNA methylation and gene expression in human hippocampus. *Nat. Commun.*, **8**, 1511.
44. Numata,S., Ye,T., Hyde,T.M., Guitart-Navarro,X., Tao,R., Wininger,M., Colantuoni,C., Weinberger,D.R., Kleinman,J.E. and Lipska,B.K. (2012) DNA methylation signatures in development and aging of the human prefrontal cortex. *Am. J. Hum. Genet.*, **90**, 260–272.

45. Cecil,C.A.M., Walton,E., Smith,R.G., Viding,E., McCrory,E.J., Relton,C.L., Suderman,M., Pingault,J.-B., McArdle,W., Gaunt,T.R. *et al.* (2016) DNA methylation and substance-use risk: a prospective, genome-wide study spanning gestation to adolescence. *Transl. Psychiatry*, **6**, e976–e976.
46. Bachman,M., Uribe-Lewis,S., Yang,X., Burgess,H.E., Iurlaro,M., Reik,W., Murrell,A. and Balasubramanian,S. (2015) 5-Formylcytosine can be a stable DNA modification in mammals. *Nat. Chem. Biol.*, **11**, 555–557.
47. Kellinger,M.W., Song,C.-X., Chong,J., Lu,X.-Y., He,C. and Wang,D. (2012) 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription. *Nat. Struct. Mol. Biol.*, **19**, 831–833.
48. de la Rica,L., Deniz,Ö., Cheng,K.C.L., Todd,C.D., Cruz,C., Houseley,J. and Branco,M.R. (2016) TET-dependent regulation of retrotransposable elements in mouse embryonic stem cells. *Genome Biol.*, **17**, 234.
49. Hu,L., Lu,J., Cheng,J., Rao,Q., Li,Z., Hou,H., Lou,Z., Zhang,L., Li,W., Gong,W. *et al.* (2015) Structural insight into substrate preference for TET-mediated oxidation. *Nature*, **527**, 118–122.