ARTICLES
https://doi.org/10.1038/s42256-020-0211-4
nature machine intelligence

Check for updates

Elucidation of DNA methylation on N^6-adenine with deep learning

Fei Tan[1,2,8], Tian Tian[1,3,8], Xiurui Hou[1], Xiang Yu[4], Lei Gu[5,6], Fernanda Mafra[3], Brian D. Gregory[4], Zhi Wei[1]✉ and Hakon Hakonarson[3,7]✉

Research on DNA methylation on N^6-adenine (6mA) in eukaryotes has received much recent attention. Recent studies have generated a large amount of 6mA genomic data, yet the role of DNA 6mA in eukaryotes remains elusive, or even controversial. We argue that the sparsity of DNA 6mA in eukaryotes, the limitations of current biotechnologies for 6mA detection and the sophistication of the 6mA regulatory mechanism together pose great challenges for elucidation of DNA 6mA. To exploit existing 6mA genomic data and address this challenge, here we develop a deep-learning-based algorithm for predicting potential DNA 6mA sites de novo from sequence at single-nucleotide resolution, with application to three representative model organisms, *Arabidopsis thaliana*, *Drosophila melanogaster* and *Escherichia coli*. Extensive experiments demonstrate the accuracy of our algorithm and its superior performance compared with conventional *k*-mer-based approaches. Furthermore, our saliency maps-based context analysis protocol reveals interesting *cis*-regulatory patterns around the 6mA sites that are missed by conventional motif analysis. Our proposed analytical tools and findings will help to elucidate the regulatory mechanisms of 6mA and benefit the in-depth exploration of their functional effects. Finally, we offer a complete catalogue of potential 6mA sites based on in silico whole-genome prediction.

DNA methylation is extensively involved in epigenetic settings and exerts different regulatory roles in multiple species[1–3]. It is traditionally acknowledged that 5-methylcytosine (5mC) presents a dominant modification in eukaryotes, while N^6-methyladenine (6mA) is mostly prevalent in prokaryotes[4]. Recently, thanks to the development of high-throughput sequencing (6mA-IP-seq) and single-molecule real-time (SMRT) sequencing technology, the prevalence and significance of DNA 6mA in eukaryotes (for example, *Arabidopsis thaliana* and *Drosophila melanogaster*) has been revealed[5–9]. However, DNA 6mA is a dynamic process, which can be developmental and tissue-specific[2]. In addition, many 6mA sites may be methylated at very low levels, making them very hard to capture. Consequently, current experimental approaches, although precise, are unable to provide a complete catalogue of all 6mA sites.

It has been long recognized that 6mA plays a vital role in the discrimination of host genomic DNA (gDNA) from foreign pathogenic DNA in bacteria[2,10]. Recently, it has been demonstrated that 6mA may be involved in gene activation or repression in eukaryotes[6,9]. The underlying mechanism, however, remains elusive. Methylation-associated gene regulatory motifs may shed light on understanding the mechanism. Although conventional motif analysis of 6mA has revealed some interesting *cis*-regulatory patterns, they account for only a small proportion of all methylated sites[2,11]. Therefore, we hypothesize that more sophisticated regulatory mechanisms yet to be explored may exist for 6mA formulation. Finally, the whole in vivo cataloguing procedure of 6mA is costly and laborious. Thus, in silico prediction may be an attractive alternative if we can precisely predict 6mA sites at single-nucleotide resolution based on just genomic sequence information.

It is worth noting that our proposed prediction is purely based on sequence information where only the *cis* effect will be captured. Whether a candidate is a 6mA site or not will also depend on many other exogenous *trans* effects. Therefore, what our method predicts is the candidacy or potential for being a 6mA site. Most of the methylation data we used in our model were collected without any information about developmental stages and tissue cells. We thus cannot make any developmental and tissue-specific 6mA prediction for them. When developmental stages are provided, our models are readily adapted to make the corresponding prediction. Our interest is to predict the candidacy or potential for being a 6mA site, which implies a necessary condition but not a sufficient condition. This 6mA prediction is quite similar to gene prediction or gene finding in the early bioinformatics era, which refers to the process of identifying the regions of gDNA that encode genes. Most gene prediction methods utilize DNA sequence information only, as does our method. The analogy is illustrated in Table 1. To predict 6mA candidate sites, we first developed a deep convolutional neural networks-based[12,13] end-to-end algorithmic framework (Supplementary Fig. 1 and Methods) to capture sophisticated regulatory patterns for predicting 6mA sites de novo from genomic sequences (DeepM6A). Machine learning methods have been used for genomic sequence-based prediction. Most reply on human handcrafted features, such as *k*-mer for predicting mutation effects[14] and polyadenylation sites[15], among others. Compared with *k*-mer-based methods, our proposed DeepM6A has four major advantages: automation of the sequence feature representation of different granularities, hierarchically; integration of a broad spectrum of flanking context sequences, effectively; enabling of the potential visualization of inherent sequence motifs for interpretation, naturally; facilitation

[1]Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA. [2]Yahoo! Research, New York, NY, USA. [3]The Center for Applied Genomics, Abramson Research Center, The Children's Hospital of Philadelphia, Philadelphia, PA, USA. [4]Department of Biology, University of Pennsylvania, Philadelphia, PA, USA. [5]Division of Newborn Medicine, Children's Hospital Boston, Boston, MA, USA. [6]Department of Cell Biology, Harvard Medical School, Boston, MA, USA. [7]Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, PA, USA. [8]These authors contributed equally: Fei Tan, Tian Tian. ✉e-mail: zhiwei@njit.edu; hakonarson@email.chop.edu

Table 1 | The analogy between gene prediction and 6mA prediction

Gene prediction	6mA prediction
Gene region	6mA sites
Genes get transcribed/expressed	6mA sites get methylated

of model development and prediction in large-scale genomic data, seamlessly. The multilayer perceptron (MLP) is a classical feed-forward neural network with an input layer, hidden layers and an output layer. The hidden layers are composed of fully connected nodes. DeepM6A leverages the local spatial dependency of inputs by introducing convolutional layers into MLP to capture underlying coherence patterns.

The first two desirable properties jointly contribute to the appealing predictive capacity of DeepM6A. Based on the third property, we then introduced a novel learning protocol to decode the underlying methylation patterns. Both *cis*-regulatory elements and regions are identified, which will offer useful insights into the in-depth exploration of the underlying formulating and regulatory mechanisms of 6mA. Exploiting its accurate prediction, we performed a whole-genome scan using DeepM6A to catalogue all potential 6mA sites.

DeepM6A accurately predicts 6mA candidate sites

We tested DeepM6A in three representative model organisms: *A. thaliana* (eukaryote, plant), *D. melanogaster* (eukaryote) and *Escherichia coli* (prokaryote). As a benchmark, classical *k*-mer based logistic regression (LR) was also evaluated.

The raw SMRT-seq data of *A. thaliana*, *D. melanogaster* and *E. coli* were collected from the PacBio public database. Base modification detection was performed to generate an initial set of 6mA sites following the automated data analysis workflows recommended by PacBio (see Methods for details). To reduce false positives, we further filtered out the following candidates: (1) candidates with any sequence variance located between 10 bp upstream and 5 bp downstream of the identified modification site; (2) candidates where the variation of estimated methylation level is greater than 30%. As a result, we ended up with 19,632, 10,653 and 33,700 6mA sites for *A. thaliana*, *D. melanogaster* and *E. coli*, respectively. These sites account for 0.025696% (total of 76,401,454), 0.013418% (79,393,495) and 1.475402% (2,284,124) of whole-genome adenine sites.

The above 6mA sites were used as positive samples in prediction models. To generate negative samples, we randomly sampled the same numbers of non-methylated adenine sites from the whole-genome sequences. At the same time, for the sampled negative non-methylated sites, we required that their distance to any positive methylated site be at least 200 bp away. Then, for both positive and negative samples, contextual sequences around the adenine site at each side were extracted as the input for predictive models. We considered lengths of flanking sequences from 3 bp to 200 bp.

