OXFORD

Briefings in Bioinformatics, 2025, **26(2)**, bbaf092

https://doi.org/10.1093/bib/bbaf092

Problem Solving Protocol

Predicting differentially methylated cytosines in TET and DNMT3 knockout mutants via a large language model

Saleh Sereshki and Stefano Lonardi*

Department of Computer Science and Engineering, University of California, Riverside, 900 University Ave, Riverside, CA 92521, United States

*Corresponding author. Department of Computer Science and Engineering, University of California, Riverside, 900 University Ave, Riverside, CA 92521, United States. E-mail: stelo@cs.ucr.edu

Abstract

DNA methylation is an epigenetic marker that directly or indirectly regulates several critical cellular processes. While cytosines in mammalian genomes generally maintain stable methylation patterns over time, other cytosines that belong to specific regulatory regions, such as promoters and enhancers, can exhibit dynamic changes. These changes in methylation are driven by a complex cellular machinery, in which the enzymes DNMT3 and TET play key roles. The objective of this study is to design a machine learning model capable of accurately predicting which cytosines have a fluctuating methylation level [hereafter called *differentially methylated cytosines* (DMCs)] from the surrounding DNA sequence. Here, we introduce L-MAP, a transformer-based large language model that is trained on DNMT3-knockout and TET-knockout data in human and mouse embryonic stem cells. Our extensive experimental results demonstrate the high accuracy of L-MAP in predicting DMCs. Our experiments also explore whether a classifier trained on human knockout data could predict DMCs in the mouse genome (and vice versa), and whether a classifier trained on DNMT3 knockout data could predict DMCs in TET knockouts (and vice versa). L-MAP enables the identification of sequence motifs associated with the enzymatic activity of DNMT3 and TET, which include known motifs but also novel binding sites that could provide new insights into DNA methylation in stem cells. L-MAP is available at https://github.com/ucrbioinfo/dmc_prediction.

Keywords: DNA methylation; cytosine methylation; TET; DNMT3; large language model; BERT

Introduction

DNA methylation is an epigenetic marker that directly or indirectly regulates several critical cellular processes, including gene expression, genome stability, transposon suppression, and gene imprinting (see, e.g. [1–3]). The most common form of DNA methylation, known as 5-methylcytosine (5mC), involves the attachment of a methyl group to the fifth carbon of a cytosine residue. Abnormal methylation patterns in humans have been associated with diseases, including cancer and imprinting syndromes (see, e.g. [4, 5]).

In mammals, DNA methylation primarily occurs in CpG dinucleotides, with most of them being methylated [6]. Mammalian genomes typically maintain consistent CpG methylation levels over time, except in specific regulatory regions like promoters and certain types of enhancers [7]. In these variable regions, the dynamics of methylation and demethylation are orchestrated by a complex cellular machinery, in which the enzymes DNMT3 (A/B) and TET play a major role. DNMT3A and DNMT3B are DNA methyltransferases that can add a new methyl group to cytosines, e.g. during development and cellular differentiation [8, 9]. TET is an enzyme that catalyzes the conversion of 5-methylcytosine into 5-hydroxymethylcytosine and its oxidized derivatives. The conversion of 5-hydroxymethylcytosine and its derivatives ultimately leads to active DNA demethylation [10].

Knockout experiments that disrupt DNMT3 and TET have allowed life scientists to unravel the complex dynamics of DNA methylation changes over time and space, and across cell types. During the pluripotent stages, DNMT3 and TET modulate the epigenetic landscape, thus influencing cellular differentiation [11, 12]. During postfertilization reprogramming, the embryo undergoes a two-phase process in which it loses gamete-specific DNA methylation patterns inherited from the oocyte and sperm, with the initial active demethylation of the paternal genome by TET3 followed by subsequent passive dilution of DNA methylation during cell divisions [13].

TET and DNMT3 are crucial in regulating fetal organ development and tissue generation, through DNA methylation and histone modifications [14, 15]. Their dysregulation is linked to human diseases, particularly cancers [16, 17]. Although the importance of TET is well recognized, its precise mechanisms of action are not well understood. Several studies have shown that DNMT3 and TET, both individually and in combination, influence DNA methylation patterns in human embryonic stem cell lines (e.g. [12]). Chao *et al.* [18] also studied the interactions between TET1, DNMT3A, and DNMT3B in human embryonic stem cells, and how these interactions collectively influence global methylation patterns.

Given the importance of DNMT3 and TET in developmental biology and embryogenesis, there is strong interest in

Received: August 16, 2024. **Revised:** February 3, 2025. **Accepted:** February 18, 2025

characterizing which cytosines are affected by these two classes of enzymes. A few recent studies attempted to capture the sequence preference for DNMT3 and TET. For instance, in [19] the authors showed that TET has a sequence preference for CG dinucleotides within specific transcription-factor binding sites, indicating that its activity in catalyzing DNA demethylation is influenced by the underlying sequence context. The study in [18] also reported that TET1 prefers binding to specific genomic regions. This appears to be also true for DNMT3. Jeltsch *et al.* [20] demonstrated that the enzymatic activities of DNMT3A and DNMT3B are influenced by the sequence context of their target sites. As part of these studies, several DNMT3 and TET knockout methylation data sets for human and mouse have been produced (see, e.g. [12, 21, 22]). These data sets open the possibility to investigate whether one could predict which specific cytosines are affected by DNMT3 and TET using a machine learning model.

Here, we explore for the first time the problem of predicting differentially methylated cytosines (DMCs) in TET and DNMT3 knockout mutants, using exclusively the underlying DNA sequence around the cytosines. Our classifier, called Language model-based Methyltransferases Activity Predictor (L-MAP), is a transformer-based large language model (LLM) that utilizes contextual sequence information to predict the enzymatic activity of DNMT3 and TET on cytosines.

We envision L-MAP as a tool to impute missing or uncertain DMCs from wet lab experiments, which are costly and often leave some cytosines with insufficient coverage or uncovered. In this study, we also investigate (1) whether training L-MAP on DNMT3 knockouts can be used to predict TET activities, and vice versa; (2) whether training L-MAP on human knockout data can be used to predict the enzymatic activity on mice, and vice versa; (3) whether the methylation levels of nearby cytosines can help L-MAP predict DMCs with higher accuracy; and (4) whether L-MAP has learned sequence motifs known to be associated with the enzymatic activity of DNMT3 and TET enzymes.

