


# TRAmHap: accurate prediction of transcriptional activity from DNA methylation haplotypes in bisulfite-sequencing data

Siqi Gao<sup>†</sup>, Hanwen Zhu<sup>†</sup>, Kangwen Cai, Lei Qin Liu, Zhiqiang Zhang, Yi Ding, Yaochen Xu, Xiaoqi Zheng and Jiantao Shi 

Corresponding authors. Jiantao Shi, State Key Laboratory of Molecular Biology, Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai 200031, China. E-mail: jtshi@sibcb.ac.cn; Xiaoqi Zheng, Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China. E-mail: xqzheng@shsmu.edu.cn

<sup>†</sup>Siqi Gao, Hanwen Zhu contributed equally to this work.

## Abstract

Deoxyribonucleic acid (DNA) methylation (DNAm) is an important epigenetic mechanism that plays a role in chromatin structure and transcriptional regulation. Elucidating the relationship between DNAm and gene expression is of great importance for understanding its role in transcriptional regulation. The conventional approach is to construct machine-learning-based methods to predict gene expression based on mean methylation signals in promoter regions. However, this type of strategy only explains about 25% of gene expression variation, and hence is inadequate in elucidating the relationship between DNAm and transcriptional activity. In addition, using mean methylation as input features neglects the heterogeneity of cell populations that can be reflected by DNAm haplotypes. We here developed TRAmHap, a novel deep-learning framework that predicts gene expression by utilizing the characteristics of DNAm haplotypes in proximal promoters and distal enhancers. Using benchmark data of human and mouse normal tissues, TRAmHap shows much higher accuracy than existing machine-learning based methods, by explaining 60–80% of gene expression variation across tissue types and disease conditions. Our model demonstrated that gene expression can be accurately predicted by DNAm patterns in promoters and long-range enhancers as far as 25 kb away from transcription start site, especially in the presence of intra-gene chromatin interactions.

**Keywords:** DNA methylation haplotypes, gene expression, enhancer, deep learning

## INTRODUCTION

Deoxyribonucleic acid (DNA) methylation (DNAm) is a fundamental epigenetic modification that plays a critical role in a wide range of biological processes, such as embryogenesis [1] and aging [2], through the mechanisms of transposable element repression, genetic imprinting and X chromosome inactivation [3]. DNAm refers to the addition of a methyl group to the carbon 5 position of the cytosine ring, resulting in the formation of 5-methylcytosine [4, 5], which typically occurs at 60–90% of CpG sites in the genome, except for CpG islands (CGIs) located in promoter regions that are usually unmethylated in normal cells [6]. In cancer, hypermethylation of CGIs in promoter regions can lead to the repression of tumor suppressor genes and the promotion of tumorigenesis [7, 8]. While the role of DNAm in regulating gene

expression is still a topic of ongoing research, it is known that DNAm can act not only as a repressor but also as an activator of gene expression by recruiting chromatin remodeling complexes through methyl-binding proteins [9, 10]. Motivated by these observations, a variety of quantitative models, including statistical and machine learning-based models, have been developed to predict gene expression using DNA methylation features [11, 12].

Although the relationship between DNA methylation and transcriptional activity has been widely investigated, the correlation between gene expression and DNA methylation in its promoter region is weak, with the Pearson's correlation coefficient around −0.3 [13]. This may be due to the uncharacterized role of DNA methylation in controlling gene expression or the heterogeneity effect, where mean methylation measured by array-based

**Siqi Gao** is a master student in Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Science. Her research interests include bioinformatics and epigenomics.

**Hanwen Zhu** is a master student in the Department of Mathematics, Shanghai Normal University. His research interests include biostatistics and deep learning.

**Lei Qin Liu** is a research assistant in Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Science. Her research focuses on epigenomics.

**Zhiqiang Zhang** is a research assistant in Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Science. His research focuses on computational epigenomics.

**Yi Ding** is a master student in Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Science. Her research interests include computational biology and epigenomics.

**Yaochen Xu** is a senior engineer in Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Science. His research focuses on computer science.

**Xiaoqi Zheng** is a professor in Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine. Her research interests include machine learning and computational epigenomics.

**Jiantao Shi** is a principal investigator in Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Science. His research focuses on computational biology and epigenomics.

Received: March 6, 2023. Revised: April 21, 2023. Accepted: May 18, 2023

© The Author(s) 2023. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

techniques represents an aggregated signal from a heterogeneous group of cells. On the contrary, sequencing-based techniques such as whole genome bisulfite sequencing (WGBS) and reduced represented bisulfite sequencing enable profiling DNA methylation patterns at single-nucleotide resolution. In sequencing-based experiments, a single read fragment is guaranteed to originate from a single chromosome and a single cell, and thus its methylation pattern represents a discrete DNA methylation haplotype (mHap) [14]. Based on bisulfite sequencing data, Landau et al. proposed a simple metric, i.e. the Proportion of Discordant Reads (PDR) to quantify within-sample heterogeneity in cancer [15]. Along this line, two other mHap-level summary statistics, i.e. Cell Heterogeneity-Adjusted cLonal Methylation (CHALM) [16] and methylation concurrence ratio (MCR) [17], were recently proposed. Predictive models using the above statistics are demonstrated to be more accurate for predicting gene expression than using mean methylation. Another important characteristic of DNA methylation is local correlation, i.e. in many regions, DNA methylations of nearby CpG sites tend to be highly correlated [18]. Taking this observation into consideration, two statistics, Methylation Haplotype Load (MHL) [19] and methylation block score [20] were developed to measure correlated methylations. In addition to mean methylation, the above DNA methylation summary statistics at haplotype-level enhanced our understanding on DNA methylation and its associations with gene transcriptional activity.

Besides different types of statistics to summarize DNA methylation patterns, the potential region that affects transcriptional activity is also a debatable topic. Initially, the DNAm analyses mainly focused on differentially methylated regions (DMRs) in promoters, but promoter-proximal (5 kb upstream to 1 kb downstream of transcription start site, TSS) DMRs only explain around 25% of gene expression variation when using simply a linear model [21–23]. Alternatively, higher-order features derived from mean methylation coupled with binomial probability regression achieved better results, but still only explained 25–49% of gene expression variation [11]. Recently, a deep learning framework that uses convolutional neural network (CNN) was shown to predict promoter activity landscapes from DNA methylomes in individual tumors [12]. This study included features of distal regions ( $\pm 25$  kb around the TSS) and focused on the prediction of histone modifications rather than gene expression [24].

In this study, we developed TRAmHap, a novel deep-learning framework that combines convolutional and recurrent neural networks to predict gene expression using features from both mean methylation and mHap-level summary statistics, in promoter-proximal ( $\pm 2.5$  kb around TSS) as well as distal regions ( $\pm 25$  kb around TSS). Our model significantly outperforms existing methods in terms of prediction accuracy and is capable of predicting gene expression across tissues and disease conditions. Our findings suggest that gene expression is not only influenced by DNAm at regions near TSS ( $\pm 2.5$  kb) but also by long-range enhancers, particularly in the presence of intra-gene chromatin interactions.

## MATERIALS AND METHODS

### Data processing

We utilized various public datasets to conduct our study. Normal tissue datasets were obtained from the ENCODE project consortium [25], and comprised of 15 human tissue samples, 7 mouse forebrain samples, 7 mouse heart samples and 5 mouse liver samples. These samples were profiled with both RNA-seq

and WGBS. In addition, we obtained WGBS and RNA-seq data of esophageal squamous cell carcinoma ( $n = 10$ ) and normal samples ( $n = 9$ ) from NCBI GEO under accession number GSE149612 [26] (Supplementary Table S1). Enhancer data specific to each tissue was downloaded from EnhancerAtlas 2.0 [27] and FANTOM5 [28]. Super-enhancers were downloaded from SEDb 2.0 [29] and SEA 3.0 [30]. The preprocessed ChIA-PET data were downloaded from GEO under accession number GSE90557 [31].

The adapters of bisulfite sequencing data were trimmed using Trim Galore [32] (version 0.6.2) with default parameters. The trimmed reads were mapped to the human genome version hg19 using BSMAP with the following parameters: ‘-q 20 -f 5 -r 0 -v 0.05 -s 16 -S 1’. In the case of paired-end sequencing, duplicates were masked with sambamba [33]. The methylation metrics for CpG sites were then extracted from the aligned reads using Methyl-Dackel [34].