We divided all positive and negative samples into three sets for training, validation and testing, respectively, based on their genomic locations. Specifically, for each chromosome of a species, we split it into 10 equal segments (Supplementary Fig. 2 and Methods). We then randomly picked one segment and used the samples within that segment for testing (blue). The samples on the nearest half upstream and half downstream segments were used for validation (yellow). The rest of all sites were used for training purpose (green). In this manner, we achieved a ratio of 8:1:1 among training, validation and testing datasets, with the training and testing parts strictly non-overlapping. Taking the +/− 30-bp flanking sequences as input, DeepM6A is capable of accurately predicting 6mA sites with average areas under the receiver operating characteristic curve (AUC) of 0.9564, 0.9637 and 0.9994 for *A. thaliana*, *D. melanogaster* and *E. coli*, respectively (Fig. 1a), as evaluated by the holdout testing genomic sequences (see Methods). The salient performance differences among the three model organisms indicate the more challenging task of identifying 6mA sites and associated sophisticated patterns in advanced eukaryotes than in primitive prokaryotes.

DeepM6A effectively exploits signal information from context

We varied the contextual lengths of the input sequence from 3 bp to 200 bp to demonstrate the capability of DeepM6A in exploiting signal information from contextual sequences. With varied input sequence lengths, DeepM6A consistently outperformed LR (Fig. 1b). Specifically, the performance of DeepM6A improves with increasing length and reaches a plateau after ~10 bp for both *A. thaliana* and *D. melanogaster*. By contrast, with LR, although increasing length is initially beneficial, it has the opposite effect after 7 bp. This finding confirms that the immediate up/downstream 7–10 bp region of the 6mA site is critical[16]. However, there may be additional subtle and/or sophisticated signals beyond the 10-bp position. This distant signal can be captured by DeepM6A, as indicated by its increased performance, whereas the extended region proves deleterious for the *k*-mer-based approach. We attribute the superiority and robustness of DeepM6A to its hierarchical representation of regulatory patterns and suggest that the *k*-mer based methods are suffering from their inherent drawbacks of handcrafted feature extraction.

DeepM6A is sensitive at single-nucleotide resolution

The non-N^6-methylated adenines in the control cohort are at least 200 bp away from any 6mA. To show the robustness of DeepM6A with single-nucleotide sensitivity, for each 6mA site, we selected its closest non-N^6-methylated adenine and built a new control cohort (>75% fall within 5 bp of 6mA sites; Supplementary Fig. 3). At the contextual length of 30 bp, there is a substantial overlap between cases and the new controls, making separation of 6mA from the control more challenging. We also re-evaluated previously trained models. The performance of DeepM6A drops a little, but remains consistently high, while the performance of LR deteriorates substantially (Fig. 1c). This robustness of DeepM6A advocates its application to 6mA prediction at single-nucleotide resolution.

DeepM6A outperforms standard deep-learning approaches

In addition to the classical *k*-mer-based LR, we also compared DeepM6A with a standard MLP network[17] (see Methods) that uses the same input as our method for predicting N^6-methyladenine sites. The input of MLP is also the one-hot encoding of nucleotides centred on the target adenine with different lengths of flanking sequences. The training, validation and testing procedures exactly followed the way DeepM6A was optimized. The predictive capacity under different contextual sequences and the robustness of single-nucleotide sensitivity are reported in Fig. 1b,c. DeepM6A outperforms MLP, in particular for longer flanking sequences. As with LR, initially, the performance of MLP improves with length, but it degrades slightly after 10 bp (Fig. 1b). Regarding single-nucleotide sensitivity, both DeepM6A and MLP share similar robustness. Overall, the one-hot encoding is a better feature representation, preserving the primitive sequences in comparison with the *k*-mer format. The hierarchical feature extraction of DeepM6A is more powerful than that of MLP.

We therefore conclude that DeepM6A is both precise and robust in predicting 6mA. Its superiority is—at least in part—attributed to its deep network structure, which uses several hidden layers to learn a high-level representation of the DNA sequence hierarchically. To elucidate the power of this hierarchical representation and learning, we visualized the positive and negative samples using t-SNE[18] based on the features learned at different network layers.

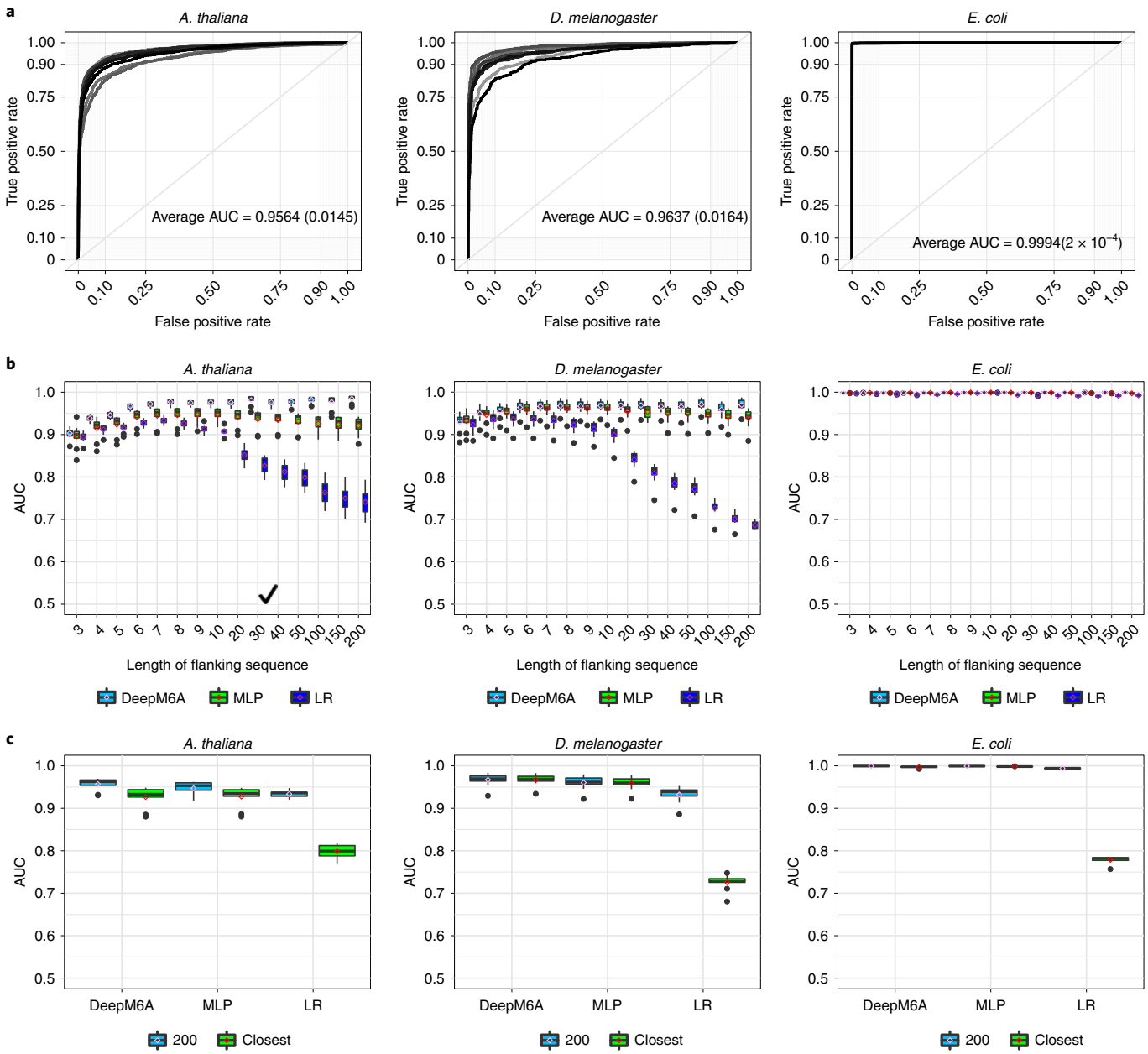

Fig. 1 | DNA 6mA site prediction. a, ROC curves for 10 independent experiments and average AUC values (mean (s.d.)). **b**, Comparison of DeepM6A, MLP and *k*-mer-based logistic regression (LR) across varied contextual sequences (flanking lengths $n = 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 100, 150, 200$). **c**, Impact comparison of remote (≥200 bp) and closest control cohorts of non-methylated adenines on the predictive capacities of DeepM6A, MLP and *k*-mer-based LR with flanking lengths of 30, 10 and 7, respectively. Model performance is measured with boxplots of AUCs for 10 independent experiments. There are 3,934, 2,138 and 6,740 sites for *A. thaliana*, *D. melanogaster* and *E. coli*, respectively.

