Predicting Differentially Methylated Cytosines in TET and DNMT3 Knockout Mutants via a Large Language Model, 2024, pp. 1–10

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

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

Saleh Sereshki[1] and Stefano Lonardi[1,*]

[1]Department of Computer Science and Engineering, University of California, Riverside, 900 University Ave, Riverside, 92521, CA, United States
[*]Corresponding author. stelo@cs.ucr.edu
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

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 in human and mouse 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, casting a new light on the understanding of DNA methylation dynamics in stem cells. Our software tool is available at `https://github.com/ucrbioinfo/dmc_prediction`.

Key words: 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., (14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26)). 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., (29; 30; 31; 32)).

In mammals, DNA methylation primarily occurs in CpG dinucleotides, with most of them being methylated (28). Mammalian genomes typically maintain consistent CpG methylation patterns over time, except in specific regulatory regions like promoters and certain types of enhancers (27). 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 (34; 33). 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 (35).

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 (36; 2). 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 (37).

TET and DNMT3s are crucial in regulating fetal organ development and tissue generation, through DNA methylation and histone modifications (55; 54; 56; 57; 58). Their dysregulation is linked to human diseases, particularly cancers (59; 60; 61; 62; 63). Although the importance of TETs is well recognized, their precise mechanisms of action is 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., (2)). Chao *et al.* also studied the interactions between TET1,

© The Author 2024. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

DNMT3A, and DNMT3B in human embryonic stem cells, and how these interactions collectively influence global methylation patterns (9).

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 (39) 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 (9) also reported that TET1 prefers binding to specific genomic regions. This appears to be also true for DNMT3. Jeltsch et al. (40) 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., (2; 4; 6)). 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.

A deeper understanding of cell functions can lead to significant advancements in medical research, therapeutic development, disease prevention, and diagnostic techniques, such as drug discoveries (64; 65; 66). Some studies have identified interacting partners for TETs and DNMT3s (67; 68; 69; 70). Here, we have identified transcription factor binding site (TFBS) 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 advancement in treatment strategies and novel drug discoveries.

Results

Seven data sets, across three studies, were used to train and test L-MAP. In the first study by Charlton *et al.* (2), 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.* (4) involved DNMT3A and DNMT3B knockout in mouse ESC. The third study by Ansari *et al.* (6) 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 (1) 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 (5) 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

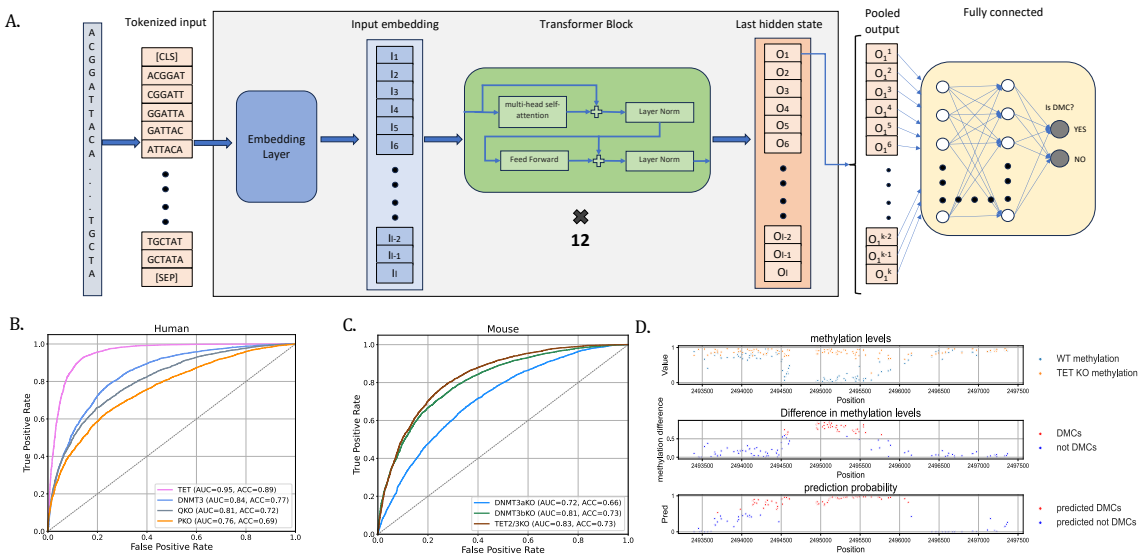


Fig. 1. **(A)** the architecture L-MAP; **(B)** ROC curves for the performance of L-MAP on human TET and DNMT3 knockout datasets (AUC=area under the curve, ACC=accuracy); **(C)** ROC curves for the performance of L-MAP on mouse TET and DNMT3 knockout datasets (AUC=area under the curve, ACC=accuracy); **(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

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 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 exceptions. 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.