A deeper understanding of cell functions can lead to significant advancements in medical research, therapeutic development, disease prevention, and diagnostic techniques [23, 24]. Some studies have identified interacting partners for TETs and DNMT3s [25, 26]. Here, we have identified transcription factor binding site motifs that may be linked to TET and DNMT3 activity in pluripotent cells. These findings can open new avenues for understanding the functions of these methyltransferases and lead to advancements in treatment strategies and novel drug discoveries.

Results

We analyzed seven knockout data sets: four human data sets (DNMT3KO, TETKO, QKO, and PKO ESC lines) and three murine data sets [DNMT3A and DNMT3B knockouts in embryonic stem cells, and TET2 and TET3 knockouts in intestinal stem cells (ISCs)]. Correspondingly, we had three wild-type datasets, namely human ESC, mouse ESC, and mouse small intestine ESC. Additional details about these data sets are reported in the Methods section. Supplemental Figure 2 reports the genome-wide methylation levels for the three wild-type and seven knockout data sets. We recall that methylation levels are expressed by a real number in the interval [0, 1], where 0 indicates that none of the cells in the sample are methylated, and 1 indicates that all the cells in the sample are methylated. The data show that the average methylation level is in the range 0.7–0.8 for all data sets, except for the DNMT3A knockout data set on the mouse ESCs.

The methylation levels for the seven knockout and three wild-type datasets were used to determine seven sets of DMCs. A cytosine was determined to be differentially methylated when its methylation level for the knockout was significantly higher or lower than its methylation level in the wild-type (details in Methods). Supplemental Figure 1 reports the number of DMCs on the seven datasets. The number of DMCs ranges from about 100 000 in the DNMT3B knockout dataset for mouse ESC, to about 1.5 million in the DNMT3A knockout for mouse ESC. Based on this, we chose a sample size of 100 000 cytosines for each dataset, half of which were differentially methylated (and the other half were not). The sample included 100 000 512 bp-long DNA sequences centered around the chosen cytosines, along with the corresponding binary label (1 indicated a DMC, 0 otherwise). We evaluated the impact of the size of the training dataset on L-MAP's accuracy in Supplemental Figure 7. The figure shows that the accuracy improves up to a sample size of 100 000. Further increases in the sample size do not significantly improve L-MAP's accuracy. We also evaluated L-MAP's performance using precision, recall, F1, and specificity score on the seven knockout datasets (see Supplemental Figure 12).

For each knockout experiment, L-MAP was trained on 90 000 cytosines (chosen uniformly at random) and tested on 10 000 cytosines (chosen uniformly at random). The cytosines in the test set were at least 256 base pairs away from any cytosine in the training set, so that the corresponding windows did not overlap. We did not mix data from different datasets. To ensure consistent results across different random train-test splits, we computed the variance of L-MAP's accuracy across five random samples of the training set for TETKO and DNMT3KO. The average and standard deviation for L-MAP's accuracy are illustrated in Supplemental Figure 5. The results indicate that the deviation in L-MAP's accuracy is very small across different random samples of the training set, which allowed us to rely on the results of a single run for the rest of the experiments.

Figure 1A shows the methylation levels of human ESC wild-type and TET knockout cytosines in the region [2493500,2497500] of chromosome 19, as a qualitative example of the training data. The middle panel shows the difference in methylation level between the two cell lines, where the red dots indicate DMCs (blue otherwise). The lower panel shows L-MAP's predictions of DMCs based on the sequence context around the cytosine. Observe how L-MAP makes accurate predictions in the middle portion of this region.

Figure 1B shows the receiver operating characteristic (ROC) curves for the binary classification performance of L-MAP on the four human knockout data set. The figure shows that the best classification performance was achieved on the TETKO dataset in which TET1, TET2, and TET3 genes were knocked out [area under the curve (AUC) 0.89, accuracy 0.79]. The second best was on the DNMT3KO dataset, in which both DNMT3A and DNMT3B genes were knocked out. The quadruple knockout (QKO) and quintuple knockout (PKO) had lower accuracy and AUC compared with TETKO and DNMT3KO. Our hypothesis is that mixing multiple enzymatic knockouts in QKO and PKO makes it harder for the classifier to capture their sequence specificity. However, the fact that L-MAP can still classify DMCs in the QKO and PKO suggests the existence of common sequence signatures between the two classes of enzymes.

Figure 1C shows the ROC curves for the binary classification performance of L-MAP on the three mouse knockout data set. Again, the figure shows that the best classification performance was achieved on the TET2/3KO dataset in which both TET2 and