As shown in Supplementary Fig. 4, the features become more and more discriminative along the layer hierarchy, with methylated and non-methylated sites mixed at the input layer, culminating with a clear separation in the output layer. Interestingly, the higher the methylation level, the better the separation, as observed in the last layer (Fig. 2a–c). This is also consistent with the observed high correlation between predicted probability and methylation level (Supplementary Fig. 5).

With appropriate design DeepM6A can predict 6mA dynamics

The aforementioned experiments were performed on the SMRT-seq data generated under general conditions. We note that DeepM6A is not limited to SMRT-seq and is also applicable in various experimental conditions. In fact, N^6-methylation is a dynamic epigenetic modification. When coupled with additional experimental design and data, DeepM6A may make more informative and valuable predictions. To illustrate this point, we applied DeepM6A to a 6mA dataset with developmental stage information available. This study captures 24,338 6mA loci in different developmental stages of *D. melanogaster* embryos[19] (see Methods). Specifically, it profiles 6mA peaks at three stages—0.75, 3 and 6 h—which account for 17,528, 4,363 and 2,447 peaks, respectively (Supplementary Fig. 6). It is noted that some loci may be observed with peaks at multiple stages. To accommodate and exploit stage information, we can simply modify DeepM6A to make multi-label (stage) predictions by introducing vectorized outputs,

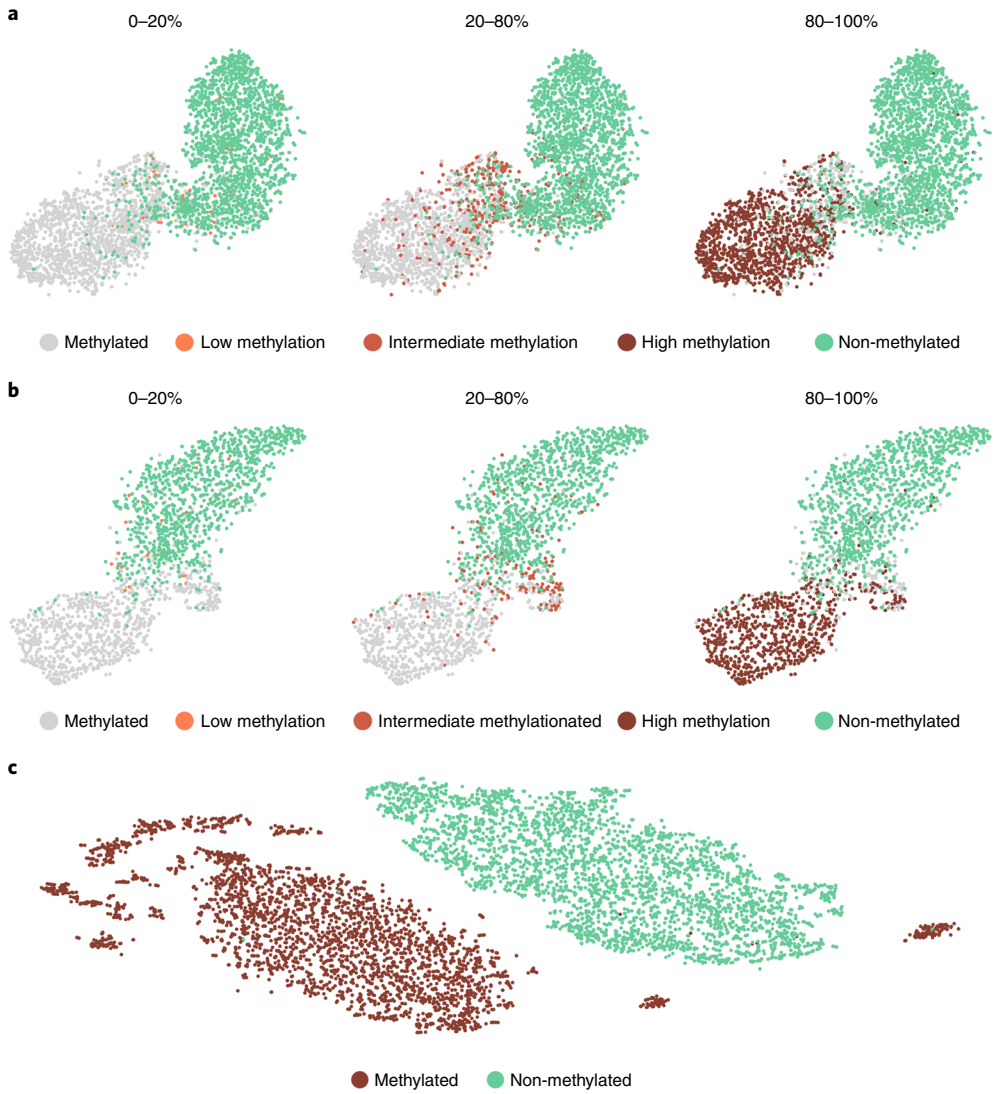

Fig. 2 | t-SNE visualization of the last hidden layer representations of methylations. a,**b**, *A. thaliana* (**a**) and *D. melanogaster* (**b**) at different methylation levels: low (0–20%), intermediate (20–80%) and high (80–100%). **c**, For *E. coli*, almost all 6mA sites lie in the highly methylated region. There are 3,934 sites for *A. thaliana*, 2,138 for *D. melanogaster* and 6,740 for *E. coli*.

which denote the probabilities of a locus being methylated at 0.75, 3 or 6 h, or not at all. Thus, DeepM6A not only predicts being methylated or not, but can also reveal at which stage(s) m6A would happen. To make a comparison, we also extended the baseline MLP and *k*-mer LR methods to the multi-label prediction task. As reported in Fig. 3, DeepM6A significantly outperforms MLP and LR, with average AUCs of 0.882, 0.939 and 0.973 for 0.75, 3 and 6 h, respectively. These superior prediction accuracies indicate that DeepM6A is able to predict methylation dynamics accurately when given relevant information for training. These findings jointly demonstrate that DeepM6A can achieve an appealing performance, regardless of whether dynamic conditions are considered.

SM-CAP can reveal advanced *cis*-regulatory patterns

After identifying 6mA, the next step, typically, is to search for regulatory sequences in the surrounding regions. Unlike conventional motif analysis[11,20,21], we developed a saliency maps-based context analysis protocol (SM-CAP). SM-CAP works by quantifying the contribution of a single base in the modelling context of all other participating bases, such as the nonlinear models DeepM6A employs. In contrast, conventional motif analysis assumes simple independence and additive effects among regulatory bases[11]. Our current knowledge suggested that genomic 6mA may not be conserved. Such conservancy status is defined conventionally as traditional motif definition. Advanced nonlinear motif patterns can exist and exhibit unconventional conservancy status. If we consider simulated data, for example, as shown in Supplementary Fig. 7, there is supposed to be a 61-bp DNA segment with adenine in the centre, and A, C, G and T distributed evenly at the remaining loci. The central adenine will become methylated if and only if a combination of {A, C, G, T} shows at four specific loci X_1, X_2, X_3 and X_4. The order A, C, G and T does not matter, for example, ACGT, TCGA or GATC can all lead to the methylation of the central adenine. We can see that the motif pattern at the four loci {X_1, X_2, X_3, X_4} is not conserved according to the conventional conservancy status definition. Conventional motif search algorithms would fail to recognize this motif. As shown in Supplementary Fig. 8, our proposed SM-CAP can successfully identify the four loci and assign appropriate importance scores for the four bases, in comparison with other irrelevant loci. Our SM-CAP can thus capture advanced patterns that are missed by traditional analysis, as verified by simulation studies (see Supplementary Information).