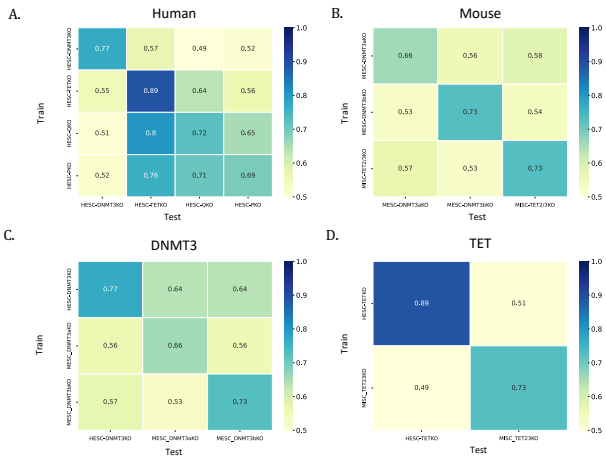


Fig. 2. (A) L-MAP's accuracy when trained on a human knockout dataset and tested on another human knockout dataset **(B)** L-MAP's accuracy when trained on one mouse knockout dataset and tested on another mouse knockout dataset **(C-D)** L-MAP's accuracy when trained on a human (mouse) knockout dataset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

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 (41), 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 (38) 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., (11; 12; 13; 3)). 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 (42). Most of the matching motif are also associated with 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 (43; 45; 46). Studies have shown that there is a link between PRDM9 activity and TET1 during meiosis in mice (44). 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 (47). 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 (48). 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 (49). 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 (50). 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 (52). 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 (51). 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 (53).

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 (67; 75; 74; 71; 72; 73). We expect that the other transcription factor 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 (76; 77; 78; 79; 80; 81). To investigate which positions in the input window are more important for the classification, we extracted the L-MAP's attention scores. Figure 10 and Figure 11 in the supplemental material show the attention scores within the input window for L-MAP on different data sets. Observe that L-MAP's strongest attention are the position flanking the center cytosine. Also observe that the attention is much stronger for the flanking positions for the DNMT3 data sets compared to 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 differentially methylated cytosines for the missing data. To test the extent of which imputation would be possible, we modified

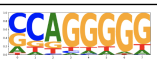
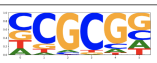
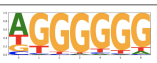
			JASPAR			
Knockout	Captured motif	p-value	Name	ID	Class	Family
HESC-TETKO		8.10E-144	PRDM9	MA1723.2	C2H2 zinc finger factors	Factors with multiple dispersed zinc fingers
HESC-DNMT3KO		7.70E-91	ZNF320	MA1976.2	C2H2 zinc finger factors	More than 3 adjacent zinc fingers
HESC-QKO		1.90E-24	ZBTB14	MA1650.2	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

Fig. 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

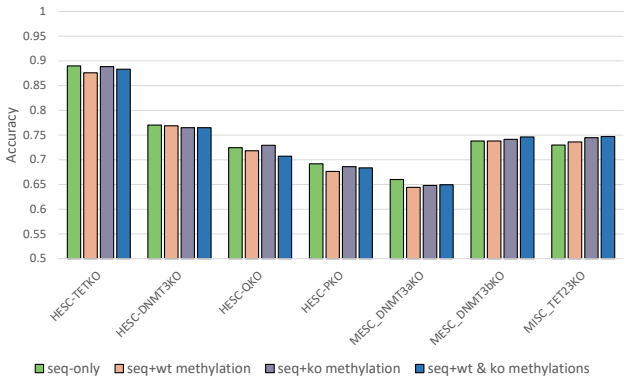


Fig. 4. Effect of including neighboring cytosine methylation levels on L-MAP's prediction accuracy for DMCs in seven knockout datasets

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. 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 discover associated sequence motifs opens up opportunities for advancing our understanding of epigenetic regulation in various cellular processes.