### DNA methylation metrics

DNA methylation haplotypes were extracted from BAM files using mHapTools (version v1.1) [14]. Subsequently, the resulting mHap files were utilized as input to calculate various DNA methylation metrics, including mean methylation, PDR, CHALM, MHL and MCR. PDR and CHALM were computed based on reads that cover a minimum of four consecutive CpGs, while MHL and MCR considered all reads passing through a specific region. To get robust signal, missing values were assigned to regions with less than 10 reads. All DNA methylation metrics were calculated using mHapSuite (<https://github.com/yoyoong/mHapSuite>), which is a Java-based implementation of mHapTk [35].

### Architecture of TRAmHap

We developed a deep learning model, named TRAmHap, for predicting gene expression profiles from DNA methylation data. TRAmHap comprises two main components, a CNN module and a recurrent neural network (RNN) module. The input to the model is a 3-dimensional tensor, representing the genomic data divided into intervals and calculated features. The CNN module performs feature extraction using two parallel convolutional layers with different kernel sizes of  $1 \times 3$  and  $1 \times 5$ , respectively. The outputs from the two convolutional layers are concatenated to form a combined feature representation, which is then processed by the RNN module, designed to recognize patterns in the genomic data, using a series of LSTM layers. The output of the RNN module is then passed through a fully connected layer to generate the final prediction. TRAmHap was implemented using Pytorch (v1.9.0) in Python 3.8.6. We adopted a leave-one-out strategy to predict the gene expression in individual sample. The training data were split into training and validation sets with a ratio of 8:2. We used mean squared error (MSE) as the loss function optimized by the Adam optimizer with a learning rate of  $5e-4$ , a decay rate of 0.98, and  $(\beta_1, \beta_2) = (0.5, 0.998)$  during each training scheme. We trained the model for 20 epochs, and the best model selected by the validation set was saved as the result. Moreover, the model's backbone can be frozen and transferred between tissues using transfer learning to improve its performance, especially for predicting gene expression levels across tissues and disease conditions.

### DNA methylation-associated regulatory potential

To assess the influence of enhancers, we compared the predictive performance of two models: the full model (50 kb and 5 kb around TSS) and the reduced model (5 kb). We defined the prediction error for each gene as the Manhattan distance between the predicted

and actual values.

$$D_5 = |\text{Predicted value}_{M_{5K}} - \text{Observed value}|,$$

$$D_{5+50} = |\text{Predicted value}_{M_{5K+50K}} - \text{Observed value}|,$$

where  $D_{5+50}$  and  $D_5$  represent prediction errors for the full model and the reduced model, respectively. A lower value indicated better prediction. The impact of enhancers was estimated by the difference between  $D_5$  and  $D_{5+50}$ , i.e.

$$V_{\text{ari}} = D_{5+50} - D_5.$$

All genes were then sorted in ascending order according to their  $V_{\text{ari}}$ . The enhanced group includes genes showing better prediction performance when the 50 kb region was included in the model, while the reduced group includes genes showing worse performance. The non-variable group comprises genes with  $V_{\text{ari}}$  near 0, i.e. no significant difference between  $D_{5+50}$  and  $D_5$ .

## RESULTS AND DISCUSSION

### Selection of DNA methylation metrics as model inputs

Our study proposes to use DNA methylation patterns around the TSS to predict of gene expression. We focus on mHap-level metrics, which not only measure the mean methylation but also the DNA methylation patterns at the read-level [14]. It is worth noting that for regions with the same mean methylation, different DNA methylation patterns reflected by mHaps can exist [36]. For instance, a region with a mean methylation of 0.5 can exist with different patterns and can be distinguished by different mHap-level metrics (Figure 1A). Several mHap-level metrics, including PDR, CHALM, MCR and MHL (Supplementary Figure S1), have been shown to be associated with gene expression. However, the calculation of mHap-level metrics may not be feasible in some genomic regions due to their low read coverage in WGBS data. For example, PDR and CHALM only count sequencing reads that cover at least four consecutive CpG sites, which cover less than 50% of the regions in a typical WGBS dataset when the window size is set to 250 bp (Figure 1B). Even with a window size of 2.5 kb, 23% of the data is still missing. On the other hand, mean methylation, MHL and MCR all cover more than 85% of the regions, regardless of the window size used, and thus were selected as possible inputs to our model. Subsequently, we explored the association between these DNA methylation metrics and gene expression using a sample with matched DNA methylation and gene expression data from ENCODE (ENCBS366XOW). As expected, the group with the lowest gene expression shows the highest mean methylation around TSS (Figure 1C). This pattern was also observed in MHL (Figure 1D), and MCR was even more effective in distinguishing between the four groups (Figure 1E, Supplementary Figure S2). Therefore, mean methylation, MHL and MCR represent effective DNA methylation metrics and thus were selected as the model inputs.

### Framework of TRAMHap

We defined two regulatory regions, namely  $\pm 2.5$  and  $\pm 25$  kb upstream and downstream of the TSS, respectively, for each gene. The shorter 5 kb region covers the proximal promoter while the longer 50 kb region covers both the promoter and long-range enhancers. These regions were further divided into 20 windows of equal size. For each window, we calculated three mHap-level metrics, namely mean methylation, MHL and MCR, resulting in

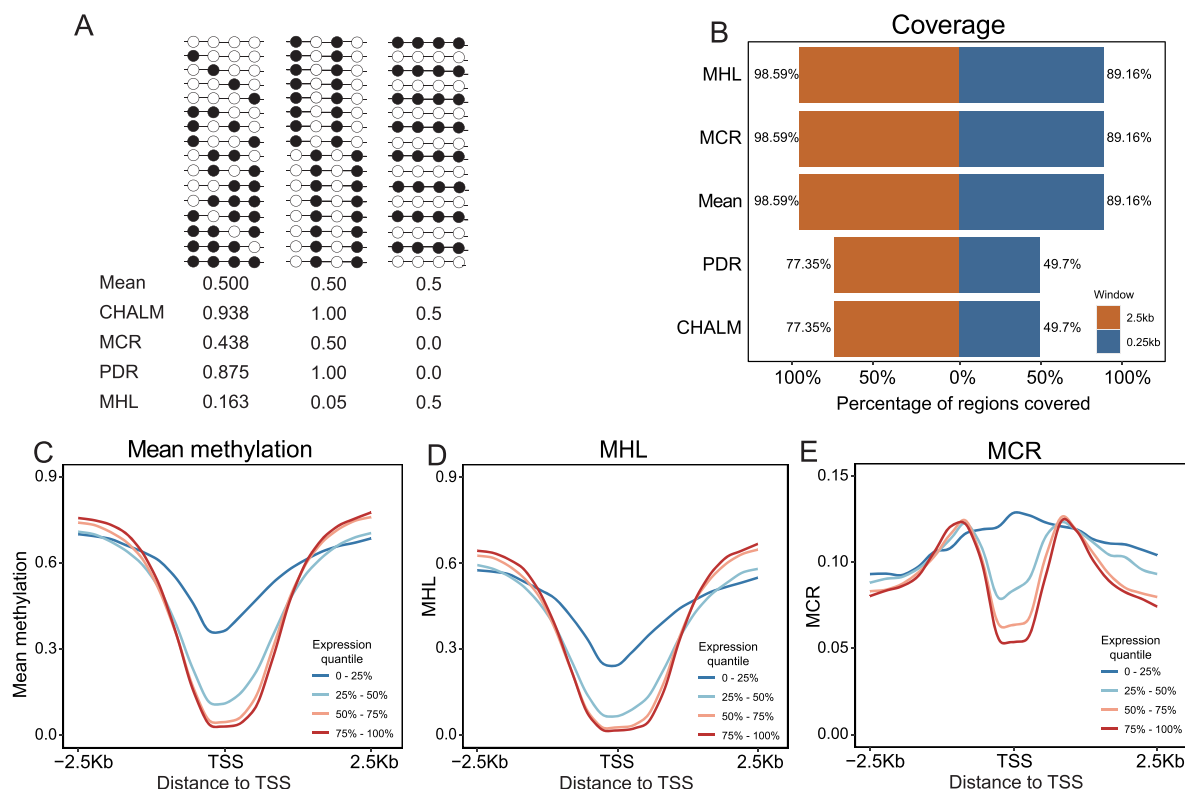
three matrices of equal dimensions of 20 by 2 (Figure 2A). Using these matrices as inputs, we developed TRAMHap, a novel deep-learning framework for predicting gene expression based on DNA methylation haplotypes (Figure 2B). TRAMHap can flexibly take one, two or all three metrics as input. The model begins with a 2-dimensional convolution layer that reshapes input matrix features into a 1-dimensional sequential array, which is then followed by two parallel convolution modules with kernel sizes of  $1 \times 5$  and  $1 \times 3$ , respectively. By using kernels of different sizes, the model can extract features at different scales. All convolution layers were linked to either a batch normalization layer or a max-pooling layer by the ReLU activation function. The convolution modules extract features based on the prior knowledge of locality and invariance. The output of the convolution layers is concatenated and fed into an LSTM module to extract features based on the sequence of windows. Finally, four fully connected layers are used to fit and predict gene expressions.