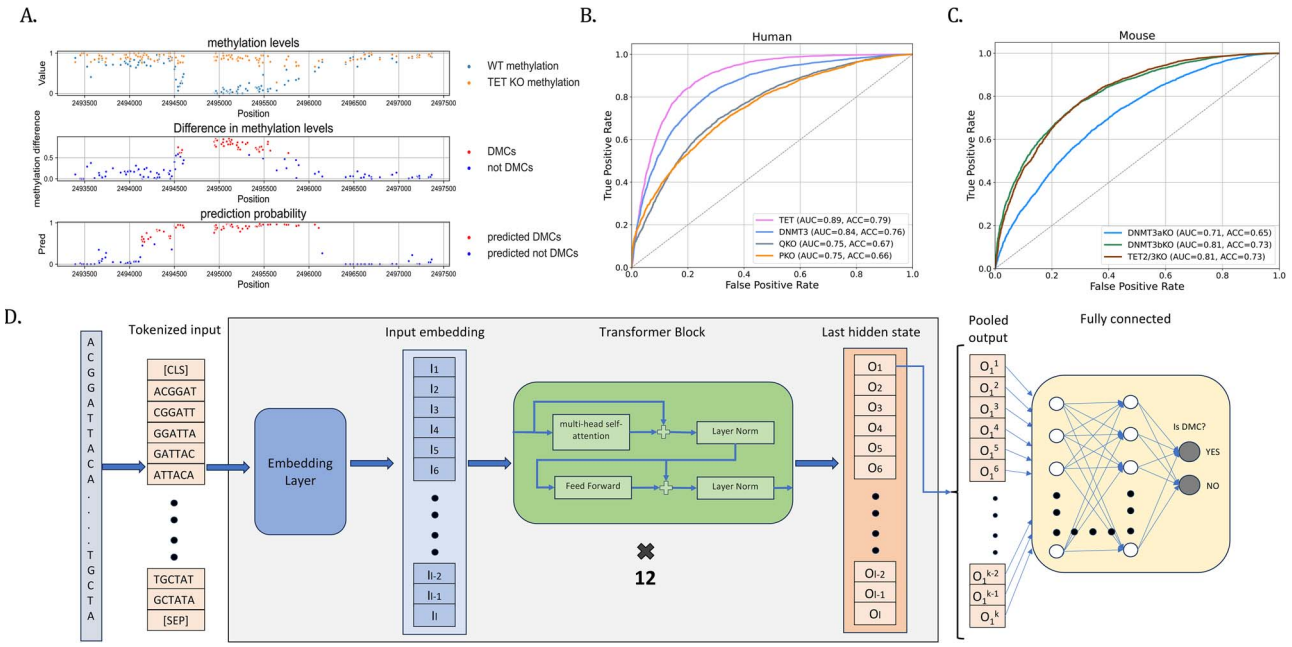


Figure 1. **(A—upper panel)** methylation levels of human ESC wild-type and TET knockout cells in the region [2493500,2497500] of chromosome 19; **(A—middle panel)** the difference in methylation level between wild-type and knockout; the red dots indicate differentially methylation cytosines; **(A—lower panel)** predictions generated by L-MAP based on contextual sequence information; the red dots indicate cytosine that are predicted to be differentially methylated; **(B)** ROC curves for the performance of L-MAP on human TET and DNMT3 knockout datasets (AUC=area under the curve, ACC=accuracy); **(C)** ROC curves for the performance of L-MAP on mouse TET and DNMT3 knockout datasets (AUC=area under the curve, ACC=accuracy); **(D)** the architecture L-MAP.

TET3 genes were knocked out (AUC 0.81, accuracy 0.73). These results also suggest that in humans and mice, the TET activity is more sequence-dependent than DNMT3. The performance of L-MAP on the DNMT3bKO dataset was the second-best.

Cross knockout prediction

In the following experiments, we carried out a set of cross-knockout and cross-species predictions. In one set of experiments, L-MAP was trained on one knockout dataset and tested on a different knockout enzyme. In the second set, L-MAP was trained on human knockout data and tested on mouse knockout data, or vice versa.

The cross-species L-MAP's accuracy is visualized in Fig. 2A and Fig. 2B, for human and mouse, respectively. The figures indicate that the highest accuracy is often observed when L-MAP is trained and tested on the same data set, as expected. However, there are some exceptions. L-MAP's accuracy is higher when trained on human PKO and QKO data sets and tested on TET data sets, compared with being tested on the same knockout dataset. This can be explained by the presence of shared patterns in the PKO and QKO cell lines, both of which include the knockout of TET.

The cross-knockout L-MAP's accuracy is illustrated in Fig. 2C and Fig. 2D. Figure 2C reports the results on three data sets: two for mouse ESC (DNMT3AKO and DNMT3BKO) and one for human ESC (DNMT3KO, which includes DNMT3A and DNMT3B knockout). Figure 2D reports the results on two data sets: one for human ESC (TETKO, corresponding to TET1, TET2, and TET3 knockout) and one for mouse ISC (TET23KO, representing TET2 and TET3 knockout). The figure shows that the highest accuracy is achieved when the model is trained and tested on the same dataset. Also, we observed that in the case of TET, the cross-species experiment yields significantly lower accuracy, suggesting that the underlying sequence contexts associated with TET activity are likely to be different in the two species.

Motif analysis

The objective of this analysis was to extract "knowledge" from the LLM to gain insights on the sequence context employed by L-MAP to make predictions about DMCs. Briefly, we used the attention layer of L-MAP to identify DNA sequences associated with DMCs and DNA sequences associated with non-DMCs. These positive and negative examples were processed using STREME [27], to obtain motifs and corresponding *P*-values (see Methods for details). Figure 3 reports the motifs with the lowest *P*-value for each of the seven knockout datasets (the lowest three *P*-value motifs are reported in Supplemental Figures 8 and 9). We utilized JASPAR [28] to search for known motifs that matched our motifs. The best matches are reported in the last four columns of Fig. 3. The results indicate that all the motifs are associated with the C2H2 zinc finger factors, which are known to have a role in methylation and demethylation processes (see, e.g. [29–32]). For instance, zinc finger protein ZNF615 plays a significant role in embryonic stem cell development through DNA methylation by facilitating the recruitment of DNA methyltransferases to specific genomic regions [33]. Most of the matching motifs are also associated with molecular mechanisms in embryonic stem cells.

The first JASPAR hit in Fig. 3 is the binding site for the PRDM9 zinc finger, which controls the location and intensity of crossovers during meiosis in humans and mice [34–36]. Studies have shown that there is a link between PRDM9 activity and TET1 during meiosis in mice [37]. The second hit is the motif associated with ZNF320, which influences the regulation of the cell cycle and immune infiltration, underscoring its significance in the molecular pathways of hepatocellular carcinoma progression [38]. The third hit is the binding site for ZBTB14, which is a key protein in Xenopus embryonic development, influencing neural induction and differentiation by modulating BMP and WNT signaling pathways [39]. ZBTB14 is also known as a regulator that binds to non-methylated CpG islands, playing a crucial role in controlling gene