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 vs. mouse, or TET vs DNMT3)
- L-MAP can be used to discover new transcription factor associated with TET and DNMT3

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 (2); (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 (4); (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 (6). All the datasets are publicly available from NCBI. We note that in all these datasets, due to the choice of the protocol used to carry out bisufite-treated sequencing, only the methylation levels for the forward strand is available.

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 (2). 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 (10) 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 (41) 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 (38).

Data access

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`

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

Competing interests

The authors declare that they have no competing interests.

Funding

This project was supported in part by NIH 1R01AI169543-01 and NSF CBET 2225878.

Acknowledgements

The authors wish to thank Daniel Koenig (UC Riverside) and Jikui Song (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

1. Yuanxin Xi and Wei Li. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, 10(1):232, 2009.
2. Jocelyn Charlton, Eunmi J Jung, Alexandra L Mattei, Nina Bailly, Jing Liao, Eric J Martin, Pay Giesselmann, Björn Brändl, Elena K Stamenova, Franz-Josef Müller, and others. TETs compete with DNMT3 activity in pluripotent cells at thousands of methylated somatic enhancers. *Nature Genetics*, 52(8):819–827, 2020.
3. Saleh Sereshki, Nathan Lee, Michalis Omirou, Dionysia Fasoula, and Stefano Lonardi. On the prediction of non-CG DNA methylation using machine learning. *NAR Genomics and Bioinformatics*, 5(2):lqad045, 2023.
4. Tianpeng Gu, Xueqiu Lin, Sean M Cullen, Min Luo, Mira Jeong, Marcos Estecio, Jianjun Shen, Swanand Hardikar, Deqiang Sun, Jianzhong Su, and others. DNMT3A and TET1 cooperate to regulate promoter epigenetic landscapes in mouse embryonic stem cells. *Genome Biology*, 19:1–15, 2018.
5. Felix Krueger and Simon R Andrews. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11):1571–1572, 2011.
6. Ihab Ansari, Llorenç Solé-Boldo, Meshi Ridnik, Julian Gutekunst, Oliver Gilliam, Maria Korshko, Timur Liwinski, Birgit Jickeli, Noa Weinberg-Corem, Michal Shoshkes-Carmel, and others. TET2 and TET3 loss disrupts small intestine differentiation and homeostasis. *Nature Communications*, 14(1):4005, 2023.
7. Paul Adrian Ginno, Dimos Gaidatzis, Angelika Feldmann, Leslie Hoerner, Dilek Imanci, Lukas Burger, Frederic Zilbermann, Antoine HFM Peters, Frank Edenhofer, Sébastien A Smallwood, and others. A genome-scale map of DNA methylation turnover identifies site-specific dependencies of DNMT and TET activity. *Nature Communications*, 11(1):2680, 2020.
8. Charalampos Kyriakopoulos, Karl Nordström, Paula Linh Kramer, Judith Yumiko Gottfreund, Abdulrahman Salhab, Julia Arand, Fabian Müller, Ferdinand von Meyenn, Gabriella Ficz, Wolf Reik, and others. A comprehensive approach for genome-wide efficiency profiling of DNA modifying enzymes. *Cell Reports Methods*, 2(3), 2022.
9. Lemuge Chao, Siqi Yang, Hanshuang Li, Chunshen Long, Qilemuge Xi, and Yongchun Zuo. Competitive binding of TET1 and DNMT3A/B cooperates the DNA methylation pattern in human embryonic stem cells. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1865(7):194861, 2022.
10. Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.
11. Frank W Schmitges, Ernest Radovani, Hamed S Najafabadi, Marjan Barazandeh, Laura F Campitelli, Yimeng Yin, Arttu Jolma, Guoqing Zhong, Hongbo Guo, Tharsan Kanagalingam, and others. Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Research*, 26(12):1742–1752, 2016.
12. Michaël Imbeault, Pierre-Yves Helleboid, and Didier Trono. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*, 543(7646):550–554, 2017.
13. A Patel, H Hashimoto, X Zhang, and X Cheng. Characterization of how DNA modifications affect DNA binding by C2H2 zinc finger proteins. In *Methods in Enzymology*, volume 573, pages 387–401, 2016.
14. Maxim VC Greenberg and Deborah Bourc'his. The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology*, 20(10):590–607, 2019.
15. Wanxue Xu, Mengyao Xu, Longlong Wang, Wei Zhou, Rong Xiang, Yi Shi, Yunshan Zhang, and Yongjun Piao. Integrative analysis of DNA methylation and gene expression identified cervical cancer-specific diagnostic biomarkers. *Signal Transduction and Targeted Therapy*, 4(1):55, 2019.
16. Michael Ackah, Liangliang Guo, Shaocong Li, Xin Jin, Charles Asakiya, Evans Tawiah Aboagye, Feng Yuan, Mengmeng Wu, Lionnelle Gyllye Essoh, Daniel Adjibolosoo, and others. DNA methylation changes and its associated genes in mulberry (Morus alba L.) Yu-711 response to drought stress using MethylRAD sequencing. *Plants*, 11(2):190, 2022.
17. John CG Spainhour, Hong Seo Lim, Soojin V Yi, and Peng Qiu. Correlation patterns between DNA methylation and gene expression in the cancer genome atlas. *Cancer Informatics*, 18:1176935119828776, 2019.
18. Markus Kuhlmann, Hua Jiang, Marco Catoni, and Frank Johannes. DNA methylation in plants associated with abiotic stress. *Frontiers in Plant Science*, 12:778004, 2021.
19. Lauren E Blake, Julien Roux, Irene Hernando-Herraez, Nicholas E Banovich, Raquel Garcia Perez, Chiaowen Joyce Hsiao, Ittai Eres, Claudia Cuevas, Tomas Marques-Bonet, and Yoav Gilad. A comparison of gene expression and DNA methylation patterns across tissues and species. *Genome Research*, 30(2):250–262, 2020.
20. Suresh Kumar and Trilochan Mohapatra. Dynamics of DNA methylation and its functions in plant growth and development. *Frontiers in Plant Science*, 12:596236, 2021.
21. Xena Giada Pappalardo and Viviana Barra. Losing DNA methylation at repetitive elements and breaking bad. *Epigenetics & Chromatin*, 14(1):1–21, 2021.
22. Reza Bozorgpour. Computational explorations in biomedicine: Unraveling molecular dynamics for cancer, drug delivery, and biomolecular insights using LAMMPS simulations. *arXiv preprint arXiv:2311.13000*, 2023.
23. Junxing Zhang, Hui Sheng, Chunli Hu, Fen Li, Bei Cai, Yanfen Ma, Yachun Wang, and Yun Ma. Effects of DNA methylation on gene expression and phenotypic traits in