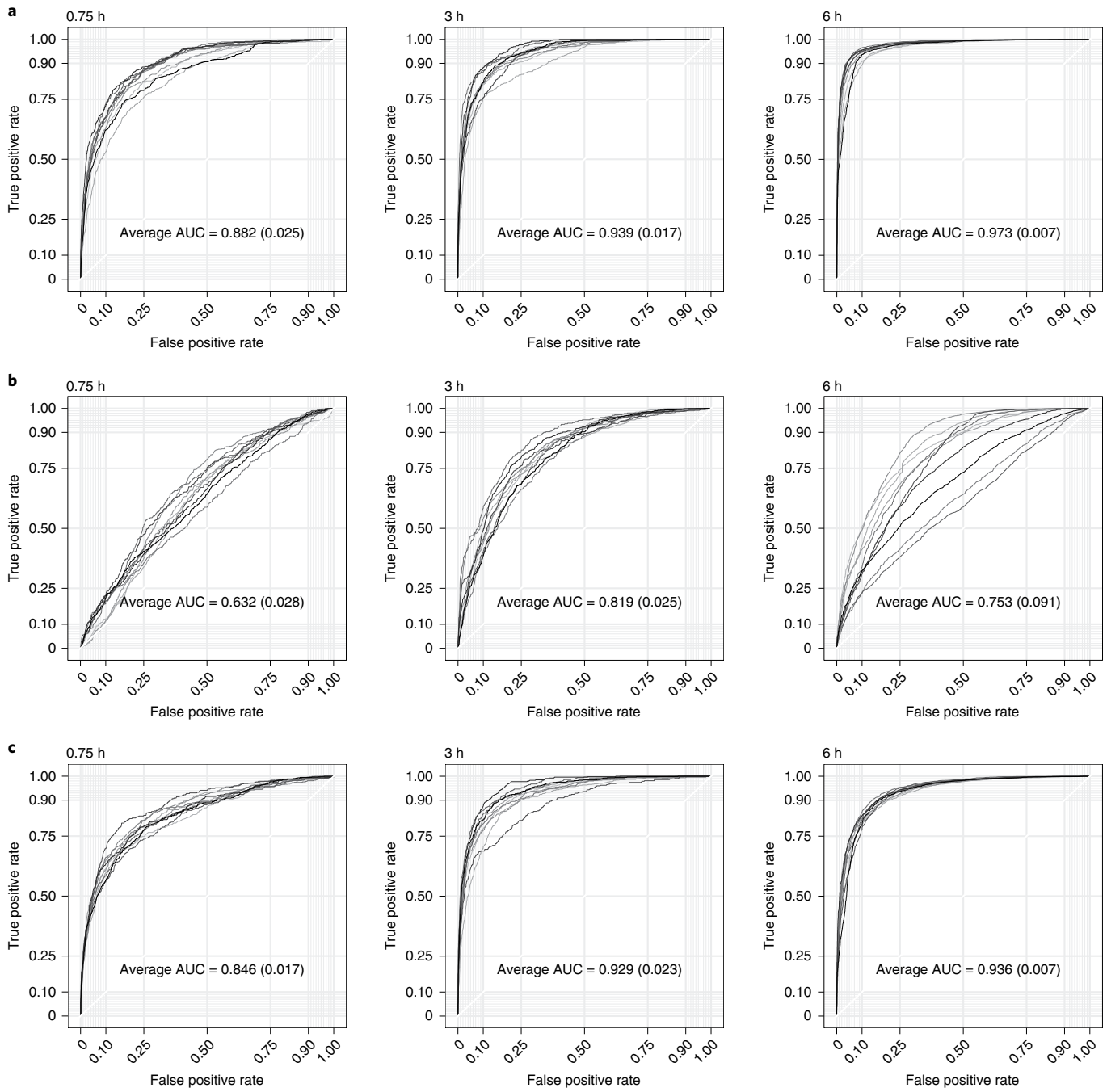

Fig. 3 | ROC curves for *D. melanogaster* embryos in different developmental stages. a–**c**, ROC curves for DeepM6A (**a**), MLP (**b**) and *k*-mer-based LR (**c**). AUC values (mean (s.d.)) of 6mA sites based on 10 independent experiments are shown. There are 17,528, 4,363 and 2,447 peaks and the same number of non-methylated regions for 0.75-, 3- and 6-h stages.

cis-regulatory patterns of 6mA revealed by SM-CAP

We used SM-CAP to analyse and visualize the contextual region of 6mA for the three species (Fig. 4a–c). We can see that the central region is the most critical and that the eukaryotes exhibit more sophisticated patterns than the prokaryote. Interestingly, we observe asymmetrical contributions of the flanking contextual sequences, with the downstream more predominant than the upstream sequences in terms of both strength and length. To further quantify and confirm their contributions, we perturbed nearby regions alternately and evaluate their impact on the prediction performance (Fig. 4d–f and Supplementary Fig. 9). We observed that the central M0 [±2 bp] and downstream D1 [+3 bp, +7 bp] regions play the most important role in predicting 6mA across different species, which is in line with the *cis*-regulatory patterns elucidated by SM-CAP. We noticed that some patterns are shared by the two eukaryotes, for example, GAGG [−1 bp, +2 bp], as shown in Fig. 4a,b. The normalized scoring maps thus present a good summary of the underlying conventional conserved motifs, which might co-regulate 6mA.

Further examination of the salient patterns revealed by SM-CAP shows that these patterns are more discriminative and account for more than 45% of the 6mA sites with an odds ratio of more than 20 for *A. thaliana* (Table 2) and *D. melanogaster* (Table 3), respectively (see Methods). As a comparison, we also tried to elucidate conventional motif patterns using HOMER[11] (Supplementary

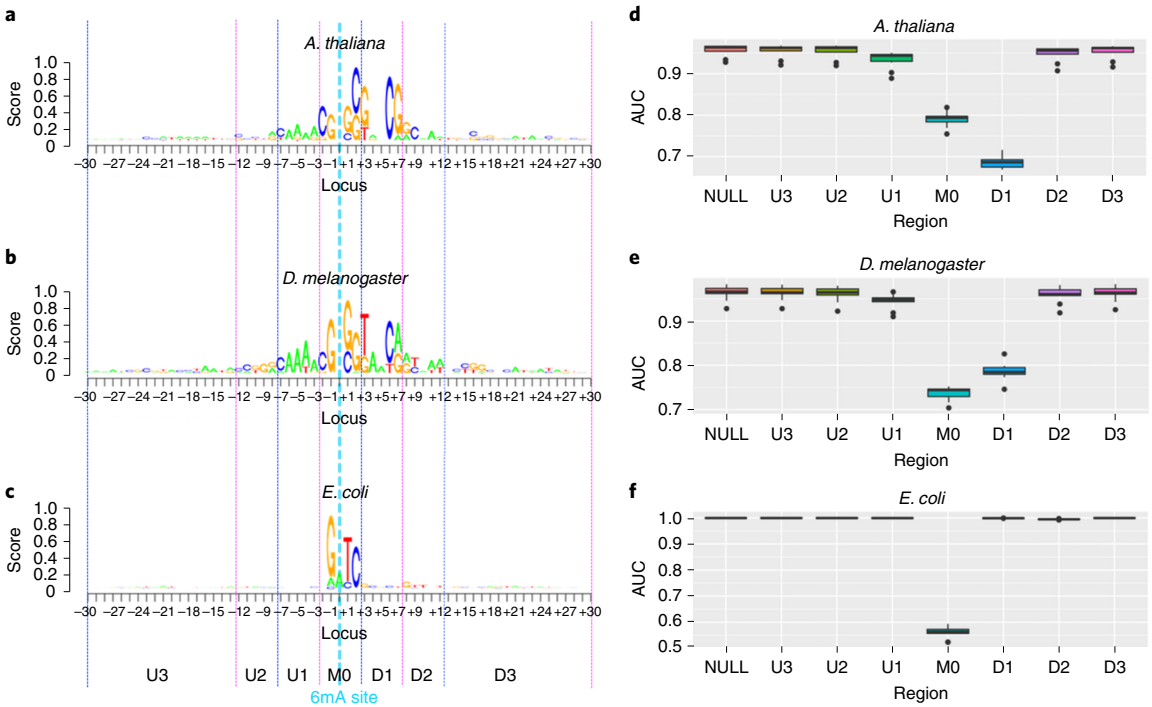

Fig. 4 | DNA motifs and loci revealed by SM-CAP. a–c, *cis*-regulatory patterns (6mA site based in locus 0) of *A. thaliana* (**a**), *D. melanogaster* (**b**) and *E. coli* (**c**). **d–f**, Evolution of the predictive capacity of DeepM6A when perturbing nearby regions for *A. thaliana* (**d**), *D. melanogaster* (**e**) and *E. coli* (**f**). M0 and D1 are critical regions. There are 3,934 sites for *A. thaliana*, 2,138 for *D. melanogaster* and 6,740 for *E. coli*.

