2020

DeepH&M: Estimating single-CpG hydroxymethylation and methylation levels from enrichment and restriction enzyme sequencing methods

Yu He

Hyo Sik Jang

Xiaoyun Xing

Daofeng Li

Michael J Vasek

See next page for additional authors

Authors

Yu He, Hyo Sik Jang, Xiaoyun Xing, Daofeng Li, Michael J Vasek, Joseph D Dougherty, and Ting Wang

GENETICS

DeepH&M: Estimating single-CpG hydroxymethylation and methylation levels from enrichment and restriction enzyme sequencing methods

Yu He[1,2]*, Hyo Sik Jang[1,2]*, Xiaoyun Xing[1,2], Daofeng Li[1,2], Michael J. Vasek[1,3], Joseph D. Dougherty[1,3], Ting Wang[1,2,4]†

Increased appreciation of 5-hydroxymethylcytosine (5hmC) as a stable epigenetic mark, which defines cell identity and disease progress, has engendered a need for cost-effective, but high-resolution, 5hmC mapping technology. Current enrichment-based technologies provide cheap but low-resolution and relative enrichment of 5hmC levels, while single-base resolution methods can be prohibitively expensive to scale up to large experiments. To address this problem, we developed a deep learning–based method, "DeepH&M," which integrates enrichment and restriction enzyme sequencing methods to simultaneously estimate absolute hydroxymethylation and methylation levels at single-CpG resolution. Using 7-week-old mouse cerebellum data for training the DeepH&M model, we demonstrated that the 5hmC and 5mC levels predicted by DeepH&M were in high concordance with whole-genome bisulfite–based approaches. The DeepH&M model can be applied to 7-week-old frontal cortex and 79-week-old cerebellum, revealing the robust generalizability of this method to other tissues from various biological time points.

INTRODUCTION

A single genome can derive phenotypically unique cell types through various epigenetic modifications that instruct specific gene expression patterns (*1*, *2*). DNA modifications, such as methylation of five positions of cytosines (5mC) at the CpG dinucleotide context, play a vital role in gene regulation, genomic imprinting, X-chromosome inactivation, and repression of transposable elements (*3*–*6*). The recent discovery that Ten-eleven translocation (TET) oxidase proteins can oxidize 5mC to 5-hydroxymethylcytosine (5hmC) has spurred an effort at characterizing the landscape of 5hmC in normal and diseased tissues and deciphering its potential functional role in gene regulation (*7*–*12*). Genome-wide profiling of 5hmC has found that 5hmC is not only just an intermediate product of the active DNA demethylation process but also a stable epigenetic mark correlated with gene expression. 5hmC abundance varies considerably across different tissues (*13*). 5hmC is present as high as 40% of 5mC levels in Purkinje neurons (*14*) and 5% of 5mC levels in embryonic stem cells (*15*), and is low (less than 1% of 5mC level) in other cell types (*16*). 5hmC is enriched in promoters, gene bodies, and enhancers; 5hmC levels in promoters and gene bodies are positively correlated with gene expression (*16*–*18*). 5hmC levels in enhancers are often cell type specific and are positively correlated with active enhancer histone marks, such as H3K4me1 and H3K27ac (*19*). However, the molecular mechanism by which 5hmC might regulate the genome has yet to be fully elucidated (*20*).

Rapid technological innovations for mapping 5mC have cemented 5mC as a crucial epigenetic mark for cell fate. Technologies for mapping 5mC include bisulfite conversion of unmethylated cytosine to uracil, such as whole-genome bisulfite sequencing (WGBS); enrichment of methylated DNA using methylcytosine-specific antibodies, such as methylated DNA immunoprecipitation sequencing (MeDIP-seq); and enrichment of unmethylated regions using methylation-sensitive restriction enzymes, such as methylation-sensitive restriction enzyme sequencing (MRE-seq) (*21*). The gold standard method WGBS can measure methylation genome-wide at single-base resolution but requires high coverage of the genome (at least 10× coverage for each cytosine) and therefore can be 10 times more expensive than enrichment or restriction enzyme sequencing methods (*22*). MeDIP-seq enriches for methylated regions but has low resolution [~150 base pairs (bp)] (*23*, *24*). MRE-seq provides CpG resolution, but can only interrogate methylation status at restriction enzyme sites (~30% of the genome) (*24*).

Similarly, 5hmC profiling technologies advanced from immunoprecipitation/enrichment-based methods to whole-genome single-base resolution. Because WGBS cannot distinguish 5hmC from 5mC, Yu *et al.* developed a method called TET-assisted bisulfite sequencing (TAB-seq), where 5hmCs are first protected by glucosylation and then 5mC is completely oxidized to 5caC with TET enzyme (*18*). The following bisulfite treatment can reveal which CpGs are protected and infer hydroxymethylation levels. TAB-seq can measure genome-wide 5hmC at single-base resolution but requires very high coverage to confidently call 5hmC at all cytosines. For example, for 5% 5hmC, based on binomial test with a probability of 2.22% for 5mC nonconversion rate, a coverage of 120 is required to call 5hmC at 95% confidence level (see Materials and Methods). The study from Yu *et al.* could only confidently call 20% or higher 5hmC at an average coverage of 27. Often in TAB-seq experiments, both WGBS and TAB-seq libraries are deeply sequenced to parse out 5mC and 5hmC levels in a single sample. Achieving high-confidence, single-base resolution of 5hmC can be a heavy financial strain for large experimental designs due to the necessary sequencing depth. Therefore, many adopted the cheaper alternative of using antibody-based enrichment method, such as hydroxymethylated DNA immunoprecipitation sequencing (hMeDIP-seq), which can reveal hydroxymethylated

[1]Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA. [2]The Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63110, USA. [3]Department of Psychiatry, Washington University School of Medicine, St. Louis, MO 63110, USA. [4]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO 63108, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: twang@wustl.edu

regions with limited sensitivity (*17*). hMeDIP-seq can also provide relative hydroxymethylation over controls, but at the cost of low resolution. Similar to antibody-based enrichment methods such as hMeDIP-seq, hmC-Seal chemically tags hydroxymethylated cytosine and enriches hydroxymethylated regions by pulling down tagged 5hmC (*16*, *19*). hmC-Seal can pull down regions with extremely low 5hmc content and, thus, have higher sensitivity than hMeDIP-seq.

Because of the high cost of single-base resolution profiling methods for 5hmC and 5mC, several computational methods were developed to estimate 5hmC and 5mC at single-base resolution. Xiao *et al.* developed a random forest regression–based method MeSiC (prediction from MeDIP-seq data at single-CpG resolution) to estimate single-CpG 5mC from MeDIP-seq data (*25*). Stevens *et al.* took advantage of the complementary properties of MeDIP-seq and MRE-seq and developed a conditional random field–based algorithm methylCRF to effectively predict single-CpG 5mC from MeDIP-seq and MRE-seq data (*26*). However, the two aforementioned algorithms cannot predict 5hmC levels. Pavlovic *et al.* developed a support vector machine (SVM)/random forest–based method DIRECTION to predict single-CpG 5mC or 5hmC from histone modification and transcription factor chromatin immunoprecipitation sequencing (ChIP-seq) data (*27*). This method can only predict binary values, either high or low 5mC/5hmC, but not the absolute quantitative level. To address these limitations, we developed a deep learning–based method, DeepH&M, which integrates enrichment and restriction enzyme sequencing methods to estimate absolute single-CpG resolution hydroxymethylation and methylation levels simultaneously.

RESULTS

Description of DeepH&M model

To estimate single-CpG hydroxymethylation and methylation, we developed a deep learning–based algorithm, DeepH&M, to integrate MeDIP-seq, MRE-seq, and hmC-Seal data (Fig. 1A). The core of DeepH&M is to model the relationship between MeDIP-seq/MRE-seq/hmC-Seal data and TAB-seq/WGBS data using deep learning networks. The relationship between MeDIP-seq/MRE-seq data and WGBS data was well characterized previously in a conditional random field–based algorithm, methylCRF, which was used to integrate MeDIP-seq and MRE-seq data to predict absolute methylation levels at single-CpG resolution (*26*). hmC-Seal data are positively correlated with TAB-seq data, while MeDIP-seq and MRE-seq data present a complex relationship with TAB-seq data (fig. S1A). The DeepH&M model is composed of three modules: a regular neural network–based CpG module, a convolutional neural network–based DNA module, and a regular neural network–based joint module (Fig. 1B). The inputs for the CpG module are genomic features and methylation features (table S1) for each CpG. Genomic features include GC percent, CpG density, and distance to the nearest CpG island (CGI). Methylation features include MeDIP-seq, MRE-seq, and hmC-Seal signal. Because CpG in proximity tends to have similar 5hmC and 5mC levels (fig. S1B), we also include average signal for the above features in neighboring windows around the target CpG. The DNA module takes DNA sequence around a CpG as inputs and uses convolutional neural network to extract information from the DNA sequence. The joint module combines outputs from the CpG module and DNA module and predicts 5hmC and 5mC levels simultaneously.