cattle: A review. *International Journal of Molecular Sciences*, 24(15):11882, 2023.

24. Wei Guo, Dafang Wang, and Damon Lisch. RNA-directed DNA methylation prevents rapid and heritable reversal of transposon silencing under heat stress in Zea mays. *PLoS Genetics*, 17(6):e1009326, 2021.
25. Wanding Zhou, Gangning Liang, Peter L Molloy, and Peter A Jones. DNA methylation enables transposable element-driven genome expansion. *Proceedings of the National Academy of Sciences*, 117(32):19359–19366, 2020.
26. Elena Ivanova, Sebastian Canovas, Soledad Garcia-Martínez, Raquel Romar, Jordana S Lopes, Dimitrios Rizos, Maria J Sanchez-Calabuig, Felix Krueger, Simon Andrews, Fernando Perez-Sanz, and others. DNA methylation changes during preimplantation development reveal inter-species differences and reprogramming events at imprinted genes. *Clinical Epigenetics*, 12:1–18, 2020.
27. Alexander Meissner, Tarjei S Mikkelsen, Hongcang Gu, Marius Wernig, Jacob Hanna, Andrey Sivachenko, Xiaolan Zhang, Bradley E Bernstein, Chad Nusbaum, David B Jaffe, and others. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770, 2008.
28. En Li and Yi Zhang. DNA methylation in mammals. *Cold Spring Harbor Perspectives in Biology*, 6(5):a019133, 2014.
29. Ksenia Skvortsova, Clare Stirzaker, and Phillippa Taberlay. The DNA methylation landscape in cancer. *Essays in Biochemistry*, 63(6):797–811, 2019.
30. Yongzheng Li, Zhiyao Fan, Yufan Meng, Shujie Liu, and Hanxiang Zhan. Blood-based DNA methylation signatures in cancer: A systematic review. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1869(1):166583, 2023.
31. Shuang G Zhao, William S Chen, Haolong Li, Adam Foye, Meng Zhang, Martin Sjöström, Rahul Aggarwal, Denise Playdle, Arnold Liao, Joshi J Alumkal, and others. The DNA methylation landscape of advanced prostate cancer. *Nature Genetics*, 52(8):778–789, 2020.
32. Zahra Anvar, Imen Chakchouk, Hannah Demond, Momal Sharif, Gavin Kelsey, and Ignatia B Van den Veyver. DNA methylation dynamics in the female germline and maternal-effect mutations that disrupt genomic imprinting. *Genes*, 12(8):1214, 2021.
33. Linfeng Gao, Max Emperle, Yiran Guo, Sara A Grimm, Wendan Ren, Sabrina Adam, Hidetaka Uryu, Zhi-Min Zhang, Dongliang Chen, Jiekai Yin, and others. Comprehensive structure-function characterization of DNMT3B and DNMT3A reveals distinctive de novo DNA methylation mechanisms. *Nature Communications*, 11(1):3355, 2020.
34. Masaki Okano, Daphne W Bell, Daniel A Haber, and En Li. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257, 1999.
35. Hao Wu and Yi Zhang. Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. *Genes & Development*, 25(23):2436–2452, 2011.
36. Riccardo M Betto, Linda Diamante, Valentina Perrera, Matteo Audano, Stefania Rapelli, Andrea Lauria, Danny Incarnato, Mattia Arboit, Silvia Pedretti, Giovanni Rigoni, and others. Metabolic control of DNA methylation in naive pluripotent cells. *Nature Genetics*, 53(2):215–229, 2021.