Table 2 | SM-CAP results for *A. thaliana*

Motif	CCGA	GGTA	ACCG	GAGG	TCGA	GAAC	CGAG	CGAA	AGGT	AGGC	GACC	Total
Loci	[1, 4]	[1, 4]	[0, 3]	[−1, 2]	[5, 8]	[3, 6]	[−2, 1]	[2, 5]	[0, 3]	[0, 3]	[−1, 2]	[−2, 8]
Coverage	1,606	1,315	2,439	3,007	1,074	1098	1,642	1,098	2,444	1,201	1,272	9,154
Percentage (%)	8.18	6.70	12.42	15.32	5.47	5.60	8.36	5.59	12.45	6.12	6.48	46.63
Odds ratio	58.21	46.91	28.28	24.48	22.22	18.40	18.01	17.83	16.38	13.84	12.64	21.40

Table 3 | SM-CAP results for *D. melanogaster*

Motif	GGTA	GAGG	AGGT	CGAG	GGAG	Total
Loci	[1, 4]	[−1, 2]	[0, 3]	[−2, 1]	[−2, 1]	[−2, 4]
Coverage	1,069	3,534	2,335	1,541	1,631	4,882
Percentage (%)	10.03	33.17	21.92	14.47	15.31	45.83
Odds ratio	74.15	44.32	31.87	16.21	14.30	23.38

Table 4 | SM-CAP results for *E. coli*

Motif	GATC	Total
Loci	[−1, 2]	[−1, 2]
Coverage	32,622	32,622
Percentage (%)	96.80	96.80
Odds ratio	92,680.26	92,680.26

Figs. 10–12) (all with odds ratios of less than 2 for the two eukaryotes). It is noted that conventional conserved motifs could be a special case of the patterns SM-CAP can capture. An example in point is the well-known motif GATC (−1 bp, +2 bp) in *E. coli*[10], as also identified successfully by SM-CAP (Table 4).

DeepM6A is robust to false-positive samples

We also performed two additional experiments to show that DeepM6A is robust to false-positive samples, with consistent results. First, we performed simulation studies to illustrate the impact of the percentage of false-positive sites for different motifs (Supplementary Tables 11 and 12). We intentionally mixed negative samples into positive samples to mimic false positives, with the aim of confusing DeepM6A. We show that, even with 30% false-positive sites, the AUCs of DeepM6A were reduced by only 0.0141, 0.0126 and 0.0001 for *A. thaliana*, *D. melanogaster* and *E. coli*, respectively (Supplementary Table 13). For motif identification, our method can still capture the sequence patterns precisely when the percentage of false-positive sites is less than 50% (Supplementary Figs. 13–15). Interestingly, albeit not surprisingly, we observed that the complicated motifs in the eukaryotes *A. thaliana* and *D. melanogaster* identified by our method were more easily distorted by false-positive sites than the simple motifs for prokaryote *E. coli*. Second, we used DeepM6A trained on the SMRT-seq 6mA data to predict the peaks of 6mA regions detected using 6mA-DNA-IP-Seq for *D. melanogaster* (see Methods) from another independent study. 6mA-DNA-IP-Seq is the current state-of-the-art sequencing technology, with a low error rate, although it could not pinpoint the

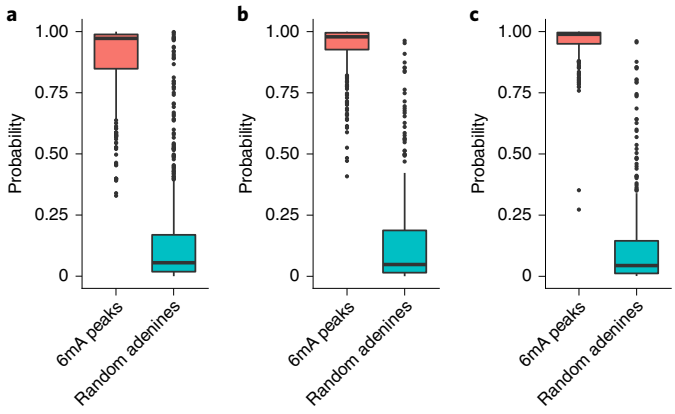

Fig. 5 | Comparison of DeepM6A-based predictive probability of peak regions and randomly selected adenines. a, 801 peaks and 801 non-methylated regions of *D. melanogaster*. **b**,**c**, 297 peaks and 297 non-methylated regions of replication 1 (**b**) and 376 peaks and 376 non-methylated regions of replication 2 (**c**) for *A. thaliana*.

exact methylation sites of 6mA (see Methods). The predictive scores of our model are reported in Fig. 5a. Of note, our predictive models trained on the SMRT 6mA data can capture almost all peaks with high probability generated from another independent study.

These two experiments jointly demonstrate the robustness of DeepM6A to false-positive sites and lend support to the predicted 6mA sites and identified sequence patterns outputted from DeepM6A. To extend these observations, we also performed a whole-genome 6mA prediction using the proposed DeepM6A (Supplementary Section 'A whole-genome 6mA catalogue made by DeepM6A').

Experimental validation

We have used public data to demonstrate the robustness and prediction performance of DeepM6A. We also conducted our own experimental validation to strengthen the work. DNA immunoprecipitation-based genomic profiling of 6mA has proven to be accurate and robust in *A. thaliana*[22]. We thus chose *A. thaliana* as the model organism and validated our prediction using DNA immunoprecipitation followed by a next-generation sequencing experiment (see Methods '6mA-DNA-IP-Seq and sequencing of *A. thaliana*' for details). Briefly, we grew the Columbia-0 ecotype of *A. thaliana* in a chamber under long-day conditions and extracted gDNA from nine-day-old seedlings. We produced two biological replicates and sonicated the gDNA into ~300-bp fragments. Following ref. [22], we then used the same 6mA antibody to pull down 6mA-methylated DNA fragments, followed by high-throughput next-generation sequencing. We mapped the sequencing data (GEO accession no. GSE149060) to the plant reference genome and used MACS2[23,24] to call peaks (enriched regions) that contained 6mA sites. Under a false discovery rate (FDR) cutoff of 0.05, MACS2 reported 297 and 376 peaks for the two replicates, respectively. As shown in Fig. 5b,c, DeepM6A predicted most of those DNA-IP-Seq validated 6mA sites with very high probability (>0.95), while as negative control, DeepM6A predicted most randomly selected adenines with very low probability (<0.05). This cross-platform independent experimental validation confirms the accuracy of DeepM6A.

Discussion and conclusion

It is noted that our prediction is purely based on sequence information. In other words, only the *cis* effect has been captured. This means that what we predict is the candidacy or potential for being a 6mA site. Whether a candidate is 6mA-ed or not will also depend on many other exogenous *trans* effects. The origin of the *trans* effects has remained largely elusive so far for eukaryotes. Conventionally, epigenetic modifications on individual bases in DNA are regulated by certain methyltransferases and can encode heritable genetic information. In eukaryotes, only the DNA N^6 adenine methyltransferase 1 (DAMT-1) in *Caenorhabditis elegans* has been reported to execute DNA adenine methyltransferase activities[25]. A recent study shows that the majority of DNA 6mA in mammalian cells, instead of originating from direct methylation, may be incorporated by DNA polymerases via a nucleotide-salvage pathway[26]. This finding is interesting because it delineates another pathway of generating 6mA that results from an environmental effect instead of a heritable effect. Using only DNA sequence information, our prediction may not directly distinguish the origins of 6mA, for example nucleotide salvage or direct DNA methylation. Nevertheless, the good prediction accuracy suggests that the presence of genomic 6mA is non-random and may have biological consequence. Even if the 6mA is incorporated into the genome during DNA replication, the frequency of the T base in DNA template strand pairing with adenine or 6mA may be affected by its flanking sequences. Regardless of whether this 6mA site originates from direct methylation by writer proteins or from DNA replication, its interaction with 6mA binding proteins (readers) can trigger downstream biological consequence. In the event 6mA is released from the microbiome or bacteria and is absorbed by the host cell and integrated into its own genome, this would be an interesting mechanism by which the microbiome or bacteria in the environment communicate with the host (via the releasing/recycling of bacterial DNA bases). Our in silico 6mA candidacy map helps to provide a global view of 6mA candidacy (Supplementary Tables 1–3, Supplementary Fig. 16 and Supplementary Information) and may shed light on host–microbiome interactions. It is recognized that genomic 6mA methylation is dynamically regulated, with methylation status probably changed over time (developmental stages) or location (tissue-specific). We note that DeepM6A is a general 6mA prediction algorithm, which is not limited to the conditions, time or tissues for which data are collected. If 6mA methylation data are collected and provided with development-stage and/or tissue information, our model can be trained over those data and make predictions of tissue- and/or development-stage-specific 6mA sites accordingly. For example, if trained on 6mA methylation data collected at the embryonic development stage, DeepM6A would be able to predict embryo-specific 6mA candidacy.