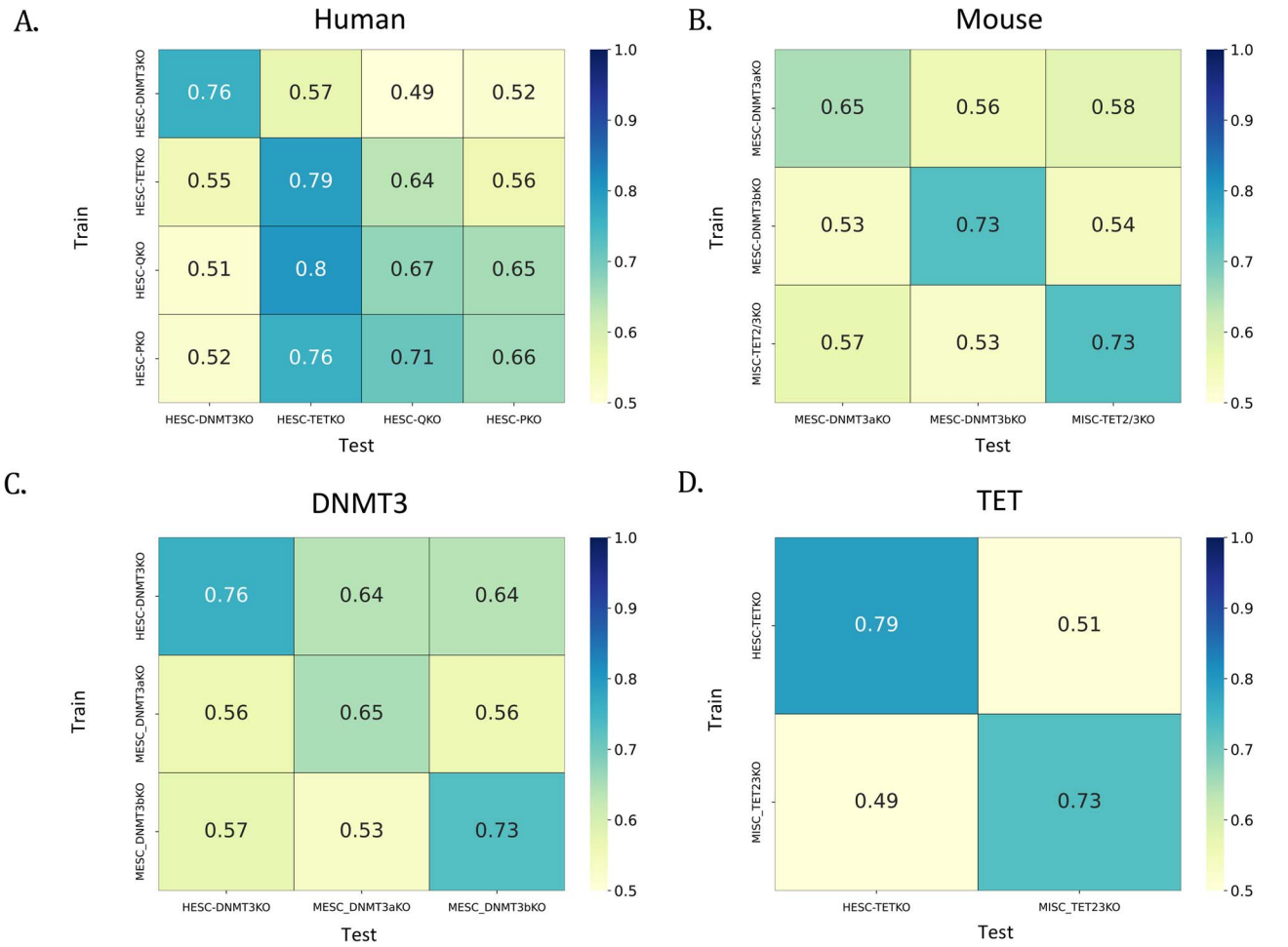


Figure 2. **(A)** L-MAP's accuracy when trained on a human knockout dataset and tested on another human knockout dataset; **(B)** L-MAP's accuracy when trained on one mouse knockout dataset and tested on another mouse knockout dataset; **(C and D)** L-MAP's accuracy when trained on a human (mouse) knockout dataset and tested on a mouse (human) knockout datasets.

expression associated with the two-cell-like state [40]. The fourth hit is the motif associated with GLIS2, which has been identified as a transcriptional activator and is implicated as an epigenetically defined biomarker of a pluripotent phenotype in human ESCs [41]. The fifth hit is the binding site for ZNF740, which plays a crucial role in cell differentiation by modulating the expression of MEF2C and its target genes, influencing the transition of pluripotent stem cells into trophoblasts through its interaction with a specific genomic variation [42]. The sixth hit is the motif associated with ZNF343, which is involved in the early stages of human embryonic development and influences embryo quality and developmental potential [43]. The last hit on Fig. 3 is the KLF17 binding site, which plays a significant role in the establishment of naive pluripotency in human ESCs [44].

L-MAP's high accuracy in the prediction of DMCs for the TET knockout samples can be leveraged for a deeper analysis of the related motifs. In Supplemental File 1, we collected the 20 motifs with the lowest *P*-values and searched the JASPAR database for corresponding transcription factor binding motifs. These transcription factors can be further analyzed for potential interaction with TET. Notably, CTCF had the highest occurrence in the JASPAR hits. The interaction between CTCF and TET is well studied [25, 45–49]. We expect that the other transcription factors in this list are also interacting with TET, but most of them are unexplored in the literature. This is an opportunity for research in functional determinants of TET proteins.

DNMT3a and DNMT3b *de novo* DNA methyltransferases are known to have strong sequence preferences, particularly in the sequences surrounding the CpG dinucleotides [50–55]. To investigate which positions in the input window are more important for the classification, we extracted L-MAP's attention scores. Figures 10 and 11 in the supplemental material show the attention scores within the input window for L-MAP on different data sets. The results indicate that L-MAP's strongest attention is on the positions flanking the center cytosine. Also, the figures show that the attention is much stronger on the flanking positions for the DNMT3 data sets compared with TET data sets, consistent with the literature.

Predictions using sequence and methylation levels

Within the scope of data imputation, one could assume to have the methylation levels of some cytosines and want to predict DMCs for the missing data. To test the extent to which imputation would be possible, we modified the input to L-MAP to allow the use of nearby methylation levels for the wild-type sample, the knockout sample, or both (in addition to the primary DNA sequence surrounding the cytosine of interest). Details about the architecture of this variant of L-MAP can be found in the Methods section.

Figure 4 illustrates the performance of L-MAP using various input combinations. Except for HESC-DNMT3KO, the figure shows

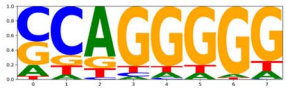
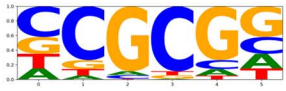
			JASPAR			
Knockout	Captured motif	p-value	Name	ID	Class	Family
HESC-TETKO		8.10E-144	PRDM9	MA1723.2	C2H2 zinc finger factors	Factors with multiple dispersed zinc fingers
HESC-DNMT3KO		7.70E-91	ZNF320	MA1976.2	C2H2 zinc finger factors	More than 3 adjacent zinc fingers
HESC-QKO		1.90E-24	ZBTB14	MA1650.2	C2H2 zinc finger factors	More than 3 adjacent zinc fingers
HESC-PKO		2.70E-61	GLIS2	MA0736.1	C2H2 zinc finger factors	More than 3 adjacent zinc fingers
MESC-DNMT3aKO		9.80E-37	ZNF740	MA0753.3	C2H2 zinc finger factors	Other factors with up to three adjacent zinc fingers
MESC-DNMT3aKO		2.10E-94	ZNF343	MA1711.2	C2H2 zinc finger factors	More than 3 adjacent zinc fingers
MISC-TET23KO		1.20E-194	KLF17	MA1514.2	C2H2 zinc finger factors	Three-zinc finger Kruppel-related

Figure 3. Sequence motifs (extracted from the attention layer of L-MAP) that achieved the lowest *P*-values in each knockout dataset and the corresponding the best hits from the JASPAR motif database.