37. Khursheed Iqbal, Seung-Gi Jin, Gerd P Pfeifer, and Piroska E Szabó. Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *Proceedings of the National Academy of Sciences*, 108(9):3642–3647, 2011.
38. Jaime A Castro-Mondragon, Rafael Riudavets-Puig, Ieva Rauluseviciute, Roza Berhanu Lemma, Laura Turchi, Romain Blanc-Mathieu, Jeremy Lucas, Paul Boddie, Aziz Khan, Nicolás Manosalva Pérez, Oriol Fornes, Tiffany Y Leung, Alejandro Aguirre, Fayrouz Hammal, Daniel Schmelter, Damir Baranasic, Benoit Ballester, Albin Sandelin, Boris Lenhard, Klaas Vandepoele, Wyeth W Wasserman, François Parcy, and Anthony Mathelier. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 50(D1):D165–D173, 2021.
39. Mirunalini Ravichandran, Dominik Rafalski, Claudia I Davies, Oscar Ortega-Recalde, Xinsheng Nan, Cassandra R Glanfield, Annika Kotter, Katarzyna Misztal, Andrew H Wang, Marek Wojciechowski, and others. Pronounced sequence specificity of the TET enzyme catalytic domain guides its cellular function. *Science Advances*, 8(36):eabm2427, 2022.
40. Albert Jeltsch, Sabrina Adam, Michael Dukatz, Max Emperle, and Pavel Bashtrykov. Deep enzymology studies on DNA methyltransferases reveal novel connections between flanking sequences and enzyme activity. *Journal of Molecular Biology*, 433(19):167186, 2021.
41. Timothy L Bailey. STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, 37(18):2834–2840, 2021.
42. Xiaopan Zuo, Jipo Sheng, Ho-Tak Lau, Carol M McDonald, Monica Andrade, Dana E Cullen, Fong T Bell, Michelina Iacovino, Michael Kyba, Guoliang Xu, and others. 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. *Journal of Biological Chemistry*, 287(3):2107–2118, 2012.
43. Laure Ségurel, Ellen Miranda Leffler, and Molly Przeworski. The case of the fickle fingers: how the PRDM9 zinc finger protein specifies meiotic recombination hotspots in humans. *PLoS Biology*, 9(12):e1001211, 2011.
44. Shinpei Yamaguchi, Kwonho Hong, Rui Liu, Li Shen, Azusa Inoue, Dinh Diep, Kun Zhang, and Yi Zhang. Tet1 controls meiosis by regulating meiotic gene expression. *Nature*, 492(7429):443–447, 2012.
45. Ingrid L Berg, Rita Neumann, Kwan-Wood G Lam, Shriparna Sarbajna, Linda Odenthal-Hesse, Celia A May, and Alec J Jeffreys. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nature Genetics*, 42(10):859–863, 2010.
46. Anjali G Hinch, Arti Tandon, Nick Patterson, Yunli Song, Nadin Rohland, Cameron D Palmer, Gary K Chen, Kai Wang, Sarah G Buxbaum, Ermeg L Akylbekova, and others. The landscape of recombination in African Americans. *Nature*, 476(7359):170–175, 2011.
47. Jing Zhen, Yun Ke, Jingying Pan, Minqin Zhou, Hong Zeng, Gelin Song, Zichuan Yu, Bidong Fu, Yue Liu, Da Huang, and others. ZNF320 is a hypomethylated prognostic biomarker involved in immune infiltration of hepatocellular carcinoma and associated with cell cycle. *Aging (Albany NY)*, 14(20):8411, 2022.
48. Kimiko Takebayashi-Suzuki, Hidenori Konishi, Tatsuo Miyamoto, Tomoko Nagata, Misa Uchida, and Atsushi Suzuki. 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(3):158–173, 2018.