### TRAMHap accurately predicts transcriptional activity within the same tissue

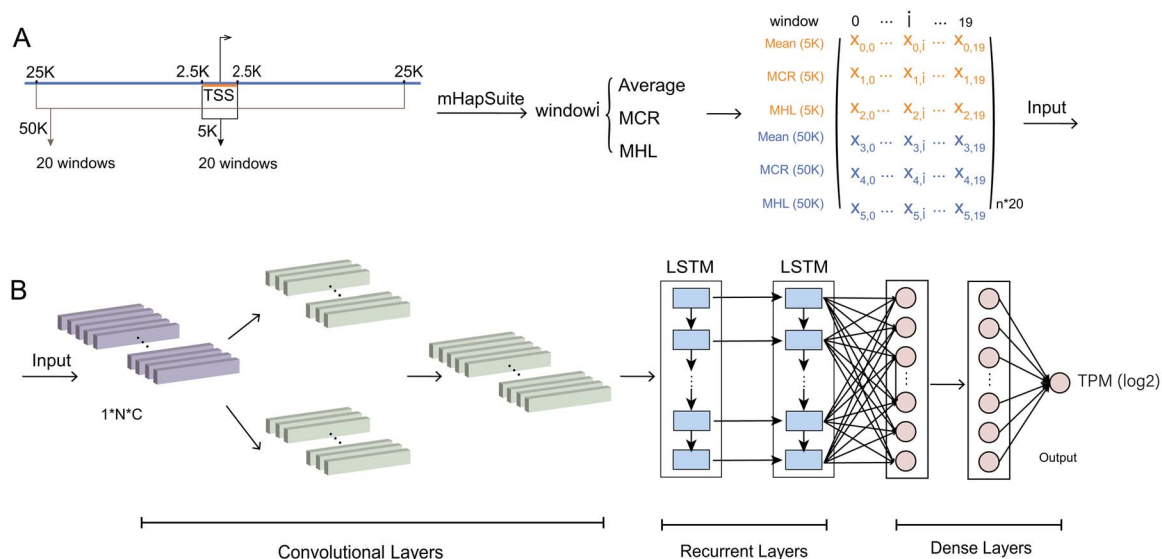
We assessed the performance of TRAMHap using various DNA methylation metrics as input. To achieve this, we curated a dataset with matched gene expression and DNA methylation profiles from three tissue types, including seven heart samples, seven forebrain samples and five liver samples (Supplementary Table S1). For each tissue type, we used a leave-one-out cross-validation scheme, where we cyclically chose one sample as the test set, and combined the remaining samples into training and validation sets at an 8:2 ratio. We assessed the model performance by calculating the Pearson's correlation coefficient (PCC) between measured and predicted gene expressions. Our results showed that TRAMHap models based on mean methylation and MCR outperformed those based on other metrics or all three metrics in all three mouse tissue types, with median PCC values of 0.91, 0.87 and 0.86 for heart, forebrain and liver tissues, respectively (Figure 3A).

We subsequently assessed the performance of TRAMHap when the genomic regions were partitioned into different numbers of windows. Larger window size results in fewer windows and a higher proportion of covered windows, while smaller window size produces more windows and a lower proportion of covered windows. For each gene, we required at least 90% of covered windows, or else that gene would be excluded from further analyses. In human normal tissue dataset, for instance, approximately 10 000 genes were retained when the regions were split into 20 windows, whereas only 200 genes were retained when the regions were split into 30 windows (Figure 3B, left panel). Regarding the overall prediction accuracy, the model with intervals partitioned into 20 windows outperformed that partitioned into 10 windows (paired Wilcoxon rank-sum test,  $P$ -value=0.018) and 30 windows (paired Wilcoxon rank-sum test,  $P$ -value=0.011) (Figure 3B, right panel).

In conclusion, these findings demonstrate that mean methylation and MCR are the optimal inputs for TRAMHap for predicting gene expression. As an example, for a mouse heart sample (ENCBS004ZLN), the PCC between the mean methylation in the promoter region and gene expression was  $-0.32$  ( $P < 2.2 \times 10^{-16}$ ), consistent with previous studies (Supplementary Figure S3, left panel). With the TRAMHap model, the PCC between the observed and predicted values was 0.94 (Figure 3C, left panel), indicating an accurate prediction of gene expression. A similar result was obtained with a brain tissue sample (ENCBS273RVL) (Figure 3C, right panel).



**Figure 1.** Selection of DNA methylation metrics as the model input. **(A)** For a region with the same mean methylation, three hypothetical sets of DNA methylation haplotypes are shown. For each set, DNA methylation is quantified by five metrics, including mean methylation, CHALM, MCR, PDR and MHL. **(B)** Genome-wide coverage of different DNA methylation metrics. For each gene, 10 windows upstream and 10 windows downstream of TSS were assessed with window sizes of 0.25 and 2.5 kb. The percentages of windows covered in 13 representative human WGBS samples were shown for 5 metrics. **(C-E)** Association of gene expression with DNA methylation metrics. For an ENCODE sample (ENCBS366XOW) with matched RNA-seq and WGBS data, genes were assigned to four equal-sized groups based on expression quantiles (0–25%, 25–50%, 50–75% and 75–100%). The average profiles of DNA methylation metrics, including mean methylation **(C)**, MHL **(D)**, MCR **(E)** are shown.