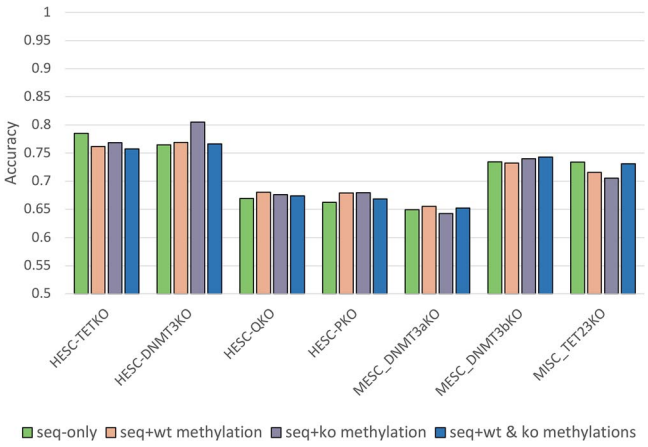


Figure 4. L-MAP's prediction accuracy in seven knockout datasets when neighboring cytosine methylation levels are used (in addition to the primary DNA sequence).

that providing the methylation levels of neighboring cytosines does not significantly improve L-MAP's accuracy. In fact, in five out of seven cases, L-MAP performs slightly better when neighboring cytosine methylation levels were not provided.

Discussion

Here, we introduced L-MAP, an LLM capable of predicting DMCs for TET and DNMT3 knockouts from the DNA sequence surrounding the cytosine of interest. To the best of our knowledge, we are the first to explore this prediction problem. Our experimental results show L-MAP can accurately predict DMCs in human and mouse ESCs.

Our cross-species and cross-enzyme experiments highlight the potential of L-MAP to predict DMCs even when trained on different knockout datasets, with the exception of the model trained on the human TETKO dataset and tested on mouse ESC TET23KO and vice versa. This observation suggests distinct TET activity domains in ESCs between mouse and human species.

In practice, L-MAP can be used to obtain the methylation levels for cytosines that have insufficient sequencing coverage, enabling researchers to include more cytosines in their downstream analyses. This increase in the methylation signal could make a difference in discovering new relationships between methylation enzymes like DNMT3 and TET and specific genes or regulatory elements.

By analyzing the attention layer of L-MAP, our study identified DNA sequence motifs associated with TET and DNMT3 activity in human ESC, mouse ESC, and mouse ISC, which were validated by comparing them with known motifs. Our work represents the first attempt in addressing this challenging problem, and it provides a tool to gain new insights into the role of TET and DNMT3 activity in cell processes, particularly during cell differentiation. The ability to predict DMCs and discover associated sequence motifs opens up opportunities for advancing our understanding of epigenetic regulation in various cellular processes.

Although we assessed the capability of L-MAP for predicting DMCs on a limited set of knockouts in human and mouse ESCs, it remains uncertain whether these findings are applicable to other knockouts or cell lines. The biggest obstacle to expanding L-MAP

to other cell lines or other species is the lack of knockout data sets, which are laborious and expensive to produce. Currently, L-MAP relies solely on the DNA sequence surrounding the cytosines. In order to extend L-MAP to other cell lines or species, it is possible that the classifier might need additional genomic features. The sequence alone may not be sufficient to capture the influence of broader epigenetic contexts or chromatin structures. In addition, our choices for L-MAP's main hyperparameters (i.e. token size and window size) were determined experimentally by maximizing the predictive performance of L-MAP on our data sets. It is possible that other knockout data sets would require a different choice of these hyperparameters.

We observed that a significant proportion of DMCs were located in close proximity to each other, possibly connected to the presence of CpG islands. A failure to account for this phenomenon could result in overlaps between training windows and test windows, biasing the model performance. Future studies could assess the performance of predictive models separately for CpG islands and non-CpG islands to provide a deeper understanding of their capabilities.

Methods

Data sources and pre-processing

We used a total of seven knockout data sets and three wild-type datasets from three studies, with each wild-type data set associated with one of these studies. In the first study, Charlton *et al.* [12] engineered HUES8 human embryonic stem cell lines using CRISPR-Cas9 to selectively inactivate DNMT3A, DNMT3B, TET1, TET2, and TET3 genes, producing variants with multiple genetic knockouts. Specifically, they created the DNMT3KO line (DNMT3A and DNMT3B were deactivated), the TETKO line (TET1, TET2, and TET3 were knocked out), the QKO line (TET1, TET2, TET3, and DNMT3B were deactivated), and the PKO line (TET1, TET2, TET3, DNMT3A, and DNMT3B were knocked out). In the second study, Gu *et al.* [21] generated mouse embryonic stem cell knockout data sets by inactivating DNMT3A and DNMT3B to analyze their roles in DNA methylation. In the third study, Ansari *et al.* [22] generated mouse ISC knockout data sets by creating double knockout mice lacking TET2 and TET3 to investigate their roles in the small intestine.

All 10 datasets (seven knockout and three wild-type) (i) were obtained using whole-genome bisulfite sequencing using Illumina sequencing instruments and (ii) were processed using the BSMAP pipeline [56] for mapping bisulfite-treated reads to the reference genome. In all these datasets, due to the choice of the protocol used to carry out the bisufite-treated sequencing, only the methylation levels for the forward strand are available. In our experiments, we used the methylation levels provided by the authors. However, to ensure that we could compare methylation levels across different studies, we re-analyzed the three wild-type samples using a common software pipeline. We processed the three sets of Illumina reads through Bismark [57] using default parameters. The methylation levels obtained from our pipeline matched almost exactly the methylation levels provided by the authors: the mean square difference between our levels and those provided by the authors was ¡2%.

Given a pair of (wild-type, knockout) data sets, we compared the difference in methylation levels for the same cytosine in the two experiments. We defined a cytosine to be differentially methylated (DMC) if the absolute value of the difference between the methylation level in the wild-type and the methylation level in the knockout from the same study was at least 0.6, as proposed by Charlton *et al.* [12]. We only called DMC for cytosines that were covered by at least 10 reads in both wild-type and knockout experiments. Cytosines that were not covered by at least 10 reads in either experiment were considered undetermined and ignored in our study.