Benchmarking DeepH&M model

To examine the performance of DeepH&M, we generated WGBS, TAB-seq, MeDIP-seq, MRE-seq, and hmC-Seal data for 7-week-old mouse cerebellum and trained DeepH&M model with these datasets. Because DeepH&M requires 5hmC and 5mC as the labels, we used a statistical method MLML (maximum likelihood methylation levels) (*28*) to integrate TAB-seq and WGBS data to get consistent 5hmC, 5mC, and total methylation. MLML can prevent obtaining negative

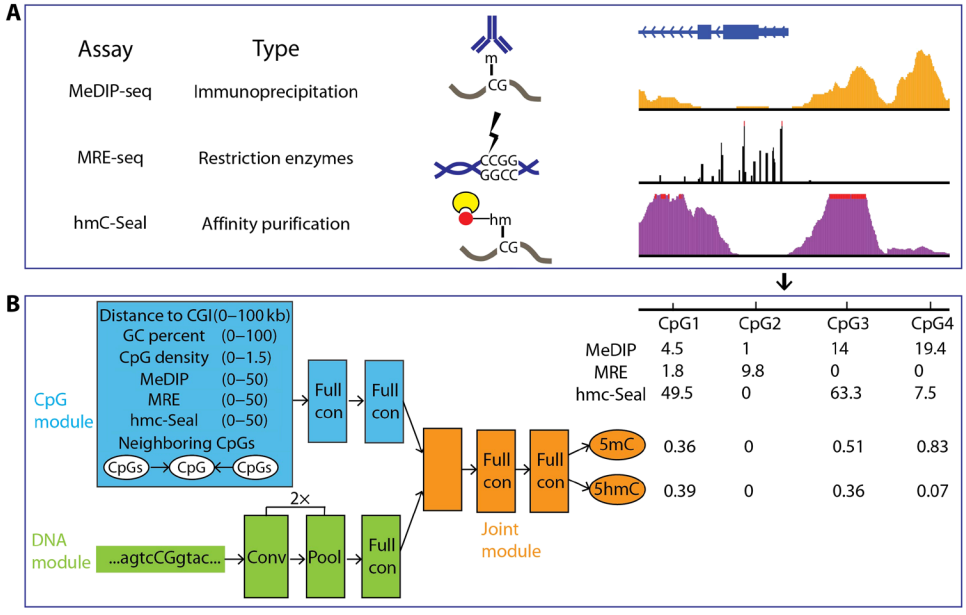

Fig. 1. DeepH&M model. (**A**) Schematic explanations for the three main assays used for the DeepH&M model. (**B**) Structure of the DeepH&M model. DeepH&M is composed of three modules. CpG module takes inputs of genomic features and methylation features. DNA module processes raw DNA sequence data using a convolutional neural network. Joint module combines outputs from the CpG module and DNA module to predict 5hmC and 5mC simultaneously. Examples were given to show how 5hmC and 5mC were predicted from the three main assays. Conv is convolutional layer. Pool is pooling layer. Full con is full connected layer.

5mC values by subtracting TAB-seq data directly from the WGBS data and also prevent the contradiction of TAB-seq and WGBS data at some CpG sites. As a reference, we called 5hmC, 5mC, and total methylation derived from MLML as "gold standard" data and evaluated our predictions against them. However, we recognize that even the gold standard data might not represent the true hydroxymethylation and methylation levels of a sample due to intrinsic limitations of profiling methods as described previously (*29*, *30*).

Our predicted 5hmC, 5mC, and total methylation levels are in high concordance with gold standard results. DeepH&M recapitulates the distribution of gold standard 5hmC, 5mC, and total methylation (Fig. 2, A and B). The genome-wide correlation across our predictions and gold standard data for 5hmC, 5mC, and total methylation is 0.8, 0.85, and 0.85, respectively (Fig. 2A). Using a previously developed concordance metric (defined as the percentage of CpGs with a methylation proportion difference less than 0.1 or 0.25) (*31*), 5hmC predictions are 86% concordant with gold standard data within 0.1 difference, 5mC predictions are 90% concordant within 0.25 difference, and total methylation predictions are 91% concordant within 0.25 difference. To examine whether the concordance is high only at particular 5hmC/5mC/total methylation levels, we examined the concordance at differing 5hmC/5mC/total methylation windows (Fig. 2C). 5hmC concordance is over 80% for 5hmC levels less than 0.4, and 45% for 5hmC levels higher than 0.4. We report that less than 1% of the CpGs in mouse cerebellum have 5hmC levels higher than 0.4. One explanation for the low concordance could be the paucity of high hmC CpGs in the training set (2 million CpGs); thus, DeepH&M might have difficultly learning the rules for high 5hmC CpGs. The concordance for 5mC is relatively lower for 5mC at the 0.2 to 0.4 window, and the concordance for total methylation is low for total methylation at the 0.2 to 0.6 window. This may be due to the difficulty in predicting intermediate methylation, as the problem also existed in predictions by methylCRF (*26*). The high concordance can be appreciated in the WashU Epigenome browser view of the *Slc22a17* and *Efs* locus, where 5hmC, 5mC, and total methylation levels of predicted and gold standard data are visualized (Fig. 2D). Furthermore, as a positive control for evaluating our predictions against gold standard data, we examined the concordance of two 7-week-old cerebellum replicates (fig. S2). The genome-wide correlation for 5hmC, 5mC, and total methylation between the two replicates is 0.82, 0.89, and 0.91, respectively, and the concordance is 88, 92, and 94%, respectively. The concordance of our predictions with gold standard data is very close to the concordance of the two replicates. These results confirm that DeepH&M can estimate single-CpG hydroxymethylation and methylation with high accuracy.

Because it has been shown that 5hmC is enriched at enhancers and 5hmC levels at the gene body are positively correlated with gene expression (*16*–*18*), we investigated whether our 5hmC predictions can reveal these relationships. To examine the enrichment of 5hmC in genomic features, we divided CpGs into four categories based on their 5hmC levels and calculated the enrichment fold of the four CpG categories in genomic features. We found that the enrichment of DeepH&M-predicted 5hmC in genomic features was similar to that of gold standard 5hmC (fig. S3A). CpGs with high 5hmC levels by predictions or gold standard data were highly enriched for enhancers and depleted for promoters. To examine the relationship between 5hmC and gene expression, we grouped genes into four categories based on expression levels and profiled average 5hmC levels at the gene body of the four categories of genes. We observed that similar to the gold standard 5hmC, the predicted 5hmC levels were positively correlated with gene expression (fig. S3B).

Factors affecting DeepH&M performance

Next, we wanted to investigate factors that may affect DeepH&M's performance. First, we examined DeepH&M's performance across different genomic features, as DNA methylation and hydroxymethylation were known to be highly nonrandom across the genome. The concordance is over 93% at CGIs and promoters for 5hmC and 5mC (Fig. 3A). The concordance for other genomic features is over 80% for 5hmC and over 87% for 5mC. Because most CGIs are lowly methylated and only a small portion of CGIs are highly methylated, we wanted to see whether DeepH&M can distinguish highly methylated CGIs from lowly methylated CGIs. We divided CGIs into lowly methylated CGI and highly methylated CGIs based on total methylation levels and then examined the concordance of predictions and gold standard data in these two types of CGIs. At lowly methylated CGIs, the concordance for 5hmC and 5mC is 99.9 and 99.8%, respectively (Fig. 3B). At highly methylated CGIs, the concordance for 5hmC and 5mC is 95 and 98%, respectively. These results indicate that DeepH&M's predictions are determined by experimental data instead of a learned assumption that all CGIs are lowly methylated. Second, because the accuracy of methylation levels from TAB-seq and WGBS data is substantially influenced by sequencing coverage, we examined DeepH&M's performance across differing CpG coverage from TAB-seq and WGBS data. The concordance for 5hmC and 5mC increases steadily from less than 10× coverage to over 10× coverage (85 to 88% for 5hmC, 78 to 89% for 5mC) (Fig. 3C). Thus, the lower concordance at lower coverage is likely a consequence of lower confidence in gold standard data, underscoring the robustness of our algorithm. Third, we examined DeepH&M's performance across regions with differing CpG density, as CpG density is a confounding factor for our enrichment-based sequencing methods, MeDIP-seq and hmC-Seal, which do not work optimally for regions with low CpG density. We observed increasing concordance for 5hmC and 5mC with increasing CpG density. Note that the concordance was greater than 0.8 even at the lowest CpG density; it increased to over 88% (5hmC) and 92% (5mC) for high CpG density regions that most of the current investigations focus on (Fig. 3D).