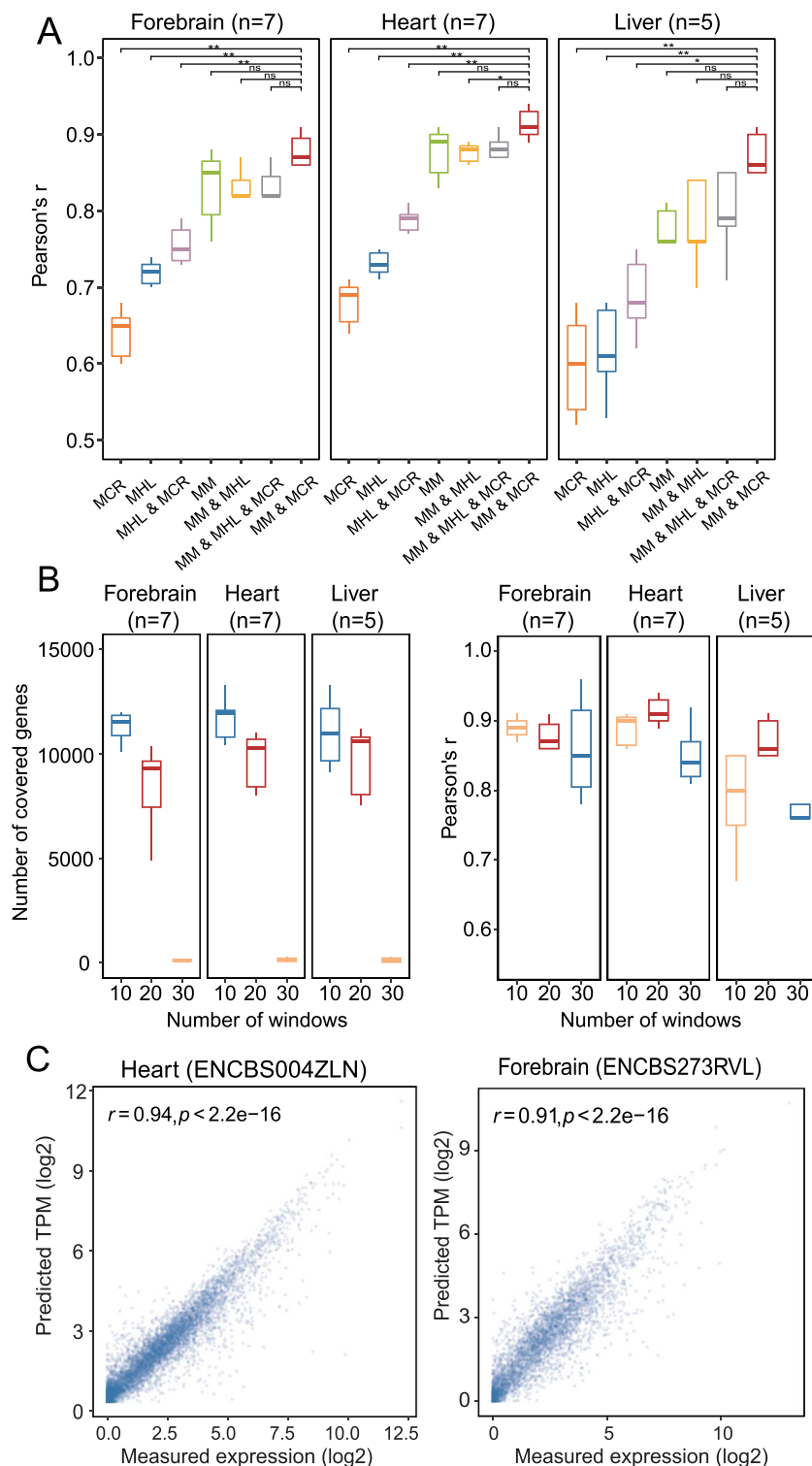
**Figure 2.** The architecture of TRAMHap. **(A)** Two regions, namely 5 and 50 kb around the transcription start site (TSS) of each gene, were divided into 20 non-overlapping windows of equal size, resulting in three sets of 2 by 20 matrices for the calculation of three DNA methylation metrics per window. These matrices are used either individually or combined as the model input. **(B)** The framework of TRAMHap is shown, including the pre-processing of DNA methylation data, feature extraction, model training, and prediction. Details can be found in Materials and methods.

## Comparison of TRAMHap with existing methods

We conducted a comparative analysis of TRAMHap against several existing methods for predicting gene expression from DNA methylation. The methods included linear regression (LR),

support vector regression (SVR), random forest (RF) and CNN. All above methods used the same datasets from three tissue types: seven heart tissue samples, seven forebrain tissue samples and five liver tissue samples. To ensure a fair comparison, all methods

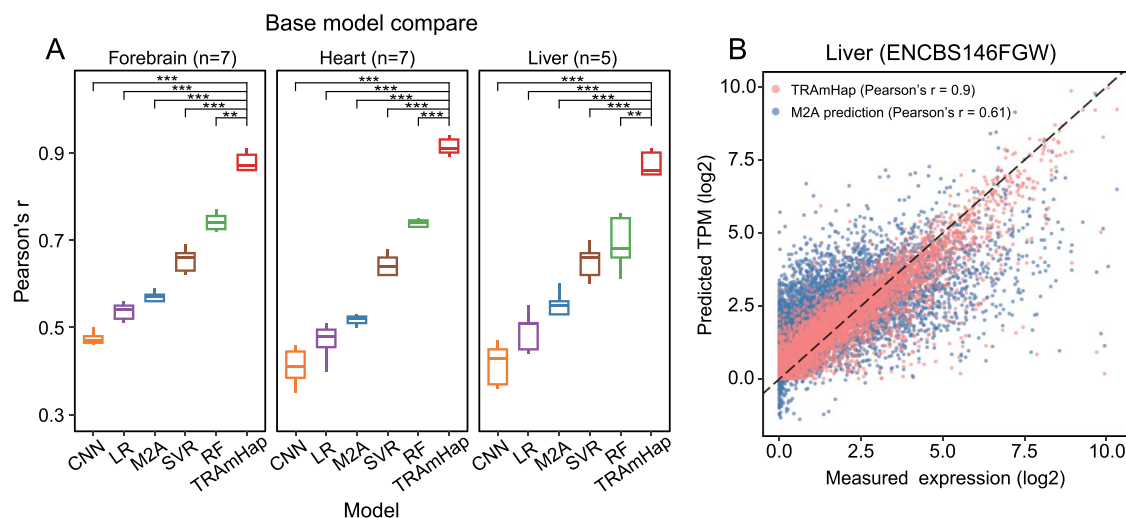




**Figure 3.** Refinement of model input and performance evaluation. **(A)** Selection the optimal input for TRAmHap. To select the optimal DNA methylation metrics as model input, the performance of TRAmHap was evaluated with different inputs, including mean methylation, MHL, MCR, combination of two metrics, and all three together. The performance was assessed by the Pearson's correlation coefficient between measured and predicted gene expressions. The significance between different groups were evaluated by Wilcoxon rank sum test  $*P < 0.05$ ;  $**P < 0.01$ ; ns not significant. **(B)** The left panel displays the number of covered genes when the regions surrounding TSS were divided into varying numbers of windows. The corresponding predictive performance is presented in the right panel. **(C)** Scatter plots of measured and predicted gene expression by the TRAmHap model are shown for two representative samples. Pearson's correlation coefficient  $r$  and associated  $P$ -value are shown for each sample.

were evaluated on the same training and validation sets. For SVR and RF, we conducted a grid search over a wide range of parameter combinations and chose the optimal parameters for modeling and comparison in subsequent results. Detailed information

about the grid search can be found in the supplementary tables (Supplementary Table S2). TRAmHap outperformed all other methods on all datasets, as measured by the PCC between predicted and true gene expression (Figure 4A). Specifically,



**Figure 4.** Comparison of TRAmHap with existing methods. **(A)** The performance of TRAmHap was evaluated using three mouse tissue datasets and compared with five existing models, including CNN, LR, M2A, SVR and RF. The performance was assessed by the Pearson's correlation coefficient between measured and predicted gene expressions. The significance between different groups were evaluated by Wilcoxon rank sum test \* $P < 0.05$ ; \*\* $P < 0.01$ ; ns not significant. **(B)** Scatter plot of measured and predicted gene expression values by TRAmHap and M2A, respectively, in a mouse liver sample (ENCBS146FGW). Pearson's correlation coefficient  $r$  and associated  $P$ -value were shown for each method.

TRAmHap achieved a median PCC of 0.85, explaining approximately 72% of gene expression variation through DNA methylation. In contrast, all other methods achieved median PCC values below 0.8, with classical CNN performing the worst with median PCC values of 0.47, 0.41 and 0.36 in mouse heart, brain and liver tissues, respectively. RF showed a more robust performance than LR, SVR and CNN, but only explained 50% of gene expression variation (median PCC  $\approx 0.74$ ).

In a recent study, the authors developed a deep-learning framework called MethylationToActivity (M2A) [12] that predicts promoter activity by measuring the enrichment of H3K4me3 and H3K27ac in the  $\pm 1$  kb region of the TSS (Supplementary Figure S4). We then tested M2A on samples used in this study by replacing the histone modification signal with gene expression as the output. However, M2A exhibited poor accuracy in predicting gene expression, with median PCC values of 0.57, 0.52 and 0.55 in mouse heart, brain and liver tissues, respectively. For example, for a mouse liver sample (ENCBS146FGW), TRAmHap achieved a PCC of 0.90, while M2A only achieved a PCC of 0.61 (Figure 4B). These findings suggest that TRAmHap outperforms both basic machine learning models as well as existing deep-learning models for predicting gene expression using DNA methylation data.

