Predicting Differentially Methylated Cytosines in TET and DNMT3 Knockout Mutants via a Large Language Model

Saleh Sereshki[1,a] and Stefano Lonardi[1,b]

[1]*Department of Computer Science and Engineering, University of California, Riverside, 92521, USA*

[a]E-mail: ssere004@ucr.edu
[e]E-mail: stelo@cs.ucr.edu

Abstract

DNA cytosine methylation is an epigenetic marker which regulates many cellular processes. Mammalian genomes typically maintain consistent methylation patterns over time, except in specific regulatory regions like promoters and certain types of enhancers. The dynamics of DNA methylation is controlled by a complex cellular machinery, in which the enzymes DNMT3 and TET play a major role. This study explores the identification of differentially methylated cytosines (DMCs) in TET and DNMT3 knockout mutants in mice and human embryonic stem cells. We investigate (i) whether a large language model can be trained to recognize DMCs from the sequence surrounding the cytosine of interest, (ii) whether a classifier trained on human knockout data can predict DMCs in the mouse genome (and vice versa), (iii) whether a classifier trained on DNMT3 knockout can predict DMCs for TET knockout (and vice versa). Our study identifies statistically significant motifs associated with the prediction of DMCs each mutant, shedding new light on the understanding of DNA methylation dynamics in stem cells. Our software tool 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 which directly or indirectly regulates several critical cellular processes, including gene expression, genome stability, transposon suppression, and gene imprinting (see, e.g., Greenberg and Bourc'his 2019; W Xu et al. 2019; Ackah et al. 2022; Spainhour et al. 2019; Kuhlmann et al. 2021; Blake et al. 2020; Kumar and Mohapatra 2021; Pappalardo and Barra 2021; J Zhang et al. 2023; W Guo et al. 2021; W Zhou et al. 2020; Ivanova et al. 2020). 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., (Skvortsova et al. 2019; Y Li et al. 2023; SG Zhao et al. 2020; Anvar et al. 2021)).

In mammals, DNA methylation primarily occurs in CpG dinucleotides, with most of them being methylated (E Li and Y Zhang 2014). Mammalian genomes typically maintain consistent CpG methylation patterns over time, except in specific regulatory regions like promoters and certain types of enhancers (Meissner et al. 2008). In these variable regions, the dynamics of methylation and demethylation is 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 (Okano et al. 1999; L Gao et al. 2020). TET is an enzyme that catalyzes the conversion of 5-methylcytosine to 5-hydroxymethylcytosine and its oxidized derivatives. The conversion of 5-hydroxymethylcytosine and its derivatives ultimately leads to active DNA demethylation (H Wu and Y Zhang 2011).

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 pluripotency stages, DNMT3 and TET modulate the epigenetic landscape, thus influencing cellular differentiation (Betto et al. 2021; Charlton et al. 2020). During post-fertilization 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 (Iqbal et al. 2011). 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., (Charlton et al. 2020)). Chao *et al.* also studied the interactions between TET1, DNMT3A, and DNMT3B in human embryonic stem cells, and how these interactions collectively influence global methylation patterns (Chao et al. 2022).

Given the importance of DNMT3 and TET in developmental biology and embryogenesis, there is strong interest in 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 (Ravichandran et al. 2022) 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 (Chao et al. 2022) also reported that TET1 prefers binding to specific genomic regions. This appears to be also true for DNMT3. Jeltsch et al. (Jeltsch et al. 2021) 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., (Charlton et al. 2020; T Gu et al. 2018; Ansari et al. 2023)). 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 (DMC) in TET and DNMT3 knockout mutants, using exclusively the underlying DNA sequence around the cytosines. Our predictor, called L-MAP (Language model-based Methyltransferases Activity Predictor), is transformer-based large language model that utilizes contextual sequence information to predict the enzymatic activity of DNMT3 and TET on cytosines.

We envision the main use of L-MAP as a tool to impute missing and/or uncertain DMCs obtained from wet lab experiments. 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 enzymatic activity on mice, and vice versa; (3) whether the methylation levels of nearby cytosines can help L-MAP predicting DMCs with higher accuracy; (4) whether L-MAP has learned sequence motifs known to be associated with the activity of DNMT3 and TET enzymes.