49. Nikhil Gupta, Lounis Yakhou, Julien Richard Albert, Anaelle Azogui, Laure Ferry, Olivier Kirsh, Fumihito Miura, Sarah Battault, Kosuke Yamaguchi, Marthe Laisné, and others. A genome-wide screen reveals new regulators of the 2-cell-like cell state. *Nature Structural & Molecular Biology*, 30(8):1105–1118, 2023.
50. Steve Pells, Eirini Koutsouraki, Sofia Morfopoulou, Sara Valencia-Cadavid, Simon R Tomlinson, Ravi Kalathur, Matthias E Futschik, and Paul A De Sousa. Novel human embryonic stem cell regulators identified by conserved and distinct CpG island methylation state. *PLoS One*, 10(7):e0131102, 2015.
51. Weizhou Wang, Mengmeng Zhao, Haiyang Zuo, Jingyao Zhang, Bin Liu, Fu Chen, Pengyun Ji, Guoshi Liu, Shuai Gao, Wei Shang, and others. Evaluate the developmental competence of human 8-cell embryos by single-cell RNA sequencing. *Reproduction and Fertility*, 4(2), 2023.
52. Hai-tao Li, Yajun Liu, Hongde Liu, and Xiao Sun. Effect for human genomic variation during the BMP4-induced conversion from pluripotent stem cells to trophoblast. *Frontiers in Genetics*, 11:230, 2020.
53. Rebecca A Lea, Afshan McCarthy, Stefan Boeing, Todd Fallesen, Kay Elder, Phil Snell, Leila Christie, Sarah Adkins, Valerie Shaikly, Mohamed Taranissi, and others. KLF17 promotes human naïve pluripotency but is not required for its establishment. *Development*, 148(22):dev199378, 2021.
54. Daniel N Weinberg, Simon Papillon-Cavanagh, Haifen Chen, Yuan Yue, Xiao Chen, Kartik N Rajagopalan, Cynthia Horth, John T McGuire, Xinjing Xu, Hamid Nikbakht, and others. The histone mark H3K36me2 recruits DNMT3A and shapes the intergenic DNA methylation landscape. *Nature*, 573(7773):281–286, 2019.
55. Chan-Wang Lio, Jiayuan Zhang, Edahí González-Avalos, Patrick G Hogan, Xing Chang, and Anjana Rao. Tet2 and Tet3 cooperate with B-lineage transcription factors to regulate DNA modification and chromatin accessibility. *Elife*, 5:e18290, 2016.
56. Simon Andrews, Christel Krueger, Maravillas Mellado-Lopez, Myriam Hemberger, Wendy Dean, Vicente Perez-Garcia, and Courtney W Hanna. Mechanisms and function of de novo DNA methylation in placental development reveals an essential role for DNMT3B. *Nature Communications*, 14(1):371, 2023.
57. Xin Huang, Sophie Balmer, Cong Lyu, Yunlong Xiang, Vikas Malik, Hailin Wang, Yu Zhang, Bishuang Cai, Wei Xie, Anna-Katerina Hadjantonakis, and others. ZFP281 controls transcriptional and epigenetic changes promoting mouse pluripotent state transitions via DNMT3 and TET1. *Developmental Cell*, 59(4):465–481, 2024.
58. Ageliki Tsagaratou. TET proteins in the spotlight: emerging concepts of epigenetic regulation in T cell biology. *Immunohorizons*, 7(1):106–115, 2023.
59. Eric Genaro Salmerón-Bárcenas, Ana Elvira Zacapala-Gómez, Francisco Israel Torres-Rojas, Verónica Antonio-Véjar, Pedro Antonio Ávila-López, Christian Johana Baños-Hernández, Hober Nelson Núñez-Martínez, Roberto Dircio-Maldonado, Dinorah Nashely Martínez-Carrillo, Julio Ortiz-Ortiz, and others. TET enzymes and 5hmC levels in carcinogenesis and progression of breast cancer: potential therapeutic targets. *International Journal of Molecular Sciences*, 25(1):272, 2023.