### TRAmHap is predictive of transcriptional activity across tissues and disease conditions

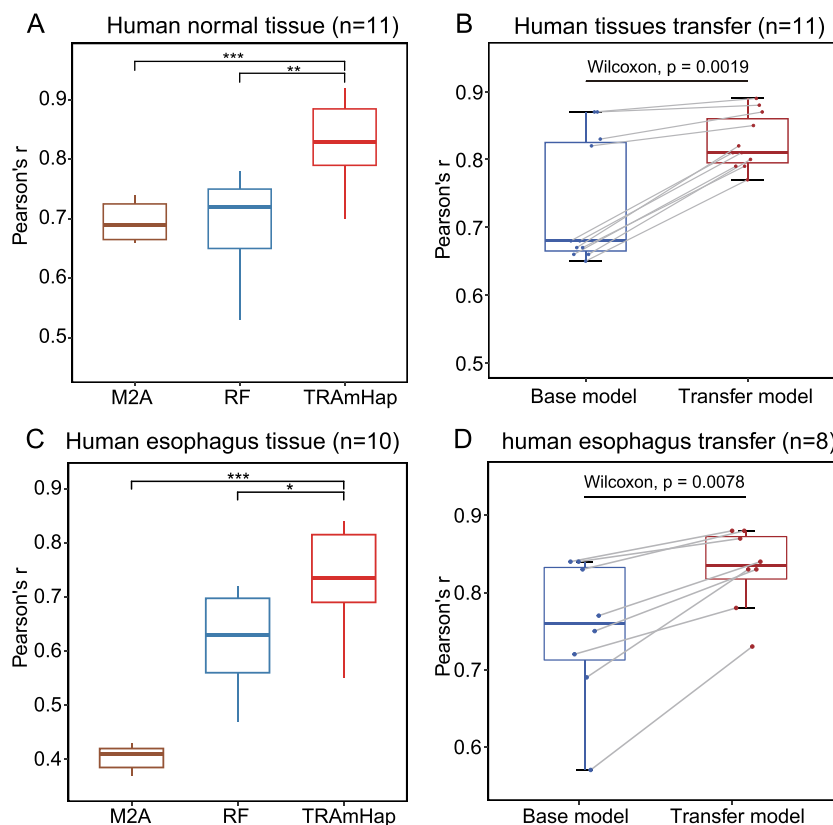
We further explored whether the proposed model maintains predictive power when training and testing samples come from different tissues. We curated a WGBS dataset covering 11 types of tissues with matched gene expression data (Supplementary Table S1). As an example, we tested our method on lung tissue samples using a model trained with all other tissues. Although the prediction accuracy for gene expression was slightly lower than that for within-tissue prediction, our model still achieved reasonably high accuracy (PCC = 0.89). Similar results are observed when other tissues are considered as test samples, with a median PCC of 0.83 (Figure 5A). When using the same training data and input features, TRAmHap significantly outperforms all other machine learning methods for the prediction of gene expression (Figure 5A,  $P < 0.01$ , Wilcoxon rank sum test). For example, while random

forest (RF) outperforms other existing methods, it only achieves a median PCC of 0.72, and the median PCC for the M2A model is 0.69. We also tested whether transfer learning by adding a small dataset that shares characteristics with the target dataset would improve the predictive performance of the model. We first trained a baseline model using samples from 10 different tissue types. Then, the model was optimized by fine-tuning with a small number of samples (1 or 2) from four additional tissue types, which included heart left ventricle, psoas muscle, esophagus and spleen. The fine-tuned model was then used to make predictions for these tissue types. The results showed that the transfer learning approach significantly improved the model's ability to predict gene expression across tissue types (paired Wilcoxon rank sum test,  $P$ -value = 0.0019) (Figure 5B).

Subsequently, we assessed the generalizability of our model by testing samples from the same tissue but in different disease states. We obtained a dataset from the GEO database (GSE149612) comprising normal ( $n = 9$ ) and cancer ( $n = 10$ ) tissue samples from the esophagus. Consistent with previous results, the model trained with normal samples demonstrated accurate prediction of gene expression in other normal samples, with a median PCC of 0.94. Similarly, the model trained with cancer samples achieved high accuracy in predicting gene expression in other cancer samples, with a median PCC of 0.88 (Supplementary Figure S5). Remarkably, the models trained from normal samples were capable of reasonably predicting tumor samples, despite different regulatory mechanisms involved, with a median PCC of 0.74. In comparison to RF and M2A ( $P < 0.01$ , Wilcoxon rank sum test), TRAmHap substantially outperformed them (Figure 5C). The performance of TRAmHap can be further improved by transfer learning and reach a median PCC of 0.80 (paired Wilcoxon rank sum test,  $P$ -value = 0.00078) (Figure 5D).

### Exploring DNA methylation-associated regulatory potential

The robustness and accuracy of TRAmHap enable us to explore DNAm-associated regulatory potential in regions surrounding TSSs. To measure the regulatory potential of a specific region, we perturbed the test data by replacing summary statistics within the



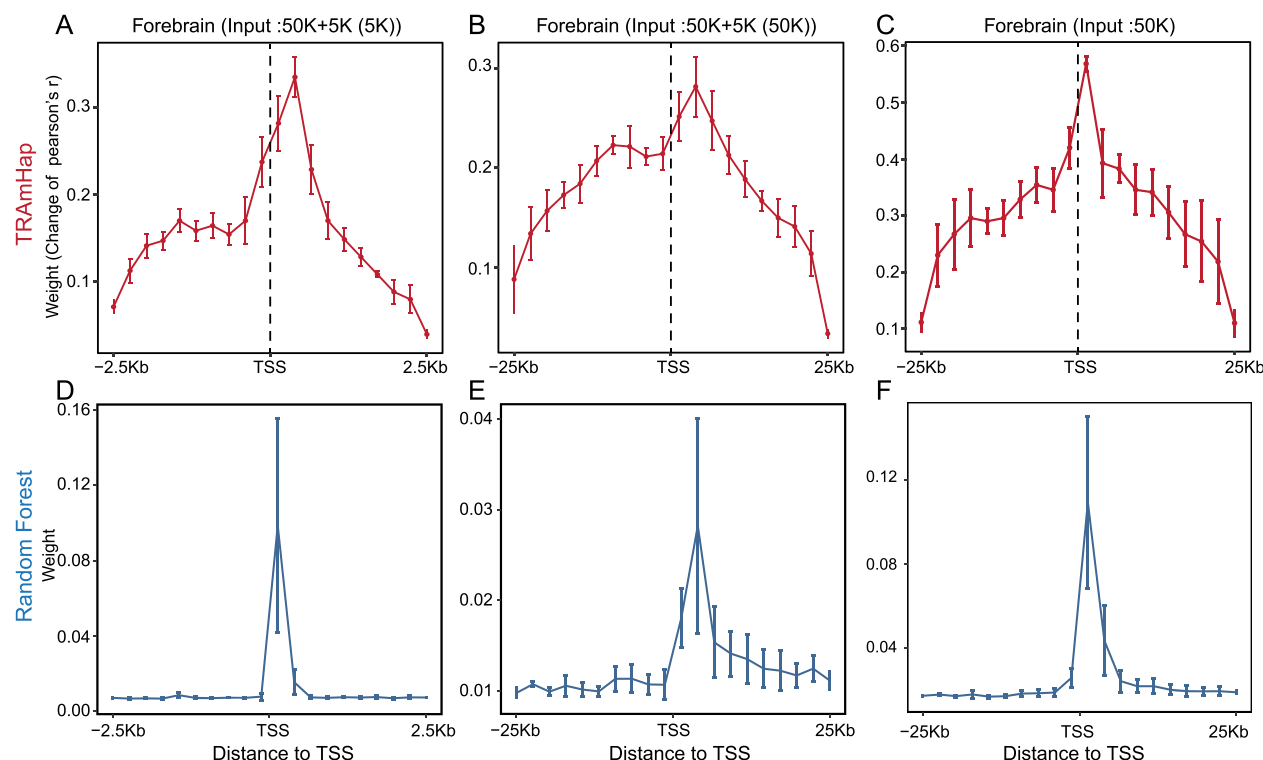
**Figure 5.** TRAmHap predicts transcriptional activity across tissue types and disease conditions. **(A)** Comparison of TRAmHap with M2A and RF for prediction gene expression across tissue types. We performed a leave-one-out cross-validation to ensure that the testing tissue was not included in the training data. **(B)** Transfer learning improves tissue prediction. We first trained a baseline model using samples from 10 different tissue types. Then, the model was optimized by fine-tuning with a small number of samples (1 or 2) from four additional tissue types, which included heart left ventricle, psoas muscle, esophagus and spleen. The optimized model was then used to make predictions for these tissue types. **(C)** Comparison of TRAmHap with M2A and RF for prediction gene expression across disease conditions. Esophagus tumor samples were tested with the model that was trained with nine esophagus normal samples. **(D)** Transfer learning improves across disease condition prediction. The baseline model was trained with all nine esophagus normal samples, which was further updated with two esophagus tumor samples, the resulting model was used to predict the other eight tumor samples. **(A–D)** The performance was assessed by the Pearson's correlation coefficient between measured and predicted gene expressions. The significance between different groups were evaluated by the Wilcoxon rank sum test. \* $P < 0.05$ ; \*\* $P < 0.01$ .