Generalizability of DeepH&M model to explore hydroxymethylation and methylation dynamics

Last, we wanted to test whether the DeepH&M model, trained on data from 7-week-old mouse cerebellum, can be generalized to data of other samples. This includes whether DeepH&M can predict differentially hydroxymethylated regions (DHMRs) and differentially methylated regions (DMRs) between two samples. We generated WGBS, TAB-seq, MeDIP-seq, MRE-seq, and hmC-Seal data for 79-week-old mouse cerebellum as we wanted to explore 5hmC changes during aging. Using the DeepH&M model from 7-week-old mouse cerebellum, we predicted 5hmC and 5mC for 79-week-old mouse cerebellum. We performed similar concordance analysis between predictions and gold standard data for 79-week-old mouse cerebellum. The overall performance of the DeepH&M model in 79-week-old mouse cerebellum is similarly high as that in 7-week-old mouse cerebellum (Fig. 4, A to C). The genome-wide correlation for 5hmC, 5mC, and total methylation between predictions and gold standard data is 0.81, 0.86, and 0.86, respectively, and the concordance is 84, 91, and 92%, respectively. As illustrated by the WashU Epigenome browser

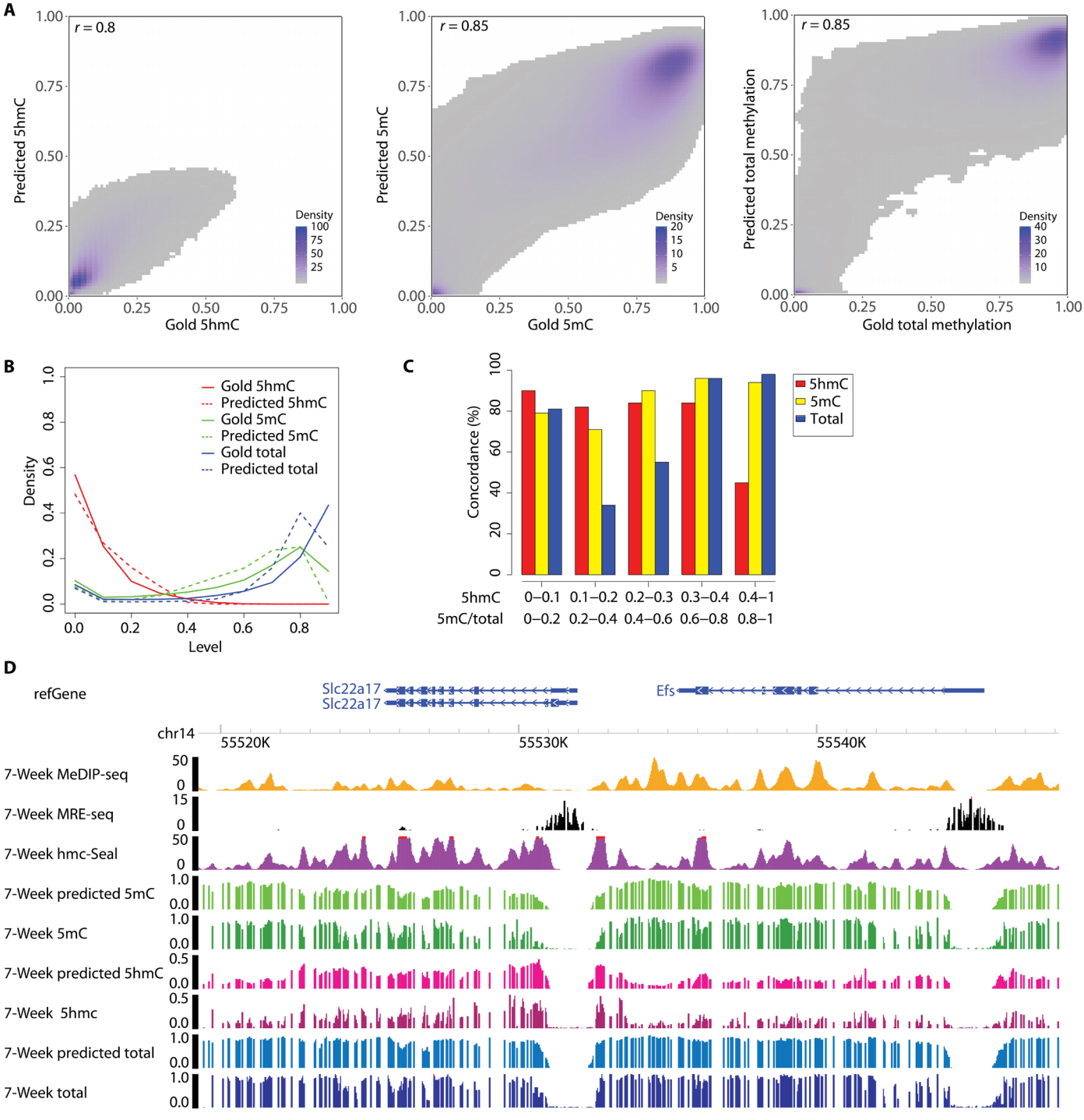

Fig. 2. Performance of the DeepH&M model in 7-week-old mouse cerebellum. (**A**) Density plots of predictions and gold standard data for 5hmC, 5mC, and total methylation. Pearson correlation coefficient is used as correlation metric. (**B**) Global distribution comparison of predictions and gold standard data for 5hmC, 5mC, and total methylation. (**C**) Concordance between predictions and gold standard data for 5hmC, 5mC, and total methylation at CpGs with differing 5hmC/5mC/total methylation levels. For 5hmC, 0.1 difference is used to calculate concordance. For 5mC and total methylation, 0.25 difference is used. Concordance for five ascending 5hmC windows and five ascending 5mC/total methylation windows is calculated to see how concordance distributes in differing 5hmC/5mC/total methylation levels. (**D**) Genome browser view of predictions and gold standard data for 7-week-old cerebellum at a representative locus.

view, there is high concordance between DeepH&M prediction and gold standard data across 5hmC, 5mC, and total methylation levels in the 5′ untranslated region (5′UTR) and first exon of the *Kcnd2* gene (Fig. 4D).

Recent research suggests that epigenetic mechanisms, DNA methylation in particular, play a central role in the aging process (*32*). Using antibody-based methods to quantify 5hmC levels, several studies reported global levels of 5hmC increase in mouse cerebellum during

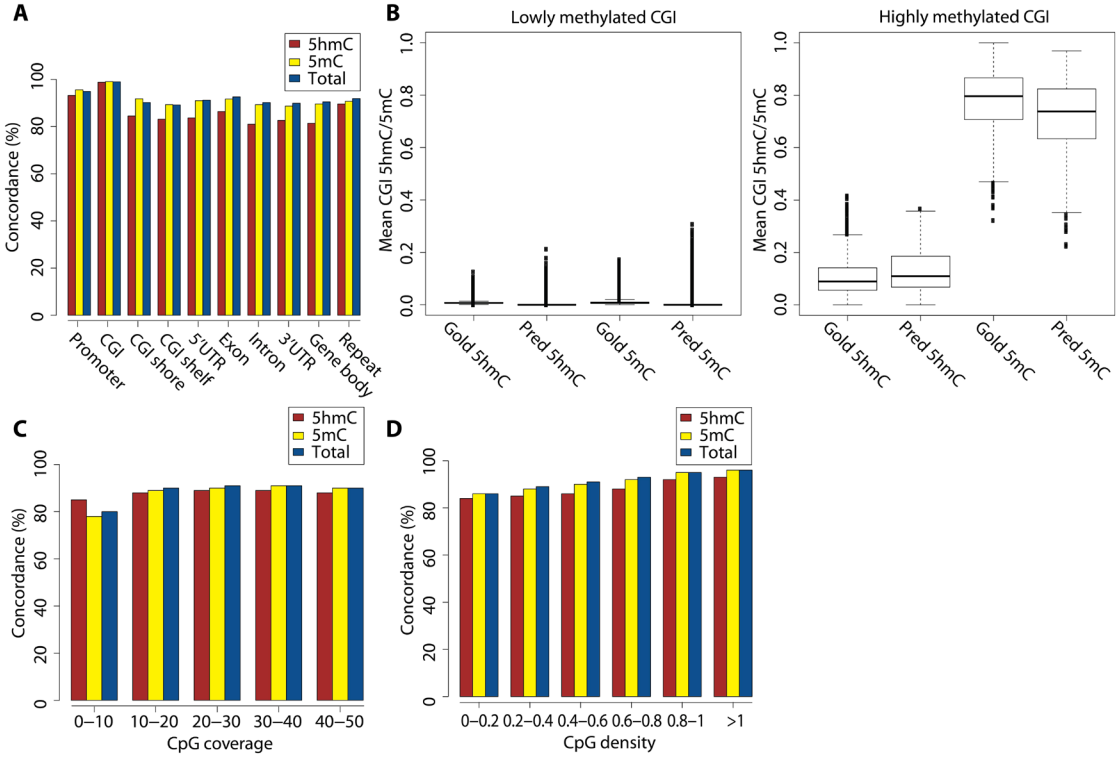

Fig. 3. Factors affecting concordance between gold standard data and predictions. (**A**) Concordance for 5hmC/5mC/total methylation at different genomic features. (**B**) Comparison of gold standard 5hmC/5mC and predicted 5hmC/5mC at lowly methylated CGIs and highly methylated CGIs. CGIs are divided into lowly methylated CGIs (<0.2) and highly methylated CGIs (>0.7) based on their average total methylation levels. (**C**) Concordance for 5hmC/5mC/total methylation as a function of CpG coverage. For 5hmC concordance, CpG coverage is from TAB-seq data. For 5mC/total methylation concordance, CpG coverage is from WGBS data. (**D**) Concordance for 5hmC/5mC/total methylation as a function of CpG density.

aging but remain stable in mouse hippocampus (*33*, *34*). Furthermore, a recent study used single-base resolution sequencing method [oxidative-bisulfite sequencing (oxBS-seq)] to measure 5hmC at single sites in mouse hippocampus and found no global 5hmC changes (*35*). However, because of low sequencing depth (2×), the study only examined 5hmC changes at the chromosome level and genomic element level, such as CGIs and promoters, and could not provide single-base resolution 5hmC dynamics at local regions.