Although we have conducted a rigorous filtering protocol on the 6mA sites as determined by SMRT and have generated the final datasets for our research, some false-positive sites might still exist. It is therefore good to assess the impact of the error rate on our study. It is interesting to see whether DeepM6A is robust to false-positive samples and our results remain convincing. To this end, we performed twofold experiments. First, we used simulation studies to show the effectiveness even when the dataset for model training contained a significant proportion of false-positive samples. Second, we applied the models to 6mA regions generated using 6mA-DNA-IP-Seq. This is the current state-of-the-art sequencing technology, with a low error rate, but it could not pinpoint the exact methylation sites of 6mA. Interestingly—and impressively—our predictive models, trained on the SMRT 6mA data, captured almost all peaks with high probability. These two independent explorations indicate that our approaches are robust to false-positive samples and their outputs are reliable.

In conclusion, we have proposed a set of methodologies based on deep-learning techniques for elucidating DNA methylation on N^6-adenine. This work is a comprehensive study across three representative model organisms from a computational perspective. The proposed DeepM6A can identify 6mA sites at single-nucleotide resolution with high accuracy and reliability. Exploiting the predictive

power of DeepM6A, we have constructed a whole-genome complete catalogue of 6mA sites. We have also developed SM-CAP in an attempt to elucidate the regulatory mechanisms of 6mA methylation. This has uncovered quite a few interesting regulatory patterns that are missed by conventional motif analysis. These unconventional motif patterns are worthy of further investigation in epigenetics. Taken together, these computational approaches (DeepM6A and SM-CAP) and their findings may shed new light on the functional effects of 6mA methylation.

Methods

Deep learning techniques. The deep convolutional neural network, one of most popular deep-learning architectures, has set many records in various fields due to its powerful feature representation learning. A typical architecture of deep convolutional learning is composed of alternate convolution, pooling and nonlinear activation layers that learn the intermediate feature representation hierarchically, followed by fully connected layers and one sigmoid output layer for fusing internal features learned by convolutional layers together and delivering the output probability. Our learning architecture mainly involves convolutional network layers (ConvNet), leaky rectified linear activation (LeakyReLU) unit, dropout regularization, a fully connected layer and a sigmoid activation layer (Supplementary Fig. 1). Each convolutional neural network layer computes

$$\mathrm{ConvNet}(X)_{if} = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} W^{f}_{mn} X_{i+m,n}$$

where X is the input sequence, i and f are the indices of output position and filter kernel, respectively. Convolutional filter W^f is the $M \times N$ weight matrix with M and N being the window size and input channel (dimension), respectively. For example, N is 4 for the first convolutional layer and equals the filter number of the previous convolutional layer for higher-level convolutional layers.

LeakyReLU[27] is defined as

$$\text{Leaky ReLU}(x) = \begin{cases} x \text{ if } x \geq 0 \\ \alpha x \text{ if } x < 0 \end{cases}$$

where α is a constant. If α is 0, LeakyReLU is equivalent to ReLU. Each combination of convolutional layer and LeakyReLU is followed by a dropout regularization layer[28]. Dropout is utilized here to occasionally drop intermediate values by randomly setting them to be zero during training. The mask value m_k is dependent on a hyper-parameter dropout probability $p \in [0, 1]$. More formally

$$S_{\text{train}} = b + \sum_{k=0}^{K-1} m_k w_k z_k, \text{ where } m_k \sim B(1, p)$$

The stochasticity of dropout, however, is eliminated for validation and testing to guarantee the consistency of prediction. Furthermore, the predicted values are scaled by the expected value of the masks. Formally

$$S_{\text{prediction}} = b + p\sum_{k=0}^{K-1} w_k z_k$$

The sigmoid layer is given by

$$\text{Sigmoid}(x) = \frac{1}{1+\mathrm{e}^{-x}}$$

Altogether, our model has five convolutional neural networks, each with 80 filters ($f = 1, \ldots, 80$), one fully connected layer of 100 units and the final sigmoid classification layer. The dropout probability p is [0.2, 0.2, 0.2, 0.2, 0.5, 0.5] for the respective layers. The window size M of filters is set to be [4, 2, 4, 4, 4], respectively. The α of LeakyReLU is tuned to be 0.001 for all layers.

Data for DeepM6A development. The raw data of *A. thaliana*, *D. melanogaster* and *E. coli* were taken from the PacBio public database (https://github.com/PacificBiosciences/DevNet/wiki/Arabidopsis-P5C3, https://github.com/PacificBiosciences/DevNet/wiki/Drosophila-sequence-and-assembly and https://s3.amazonaws.com/files.pacb.com/datasets/secondary-analysis/e-coli-k12-P6C4/p6c4_ecoli_RSII_DDR2_with_15kb_cut_E01_1.tar.gz). Raw reads were first filtered based on a minimum subread length (50), polymerase read quality (75) and polymerase read length (50). The left reads were mapped to their reference genome by pbalign (seed=1 --nproc 8 --minAccuracy=0.75 --minLength=50 --concordant --algorithmOptions="-useQuality" --algorithmOptions=' -minMatch 12 -bestn 10 -minPctIdentity 70.0'). Additional pulse metrics were loaded into the resulting cmp.h5 file to enable downstream use of the Quiver algorithm, which identifies haploid single nucleotide polymorphisms and indels by performing a local realignment of reads using the full range of sequence quality metrics. Afterwards, the P_ModificationDetection module was applied to identify putative sites of base modification as well as common bacterial base modifications (6mA) using an in silico control consisting of expected kinetic signals. The 6mA sites were further filtered if (1) there is any sequence variance located between the upstream 10 bp and downstream 5 bp of the identified modification site and (2) the variation of estimated methylation level is greater than 30%. For *A. thaliana*, *D. melanogaster* and *E. coli* there are 19,632, 10,653 and 33,700 profiled methylated adenine sites, respectively. These sites account for 0.025696% (76,401,454), 0.013418% (79,393,495) and 1.475402% (2,284,124) proportions of the whole-genome adenine sites.

To prepare for the model development, we sampled non-methylated adenine sites (negative samples) randomly from the whole-genome sequences by preserving a distance scale of at least 200 between sampled negative non-methylated sites and the positive methylated sites. For both positive and negative samples, contextual sequences around the adenine site at each side were also extracted as input features. The above distance scale guarantees 200 non-overlapping nucleotides of contextual sequences. The length n of flanking sequences were set to be 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 100, 150 and 200, respectively. In this manner, each sample consisted of $2n+1$ nucleotides paired with methylation labels. We sampled the same number of non-methylated adenines with their counterpart methylated sites.

We utilized a one-hot encoding scheme[16] (Supplementary Fig. 1) to represent each nucleotide and then fed the input sequences into the DeepM6A model. For example, given a sequence of 61 nucleotides (30 flanking nucleotides at each side and one middle adenine site), it can be represented by a 61×4 binary matrix (with the columns being A, C, G and T). To facilitate efficient tuning of the model parameters and reasonable evaluation, we conducted a three-way data-splitting strategy. Specifically, for each chromosome of species, we split datasets into 10 different segments along the direction of 5′ to 3′, regardless of forward or reverse strands. We then picked one slice as the testing set, which is surrounded by two slices from both upstream and downstream sides. The validation dataset was set to be a combination of the nearest half upstream and half downstream segments accordingly (Supplementary Fig. 2). The remainder of the sites were used for training. In this manner, we obtained a ratio of 8:1:1 for the training, validation and testing datasets, with training and testing parts strictly non-overlapping. We used this spatial splitting to guarantee that the training and testing parts were strictly non-overlapping. Note that, for each 6mA site, we considered flanking sequences of length up to 200 bp as input. If we simply chose the 8:1:1 proportion for the training, validation and testing datasets at random, their flanking sequences would be more likely to overlap with each other. In this context, the predictive performance and obtained motif patterns might be somewhat misleading. All hyper-parameters were tuned on the basis of binary cross-entropy error on the validation dataset, which was not used for training or testing.