region with random values and evaluated the resulting change in overall predictive accuracy. Larger changes of predictive accuracy indicate higher regulatory potential of this region. We compared TRAmHap with RF, which is the best-performing existing method in mouse forebrain tissue samples ( $n = 7$ ). Our results show that, consistent with previous knowledge, the core regions around TSSs have the highest regulatory potential. Using 5 kb plus 50 kb as input, we found that masking the 250 bp window around TSSs reduces TRAmHap's PCC by 0.25 (Figure 6A). Notably, TRAmHap recovers regulatory potentials not only in proximal promoter regions but also in flanking regions as far as 25 kb away from TSSs (Figure 6B and C), whereas RF mainly relies on features in the exact window around TSSs, with regions outside 5 kb having minimal impact on overall prediction accuracy (Figure 6D and E). These findings are consistent with models using 50 kb as input (Figure 6F).

### Utilization of enhancer by TRAmHap

In our study of the TRAmHap model, we observed that the performance of the model with input intervals of both 50 and 5 kb was better than the model with reduced input intervals of only 5 kb, resulting in a significant improvement in the Pearson correlation coefficient (PCC) of the predicted effect within the same tissues. Specifically, the median PCC in mouse forebrain, liver

and heart tissues increased by 0.17, 0.16 and 0.16, respectively (Supplementary Figure S6). This improvement could be attributed to the hypothesis that larger regions contain enhancers that could improve the model's performance. To test this hypothesis, we compared the performance of our model with and without inputs of long range region (the TSS  $\pm 25$  kb). We ranked genes based on changes in prediction errors, where the top-ranked genes had enhanced prediction, bottom-ranked genes had reduced prediction, and genes in the middle had no significant change. In heart tissue, we found that 61% of the top 100 genes with enhanced prediction contained known heart-specific enhancers [27] (Figure 7A). In contrast, only 14% of the top 100 unaffected genes contained known enhancers. Interestingly, 50% of the top 100 genes with reduced prediction also contain known enhancers. Similar patterns are also observed in liver tissue (Figure 7B). In addition to the above mouse tissue specific enhancer data from the EnhancerAtlas 2.0 database, we downloaded the human enhancer data in the FANTOM database with the same pattern (Supplementary Figure S7A). We also downloaded super-Enhancer data from the SEDb 2.0 (Supplementary Figure S7B) and SEA (Supplementary Figure S7C) databases, which also had the same effect, but presenting less significant results than regular enhancers due to their sparsity. Even when 3000 genes were selected, these three groups were well separated,



**Figure 6.** DNA methylation-associated regulatory potential. For TRAMHap, to measure regulatory potential in a specific region, a set of perturbed testing data were generated by replacing summary statistics in the region with random values. The change in overall predictive accuracy compared to the original model was considered the regulatory potential of this region. For random forest, the weight of each region was directly used as regulatory potential. Two types of inputs were used, including 5 kb plus 50 kb (A and B, D and E) and 50 kb only (C and F).

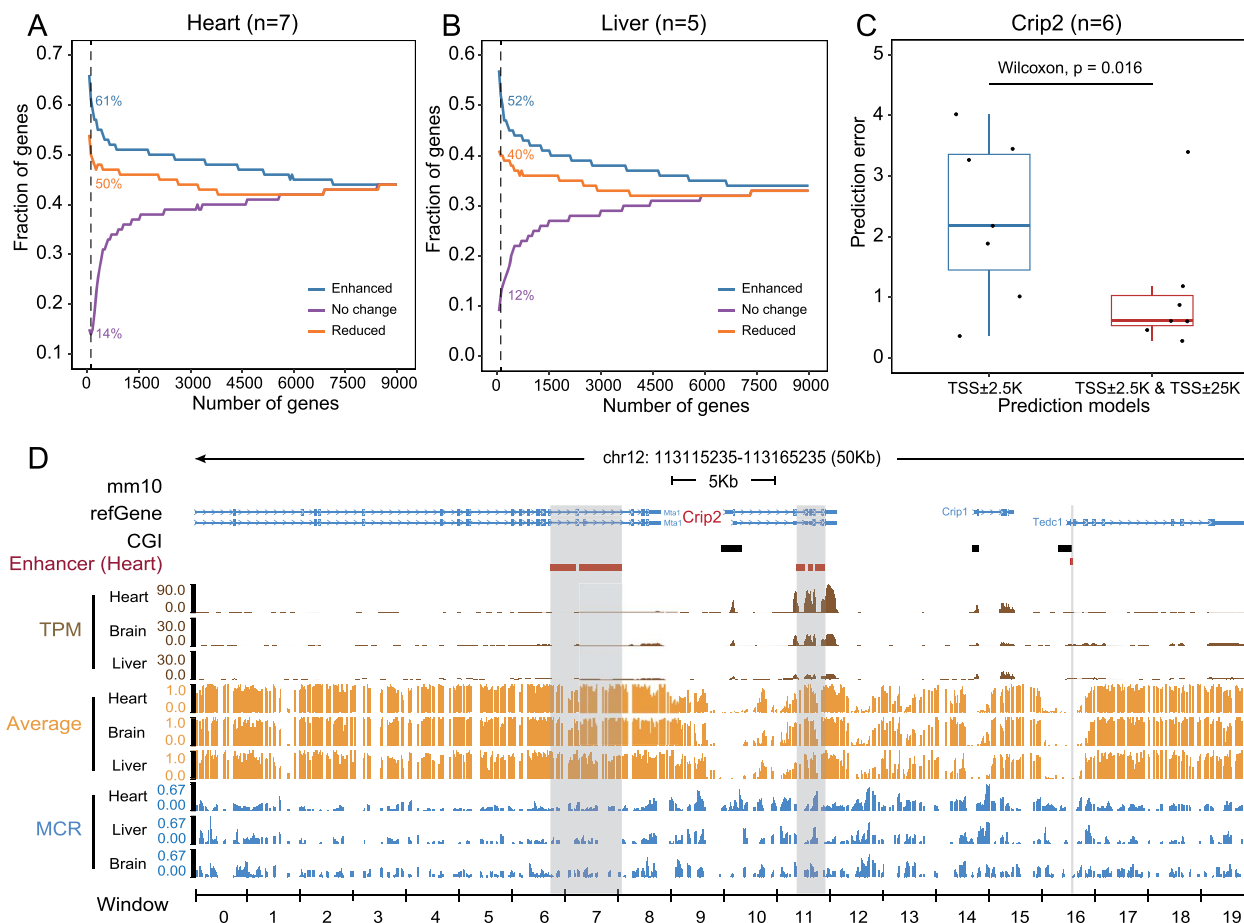
which is different from the pattern observed in RF (Supplementary Figure S8). These results demonstrate that TRAMHap can effectively capture DNAm patterns in enhancer regions and positively affect the prediction of gene expression.

As an illustrative example, we examined the prediction performance of TRAMHap on the tissue-specific gene *Crip2*, which is known to be highly expressed in mouse heart tissue. We found that the full TRAMHap model, which includes the 50-kb input region, achieved more accurate predictions of *Crip2* expression. Removing the 50-kb input region significantly increased the prediction error (paired Wilcoxon rank sum test,  $P$ -value=0.016), resulting in a mean prediction error of 2.21 for the six samples (Figure 7C). Previous studies have identified three clusters of cardiac enhancers located within the 50-kb region of the TSS of *Crip2*, all of which are located between 2.5 and 25 kb distance from the TSS locus. We speculate that the enhanced prediction performance of *Crip2* is likely due to the characteristic DNA methylation patterns in these three clustered regions (Figure 7D).