In this study, we explored whether DeepH&M could reveal how 5hmC changes globally and locally in mouse cerebellum during aging. We report that global 5hmC levels increase by 20% from 7 to 79 weeks and that global 5mC levels do not change (table S2). Next, we examined whether there are 5hmC and 5mC changes in specific regions during aging by calling DHMRs and DMRs. First, we identified 524 DHMRs between hmC-Seal data of 7- and 79-week-old mouse cerebella using DiffBind (*36*). We wanted to see whether 5hmC changes in these DHMRs are similar between predictions and gold standard data. The hyperDHMRs have significantly higher 5hmC in both gold standard data and predictions, and hypoDHMRs have significantly lower 5hmC in both gold standard data and predictions (Fig. 5A). Thus, both gold standard data and DeepH&M predictions support DHMRs defined by hmC-Seal data. Second, we defined DHMRs and DMRs by comparing TAB-seq and WGBS data between 7- and 79-week-old cerebella using the tool DSS (*37*). We examined whether these DHMRs/DMRs are supported by DeepH&M data. The differences predicted by DeepH&M are highly significant, and they are concordant with differences defined by gold standard data, although the overall magnitude tends to be smaller (Fig. 5, B and C).

We also examined enrichment of biological processes for these DHMRs and DMRs using Genomic Regions Enrichment of Annotations Tool (GREAT) (*38*). We report that hyperDHMRs are enriched near genes that regulate synaptic plasticity and transporter activity (fig. S4A) and that hyperDMRs are enriched in genes responsible for neuron axonogenesis (fig. S4B). There were no significantly enriched terms associated with hypoDMRs and hypoDHMRs, possibly due to the small number of hypoDMRs and hypoDHMRs. As an example, Fig. 4D illustrates one of the numerous DHMRs between 7- and 79-week-old cerebella. The 5hmC changes at this region are supported by changes of gold standard 5hmC, predicted 5hmC, and hmC-Seal signal between the two ages. These results suggest that DeepH&M can predict DHMRs and DMRs between two samples.

The above analysis demonstrates that the DeepH&M model, trained on data from 7-week-old mouse cerebellum, can be generalized to 79-week-old mouse cerebellum. We wanted to examine whether our DeepH&M model can be also generalized to 7-week-old mouse cortex as 5hmC levels in the cortex are much higher than that in the cerebellum. We found that the overall performance of the DeepH&M model for 5hmC is a little lower in the cortex than in the cerebellum (concordance: 72% versus 86%), and the performance for 5mC and total methylation is similar to cerebellum (Fig. 6, A to C). The genome-wide correlation for 5hmC, 5mC, and total methylation between predictions and gold standard data is 0.65, 0.82, and 0.89,

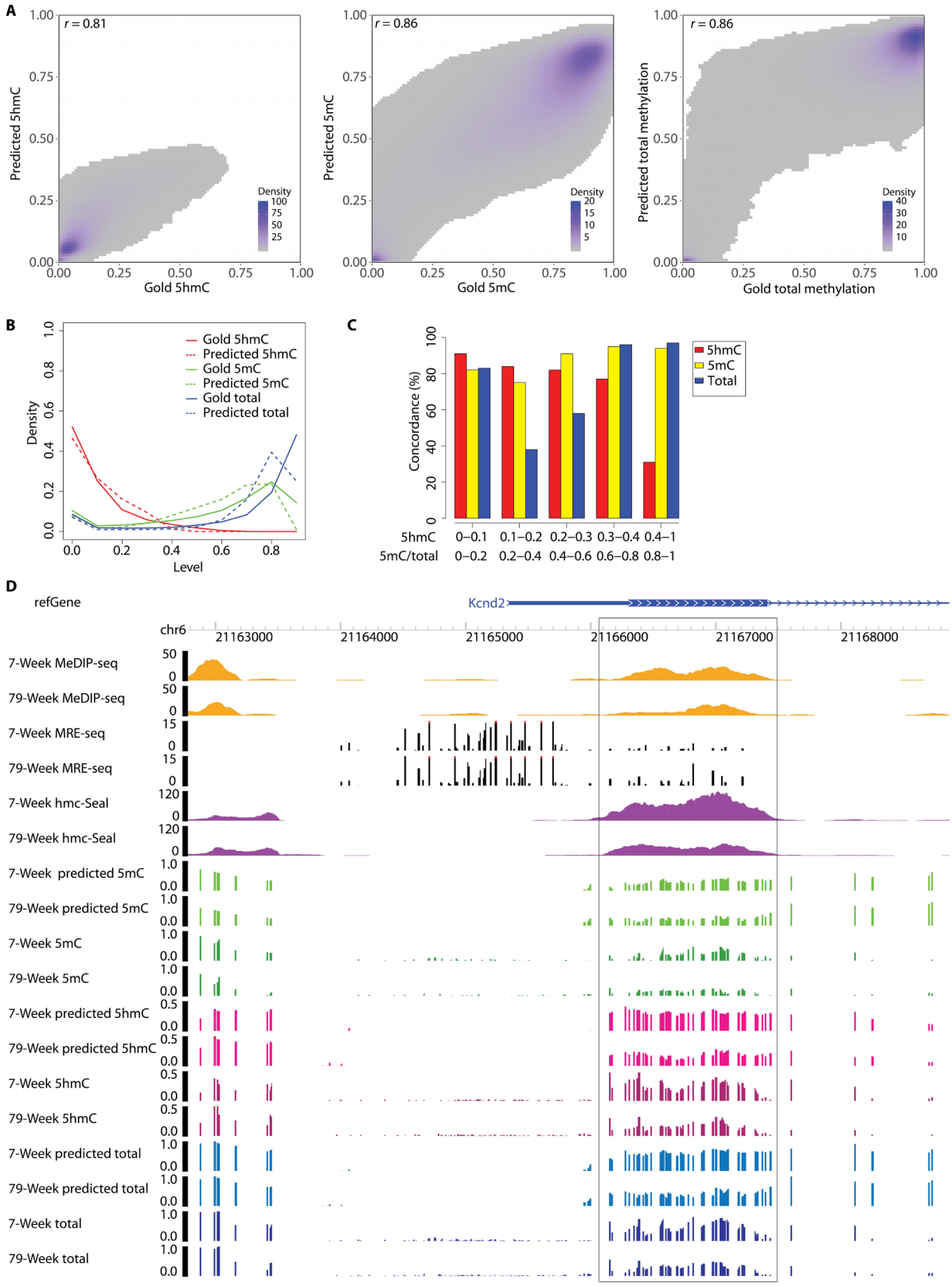

Fig. 4. Performance of the DeepH&M model in 79-week-old mouse cerebellum. (**A**) Density plots of predictions and gold standard data for 5hmC, 5mC, and total methylation. (**B**) Global distribution comparison of predictions and gold standard data for 5hmC, 5mC, and total methylation. (**C**) Concordance between predictions and gold standard data for 5hmC, 5mC, and total methylation at CpGs with differing 5hmC/5mC/total methylation levels. (**D**) Genome browser view of a DHMR between 7- and 79-week-old cerebella. The selected box is the DHMR. The 5hmC changes at this region are supported by changes of gold standard 5hmC, predicted 5hmC, and also hmC-Seal signal between the two ages.