Training procedure. We trained DeepM6A from scratch (with norm initialization[27]) on large-scale *A. thaliana* methylation data across different lengths. Afterwards, for *D. melanogaster* and *E. coli*, we fine-tuned the full architecture of trained models of *A. thaliana*. This procedure is also called transfer learning, which can speed up the training process and serve as initialization for the elucidation of regulatory patterns of other species in terms of 6mA formulation. Classical back-propagation was leveraged for training. All layers of the networks were trained with the aid of a stochastic gradient descent algorithm. Specifically, the initial learning rate was 0.01 with a decay factor of 1×10^{-6} and Nesterov momentum of 0.9. We used Keras (https://keras.io) with a backend of Theano (http://deeplearning.net/software/theano_versions/0.8.X/) to train, validate and test all network architectures.

MLP network. MLP, here, is composed of five fully connected neural networks. The numbers of hidden nodes are 200, 200, 150 and 100 and the activation functions are also LeakyReLU of parameter 0.001 for all layers. The dropout probability is 0.5 for different layers. The activation of the output layer is also a sigmoid function. The training, validation and testing datasets and optimization settings are the same as for DeepM6A, if applicable.

Perturbation procedure. For each testing sample, we partitioned the entire flanking sequences into seven regions: [−30, −13] (U3), [−12, −8] (U2), [−7, −3] (U1), [−2, +2] (M0), [+3, +7] (D1), [+8, +12] (D2) and [+13, +30] (D3), according to the normalized scoring maps (Fig. 4d–f). We then perturbed these regions alternately while keeping other regions fixed. With previously trained models, we then performed further prediction on testing data with different perturbed regions.

Data for closest non-methylated adenines prediction. Similarly, we also selected non-methylated adenine sites by minimizing the distance scale of case and control samples (Supplementary Fig. 3). In this case, the overlapped contextual sequences between two cohorts were maximized (they are more challenging to predict). They can thus further serve as alternative benchmarks to evaluate the predictive power of the proposed method.

Dataset of 6mA-DNA-IP-Seq of D. melanogaster. The 6mA-DNA-IP-Seq data are described in ref. [9]. We downloaded the raw sequencing data from https://trace.ddbj.nig.ac.jp/DRASearch/study?acc=SRP055483. We then selected IgG and DNA

6mA demethylase mutant samples to call peaks. Downloaded reads were mapped to the reference genome dm6 using the Torrent Mapping Alignment Program with parameters '-Y -u -o 2 stage1 map4'. The MACS2[23,29] callpeaks pipeline was used to filter out duplicated reads and identify enrichment peaks with an FDR cutoff of 0.05. As a result, we obtained 801 6mA peaks (Supplementary Table 4). Because of the limitations of the 6mA-DNA-IP-Seq technology, we were not able to know the exact sites of adenine that are methylated in these 6mA peaks. To evaluate the prediction performance of our DeepM6A model on these 6mA sites identified by 6mA-DNA-IP-seq, we first selected all adenine sites around the 50 bp flanking context (upstream and downstream) of both forward (Supplementary Table 5) and reverse strands (Supplementary Table 6) for each 6mA peak. Only eight out of all contextual adenine sites around peaks were covered by the dataset we used to develop our DeepM6A. Next, we generated the contextual sequences with flanking length of 30 bp for each selected adenine site. Thus, each 6mA peak will have multiple contextual sequences with adenine centred. For each 6mA peak, we applied our DeepM6A model to predict the contextual sequences from the same peak (Supplementary Tables 5 and 6) and assigned the maximum score to this peak (Supplementary Table 7). We summarize the prediction result of those 801 6mA peaks in Supplementary Table 7. The high prediction scores (Fig. 5a) indicate that our DeepM6A model is able to capture the genomic patterns of the 6mA sequences found by 6mA-DNA-IP-Seq, even though the DeepM6A was trained based on 6mA sequences from the SMRT platform.

6mA-DNA-IP-Seq and sequencing of A. thaliana. Columbia-0 (Col-0) ecotype *A. thaliana* were vertically grown in 1/2 Murashige and Skoog medium (MS) salts (Phytotech) in chambers under long-day conditions (16-h photoperiods) at 22 °C. gDNA was extracted from nine-day-old seedlings and sonicated to ~300 bp. Fragmented DNA was incubated with antibody against N^6-methyladenosine modifications of RNA and DNA (Synaptic Systems) in immunoprecipitation (IP) buffer (50 mM Tris-HCl, 750 mM NaCl and 0.5% IPEGAL) for 2 h at 4 °C. Subsequently, the mixture was incubated with protein A beads (Thermo Fisher Scientific) that was pre-bound with acetylated bovine serum albumin (Sigma-Aldrich) at 4 °C for another 2 h. After washing four times, 6mA antibody-bound DNA was eluted from beads in elution buffer (90 µl 5X IP buffer, 30 µl of 100 mM antibody against 6mA, 330 µl H_2O) at 50 °C for 45 min. The eluted DNA was purified with a 2:1 ratio of AMPure SPRI beads (Beckman Coulter) and quantified with a Qubit High Sensitivity Kit (Thermo Fisher). Library synthesis was performed using the SMARTer ThruPLEX DNA-Seq Kit (Takara), library size was determined using a Bioanalyzer High Sensitivity Kit (Agilent) and quantification was performed with a KAPA Library Quantification Kit (Roche). Sequencing was performed on an Illumina NovaSeq 6000 sequencer using an SP 300 cycles flow cell in paired-end mode and the running parameters were 150 × 8 × 8 × 150. The sequencing data are available on the GEO database under accession no. GSE149060. After sequencing, we applied Trim galore and FastQC for adapter trimming and quality control. The pair-ended reads were then aligned to the TAIR10 genome by Bowtie2[24] with default settings. The MACS2[23,29] callpeaks pipeline was used to filter out duplicate reads and identify enrichment peaks with an FDR cutoff of 0.05. As a result, we obtained 297 and 376 peaks for the two replications (Supplementary Tables 14 and 15), respectively. Following the same analysis as for 6mA-DNA-IP-Seq of *D. melanogaster*, we selected all adenine sites in the 50-bp flanking context around the peaks and applied the corresponding trained DeepM6A model to predict scores of both forward (Supplementary Tables 16 and 18) and reverse strands (Supplementary Tables 17 and 19) for each 6mA peak. One peak could have many nearby adenines, and we picked the maximum score as the score for the peak (Supplementary Tables 20 and 21). The high prediction scores are described visually in Fig. 5b,c. It is also noted that ~48% (143/297) and 21% (80/376) of peaks in two replications are covered by the SMRT dataset we use to develop our DeepM6A. The median probabilities for the uncovered parts in two replications both exceed 0.97. The above studies strongly indicate that our DeepM6A model is capable of capturing the genomic patterns of the 6mA sequences found by 6mA-DNA-IP-Seq although the DeepM6A was trained based on 6mA sequences from the SMRT platform.

Data for different stages of D. melanogaster embryos. The data were downloaded from the GEO database under accession no. GSE86795[19]. We downloaded BED files of 6mA peaks in 0.75, 3 and 6 h. Because the 6mA peaks in this dataset were detected by 6mA-DNA-IP-seq, it is impossible to identify the precise adenine sites that were methylated. To generate training data for DeepM6A, we selected the middle points (it is noted that middle points are not necessarily adenine) of 6mA peaks and padded 200 bp up- and downstream contexts (forward strand of the dm6 reference genome). The majority lengths of the 6mA peaks are less than 400 bp (the third quartile of lengths is 385 bp), so 200 bp up- and down-contexts can cover most 6mA peaks. The numbers of 6mA peaks in the 0.75-, 3- and 6-h stages were 17,528, 4,363 and 2,447, respectively (Supplementary Fig. 6). The same number (24,338) of non-methylated regions were randomly selected as negative samples. As a result, we have four different labels: 0.75 h, 3 h, 6 h and non-methylation. We modified the last fully connected layer of DeepM6A to four sigmoid outputs, enabling DeepM6A to predict four different labels. The filter size of convolutional layers in DeepM6A was set to nine. For the MLP model, we also modified the last layer to four sigmoid outputs. For *k*-mer LR, we ran it four times to predict the four labels.