Results

Seven data sets, across three studies, were used to train and test L-MAP. In the first study by Charlton *et al.* (Charlton et al. 2020), the authors utilized CRISPR-Cas9 to create an array of gene knockouts in human embryonic stem cells (ESC) involving DNMT3s and TETs, both individually and in combination. The following knockout configurations were established: (i) in the DNMT3KO ESC line, both DNMT3A and DNMT3B genes were deactivated; (ii) in the TETKO ESC line, TET1, TET2, and TET3 genes were knocked out; (ii) in the QKO ESC line, TET1, TET2, TET3 and DNMT3B were deactivated; and (iv) in the PKO ESC line, TET1, TET2, TET3, DNMT3A

and DNMT3B were knocked out. The second study by Gu *et al.* (T Gu et al. 2018) involved DNMT3A and DNMT3B knockout in mouse ESC. The third study by Ansari *et al.* (Ansari et al. 2023) involved TET2 and TET3 knockout in mouse intestinal stem cells (ISC).

All ten 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 (Xi and W Li 2009) for mapping bisulfite-treated reads to the reference genome. In our experiments, we used the methylation levels provided by the authors. However, to ensure that we could compare methylation level 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 (Krueger and SR Andrews 2011) 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 were less than 2%.

Supplemental Figure 2 reports the genome-wide methylation levels for the three wild type and seven knockout data sets. In general, the methylation level of a cytosine c ranges from 0 to 1, where 0 indicates that none the cells in the sample are methylated on c, and 1 indicates that all the cells in the sample are methylated on c. Observe 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 determined seven sets of differentially methylated cytosines (DMC). 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 differentially methylated cytosines on the seven data sets. Observe that the number of differentially methylated cytosines ranges from about 100 thousand 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 thousand cytosines for each data sets, half of which were differentially methylated (and the other half was not). The sample included 100 thousand 512 bp-long DNA sequence centered around the chosen cytosines, along with the corresponding binary label (1 indicated a DMC, 0 otherwise). We evaluated the impact of size of the training

dataset on L-MAP's accuracy in Supplemental Figure 7. Observe 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.

L-MAP was trained on 90% of the cytosines (chosen uniformly at random from each data set) and tested it on the remaining 10%. 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 is illustrated in Supplemental Figure 5. Observe that the deviation on L-MAP's accuracy is very small across different random samples for the training set, which allowed us to rely on the results of a single run for the rest of the experiments.

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

Figure 1-B show the ROC curves for the binary classification performance of L-MAP on the four human knockout data set. Observe 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 0.95, accuracy 0.89). 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 to 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 differentially methylated cytosines in the QKO and PKO suggests the existence of common sequence signatures between the two classes of enzymes.

Figure 1-C shows the ROC curves for the binary classification performance of L-MAP on the three mouse knockout data set. Again, observe that the best classification performance