60. Jiayu Zhang, Cheng Yang, Chunfu Wu, Wei Cui, and Lihui Wang. DNA methyltransferases in cancer: biology, paradox, aberrations, and targeted therapy. *Cancers*, 12(8):2123, 2020.
61. Diana L Christian, Dennis Y Wu, Jenna R Martin, J Russell Moore, Yiran R Liu, Adam W Clemens, Sabin A Nettles, Nicole M Kirkland, Thomas Papouin, Cheryl A Hill, and others. DNMT3A haploinsufficiency results in behavioral deficits and global epigenomic dysregulation shared across neurodevelopmental disorders. *Cell Reports*, 33(8), 2020.
62. Xiufei Chen, Wenjie Zhou, Ren-Hua Song, Shuang Liu, Shu Wang, Yujia Chen, Chao Gao, Chenxi He, Jianxiong Xiao, Lei Zhang, and others. Tumor suppressor CEBPA interacts with and inhibits DNMT3A activity. *Science Advances*, 8(4):eabl5220, 2022.
63. Bo Xu, Hao Wang, and Li Tan. Dysregulated TET family genes and aberrant 5mC oxidation in breast cancer: causes and consequences. *Cancers*, 13(23):6039, 2021.
64. Nessa Carey, C Joana Marques, and Wolf Reik. DNA demethylases: a new epigenetic frontier in drug discovery. *Drug Discovery Today*, 16(15-16):683–690, 2011.
65. Robert Roskoski Jr. Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes. *Pharmacological Research*, 103:26–48, 2016.
66. Tarikul I Milon, Yuhong Wang, Ryan L Fontenot, Poorya Khajouie, Francois Villinger, Vijay Raghavan, and Wu Xu. Development of a novel representation of drug 3D structures and enhancement of the TSR-based method for probing drug and target interactions. *Computational Biology and Chemistry*, page 108117, 2024.
67. Dimitris Theofilatos, Tricia Ho, Greg Waitt, Tarmo Äijö, Lucio M Schiapparelli, Erik J Soderblom, and Ageliki Tsagaratou. Deciphering the TET3 interactome in primary thymic developing T cells. *Iscience*, 27(5), 2024.
68. Falong Lu, Yuting Liu, Lan Jiang, Shinpei Yamaguchi, and Yi Zhang. Role of Tet proteins in enhancer activity and telomere elongation. *Genes & Development*, 28(19):2103–2119, 2014.
69. Jason M Foulks, K Mark Parnell, Rebecca N Nix, Suzanna Chau, Krzysztof Swierczek, Michael Saunders, Kevin Wright, Thomas F Hendrickson, Koc-Kan Ho, Michael V McCullar, and others. Epigenetic drug discovery: targeting DNA methyltransferases. *Journal of Biomolecular Screening*, 17(1):2–17, 2012.
70. Michael Brauchle, Zhiping Yao, Rishi Arora, Sachin Thigale, Ieuan Clay, Bruno Inverardi, Joy Fletcher, Paul Taslimi, Michael G Acker, Bertran Gerrits, and others. Protein complex interactor analysis and differential activity of KDM3 subfamily members towards H3K9 methylation. *PloS One*, 8(4):e60549, 2013.
71. Ryan J Marina and Shalini Oberdoerffer. Epigenomics meets splicing through the TETs and CTCF. *Cell Cycle*, 15(11):1397–1399, 2016.
72. Laura Wiehle, Graeme J Thorn, Günter Raddatz, Christopher T Clarkson, Karsten Rippe, Frank Lyko, Achim Breiling, Vladimir B Teif. DNA (de)methylation in embryonic stem cells controls CTCF-dependent chromatin boundaries. *Genome Research*, 29(5):750–761, 2019.
73. Ryan J Marina, David Sturgill, Marc A Bailly, Morgan Thenoz, Garima Varma, Maria F Prigge, Kyster K Nanan, Sanjeev Shukla, Nazmul Haque, and Shalini Oberdoerffer. TET-catalyzed oxidation of intragenic 5-methylcytosine