Based on the above results, we tried to add tissue-specific enhancer information to the model inputs, i.e. mean methylation, MCR and enhancer information for each window ('1' means that the window contains enhancers for the gene, '0' means that the window does not contain enhancers for the gene), and the results showed that the prediction of the model was slightly improved by adding this information (Supplementary Figure S9). This result also shows again that considering enhancer information in the prediction of gene expression is beneficial to improve the accuracy of the model. However, we observed that genes with reduced prediction also contain a small fraction of enhancers, but the prediction error increased when the 50-kb regions were added to the model. Enhancers are known to be located far from

promoters and regulate gene expression through long-range chromatin interactions. One possible explanation for the reduced prediction in these cases is that some enhancers may be located within the 50-kb regions of associated genes but regulate genes outside these loci (Supplementary Figure S10A). To test this hypothesis, we compared the fractions of intra- and inter-gene links, defined by chromatin interaction analysis by paired-end-tag sequencing (ChIA-PET) [31], in gene groups with enhanced, reduced, or non-variable prediction. The group with enhanced prediction tends to contain higher fraction of intra-gene links and lower fraction of inter-gene links. For mouse brain tissue, in the top 100 genes with enhanced prediction, 87% of them contain intra-gene links, and 13% of them only contain inter-gene links. In contrast, of the top 100 genes with reduced prediction, 80% of them contain intra-gene links, and 20% of them only contain inter-gene links (Figures 8A and B). For example, *Xylb* exhibited reduced prediction accuracy when the 50 kb region was included in the model (paired Wilcoxon rank sum test,  $P$ -value=0.016) (Supplementary Figure S10B). This could potentially be attributed to the absence of intra-locus loops in *Xylb* (Supplementary Figure S10C and D). Some genes contain both intra- and inter-gene links, but genes with reduced prediction tend to be dominated by inter-gene links (Figure 8C). For instance, *Hist1h4h* showed a significant increase in prediction error when the 50 kb regions were added to the model (paired rank sum test,  $P$ =0.047) (Figure 8D). Within this locus, there are 201 chromatin contact loops, with inter-gene links dominating. Specifically, out of the contact loops involving the 50 kb region but not the 5 kb region, 79% are inter-locus links (124 inter-locus loops, 32 intra-locus loops) (Figure 8E and F). However, the RF model did not show such a large difference in this case (Supplementary Figure S11).





**Figure 7.** Impact of enhancers on prediction accuracy. The performances of models with or without 50 kb input were compared. Genes were ranked based on the changes in prediction errors, so that genes ranked at the top have enhanced prediction, genes ranked at the bottom have reduced prediction, and genes in the middle were not affected. For each of the three groups, the fraction of genes with tissue-specific enhancers was shown when the different numbers of top genes were selected in datasets of the heart (A) and liver (B) tissues. When selecting the top 100 genes, the fraction of genes with enhancers was specifically shown. (C) The prediction errors of the Crip2 gene with or without 50 kb regions were compared and statistical significance was assessed by paired Wilcoxon rank sum test. (D) An IGV plot of the Crip2 gene locus. Gene expression (TPM), mean methylation and MCR are shown. Three clusters of heart tissue enhancers were also indicated.

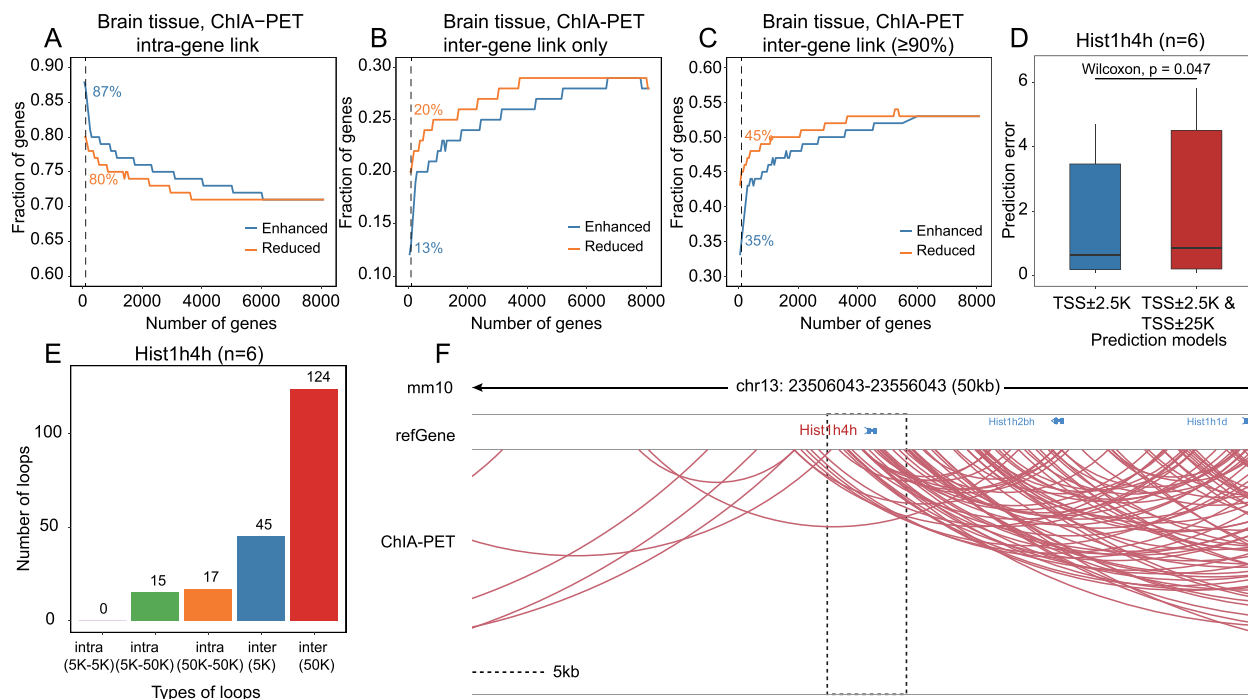
## DISCUSSION

Modeling the relationship between DNA methylation and gene expression is of great importance to understanding its role in transcriptional regulation. However, current strategies that employ linear models to characterize the mean methylation in promoter regions and gene expression are oversimplified. In this work, we addressed this challenge from three aspects. First, we recognized that DNA methylation patterns can differ in regions with similar mean methylation levels, and mHap-level summary metrics can be used to better characterize these patterns. Second, DNA methylation in enhancer regions is known to be associated with promoter activity and gene expression, making a model that encompasses both promoters and enhancers more suitable. At last, classical machine learning methods, such as linear regression, are not always effective in analyzing high-dimensional data.

Based on observations described above, we developed a novel deep-learning framework called TRAmHap, which predicts transcriptional activity by utilizing the characteristics of DNA methylation haplotypes in proximal promoters and enhancers located up to 25 kb away from TSS. Depending on the DNA methylation metrics, a variety of models is available within the framework. While traditional mean methylation provides an aggregated signal and ignores cell population heterogeneity, DNA methylation

metrics such as PDR, CHALM, MCR and MHL account for the heterogeneity and capture the patterns of DNA methylation haplotypes. However, PDR and CHALM computations only consider sequencing reads covering at least four CpG sites, resulting in over 50% of missing values in 5 kb regions around TSS for typical 30× WGBS samples. Thus, we chose mean methylation, MCR and MHL as the model inputs. In addition, the number of windows is another critical parameter to consider. A larger window size results in fewer windows and a higher proportion of covered windows, while a smaller window size produces more windows and a lower proportion of covered windows. We examined the performance of TRAmHap using various window numbers and found that a window size of 20, covering about 9000 genes, yielded the best results. Although a window size of 10 covered over 10 000 genes, the performance of TRAmHap decreased significantly. Thus, a window number of 20 balances performance and coverage.