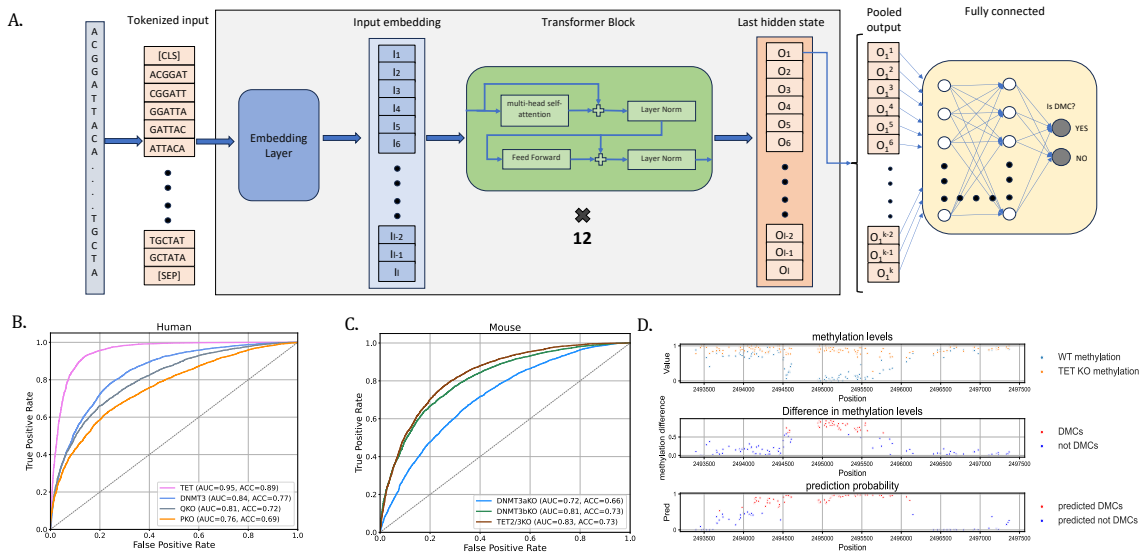


Figure 1. (A) the architecture L-MAP; **(B)** ROC curves for the performance of L-MAP on human TET and DNMT3 knockout datasets (AUC=area under the curve, ACC=accuracy); **(C)** ROC curves for the performance of L-MAP on mouse TET and DNMT3 knockout datasets (AUC=area under the curve, ACC=accuracy); **(D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

was achieved on the TET2/3KO dataset in which both TET2 and TET3 genes were knocked out (area under the curve 0.83, accuracy 0.73). These results in human and mouse suggest that 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 data, or vice versa.

The cross-species L-MAP's accuracy is visualized in Figure 2-A and Figure 2-B, for human and mouse, respectively. Observe that in most cases the highest accuracy is observed when L-MAP is trained and tested on the same data set, as expected. However, there are some ex-

ceptions. L-MAP's accuracy is higher when trained on human PKO and QKO data sets and tested on TET data sets, compared to being tested on the same knockout dataset. This can be explained by the presence of shared patterns in PKO and QKO cell lines, both of which include the knockout of TET.

The cross-knockout L-MAP's accuracy is illustrated in Figure 2-C and Figure 2-D. Figure 2-C 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 2-D 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). Observe again that the highest accuracy is achieved when the model is trained and tested on the same dataset. Also observe 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 (Bailey 2021), 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 (Castro-Mondragon et al. 2021) to search for known motifs that matched our motifs. The best matches are reported in last four columns of Figure 3. First observe that all the motifs belong to the C2H2 zinc finger factors, which are known to have a role in methylation and demethylation processes (see, e.g., (Schmitges et al. 2016; Imbeault et al. 2017; Patel et al. 2016; Sereshki et al. 2023)). 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 (X Zuo et al. 2012). Most of the matching motif are also associated with