regulates CTCF-dependent alternative splicing. *The EMBO Journal*, 35(3):335–355, 2016.

74. Julie Dubois-Chevalier, Frédérik Oger, Hélène Dehondt, François Firmin, Celine Gheeraert, Bart Staels, Philippe Lefebvre, and Jérôme Eeckhoute. A dynamic CTCF chromatin binding landscape promotes DNA hydroxymethylation and transcriptional induction of adipocyte differentiation. *Nucleic Acids Research*, 42(17):10943–10959, 2014.
75. Kyster K Nanan, David M Sturgill, Maria F Prigge, Morgan Thenoz, Allissa A Dillman, Mariana D Mandler, and Shalini Oberdoerffer. TET-catalyzed 5-carboxylcytosine promotes CTCF binding to suboptimal sequences genome-wide. *Iscience*, 19:326–339, 2019.
76. Vikas Handa and Albert Jeltsch. Profound flanking sequence preference of Dnmt3a and Dnmt3b mammalian DNA methyltransferases shape the human epigenome. *Journal of Molecular Biology*, 348(5):1103–1112, 2005.
77. Reza Bozorgpour, Sana Sheybanikashani, and Matin Mohebi. Exploring the role of molecular dynamics simulations in most recent cancer research: Insights into treatment strategies. *arXiv preprint arXiv:2310.19950*, 2023.
78. Michael Dukatz, Marianna Dittrich, Elias Stahl, Alex De Mendoza, Pavel Bashtrykov, and Albert Jeltsch. DNA methyltransferase DNMT3A forms interaction networks with the CpG site and flanking sequence elements for efficient methylation. *Journal of Biological Chemistry*, 298(10), 2022.
79. Shi-Qing Mao, Sergio Martínez Cuesta, David Tannahill, and Shankar Balasubramanian. Genome-wide DNA methylation signatures are determined by DNMT3A/B sequence preferences. *Biochemistry*, 59(27):2541–2550, 2020.
80. Zhi-Min Zhang, Rui Lu, Pengcheng Wang, Yang Yu, Dongliang Chen, Linfeng Gao, Shuo Liu, Debin Ji, Scott B Rothbart, Yinsheng Wang, and others. Structural basis for DNMT3A-mediated de novo DNA methylation. *Nature*, 554(7692):387–391, 2018.
81. Sabrina Adam, Hiwot Anteneh, Maximilian Hornisch, Vincent Wagner, Jiuwei Lu, Nicole E Radde, Pavel Bashtrykov, Jikui Song, and Albert Jeltsch. DNA sequence-dependent activity and base flipping mechanisms of DNMT1 regulate genome-wide DNA methylation. *Nature Communications*, 11(1):3723, 2020.

Saleh Sereshki is a Ph.D. candidate at the University of California, Riverside. He began his Ph.D. in Computer Science in 2019. His research focuses on applying machine learning techniques to genetic and epigenetic data.

Stefano Lonardi is Professor and Vice Chair of the Department of Computer Science and Engineering at the University of California, Riverside. His research interests include computational biology, bioinformatics, data mining, and epigenetics.