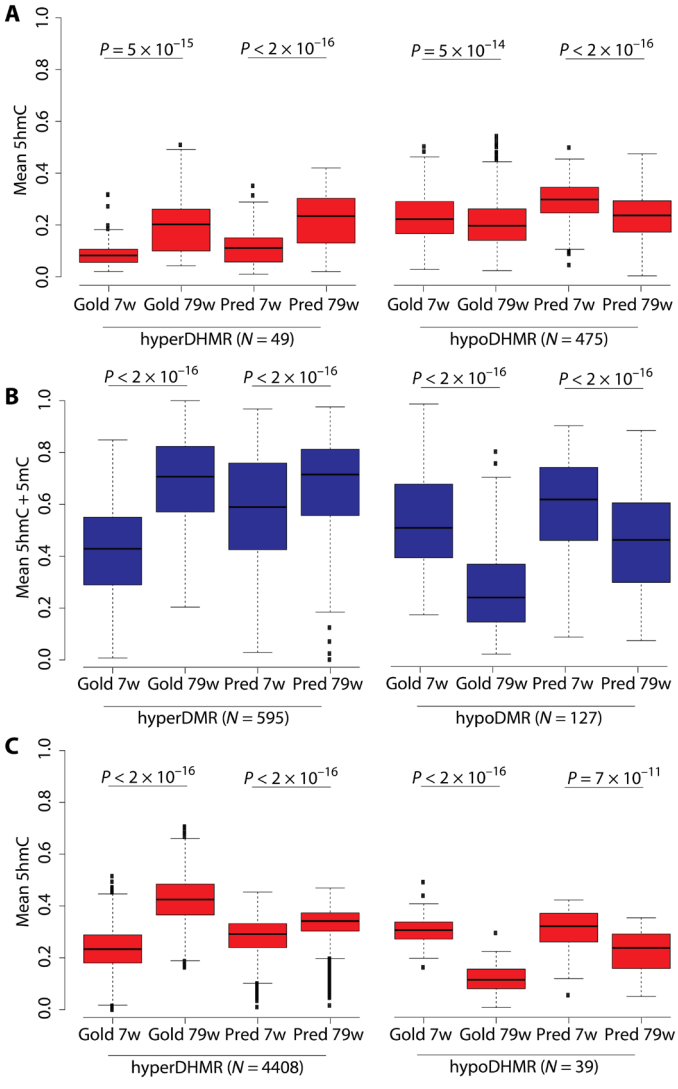

Fig. 5. DeepH&M can predict DHMRs and DMRs between 7- and 79-week-old mouse cerebella. (**A**) Distribution of mean 5hmC for gold standard data and predictions at hyperDHMRs and hypoDHMRs defined by hmC-Seal data between 7- and 79-week-old cerebella. Gold is for gold standard data. Pred is for prediction. *N* is the number. 7w, 7 weeks; 79w, 79 weeks. (**B**) Distribution of mean 5hmC + 5mC for gold standard data and predictions at hyperDMRs and hypoDMRs defined by WGBS data between 7- and 79-week-old cerebella. (**C**) Distribution of mean 5hmC for gold standard data and predictions at hyperDMRs and hypoDMRs defined by TAB-seq data between 7- and 79-week-old cerebella.

respectively, and the concordance is 72, 89, and 92%, respectively. We can see that 5hmC distribution in the cortex is distinct from that in the cerebellum (Fig. 2B versus Fig. 6B), and the mean 5hmC level in the cortex is almost twice as high as that in the cerebellum (0.19 versus 0.11). DeepH&M can still recapitulate the distribution of gold standard 5hmC and 5mC and total methylation. These results suggest that the DeepH&M model trained from cerebellum is not only generalizable to other cerebellum samples at different ages but also generalizable to adult frontal cortex. We also applied our DeepH&M model to mouse fetal cortex, which has much lower global 5hmC levels than adult cortex. The genome-wide correlation for 5hmC, 5mC, and total methylation between predictions and gold standard data provided in Lister *et al.* (*39*) is 0.44, 0.63, and 0.65, respectively, and the concordance is 61, 84, and 94%, respectively (fig. S5). The extremely low concordance for 5hmC in the fetal cortex may be explained by the rather big global differences in 5hmC distribution in adult and fetal cortices.

DISCUSSION

5hmC is known to be an intermediate, but stable, epigenetic feature of the active DNA demethylation process. However, the molecular mechanisms underlying the role of 5hmC in gene regulation remain largely unknown. Furthermore, the loss of 5hmC has been identified as a hallmark of most types of human cancers. Many cancers are characterized by down-regulation of or deleterious mutations in TET or isocitrate dehydrogenase IDH1/IDH2 (cofactors of TET enzymes) genes, which reduces the rate of oxidization of 5mC into 5hmC (*8*, *10*, *11*). Note that many of these studies use hMeDIP-seq technology to profile tumor and matched-tumor samples; therefore, there is a lack of high-resolution hydroxymethylomes of tumors.

Understanding the mechanisms underlying 5hmC's roles in development and tumorigenesis can benefit from profiling 5hmC levels at genome-wide, single-base resolution. As shown in Wen *et al.* (*40*), high-resolution 5hmC profiling of the human brain revealed intriguing 5hmC signatures, such as high hydroxymethylation levels near 5' splicing sites and transcription-correlated hmC levels on the sense strand of the gene, that hMeDIP-seq would not be able to detect due to inherent limitations of the technology. Identifying these novel signatures could hold the key in deciphering the biological machineries that 5hmC could potentiate. Currently, TAB-seq and oxBS-seq are the gold standard methods for providing single-CpG resolution DNA hydroxymethylomes (*18*, *41*). These two methods require very high coverage to confidently call 5hmC at all cytosines. The coverage required for oxBS-seq is even higher due to the fact that oxBS-seq measures 5hmC indirectly through subtracting measured 5mC from measured total methylation. The high cost associated with the high coverage is a substantial financial barrier for individual laboratories to adopt TAB-seq and oxBS-seq as a routine assay for DNA hydroxymethylomes. So far, only a few cell types have deeply sequenced hydroxymethylomes at single-base resolution (*18*, *30*, *39*, *40*, *42*–*46*).

To overcome this potential cost-barrier problem, we have developed a deep learning–based algorithm, DeepH&M, which integrates enrichment and restriction enzyme sequencing methods to estimate the absolute levels of hydroxymethylation and methylation at single-CpG resolution. The cost of the three assays combined is <5% of WGBS and TAB-seq. About 50 to 100 million MeDIP reads, 30 million MRE reads, and 50 million hmC-Seal-seq reads are sufficient for measuring a hydroxymethylome with DeepH&M, which translates to roughly 3× coverage of the human or mouse genome. In addition, TAB-seq requires ~3 μg of genomic DNA, while MeDIP-seq, MRE-seq, and hmC-Seal can be generated from 100 ng or less input, thus allowing DeepH&M to be more amenable to rare or difficult-to-procure cells or samples. Compared with 100× coverage for TAB-seq and 20× coverage for WGBS, our method can minimize the cost of generating a complete hydroxymethylome by 40-fold. Furthermore, DeepH&M can estimate for all CpGs, while WGBS and TAB-seq miss a considerable fraction of the genome due to low coverage. As mentioned previously, previous TAB-seq study on H1 cells could only confidently call 20% or higher 5hmC at a coverage of 27 and, thus, identified less than 1 million hydroxymethylated CpGs (*18*).

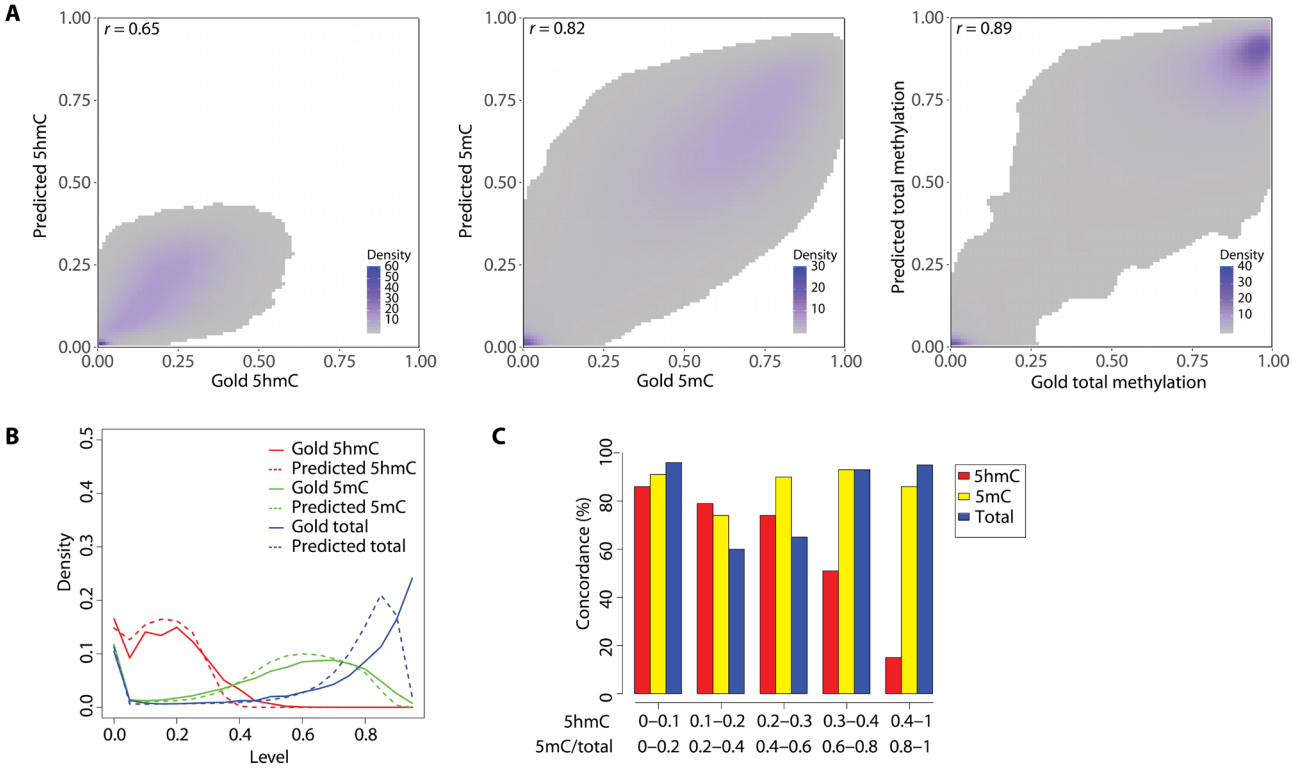

Fig. 6. Performance of the DeepH&M model in 7-week-old mouse cortex. (**A**) Density plots of predictions and gold standard data for 5hmC, 5mC, and total methylation. (**B**) Global distribution comparison of predictions and gold standard data for 5hmC, 5mC, and total methylation. (**C**) Concordance between predictions and gold standard data for 5hmC, 5mC, and total methylation at CpGs with differing 5hmC/5mC/total methylation levels.