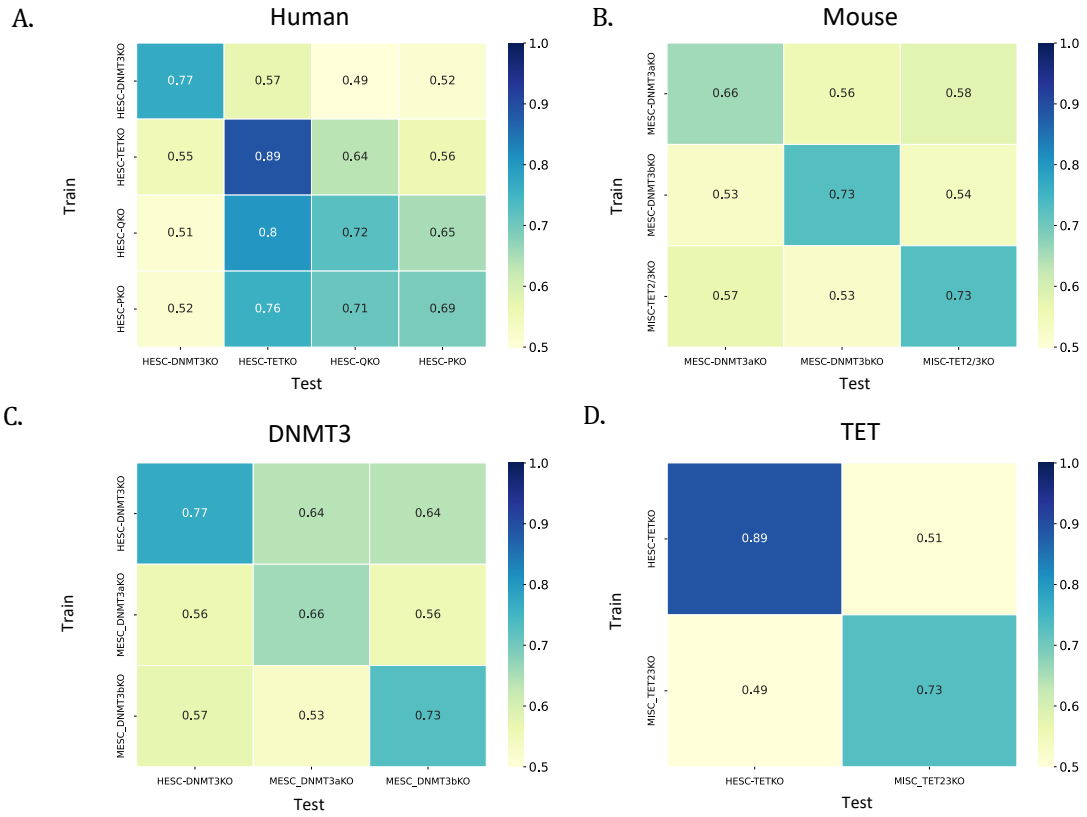


Figure 2. (A) L-MAP's accuracy when trained on a human knockout dataset and tested on another human knockout dataset **(B)** L-MAP's accuracy when trained on one mouse knockout dataset and tested on another mouse knockout dataset **(C-D)** L-MAP's accuracy when trained on a human (mouse) knockout dataset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

molecular mechanisms in embryonic stem cells.

The first JASPAR hit in Figure 3 is the binding site for the PRDM9 zinc finger, which controls the location and intensity of crossovers during meiosis in humans and mouse (Ségurel et al. 2011; Berg et al. 2010; Hinch et al. 2011). Studies have shown that there is a link between PRDM9 activity and TET1 during meiosis in mice (S Yamaguchi et al. 2012). The second hit is the motif associated with ZNF320, which influences the regulation of cell cycle and immune infiltration, underscoring its significance in the molecular pathways of hepatocellular carcinoma progression (Zhen et al. 2022). 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 (Takebayashi-Suzuki et al. 2018). ZBTB14 is also known as a regulator that binds to non-methylated CpG islands, playing a crucial role in controlling gene expression associated with the 2-cell-like state (Gupta et al. 2023). The fourth hit is the motif associate 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 (Pells et al. 2015). 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 (Ht Li et al. 2020). The sixth hit is the motif associate with ZNF343, which is involved in early stages of human embryonic development and influences embryo quality and developmental potential (W Wang et al. 2023). The last hit on Figure 3 is the binding site of KLF17 which plays a significant role in the establishment of naive pluripotency in human ESCs (Lea et al. 2021).

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 differentially methylated cytosines for the missing data. To test the extent of 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.

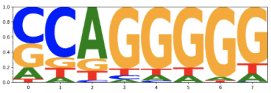
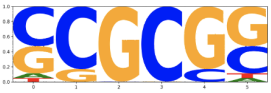
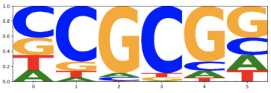
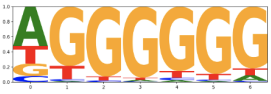
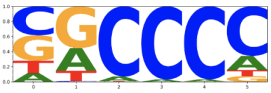
			JASPAR			
Knockout	**Captured motif**	**p-value**	**Name**	**ID**	**Class**	**Family**
HESC-TETKO		8.10E-144	PRDM9	MA1723.2	C2H2 zinc finger factors	Factors with multiple dispersed zinc fingers
HESC-DNMT3KO		7.70E-91	ZNF320	MA1976.2	C2H2 zinc finger factors	More than 3 adjacent zinc fingers
HESC-QKO		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