Training set design

We studied the effect of the size of the training set on L-MAP's accuracy in Supplemental Figure 7. The results show that L-MAP's performance improves until the training set size reaches 100 000 data points. Expanding the training set size further only increases the training time, without a significant benefit in the accuracy. Based on this analysis, for each knockout experiment, we constructed the training set by randomly sampling 45 000 cytosines that were differentially methylated (DMCs) and 45 000 cytosines that were not differentially methylated (non-DMCs) from the entire genome, resulting in a training set of 90 000 cytosines. For the test set, we randomly selected an additional 5000 DMCs and 5000 non-DMCs from the remaining data, ensuring that none of these test cytosines were closer than 256 bp to any cytosine in the training set. To clarify, we did not mix data from different datasets; each experiment was conducted using data exclusively from its specific dataset. We evaluated L-MAP's performance for various choices of the input window sizes on DNMT3 and TET knockout datasets in Supplemental Figure 4. Based on this analysis, we selected 512 bp centered around the cytosine of interest, as it yielded the best results among the tested sizes. We should note that 512 bp is the longest possible input that DNABERT allows.

The label of each sequence was binary, indicating whether the center cytosine was differentially methylated or not.

Classifiers

The architecture of L-MAP combines DNABERT [58] with a fully connected neural network as shown in Fig. 1D. For each experiment, we fine-tuned the pre-trained DNABERT and the fully connected network for five epochs. Each epoch took about 40 min on an NVIDIA GeForce RTX 3090 GPU. L-MAP required about 24GB of RAM to load all the parameters of the LLM. To evaluate the convergence and potential overfitting, we analyzed the training and test accuracy and loss over epochs on the HESC-DNMT3KO dataset (see Supplemental Figure 13).

In Supplemental Figure 6 we assessed the accuracy of other Transformer-based models. DNABERT and the nucleotide-transformer tied for the best performance on the TET knockout dataset. Between the two, we chose DNABERT because of its motif-finding module. The input sequence was first tokenized in overlapping 6-mers. In Supplemental Figure 3 we tested various sizes for the tokens on the DNMT3 dataset, and $k = 6$ produced the best performance. DNABERT's output layer was used as input to a fully connected neural network consisting of three layers with 128, 24, and 2 nodes, respectively. Each layer used a dropout rate of 0.5 and employed the ReLU activation function, with the exception of the final layer, which utilized softmax as the activation function. The model was trained utilizing the Adam optimizer, with a learning rate of 1e-5, and employed a binary class entropy as the loss function. Also, we evaluated the performance of a random forest and a support vector machine classifier using the embedding produced by the pre-trained DNABERT encoder. Supplemental Figure 14 shows that L-MAP outperformed both the RF classifier and the SVM classifier,

indicating the importance of fine-tuning DNABERT to achieve optimal performance.

In the experiments that used neighboring cytosine methylation levels, the embedding produced by DNABERT was concatenated with the vector(s) representing the methylation levels from either wild-type or knockout datasets (or both). This additional vector was -1 in all positions, except for the positions of neighboring cytosines with sufficient read coverage, where the known methylation level of the cytosine was used.

Motif analysis

We first obtained a random set of 10 000 genomic sequences of length 512 bp, where half of them were the context sequence surrounding a DMC, while the other half surrounded a non-DMC. We processed these sequences through DNABERT, then extracted the weights from DNABERT's attention layer. We used the weights to identify high-attention regions, using the DNABERT motif-finding tool. For each of these regions, we extracted the corresponding DNA sequences from the original DNA sequence dataset, resulting in two distinct sets of DNA sequences for positive and negative samples. Then, we employed STREME [27] to identify motifs (and their *P*-values) that were enriched in the positive set and depleted in the negative set, using parameters `minw=6`, `maxw=12`, and `nmotifs=100`. The position weight matrices of the three motifs with the lowest *P*-values were matched against known motifs using JASPAR [28].

Abbreviations

5mC = 5-methylcytosine
DMC = differentially methylated cytosines
LLM = large language model
L-MAP = language model-based methyltransferases activity predictor
AUC = area under the curve
TET = ten-eleven translocation (enzyme)
DNMT = DNA methyltransferase (enzyme)
ESC = embryonic stem cells
ISC = intestinal stem cells

Key Points

- L-MAP is a large language model that can predict differentially methylated cytosines (DMCs) in human and mouse when trained on TET and DNMT3 knockout data sets.
- L-MAP predicts DMCs with high accuracy exclusively based on the DNA sequence surrounding the cytosine of interest.
- L-MAP can predict DMCs even when trained on different knockout data sets (human versus mouse, or TET versus DNMT3).
- L-MAP can be used to discover new transcription factor binding sites associated with TET and DNMT3.

Acknowledgments

The authors wish to thank Daniel Koenig (UC Riverside) and Jikui Song (UC Riverside) for earlier discussions on this project.

Author contributions

Saleh Sereshki (Conceptualization, Methodology, Software, Resources, Writing original draft) and Stefano Lonardi (Conceptualization, Supervision, Writing – Review & Editing)

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: The authors declare that they have no competing interests.

Funding

This project was supported in part by US National Institutes of Health 1R01AI169543-01, US National Science Foundation CBET 2225878, and US National Science Foundation IIS 244456.

Data availability

All the datasets used in this study are publicly available from NCBI. The datasets accessions are GSE126958, GSE100956, and GSE200227. Bisulfite-treated Illumina reads were obtained from NCBI SRA, accessions SRR8611939, SRR6894127, and SRR18645747. L-MAP is available at https://github.com/ucrbioinfo/dmc_prediction