One caveat to DeepH&M is that TAB-seq and WGBS libraries must be sequenced initially to generate training data for the cell type of interest. Because creating comprehensive hydroxymethylome and methylome can be cost prohibitive, we explored alternative methods of generating training data. Currently, Infinium MethylationEPIC BeadChip Kit (Illumina, WG-317-1001) can profile the methylation levels from roughly 850,000 CpGs at single-nucleotide resolution for humans. To address whether methylation microarray results could be used as training set, we asked whether DeepH&M can be trained on 850,000 CpGs in our mouse data. Compared with 2 million CpG training data, which have 86% 5hmC and 90% 5mC concordance, DeepH&M can still predict with 83 and 89% concordance for 5hmC and 5mC, respectively. Therefore, to reduce the cost of generating training data, we can replace WGBS and TAB-seq with methylation arrays coupled with bisulfite- and TAB-treated samples, respectively (*30*). It is also feasible to supply other enrichment and restriction enzyme sequencing methods as replacement of DeepH&M inputs, such as replacing hmC-Seal with hMeDIP-seq. However, users need to retrain the DeepH&M model when using new input methods.

Using 7-week-old mouse cerebellum data for training DeepH&M model, we demonstrated that the estimated 5hmC and 5mC levels were in high concordance with those estimated by combining TAB-seq and WGBS data. DeepH&M estimated 5hmC levels at 85% concordance with TAB-seq data within 0.1 difference, and DeepH&M estimated total methylation level at 91% concordance with WGBS data within 0.25 difference. Furthermore, DeepH&M can be generalizable to other tissues and biological time points. DeepH&M model trained on 7-week-old mouse cerebellum data was able to estimate 5hmC and 5mC levels with high performance for 79-week-old mouse cerebellum (concordance for 5hmC and total methylation is 84 and 92%). DHMRs and DMRs between 7- and 79-week-old mouse cerebella can be recapitulated using the estimated 5hmC and 5mC values from DeepH&M for the two ages. However, we report relatively lower performance for 7-week-old mouse cortex (concordance for 5hmC and total methylation is 72 and 92%, respectively). The relatively lower performance for cortex may be explained by the rather big global differences of 5hmC distribution in cerebellum and cortex, as the mean 5hmC level is 0.19 in cortex and 0.11 in cerebellum. As one of the caveats of DeepH&M, these data suggest that the DeepH&M model cannot be generalized to different tissues when 5hmC levels differ greatly between tissues. When we applied our DeepH&M model to mouse fetal cortex (mean 5hmC level of 0.05), the concordance for 5hmC and total methylation is 61 and 94%, respectively. The extremely low concordance for 5hmC indicates that the mean level of 5hmC should be taken into account when applying trained models to different biological systems. Because of the dynamic range of absolute 5hmC levels in different tissues, the relationships between MeDIP-seq, MRE-seq, and hmC-Seal data and 5hmC are different in different tissues, and thus, a single DeepH&M model cannot be generalized to all tissues. One way to address this limitation is to categorize tissues into multiple classes based on their 5hmC levels and train a DeepH&M model for each group. The DeepH&M model trained for each group can then be generalized to tissues that have similar 5hmC levels.

MATERIALS AND METHODS

DeepH&M model

The DeepH&M model is derived from the DeepCpG model, which predicts single-cell DNA methylation states using deep learning (*47*). The DeepH&M model is composed of three modules: a regular neural network–based CpG module, a convolutional neural network–based DNA module, and a regular neural network–based joint module (Fig. 2). The CpG module extracts information from inputs of genomic features and methylation features of a CpG with regular neural network. The DNA module takes DNA sequence around a CpG as input and uses convolutional neural network to extract information from the DNA sequence. The joint module combines outputs from the CpG module and DNA module and predicts 5hmC and 5mC simultaneously with regular neural network.

Unlike the CpG module in DeepCpG, which is a recurrent neural network, the CpG module in DeepH&M is a regular neural network using two fully connected layers with 100 neurons and rectified linear unit (ReLU) activation function. The inputs for the CpG module are genomic features and methylation features (table S1) for each CpG. Genomic features include GC percent, CpG density, and distance to the nearest CGI. Methylation features include MeDIP-seq, MRE-seq, and hmC-Seal signal. Because CpGs in proximity tend to have similar 5hmC and 5mC levels, we also include average signal for the above features in neighboring windows (0 to 50 bp, 50 to 250 bp, 250 to 500 bp, and 500 to 1000 bp) around the target CpG.

The structure of our DNA module is the same as that of the DNA module of the DeepCpG model, except that the activation function in our DNA module is tanh function instead of ReLU function (with two connected layers: layer 1 with 120 neurons and layer 2 with 240 neurons).

Joint module uses two fully connected layers with 100 neurons and ReLU activation function to predict 5hmC and 5mC simultaneously, unlike the joint module in DeepCpG, which only predicts DNA methylation.

We used data that have at least 25× coverage from TAB-seq data and 20× coverage from WGBS data for training and validation. The feature data are normalized by Z score normalization. Because the number of high-5hmC-level CpGs was much smaller than that with low hmC levels, we balanced the training set through subsampling and oversampling. We divided CpGs into nine windows based on 5hmC levels 0 to 0.05, 0.05 to 0.1, 0.1 to 0.15, 0.15 to 0.2, 0.2 to 0.25, 0.25 to 0.3, 0.3 to 0.35, 0.35 to 0.4, and 0.4 to 1 and subsampled CpGs if the number of CpGs in the window was higher than a threshold and oversampled CpGs if the number of CpGs in the window was less than a threshold. The threshold was chosen as the median of the number of CpGs in nine windows. Data were randomly split into training set (2 million CpGs), validation set (0.5 million CpGs), and test set (the rest). Model parameters were learnt on the training set by minimizing the L2 loss function. We selected the model that had the smallest loss in the validation set and used the model to predict 5hmC and 5mC for all CpGs.

Tissue sample dissection and genomic DNA extraction

All procedures were approved by the Washington University Institutional Animal Care and Use Committee. Two male 6-week-old C57BL/6J mice and two male 78-week-old C57BL/6J mice were purchased (the Jackson laboratory, 000664) and allowed to acclimate in the mouse facility for a week. Cerebella were dissected following protocol described previously (*48*) from mice in both age groups, while the frontal cortex (from bregma +1.0 mm to the base of the olfactory bulb) was dissected as described previously (*39*) from 7-week-old mice. All tissues were snap frozen in liquid nitrogen immediately after dissection.

Each tissue was cut into two pieces with a sterile razor blade for subsequent DNA and RNA extraction immediately after. For genomic DNA extraction, we followed previously established protocol (*49*). In brief, each tissue piece was incubated in 600 μl of lysis buffer [50 mM tris-HCl (pH 8), 1 mM EDTA (pH 8), 0.5% SDS, proteinase K (1 mg/ml)] at 55°C for 4 hours. DNA was purified by phenol/chloroform/isoamyl alcohol extraction followed by ethanol extraction. DNA used for MeDIP-seq was sheared into 100- to 500-bp fragment size with the Bioruptor Pico Sonication System, while DNA for WGBS and TAB-seq was sheared into 200- to 600-bp fragment size with a Covaris E220 ultrasonicator.

MeDIP-seq, MRE-seq, and hmC-Seal library construction and data processing

MeDIP-seq libraries were generated as previously described (*49*) with few modifications. One hundred nanograms of sheared DNA was ligated with Illumina adapters, and methylation-enriched adapter-ligated DNA fragments were immunoprecipitated with 0.1 μg of anti-methylcytidine antibody (Eurogentec, BI-MECY-0100). MeDIP DNA fragments were amplified with Illumina barcodes with NEBNext High-Fidelity 2× PCR Master Mix (polymerase chain reaction) master mix (NEB, M0541). MeDIP-seq libraries were sequenced on Illumina NovaSeq 6000 platform.

MRE-seq libraries were generated as previously described (*49*) with few modifications. In brief, 50 ng of genomic DNA was digested by four restriction enzymes (HpaII, HinP1I, AciI, and HpyCH4IV) that generate a CG overhang. Adapter ligation was performed with custom Illumina adapters (5′-ACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3′ and 5′-P-CGAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3′). Adapter-ligated DNA fragments were amplified with Illumina barcodes with NEBNext High-Fidelity 2× PCR Master Mix master mix (NEB, M0541) and sequenced on Illumina NovaSeq 6000 platform.