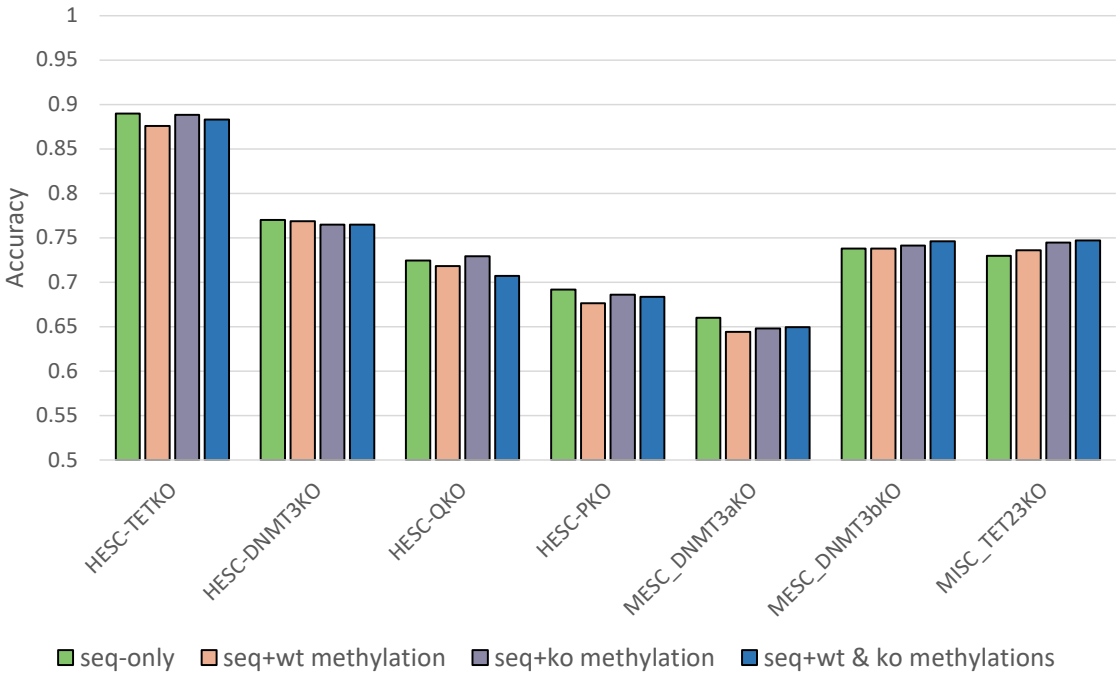


Figure 4. Effect of including neighboring cytosine methylation levels on L-MAP's prediction accuracy for DMCs in seven knockout datasets

Figure 4 illustrates the performance of L-MAP using various input combinations. Observe that providing the methylation levels of neighboring cytosines does not significant improve L-MAP's accuracy. In fact, in four out of seven cases, L-MAP performed slightly better when neighboring cytosine methylation levels were not provided.

Discussion

Here we introduced L-MAP, a large language model capable of predicting differentially methylated cytosines for TET and DNMT3 knockouts from the DNA sequence surrounding the cytosine of interest. Our findings 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.

Furthermore, 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 to

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 uncover associated sequence motifs opens up opportunities for advancing our understanding of epigenetic regulation in various biological contexts.

Methods

Data sources and pre-processing

We used seven data sets from three different studies, namely (i) the human ESC knockout data sets generated by Charlton *et al.*, who engineered several HUES8 embryonic stem cell lines using CRISPR-Cas9, producing variants with double, triple, quadruple, and quintuple genetic knockouts through the selective inactivation of DNMT3A, DNMT3B, TET1, TET2, and TET3 genes (Charlton et al. 2020); (ii) the mouse knockout data sets generated by Gu *et al.* who analyzed the roles of DNMT3A and DNMT3B in DNA methylation within mouse ESCs following the loss of those enzymes (T Gu et al. 2018); (iii) the mouse ISC knockout data sets produced by Ansari *et al.* who investigated the roles of TET2 and TET3 in the small intestine by generating double knockout mice (Ansari et al. 2023). All the datasets are publicly available from NCBI.

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 was at least 0.6, as proposed by Charlton *et al.* in (Charlton et al. 2020). We only called DMC for cytosines that were covered by at least ten reads in both wild type and knockout experiments. Cytosines that were not covered by at least ten reads in either experiments were considered undetermined and ignored in our study.

Training set design

We studied the effect of the size of the training set size on L-MAP's accuracy in Supplemental Figure 7. Observe that L-MAP's performance improves until the training set size reaches 100

thousand 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 experiment in our study, we sampled 50 thousand cytosines (uniformly at random) from all genome-wide cytosines that were differentially methylated, and another 50 thousand cytosines (uniformly at random) from all genome-wide cytosines that were not differentially methylated. 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 observe that 512 bp is the longest possible input that DNABERT allows. The sample containing 100 thousand sequences was divided into training set (90%) and test set (10%) uniformly at random. 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 (Y Ji et al. 2021) with a fully connected neural network as shown in Figure 1-A. In Supplemental Figure 6 we assessed the accuracy of other Transformer-based models. We selected DNABERT because it achieved the best performance on the TET knockout dataset. 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 the binary class entropy as the loss function.