TRAmHap outperforms existing machine-learning methods, such as LR, SVR, RF, classical CNN and M2A. TRAmHap accurately predicts transcriptional activity and explains 60–80% of the gene expression variation, a significant improvement over traditional models such as LR that only explains around 25% of gene expression variation by using mean methylation in promoter regions. Although RF can utilize the signal from



**Figure 8.** Impact of chromatin interaction on prediction accuracy. Two groups of genes were defined based on changes in prediction errors from models with or without 50 kb regions. The fraction of genes with intra-gene links (A), with only inter-gene links (B), or with 90% inter-gene links (C) are shown. (D) The prediction errors of the Hist1h4h gene with or without 50 kb regions were compared and statistical significance was assessed by paired Wilcoxon rank sum test. (E) In the Hist1h4h locus, the number of intra- and inter-gene ChIA-PET loops was shown. (F) An IGV plot of ChIA-PET links in Hist1h4h locus.

long-range regions, TRAMHap employs it more efficiently, as demonstrated by the stronger enrichment of tissue-specific enhancers and intra-gene chromatin loops in genes with enhanced prediction when 50 kb regions are included. The TRAMHap demonstrates strong predictive performance at the single-sample level, even with limited number of samples. To predict gene expression of individual genes across samples would require a significantly larger sample size, particularly when matched DNA methylation and gene expression data are considered. To improve the accuracy of individual gene expression predictions, collecting more datasets or incorporating additional gene-related features may be beneficial.

The TRAMHap model based on mean methylation and MCR outperformed other models based on DNA methylation profile indicators in terms of intra-tissue prediction. This result suggests that the characteristics of DNA methylation haplotypes contribute to the prediction of transcriptional activity. We believe that the TRAMHap framework could be further refined to include additional layers, such as those with attentional mechanisms, to improve noise tolerance.

## CONCLUSIONS

In this study, we developed a new machine-learning framework to predict gene expression using the features of DNA methylation in promoters as well as enhancers. This model outperforms existing models and can predict gene expression in different tissues and disease conditions. Our model shows that DNA methylation features can be used to accurately predict gene expression. Also, we found that gene expression is determined not only by DNAm in the region near the TSS ( $\pm 2.5$  kb) but also by long-range enhancers, especially when intra-gene chromatin interactions are present.

### Key Points

- TRAMHap is a novel deep-learning framework that predicts gene expression by utilizing the characteristics of DNA methylation haplotypes in proximal promoters and distal enhancers.
- TRAMHap shows much higher accuracy than existing machine-learning based methods, by explaining 60%~80% of gene expression variation across tissue types and disease conditions.
- Gene expression is determined not only by DNAm in the region near the transcription start site ( $\pm 2.5$  kb) but also by long-range regions that are enriched with enhancers.
- TRAMHap can also incorporate tissue-specific enhancer information in addition to DNA methylation scores, resulting in improved prediction accuracy.

## CODE AVAILABILITY

The code and documentation for TRAMHap are freely available at <https://github.com/SQ-Gao/TRAMHap>.

## SUPPLEMENTARY DATA

Supplementary data are available online at <https://academic.oup.com/bib>.

## FUNDING

Hundred Talents Program Award of the Chinese Academy of Sciences (to J.S.); National Natural Science Foundation of China (Grant IDs 32270691 and 61972257).

## REFERENCES

1. Skvortsova K, Stirzaker C, Taberlay P. The DNA methylation landscape in cancer. *Essays Biochem* 2019;**63**:797–811.
2. Unnikrishnan A, Freeman WM, Jackson J, et al. The role of DNA methylation in epigenetics of aging. *Pharmacol Ther* 2019;**195**:172–85.
3. Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev* 2002;**16**:6–21.
4. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. *Nat Rev Genet* 2013;**14**:204–20.
5. Li E, Zhang Y. DNA methylation in mammals. *Cold Spring Harb Perspect Biol* 2014;**6**:a019133.
6. Lister R, Pelizzola M, Dowen RH, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009;**462**:315–22.
7. Ehrlich M. DNA methylation in cancer: too much, but also too little. *Oncogene* 2002;**21**:5400–13.
8. Keshet I, Schlesinger Y, Farkash S, et al. Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nat Genet* 2006;**38**:149–53.
9. Moore LD, Le T, Fan G. DNA methylation and its basic function. *Neuropsychopharmacology* 2013;**38**:23–38.
10. Du Q, Luu PL, Stirzaker C, Clark SJ. Methyl-CpG-binding domain proteins: readers of the epigenome. *Epigenomics* 2015;**7**:1051–73.
11. Kapourani CA, Sanguinetti G. Higher order methylation features for clustering and prediction in epigenomic studies. *Bioinformatics* 2016;**32**:i405–12.
12. Williams J, Xu B, Putnam D, et al. MethylationToActivity: a deep-learning framework that reveals promoter activity landscapes from DNA methylomes in individual tumors. *Genome Biol* 2021;**22**:24.
13. Smith ZD, Shi J, Gu H, et al. Epigenetic restriction of extraembryonic lineages mirrors the somatic transition to cancer. *Nature* 2017;**549**:543–7.
14. Zhang Z, Dan Y, Xu Y, et al. The DNA methylation haplotype (mHap) format and mHapTools. *Bioinformatics* 2021;**37**:4892–4.
15. Landau DA, Clement K, Ziller MJ, et al. Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell* 2014;**26**:813–25.
16. Xu J, Shi J, Cui X, et al. Cellular heterogeneity-adjusted clonal methylation (CHALM) improves prediction of gene expression. *Nat Commun* 2021;**12**:400.
17. Shi J, Xu J, Chen YE, et al. The concurrence of DNA methylation and demethylation is associated with transcription regulation. *Nat Commun* 2021;**12**:5285.
18. Jenkinson G, Pujadas E, Goutsias J, Feinberg AP. Potential energy landscapes identify the information-theoretic nature of the epigenome. *Nat Genet* 2017;**49**:719–29.
19. Guo S, Diep D, Plongthongkum N, et al. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat Genet* 2017;**49**:635–42.
20. Liang N, Li B, Jia Z, et al. Ultrasensitive detection of circulating tumour DNA via deep methylation sequencing aided by machine learning. *Nat Biomed Eng* 2021;**5**:586–99.
21. Bock C, Beerman I, Lien WH, et al. DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. *Mol Cell* 2012;**47**:633–47.
22. Strahl BD, Allis CD. The language of covalent histone modifications. *Nature* 2000;**403**:41–5.
23. Musselman CA, Lalonde ME, Cote J, Kutateladze TG. Perceiving the epigenetic landscape through histone readers. *Nat Struct Mol Biol* 2012;**19**:1218–27.
24. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* 2014;**15**:272–86.
25. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74.
26. Cao W, Lee H, Wu W, et al. Multi-faceted epigenetic dysregulation of gene expression promotes esophageal squamous cell carcinoma. *Nat Commun* 2020;**11**:3675.
27. Gao T, He B, Liu S, et al. EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. *Bioinformatics* 2016;**32**:3543–51.
28. Lizio M, Harshbarger J, Shimoji H, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol* 2015;**16**:22.
29. Wang Y, Song C, Zhao J, et al. SEDb 2.0: a comprehensive super-enhancer database of human and mouse. *Nucleic Acids Res* 2023;**51**:D280–90.
30. Chen C, Zhou D, Gu Y, et al. SEA version 3.0: a comprehensive extension and update of the super-enhancer archive. *Nucleic Acids Res* 2020;**48**:D198–203.
31. Bertolini JA, Favaro R, Zhu Y, et al. Mapping the global chromatin connectivity network for Sox2 function in neural stem cell maintenance. *Cell Stem Cell* 2019;**24**:462, e466–76.
32. F K: Trim Galore. <https://github.com/FelixKrueger/TrimGalore> (last accessed, 19th November 2019) 2012.
33. Tarasov A, Vilella AJ, Cuppen E, et al. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* 2015;**31**:2032–4.
34. Ryan D: Methyl Dackel. 2017. <https://github.com/dpryan79/MethylDackel> (last accessed, 5th April 2021).
35. Ding Y, Cai K, Liu L, et al. mHapTk: a comprehensive toolkit for the analysis of DNA methylation haplotypes. *Bioinformatics* 2022;**38**:5141–3.
36. Scherer M, Nebel A, Franke A, et al. Quantitative comparison of within-sample heterogeneity scores for DNA methylation data. *Nucleic Acids Res* 2020;**48**:e46.