To identify 5hmC-enriched regions, we performed Nano-hmC-Seal (*19*) on tissue samples. In brief, 50 ng of genomic DNA was used in the tagmentation reaction. The tagmented DNA was glucosylated by incubating in a 50-μl solution containing 1× glucosylation buffer, 200 μM UDP-azide-glucose (Active Motif, 55020), and 5 U of T4 β-glucosyltransferase (Thermo Fisher Scientific, EO0831) at 37°C for 1 hour. After glucosylation, the DBCO-PEG4-biotin reaction and streptavidin C1 bead pull-down were same as the Nano-hmC-Seal (*19*). The beads were washed 10 times with 1× binding-washing buffer and twice with double-distilled water (ddH_2O) and were resuspended in 15 μl of ddH2O. The captured DNA fragments were amplified and barcoded by PCR using the NEBNext High-Fidelity 2× PCR Master Mix (NEB, M0541). hmC-Seal libraries were sequenced on Illumina NovaSeq 6000 platform.

The reads for MeDIP-seq, MRE-seq, and hmC-Seal were aligned to the mm9 reference genome with BWA (*50*) and then processed by methylQA (*49*). The signal for MeDIP-seq, MRE-seq, and hmC-Seal at each CpG was the number of reads aligned to that location divided by total reads (million). The average signal for MeDIP-seq, MRE-seq, and hmC-Seal in each window was the mean of signal at all bases in that window.

WGBS and TAB-seq library construction and data processing

WGBS and TAB-seq libraries were constructed using the 5hmC TAB-Seq Kit (Wisegene, K001) following the manufacturer's protocol with few modifications detailed below. Five micrograms of sheared gDNA was treated with β-glucosyltransferase–based reaction to glucosylate 5hmCs. Four hundred nanograms of glucosylated DNA was incubated in TET-based oxidation reaction at 37°C for 1.5 hours. Five hundred nanograms of glucosylated DNA and 250 ng of TET-oxidized DNA were bisulfite converted using EZ DNA Methylation-Gold Kit (Zymo, D5005) for subsequent WGBS and TAB-seq library construction, respectively, with Accel-NGS Methyl-Seq DNA Library Kit (Swift Biosciences, 30024). WGBS and TAB-seq libraries were sequenced on Illumina NovaSeq 6000 platform.

The reads for TAB-seq and WGBS data were aligned to mm9 reference genome and processed using Bismark (*51*). A statistical method, MLML, was used to integrate TAB-seq and WGBS data to get consistent 5hmC and 5mC and total methylation (*28*).

DHMRs and DMRs identification

DHMRs between hmC-Seal datasets were defined by DiffBind (*36*) with a *q* value of 0.01.

DHMRs between TAB-seq datasets and DMRs between WGBS datasets were defined by DSS (*37*). Two replicates and smoothing options were used for DSS. The called DHMRs and DMRs were then filtered by requiring a minimal coverage of 10 by TAB-seq and WGBS data and the absolute difference of gold standard 5hmC (for DHMRs) and total methylation (for DMRs) in two datasets over 0.15.

Coverage required to call 5% 5hmC

On the basis of the binomial test with a probability of 2.22% for 5mC nonconversion rate, the *P* value for using a coverage of 120 to call 5% 5hmC was calculated in R by binom.test(round(120*0.05), 120, $P = 0.0222$, alternative = "greater"). The resulted *P* value for the test was 0.05184. Therefore, a coverage of 120 was required to called 5% 5hmC at 95% confidence level.

Enrichment of 5hmC in genomic features

Enrichment fold = (#CpG for class A CpGs overlapping genomic feature B/#CpG in class A CpGs)/(#CpG for all classes of CpGs overlapping genomic feature B/#CpG in all classes of CpGs).

mRNA-seq library construction and data processing

Total RNA from tissue samples was extracted using TRIzol reagent as previously detailed (*52*). Five hundred nanograms of total RNA was processed with Universal Plus mRNA-seq (messenger RNA sequencing) kit (Nugen, 0508-08) to generate mRNA-seq libraries, which were sequenced on Illumina NovaSeq 6000 platform. mRNA reads were aligned to mm9 reference genome using STAR (spliced transcripts alignment to a reference) (*53*). Read counts for each gene were obtained using HTSeq (high-throughput sequencing) (*54*).

Software availability

DeepH&M tool is available in https://epigenome.wustl.edu/DeepHM/.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/27/eaba0521/DC1

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. C. M. Rivera, B. Ren, Mapping human epigenomes. *Cell* **155**, 39–55 (2013).
2. T. Chen, S. Y. Dent, Chromatin modifiers and remodellers: Regulators of cellular differentiation. *Nat. Rev. Genet.* **15**, 93–106 (2014).
3. K. D. Robertson, DNA methylation and human disease. *Nat. Rev. Genet.* **6**, 597–610 (2005).
4. M. M. Suzuki, A. Bird, DNA methylation landscapes: Provocative insights from epigenomics. *Nat. Rev. Genet.* **9**, 465–476 (2008).
5. P. W. Laird, Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.* **11**, 191–203 (2010).
6. P. A. Jones, Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
7. S. Ito, L. Shen, Q. Dai, S. C. Wu, L. B. Collins, J. A. Swenberg, C. He, Y. Zhang, Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300–1303 (2011).
8. G. P. Pfeifer, W. Xiong, M. A. Hahn, S.-G. Jin, The role of 5-hydroxymethylcytosine in human cancer. *Cell Tissue Res.* **356**, 631–641 (2014).
9. C. M. Greco, P. Kunderfranco, M. Rubino, V. Larcher, P. Carullo, A. Anselmo, K. Kurz, T. Carell, A. Angius, M. V. Latronico, R. Papait, G. Condorelli, DNA hydroxymethylation controls cardiomyocyte gene expression in development and hypertrophy. *Nat. Commun.* **7**, 12418 (2016).
10. J. Jeschke, E. Collignon, F. Fuks, Portraits of TET-mediated DNA hydroxymethylation in cancer. *Curr. Opin. Genet. Dev.* **36**, 16–26 (2016).
11. E. Smeets, A. G. Lynch, S. Prekovic, T. Van den Broeck, L. Moris, C. Helsen, S. Joniau, F. Claessens, C. E. Massie, The role of TET-mediated DNA hydroxymethylation in prostate cancer. *Mol. Cell. Endocrinol.* **462**, 41–55 (2018).
12. S. Monticelli, DNA (Hydroxy)methylation in T helper lymphocytes. *Trends Biochem. Sci.* **44**, 589–598 (2019).
13. H. Wu, Y. Zhang, Charting oxidized methylcytosines at base resolution. *Nat. Struct. Mol. Biol.* **22**, 656–661 (2015).
14. S. Kriaucionis, N. Heintz, The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**, 929–930 (2009).
15. M. Tahiliani, K. P. Koh, Y. Shen, W. A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind, A. Rao, Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).
16. C.-X. Song, K. E. Szulwach, Y. Fu, Q. Dai, C. Yi, X. Li, Y. Li, C.-H. Chen, W. Zhang, X. Jian, J. Wang, L. Zhang, T. J. Looney, B. Zhang, L. A. Godley, L. M. Hicks, B. T. Lahn, P. Jin, C. He, Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat. Biotechnol.* **29**, 68–72 (2011).
17. G. Ficz, M. R. Branco, S. Seisenberger, F. Santos, F. Krueger, T. A. Hore, C. J. Marques, S. Andrews, W. Reik, Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398–402 (2011).
18. M. Yu, G. C. Hon, K. E. Szulwach, C.-X. Song, L. Zhang, A. Kim, X. Li, Q. Dai, Y. Shen, B. Park, J.-H. Min, P. Jin, B. Ren, C. He, Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368–1380 (2012).
19. D. Han, X. Lu, A. H. Shih, J. Nie, Q. You, M. M. Xu, A. M. Melnick, R. L. Levine, C. He, A highly sensitive and robust method for genome-wide 5hmC profiling of rare cell populations. *Mol. Cell* **63**, 711–719 (2016).
20. M. Szyf, The elusive role of 5′-hydroxymethylcytosine. *Epigenomics* **8**, 1539–1551 (2016).
21. C. Bock, Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.* **13**, 705–719 (2012).
22. L. Laurent, E. Wong, G. Li, T. Huynh, A. Tsirigos, C. T. Ong, H. M. Low, K. W. K. Sung, I. Rigoutsos, J. Loring, C.-L. Wei, Dynamic changes in the human methylome during differentiation. *Genome Res.* **20**, 320–331 (2010).
23. M. Weber, J. J. Davies, D. Wittig, E. J. Oakeley, M. Haase, W. L. Lam, D. Schübeler, Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.* **37**, 853–862 (2005).
24. A. K. Maunakea, R. P. Nagarajan, M. Bilenky, T. J. Ballinger, C. D'Souza, S. D. Fouse, B. E. Johnson, C. Hong, C. Nielsen, Y. Zhao, G. Turecki, A. Delaney, R. Varhol, N. Thiessen, K. Shchors, V. M. Heine, D. H. Rowitch, X. Xing, C. Fiore, M. Schillebeeckx, S. J. Jones, D. Haussler, M. A. Marra, M. Hirst, T. Wang, J. F. Costello, Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**, 253–257 (2010).
25. Y. Xiao, F. Yu, L. Pang, H. Zhao, L. Liu, G. Zhang, T. Liu, H. Zhang, H. Fan, Y. Zhang, B. Pang, X. Li, MeSiC: A model-based method for estimating 5 mC levels at single-CpG resolution from MeDIP-seq. *Sci. Rep.* **5**, 14699 (2015).
26. M. Stevens, J. B. Cheng, D. Li, M. Xie, C. Hong, C. L. Maire, K. L. Ligon, M. Hirst, M. A. Marra, J. F. Costello, T. Wang, Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome Res.* **23**, 1541–1553 (2013).