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 (Bailey 2021) 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 (Castro-Mondragon et al. 2021).

Data access

All the datasets used in this study were 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`

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

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

Authors' contributions

SS and SL conceptualized the project. SS designed L-MAP. SS carried out the experiments. SS and SL wrote the manuscript.

References

Ackah M, Guo L, Li S, Jin X, Asakiya C, Aboagye ET, Yuan F, Wu M, Essoh LG, Adjibolosoo D, et al. 2022. DNA methylation changes and its associated genes in mulberry (Morus alba L.) Yu-711 response to drought stress using MethylRAD sequencing. *Plants*. **11**: 190.

Ansari I, Solé-Boldo L, Ridnik M, Gutekunst J, Gilliam O, Korshko M, Liwinski T, Jickeli B, Weinberg-Corem N, Shoshkes-Carmel M, et al. 2023. TET2 and TET3 loss disrupts small intestine differentiation and homeostasis. *Nature communications*. **14**: 4005.

Anvar Z, Chakchouk I, Demond H, Sharif M, Kelsey G, and Van den Veyver IB. 2021. DNA methylation dynamics in the female germline and maternal-effect mutations that disrupt genomic imprinting. *Genes*. **12**: 1214.

Bailey TL. 2021. STREME: accurate and versatile sequence motif discovery. *Bioinformatics*. **37**: 2834–2840.

Berg IL, Neumann R, Lam KWG, Sarbajna S, Odenthal-Hesse L, May CA, and Jeffreys AJ. 2010. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nature genetics*. **42**: 859–863.

Betto RM, Diamante L, Perrera V, Audano M, Rapelli S, Lauria A, Incarnato D, Arboit M, Pedretti S, Rigoni G, et al. 2021. Metabolic control of DNA methylation in naive pluripotent cells. *Nature genetics*. **53**: 215–229.

Blake LE, Roux J, Hernando-Herraez I, Banovich NE, Perez RG, Hsiao CJ, Eres I, Cuevas C, Marques-Bonet T, and Gilad Y. 2020. A comparison of gene expression and DNA methylation patterns across tissues and species. *Genome research*. **30**: 250–262.

Castro-Mondragon JA et al. 2021. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*. **50**: D165–D173.

329 Chao L, Yang S, Li H, Long C, Xi Q, and Zuo Y. 2022. Competitive binding of TET1 and DNMT3A/B
330 cooperates the DNA methylation pattern in human embryonic stem cells. *Biochimica et Bio-*
331 *physica Acta (BBA)-Gene Regulatory Mechanisms*. **1865**: 194861.
332 Charlton J, Jung EJ, Mattei AL, Bailly N, Liao J, Martin EJ, Giesselmann P, Brändl B, Stamenova
333 EK, Müller FJ, et al. 2020. TETs compete with DNMT3 activity in pluripotent cells at thousands
334 of methylated somatic enhancers. *Nature genetics*. **52**: 819–827.
335 Gao L, Emperle M, Guo Y, Grimm SA, Ren W, Adam S, Uryu H, Zhang ZM, Chen D, Yin J, et al.
336 2020. Comprehensive structure-function characterization of DNMT3B and DNMT3A reveals
337 distinctive de novo DNA methylation mechanisms. *Nature communications*. **11**: 3355.
338 Greenberg MV and Bourc'his D. 2019. The diverse roles of DNA methylation in mammalian de-
339 velopment and disease. *Nature reviews Molecular cell biology*. **20**: 590–607.
340 Gu T, Lin X, Cullen SM, Luo M, Jeong M, Estecio M, Shen J, Hardikar S, Sun D, Su J, et al.
341 2018. DNMT3A and TET1 cooperate to regulate promoter epigenetic landscapes in mouse
342 embryonic stem cells. *Genome biology*. **19**: 1–15.
343 Guo W, Wang D, and Lisch D. 2021. RNA-directed DNA methylation prevents rapid and heri-
344 table reversal of transposon silencing under heat stress in Zea mays. *PLoS Genetics*. **17**:
345 e1009326.
346 Gupta N, Yakhou L, Albert JR, Azogui A, Ferry L, Kirsh O, Miura F, Battault S, Yamaguchi K,
347 Laisné M, et al. 2023. A genome-wide screen reveals new regulators of the 2-cell-like cell
348 state. *Nature Structural & Molecular Biology*. **30**: 1105–1118.
349 Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, Chen GK, Wang K, Buxbaum
350 SG, Akylbekova EL, et al. 2011. The landscape of recombination in African Americans. *Na-*
351 *ture*. **476**: 170–175.
352 Imbeault M, Helleboid PY, and Trono D. 2017. KRAB zinc-finger proteins contribute to the evo-
353 lution of gene regulatory networks. *Nature*. **543**: 550–554.
354 Iqbal K, Jin SG, Pfeifer GP, and Szabó PE. 2011. Reprogramming of the paternal genome upon
355 fertilization involves genome-wide oxidation of 5-methylcytosine. *Proceedings of the National*
356 *Academy of Sciences*. **108**: 3642–3647.