References

1. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* 2019;**20**:590–607. https://doi.org/10.1038/s41580-019-0159-6
2. Wanxue X, Mengyao X, Wang L. *et al.* Integrative analysis of DNA methylation and gene expression identified cervical cancer-specific diagnostic biomarkers. *Signal Transduct Target Ther* 2019;**4**:55. https://doi.org/10.1038/s41392-019-0081-6
3. Bozorgpour R. Computational explorations in biomedicine: Unraveling molecular dynamics for cancer, drug delivery, and biomolecular insights using LAMMPS simulations. arXiv preprint. 2023;arXiv:2311.13000. https://arxiv.org/abs/2311.13000
4. Skvortsova K, Stirzaker C, Taberlay P. The DNA methylation landscape in cancer. *Essays Biochem* 2019;**63**:797–811. https://doi.org/10.1042/EBC20190037
5. Anvar Z, Chakchouk I, Demond H. *et al.* DNA methylation dynamics in the female germline and maternal-effect mutations that disrupt genomic imprinting. *Genes* 2021;**12**:1214.
6. Li E, Zhang Y. DNA methylation in mammals. *Cold Spring Harb Perspect Biol* 2014;**6**:a019133. https://doi.org/10.1101/cshperspect.a019133
7. Meissner A, Mikkelsen TS, Hongcang G. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 2008;**454**:766–70. https://doi.org/10.1038/nature07107
8. Okano M, Bell DW, Haber DA. *et al.* DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* 1999;**99**:247–57. https://doi.org/10.1016/S0092-8674(00)81656-6
9. Gao L, Emperle M, Guo Y. *et al.* Comprehensive structure-function characterization of DNMT3B and DNMT3A reveals distinctive de novo DNA methylation mechanisms. *Nat Commun* 2020;**11**:3355. https://doi.org/10.1038/s41467-020-17109-4

10. Hao W, Zhang Y. Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. *Genes Dev* 2011;**25**: 2436–52. https://doi.org/10.1101/gad.179184.111
11. Betto RM, Diamante L, Perrera V. *et al*. Metabolic control of DNA methylation in naive pluripotent cells. *Nat Genet* 2021;**53**:215–29. https://doi.org/10.1038/s41588-020-00770-2
12. Charlton J, Jung EJ, Mattei AL. *et al*. TETs compete with DNMT3 activity in pluripotent cells at thousands of methylated somatic enhancers. *Nat Genet* 2020;**52**:819–27. https://doi.org/10.1038/s41588-020-0639-9
13. Iqbal K, Jin S-G, Pfeifer GP. *et al*. Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *Proc Natl Acad Sci* 2011;**108**:3642–7.
14. Lio C-W, Zhang J, González-Avalos E. *et al*. Tet2 and Tet3 cooperate with B-lineage transcription factors to regulate DNA modification and chromatin accessibility. *eLife* 2016;**5**:e18290. https://doi.org/10.7554/eLife.18290
15. Huang X, Balmer S, Lyu C. *et al*. ZFP281 controls transcriptional and epigenetic changes promoting mouse pluripotent state transitions via DNMT3 and TET1. *Dev Cell* 2024;**59**:465–481.e6. https://doi.org/10.1016/j.devcel.2023.12.018
16. Salmerón-Bárcenas EG, Zacapala-Gómez AE, Torres-Rojas FI. *et al*. TET enzymes and 5hmC levels in carcinogenesis and progression of breast cancer: Potential therapeutic targets. *Int J Mol Sci* 2023;**25**:272.
17. Chen X, Zhou W, Song R-H. *et al*. Tumor suppressor CEBPA interacts with and inhibits DNMT3A activity. *Sci Adv* 2022;**8**:eabl5220. https://doi.org/10.1126/sciadv.abl5220
18. Chao L, Yang S, Li H. *et al*. Competitive binding of TET1 and DNMT3A/B cooperates the DNA methylation pattern in human embryonic stem cells. *Biochim Biophys Acta Gene Regul Mech* 2022; **1865**:194861. https://doi.org/10.1016/j.bbagrm.2022.194861
19. Ravichandran M, Rafalski D, Davies CI. *et al*. Pronounced sequence specificity of the TET enzyme catalytic domain guides its cellular function. *Sci Adv* 2022;**8**:eabm2427. https://doi.org/10.1126/sciadv.abm2427
20. Jeltsch A, Adam S, Dukatz M. *et al*. Deep enzymology studies on DNA methyltransferases reveal novel connections between flanking sequences and enzyme activity. *J Mol Biol* 2021;**433**:167186. https://doi.org/10.1016/j.jmb.2021.167186
21. Tianpeng G, Lin X, Cullen SM. *et al*. Deqiang Sun, Jianzhong Su. DNMT3A and TET1 cooperate to regulate promoter epigenetic landscapes in mouse embryonic stem cells. *Genome Biol* 2018;**19**: 1–15. https://doi.org/10.1186/s13059-018-1464-7
22. Ansari I, Solé-Boldo L, Ridnik M. *et al*. TET2 and TET3 loss disrupts small intestine differentiation and homeostasis. *Nat Commun* 2023;**14**:4005. https://doi.org/10.1038/s41467-023-39512-3
23. Nessa Carey C, Marques J, Reik W. DNA demethylases: A new epigenetic frontier in drug discovery. *Drug Discov Today* 2011;**16**: 683–90. https://doi.org/10.1016/j.drudis.2011.05.004
24. Milon TI, Wang Y, Fontenot RL. *et al*. Development of a novel representation of drug 3D structures and enhancement of the TSR-based method for probing drug and target interactions. *Comput Biol Chem* 2024;**112**:108117.
25. Theofilatos D, Ho T, Waitt G. *et al*. Deciphering the TET3 interactome in primary thymic developing T cells. *Iscience* 2024;**27**:109782. https://doi.org/10.1016/j.isci.2024.109782
26. Brauchle M, Yao Z, Arora R. *et al*. Protein complex interactor analysis and differential activity of KDM3 subfamily members towards H3K9 methylation. *PloS One* 2013;**8**:e60549. https://doi.org/10.1371/journal.pone.0060549
27. Bailey TL. STREME: Accurate and versatile sequence motif discovery. *Bioinformatics* 2021;**37**:2834–40. https://doi.org/10.1093/bioinformatics/btab203
28. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I. *et al*. François Parcy, and Anthony Mathelier. JASPAR 2022: The 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2021;**50**:D165–73.
29. Schmitges FW, Radovani E, Najafabadi HS. *et al*. Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res* 2016;**26**:1742–52. https://doi.org/10.1101/gr.209643.116
30. Imbeault M, Helleboid P-Y, Trono D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* 2017;**543**:550–4. https://doi.org/10.1038/nature21683
31. Patel A, Hashimoto H, Zhang X. *et al*. Characterization of how DNA modifications affect DNA binding by C2H2 zinc finger proteins. *Methods Enzymol* 2016;**573**:387–401. https://doi.org/10.1016/bs.mie.2016.01.019
32. Sereshki S, Lee N, Omirou M. *et al*. On the prediction of non-CG DNA methylation using machine learning. *NAR Genomics Bioinf* 2023;**5**:lqad045. https://doi.org/10.1093/nargab/lqad045
33. Zuo X, Sheng J, Lau H-T. *et al*. Zinc finger protein ZFP57 requires its co-factor to recruit DNA methyltransferases and maintains DNA methylation imprint in embryonic stem cells via its transcriptional repression domain. *J Biol Chem* 2012;**287**:2107–18. https://doi.org/10.1074/jbc.M111.322644
34. Ségurel L, Leffler EM, Przeworski M. The case of the fickle fingers: How the PRDM9 zinc finger protein specifies meiotic recombination hotspots in humans. *PLoS Biol* 2011;**9**:e1001211. https://doi.org/10.1371/journal.pbio.1001211
35. Berg IL, Neumann R, Lam K-WG. *et al*. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nat Genet* 2010;**42**:859–63. https://doi.org/10.1038/ng.658
36. Hinch AG, Tandon A, Patterson N. *et al*. The landscape of recombination in African Americans. *Nature* 2011;**476**:170–5. https://doi.org/10.1038/nature10336
37. Yamaguchi S, Hong K, Liu R. *et al*. Tet1 controls meiosis by regulating meiotic gene expression. *Nature* 2012;**492**:443–7. https://doi.org/10.1038/nature11709
38. Zhen J, Ke Y, Pan J. *et al*. ZNF320 is a hypomethylated prognostic biomarker involved in immune infiltration of hepatocellular carcinoma and associated with cell cycle. *Aging (Albany NY)* 2022;**14**:8411. https://doi.org/10.18632/aging.204350
39. Takebayashi-Suzuki K, Konishi H, Miyamoto T. *et al*. Coordinated regulation of the dorsal-ventral and anterior-posterior patterning of Xenopus embryos by the BTB/POZ zinc finger protein Zbtb14. *Dev Growth Differ* 2018;**60**:158–73. https://doi.org/10.1111/dgd.12431
40. Gupta N, Yakhou L, Albert JR. *et al*. A genome-wide screen reveals new regulators of the 2-cell-like cell state. *Nat Struct Mol Biol* 2023;**30**:1105–18. https://doi.org/10.1038/s41594-023-01038-z
41. Pells S, Koutsouraki E, Morfopoulou S. *et al*. Novel human embryonic stem cell regulators identified by conserved and distinct CpG island methylation state. *PloS One* 2015;**10**:e0131102. https://doi.org/10.1371/journal.pone.0131102
42. Li H-T, Liu Y, Liu H. *et al*. Effect for human genomic variation during the BMP4-induced conversion from pluripotent stem cells to trophoblast. *Front Genet* 2020;**11**:230. https://doi.org/10.3389/fgene.2020.00230
43. Wang W, Zhao M, Zuo H. *et al*. Evaluate the developmental competence of human 8-cell embryos by single-cell RNA sequencing. *Reprod Fertil* 2023;**4**:e220119.