27. M. Pavlovic, P. Ray, K. Pavlovic, A. Kotamarti, M. Chen, M. Q. Zhang, DIRECTION: A machine learning framework for predicting and characterizing DNA methylation and hydroxymethylation in mammalian genomes. *Bioinformatics* **33**, 2986–2994 (2017).
28. J. Qu, M. Zhou, Q. Song, E. E. Hong, A. D. Smith, MLML: Consistent simultaneous estimates of DNA methylation and hydroxymethylation. *Bioinformatics* **29**, 2645–2646 (2013).
29. D. Barros-Siva, C. J. Marques, H. Rui, C. Jerónimo, Profiling DNA methylation based on next-generation sequencing approaches: New insights and clinical Applications. *Genes* **9**, 429 (2018).
30. K. Skvortsova, E. Zotenko, P.-L. Luu, C. M. Gould, S. S. Nair, S. J. Clark, C. Stirzaker, Comprehensive evaluation of genome-wide 5-hydroxymethylcytosine profiling approaches in human DNA. *Epigenetics Chromatin* **10**, 16 (2017).
31. R. A. Harris, T. Wang, C. Coarfa, R. P. Nagarajan, C. Hong, S. L. Downey, B. E. Johnson, S. D. Fouse, A. Delaney, Y. Zhao, A. Olshen, T. Ballinger, X. Zhou, K. J. Forsberg, J. Gu, L. Echipare, H. O'Geen, R. Lister, M. Pelizzola, Y. Xi, C. B. Epstein, B. E. Bernstein, R. D. Hawkins, B. Ren, W.-Y. Chung, H. Gu, C. Bock, A. Gnirke, M. Q. Zhang, D. Haussler, J. R. Ecker, W. Li, P. J. Farnham, R. A. Waterland, A. Meissner, M. A. Marra, M. Hirst, A. Milosavljevic, J. F. Costello, Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.* **28**, 1097–1105 (2010).
32. A. Unnikrishnan, W. M. Freeman, J. Jackson, J. D. Wren, H. Porter, A. Richardson, The role of DNA methylation in epigenetics of aging. *Pharmacol. Ther.* **195**, 172–185 (2019).
33. K. E. Szulwach, X. Li, Y. Li, C.-X. Song, H. Wu, Q. Dai, H. Irier, A. K. Upadhyay, M. Gearing, A. I. Levey, A. Vasanthakumar, L. A. Godley, Q. Chang, X. Cheng, C. He, P. Jin, 5-hmC–mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nat. Neurosci.* **14**, 1607–1616 (2011).
34. H. Chen, S. Dzitoyeva, H. Manev, Effect of aging on 5-hydroxymethylcytosine in the mouse hippocampus. *Restor. Neurol. Neurosci.* **30**, 237–245 (2012).
35. N. Hadad, D. R. Masser, S. Logan, B. Wronowski, C. A. Mangold, N. Clark, L. Otalora, A. Unnikrishnan, M. M. Ford, C. B. Giles, J. D. Wren, A. Richardson, W. E. Sonntag, D. R. Stanford, W. Freeman, Absence of genomic hypomethylation or regulation of cytosine-modifying enzymes with aging in male and female mice. *Epigenetics Chromatin* **9**, 30 (2016).
36. R. Stark, G. Brown, DiffBind: Differential binding analysis of ChIP-Seq peak data, (2011); *http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf*.
37. H. Wu, C. Wang, Z. Wu, A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* **14**, 232–243 (2013).
38. C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, G. Bejerano, GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
39. R. Lister, E. A. Mukamel, J. R. Nery, M. Urich, C. A. Puddifoot, N. D. Johnson, J. Lucero, Y. Huang, A. J. Dwork, M. D. Schultz, M. Yu, J. Tonti-Filippini, H. Heyn, S. Hu, J. C. Wu, A. Rao, M. Esteller, C. He, F. G. Haghighi, T. J. Sejnowski, M. M. Behrens, J. R. Ecker, Global epigenomic reconfiguration during mammalian brain development. *Science* **341**, 1237905 (2013).
40. L. Wen, X. Li, L. Yan, Y. Tan, R. Li, Y. Zhao, Y. Wang, J. Xie, Y. Zhang, C. Song, M. Yu, X. Liu, P. Zhu, X. Li, Y. Hou, H. Guo, X. Wu, C. He, R. Li, F. Tang, J. Qiao, Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. *Genome Biol.* **15**, R49 (2014).
41. M. J. Booth, M. R. Branco, G. Ficz, D. Oxley, F. Krueger, W. Reik, S. Balasubramanian, Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934–937 (2012).
42. M. Mellén, P. Ayata, N. Heintz, 5-hydroxymethylcytosine accumulation in postmitotic neurons results in functional demethylation of expressed genes. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E7812–E7821 (2017).
43. X. Li, Y. Liu, T. Salz, K. D. Hansen, A. Feinberg, Whole-genome analysis of the methylome and hydroxymethylome in normal and malignant lung and liver. *Genome Res.* **26**, 1730–1741 (2016).
44. L. Wang, J. Zhang, J. Duan, X. Gao, W. Zhu, X. Lu, L. Yang, J. Zhang, G. Li, W. Ci, W. Li, Q. Zhou, N. Aluru, F. Tang, C. He, X. Huang, J. Liu, Programming and inheritance of parental DNA methylomes in mammals. *Cell* **157**, 979–991 (2014).
45. Q. Ma, H. Lu, Z. Xu, Y. Zhou, W. Ci, Mouse olfactory bulb methylome and hydroxymethylome maps reveal noncanonical active turnover of DNA methylation. *Epigenetics* **12**, 708–714 (2017).
46. A. Kozlenkov, J. Li, P. Apontes, Y. L. Hurd, W. M. Byne, E. V. Koonin, M. Wegner, E. A. Mukamel, S. Dracheva, A unique role for DNA (hydroxy)methylation in epigenetic regulation of human inhibitory neurons. *Sci. Adv.* **4**, eaau6190 (2018).
47. C. Angermueller, H. J. Lee, W. Reik, O. Stegle, DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* **18**, 67 (2017).
48. S. Spijker, Dissection of rodent brain regions, in *Neuroproteomics,* K.W. Li, Ed. (Springer, 2011), pp. 13–26.
49. D. Li, B. Zhang, X. Xing, T. Wang, Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation. *Methods* **72**, 29–40 (2015).
50. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler Transform. *Bioinformatics* **25**, 1754–1760 (2009).
51. F. Krueger, S. R. Andrews, Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
52. H. S. Jang, N. M. Shah, A. Y. Du, Z. Z. Dailey, E. C. Pehrsson, P. M. Godoy, D. Zhang, D. Li, X. Xing, S. Kim, D. O'Donnell, J. I. Gordon, T. Wang, Transposable elements drive widespread expression of oncogenes in human cancers. *Nat. Genet.* **51**, 611–617 (2019).
53. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
54. S. Anders, P. T. Pyl, W. Huber, HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

Acknowledgments: We thank all members in T.W.'s laboratory for suggestions on the manuscript and the tool. **Funding:** This work was supported by the NIH (R01HG007354, R01HG007175, R01ES024992, U01CA200060, U24ES026699, and U01HG009391) and the American Cancer Society Research Scholar Grant (RSG-14-049-01-DMC), and H.S.J. was supported by NIGMS (T32 GM007067). **Author contributions:** Y.H., H.S.J., and T.W. conceptualized and designed the study. Y.H. designed the DeepH&M algorithm and performed all computational analysis. H.S.J. and X.X. conducted the experiments. M.V. and J.D.D. dissected mouse tissues. D.L. made the DeepH&M website. Y.H., H.S.J., and T.W. wrote and revised the manuscript with input from all authors. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The datasets generated and analyzed in this study are available in the NCBI's Gene Expression Omnibus (GEO) repository: GSE140125. DeepH&M tool is available in https://epigenome.wustl.edu/DeepHM/. The code is available in https://github.com/hcharles14/DeepHM.git. All data needed to evaluate the conclusions i the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 31 October 2019
Accepted 18 May 2020
Published 1 July 2020
10.1126/sciadv.aba0521

Citation: Y. He, H. S. Jang, X. Xing, D. Li, M. J. Vasek, J. D. Dougherty, T. Wang, DeepH&M: Estimating single-CpG hydroxymethylation and methylation levels from enrichment and restriction enzyme sequencing methods. *Sci. Adv.* **6**, eaba0521 (2020).

Science Advances

DeepH&M: Estimating single-CpG hydroxymethylation and methylation levels from enrichment and restriction enzyme sequencing methods

Yu He, Hyo Sik Jang, Xiaoyun Xing, Daofeng Li, Michael J. Vasek, Joseph D. Dougherty and Ting Wang

Sci Adv **6** (27), eaba0521.
DOI: 10.1126/sciadv.aba0521

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.