Ivanova E, Canovas S, Garcia-Martínez S, Romar R, Lopes JS, Rizos D, Sanchez-Calabuig MJ, Krueger F, Andrews S, Perez-Sanz F, et al. 2020. DNA methylation changes during preimplantation development reveal inter-species differences and reprogramming events at imprinted genes. *Clinical epigenetics*. **12**: 1–18.

Jeltsch A, Adam S, Dukatz M, Emperle M, and Bashtrykov P. 2021. Deep enzymology studies on DNA methyltransferases reveal novel connections between flanking sequences and enzyme activity. *Journal of molecular biology*. **433**: 167186.

Ji Y, Zhou Z, Liu H, and Davuluri RV. 2021. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*. **37**: 2112–2120.

Krueger F and Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *bioinformatics*. **27**: 1571–1572.

Kuhlmann M, Jiang H, Catoni M, and Johannes F 2021. DNA methylation in plants associated with abiotic stress.

Kumar S and Mohapatra T. 2021. Dynamics of DNA methylation and its functions in plant growth and development. *Frontiers in Plant Science*. **12**: 596236.

Lea RA, McCarthy A, Boeing S, Fallesen T, Elder K, Snell P, Christie L, Adkins S, Shaikly V, Taranissi M, et al. 2021. KLF17 promotes human naı□ve pluripotency but is not required for its establishment. *Development*. **148**: dev199378.

Li E and Zhang Y. 2014. DNA methylation in mammals. *Cold Spring Harbor perspectives in biology*. **6**: a019133.

Li Ht, Liu Y, Liu H, and Sun X. 2020. Effect for human genomic variation during the BMP4-induced conversion from pluripotent stem cells to trophoblast. *Frontiers in Genetics*. **11**: 230.

Li Y, Fan Z, Meng Y, Liu S, and Zhan H. 2023. Blood-based DNA methylation signatures in cancer: A systematic review. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*. **1869**: 166583.

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, et al. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*. **454**: 766–770.

Okano M, Bell DW, Haber DA, and Li E. 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*. **99**: 247–257.

Pappalardo XG and Barra V. 2021. Losing DNA methylation at repetitive elements and breaking bad. *Epigenetics & chromatin*. **14**: 1–21.

Patel A, Hashimoto H, Zhang X, and Cheng X 2016. Characterization of how DNA modifications affect DNA binding by C2H2 zinc finger proteins. In: *Methods in enzymology*. Vol. 573. Elsevier, pp. 387–401.

Pells S, Koutsouraki E, Morfopoulou S, Valencia-Cadavid S, Tomlinson SR, Kalathur R, Futschik ME, and De Sousa PA. 2015. Novel human embryonic stem cell regulators identified by conserved and distinct CpG island methylation state. *PLoS One*. **10**: e0131102.

Ravichandran M, Rafalski D, Davies CI, Ortega-Recalde O, Nan X, Glanfield CR, Kotter A, Misztal K, Wang AH, Wojciechowski M, et al. 2022. Pronounced sequence specificity of the TET enzyme catalytic domain guides its cellular function. *Science Advances*. **8**: eabm2427.