44. Lea RA, McCarthy A, Boeing S. *et al*. KLF17 promotes human naïve pluripotency but is not required for its establishment. *Development* 2021;**148**:dev199378. https://doi.org/10.1242/dev.199378
45. Nanan KK, Sturgill DM, Prigge MF. *et al*. TET-catalyzed 5-carboxylcytosine promotes CTCF binding to suboptimal sequences genome-wide. *Iscience* 2019;**19**:326–39. https://doi.org/10.1016/j.isci.2019.07.041
46. Dubois-Chevalier J, Oger F, Dehondt H. *et al*. A dynamic CTCF chromatin binding landscape promotes DNA hydroxymethylation and transcriptional induction of adipocyte differentiation. *Nucleic Acids Res* 2014;**42**:10943–59. https://doi.org/10.1093/nar/gku780
47. Marina RJ, Oberdoerffer S. Epigenomics meets splicing through the TETs and CTCF. *Cell Cycle* 2016;**15**:1397–9. https://doi.org/10.1080/15384101.2016.1171650
48. Wiehle L, Thorn GJ, Raddatz G. *et al*. DNA (de)methylation in embryonic stem cells controls CTCF-dependent chromatin boundaries. *Genome Res* 2019;**29**:750–61. https://doi.org/10.1101/gr.239707.118
49. Marina RJ, Sturgill D, Bailly MA. *et al*. TET-catalyzed oxidation of intragenic 5-methylcytosine regulates CTCF-dependent alternative splicing. *EMBO J* 2016;**35**:335–55. https://doi.org/10.15252/embj.201593235
50. Handa V, Jeltsch A. Profound flanking sequence preference of Dnmt3a and Dnmt3b mammalian DNA methyltransferases shape the human epigenome. *J Mol Biol* 2005;**348**:1103–12. https://doi.org/10.1016/j.jmb.2005.02.044
51. Bozorgpour R, Sheybanikashani S, Mohebi M. Exploring the role of molecular dynamics simulations in most recent cancer research: Insights into treatment strategies. arXiv preprint 2023;arXiv:2310.19950. https://arxiv.org/abs/2310.19950
52. Dukatz M, Dittrich M, Stahl E. *et al*. DNA methyltransferase DNMT3A forms interaction networks with the CpG site and flanking sequence elements for efficient methylation. *J Biol Chem* 2022;**298**:102462. https://doi.org/10.1016/j.jbc.2022.102462
53. Mao S-Q, Cuesta SM, Tannahill D. *et al*. Genome-wide DNA methylation signatures are determined by DNMT3A/B sequence preferences. *Biochemistry* 2020;**59**:2541–50. https://doi.org/10.1021/acs.biochem.0c00339
54. Zhang Z-M, Rui L, Wang P. *et al*. Structural basis for DNMT3A-mediated de novo DNA methylation. *Nature* 2018;**554**:387–91. https://doi.org/10.1038/nature25477
55. Adam S, Anteneh H, Hornisch M. *et al*. DNA sequence-dependent activity and base flipping mechanisms of DNMT1 regulate genome-wide DNA methylation. *Nat Commun* 2020;**11**:3723.
56. Xi Y, Li W. BSMAP: Whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* 2009;**10**:232. https://doi.org/10.1186/1471-2105-10-232
57. Krueger F, Andrews SR. Bismark: A flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics* 2011;**27**:1571–2. https://doi.org/10.1093/bioinformatics/btr167
58. Ji Y, Zhou Z, Liu H. *et al*. DNABERT: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* 2021;**37**:2112–20. https://doi.org/10.1093/bioinformatics/btab083

Briefings in Bioinformatics, 2025, **26(2)**, bbaf092
https://doi.org/10.1093/bib/bbaf092
Problem Solving Protocol