Schmitges FW, Radovani E, Najafabadi HS, Barazandeh M, Campitelli LF, Yin Y, Jolma A, Zhong G, Guo H, Kanagalingam T, et al. 2016. Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome research*. **26**: 1742–1752.

Ségurel L, Leffler EM, and Przeworski M. 2011. The case of the fickle fingers: how the PRDM9 zinc finger protein specifies meiotic recombination hotspots in humans. *PLoS biology*. **9**: e1001211.

Sereshki S, Lee N, Omirou M, Fasoula D, and Lonardi S. 2023. On the prediction of non-CG DNA methylation using machine learning. *NAR genomics and bioinformatics*. **5**: lqad045.

Skvortsova K, Stirzaker C, and Taberlay P. 2019. The DNA methylation landscape in cancer. *Essays in biochemistry*. **63**: 797–811.

Spainhour JC, Lim HS, Yi SV, and Qiu P. 2019. Correlation patterns between DNA methylation and gene expression in the cancer genome atlas. *Cancer informatics*. **18**: 1176935119828776.

Takebayashi-Suzuki K, Konishi H, Miyamoto T, Nagata T, Uchida M, and Suzuki A. 2018. Coordinated regulation of the dorsal-ventral and anterior-posterior patterning of Xenopus embryos by the BTB/POZ zinc finger protein Zbtb14. *Development, Growth & Differentiation*. **60**: 158–173.

415 Wang W, Zhao M, Zuo H, Zhang J, Liu B, Chen F, Ji P, Liu G, Gao S, Shang W, et al. 2023. Evalu-
416 ate the developmental competence of human 8-cell embryos by single-cell RNA sequencing.
417 *Reproduction and Fertility*. **4**:
418 Wu H and Zhang Y. 2011. Mechanisms and functions of Tet protein-mediated 5-methylcytosine
419 oxidation. *Genes & development*. **25**: 2436–2452.
420 Xi Y and Li W. 2009. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioin-*
421 *formatics*. **10**: 232.
422 Xu W, Xu M, Wang L, Zhou W, Xiang R, Shi Y, Zhang Y, and Piao Y. 2019. Integrative anal-
423 ysis of DNA methylation and gene expression identified cervical cancer-specific diagnostic
424 biomarkers. *Signal transduction and targeted therapy*. **4**: 55.
425 Yamaguchi S, Hong K, Liu R, Shen L, Inoue A, Diep D, Zhang K, and Zhang Y. 2012. Tet1
426 controls meiosis by regulating meiotic gene expression. *Nature*. **492**: 443–447.
427 Zhang J, Sheng H, Hu C, Li F, Cai B, Ma Y, Wang Y, and Ma Y. 2023. Effects of DNA methylation
428 on gene expression and phenotypic traits in cattle: A review. *International Journal of Molecular*
429 *Sciences*. **24**: 11882.
430 Zhao SG, Chen WS, Li H, Foye A, Zhang M, Sjöström M, Aggarwal R, Playdle D, Liao A, Alumkal
431 JJ, et al. 2020. The DNA methylation landscape of advanced prostate cancer. *Nature genetics*.
432 **52**: 778–789.
433 Zhen J, Ke Y, Pan J, Zhou M, Zeng H, Song G, Yu Z, Fu B, Liu Y, Huang D, et al. 2022. ZNF320
434 is a hypomethylated prognostic biomarker involved in immune infiltration of hepatocellular
435 carcinoma and associated with cell cycle. *Aging (Albany NY)*. **14**: 8411.
436 Zhou W, Liang G, Molloy PL, and Jones PA. 2020. DNA methylation enables transposable element-
437 driven genome expansion. *Proceedings of the National Academy of Sciences*. **117**: 19359–
438 19366.
439 Zuo X, Sheng J, Lau HT, McDonald CM, Andrade M, Cullen DE, Bell FT, Iacovino M, Kyba M,
440 Xu G, et al. 2012. Zinc finger protein ZFP57 requires its co-factor to recruit DNA methyltrans-
441 ferases and maintains DNA methylation imprint in embryonic stem cells via its transcriptional
442 repression domain. *Journal of Biological Chemistry*. **287**: 2107–2118.