Table 5 | Contingency table

	Methylated	Non-methylated
Exposed	M_E	N_E
Not exposed	M_N	N_N

Whole-genome sequence prediction. To prioritize the whole-genome adenine sites, we extracted all adenine sites and their associated contextual sequences (30 flanking nucleotides at each side) for each species, giving a total of 76,401,454, 79,393,495 and 2,284,124 adenine sites of interest for *A. thaliana*, *D. melanogaster* and *E. coli*, respectively.

SM-CAP. The saliency map is a topographically arranged map that represents the visual saliency of a corresponding visual scene[30]. To interpret the proposed DeepM6A and elucidate the regulatory motifs of 6mA for different species, we propose the following saliency maps-based context analysis protocol (SM-CAP). (1) Given a methylated adenine site with contextual sequence, the corresponding first-order derivative with respect to the input sequence is first computed by utilizing back-propagation. In this manner, the protocol is able to highlight most of its signals and ignores irrelative contextual sequences. (2) The derivative is then multiplied by the encoding representation of the input sequence to obtain sequence-specific saliency maps. (3) The derivatives of all sequence samples are averaged accordingly. This leads to the overall scoring maps with respect to both sequence loci and nucleotides. (4) The scoring maps are finally normalized based on the min–max rule in the range from 0 to 1. SM-CAP can quantify the contribution of a single base in the modelling context of all other participating bases.

k-mer based logistic regression. *k*-mer refers to all the possible sub-sequences (of length *k*) from a sequence. In this Article, we set *k* to be ~1–6, and a total of 5,460 input *k*-mer based features were generated. Here, a *k*-mer coding scheme was utilized to represent features of contextual sequences. We first normalized the *k*-mer counts on training datasets. Both validation and testing datasets were then normalized based on the centre and scale of the training data accordingly. Finally we took as input the obtained normalized *k*-mer patterns to train the logistic regression with regularization by using the Python module scikit-learn (http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegressionCV.html). The optimal hyper-parameters were tuned and gained through fivefold cross-validation to the best of our efforts. The regularization parameters were searched in a grid of values between 1×10^{-4} and 1×10^{4}.

Naive scanning of sequence motifs. For all positive and negative sample sequences, we scanned all contextual nucleotides from 5′ to 3′ with a sliding window of 4 and stride of 1. There are four nucleotides for each position, thus 256 motifs are expected for each stride. In motif analysis, the sequence length was set to be 61, thus we have 58 strides. We used two metrics to quantify the discriminating power of stride-specific motifs, that is, coverage and odds ratio. The coverage with respect to a stride-specific motif is the absolute number of associated methylated adenine sites. The odds ratio is defined as

$$\mathrm{OR} = \frac{M_E/M_N}{N_E/N_N}$$

where M_E, M_N, N_E and N_N are defined in Table 5.

The coverage and odds ratio of stride-specific motifs are able to quantify how strongly the presence or absence of a specific motif is associated with methylation or non-methylation of adenine sites for a given species.

Examination of salient patterns revealed by SM-CAP. The motifs we identify summarize the aggregate importance of each base, which may be composed of the contribution of several individual patterns. To untangle their effects, we scrutinize the whole vital contextual sequences and examine each individual pattern. Specifically, we compute the coverage and odds ratio for all 4-mers at each position. As shown in Supplementary Tables 8–10, we observe around 10 powerful (high coverage) and discriminative (high odds ratio) patterns. For example, some major salient patterns, when combined together, account for 9,154 (46.63%) and 4,882 (45.83%) methylated sites, with odds ratios of 21.40 and 23.38 for *A. thaliana* and *D. melanogaster*, respectively (Tables 2 and 3). The well-known motif GATC [−1 bp, +2 bp] in *E. coli*[10] is identified successfully by SM-CAP as well (Table 4).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The SMART-seq data that support the findings of this study are available from GitHub (https://github.com/tanfei2007/DeepM6A/tree/master/Data). The sequencing data for *A. thaliana* are available on the GEO database under accession

no. GSE149060. The data for different stages of *D. melanogaster* embryos are available on the GEO database under accession no. GSE86795. The raw data for 6mA-DNA-IP-Seq of *D. melanogaster* are available from https://trace.ddbj.nig.ac.jp/DRASearch/study?acc=SRP055483.

Code availability

The custom computer code is available from GitHub (https://github.com/tanfei2007/DeepM6A/tree/master/Code) under https://doi.org/10.5281/zenodo.3887349.

Received: 5 April 2019; Accepted: 3 July 2020;
Published online: 3 August 2020

References

1. Heyn, H. & Esteller, M. An adenine code for DNA: a second life for N^6-methyladenine. *Cell* **161**, 710–713 (2015).
2. Luo, G.-Z., Blanco, M. A., Greer, E. L., He, C. & Shi, Y. DNA N^6-methyladenine: a new epigenetic mark in eukaryotes? *Nat. Rev. Mol. Cell Biol.* **16**, 705–710 (2015).
3. Zeng, H. & Gifford, D. K. Predicting the impact of non-coding variants on DNA methylation. *Nucleic Acids Res* **45**, e99 (2017).
4. Feng, S. et al. Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl Acad. Sci. USA* **107**, 8689–8694 (2010).
5. Wu, T. P. et al. DNA methylation on N^6-adenine in mammalian embryonic stem cells. *Nature* **532**, 329–333 (2016).
6. Fu, Y. et al. N^6-methyldeoxyadenosine marks active transcription start sites in *Chlamydomonas*. *Cell* **161**, 879–892 (2015).
7. Greer, E. L. et al. DNA methylation on N^6-adenine in *C. elegans*. *Cell* **161**, 868–878 (2015).
8. Liu, J. et al. Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig. *Nat. Commun.* **7**, 13052 (2016).
9. Zhang, G. et al. N^6-methyladenine DNA modification in *Drosophila*. *Cell* **161**, 893–906 (2015).
10. Barras, F. & Marinus, M. G. The great GATC: DNA methylation in *E. coli*. *Trends Genet.* **5**, 139–143 (1989).
11. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
12. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
13. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems* 1097–1105 (NIPS, 2012).
14. Lee, D. et al. A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).
15. Ji, G., Wu, X., Shen, Y., Huang, J. & Li, Q. Q. A classification-based prediction model of messenger RNA polyadenylation sites. *J. Theor. Biol.* **265**, 287–296 (2010).
16. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
17. Rosenblatt, F. *Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms* (Cornell Aeronautical Lab, 1961).
18. Maaten, Lvd & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
19. He, S. et al. 6mA-DNA-binding factor Jumu controls maternal-to-zygotic transition upstream of Zelda. *Nat. Commun.* **10**, 2219 (2019).
20. D'haeseleer, P. What are DNA sequence motifs? *Nat. Biotechnol.* **24**, 423–425 (2006).
21. Bailey, T. L. & Elkan, C. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach. Learn.* **21**, 51–80 (1995).
22. Liang, Z. et al. DNA N^6-adenine methylation in *Arabidopsis thaliana*. *Dev. Cell* **45**, 406–416 (2018).
23. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).
24. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
25. Li, Z., Zhao, P. & XiaQ. Epigenetic methylations on N^6-adenine and N^6-adenosine with the same input but different output. *Int. J. Mol. Sci.* **20**, 2931 (2019).
26. Musheev, M. U., Baumgartner, A., Krebs, L. & Niehrs, C. The origin of genomic N6-methyl-deoxyadenosine in mammalian cells. *Nat. Chem. Biol.* **16**, 630–634 (2020).
27. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In *2015 International Conference on Computer Vision* 1026–1034 (IEEE, 2015)
28. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
29. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
30. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: visualising image classification models and saliency maps. *Workshop at International Conference on Learning Representations* (2014).

Acknowledgements

We thank H. Liu for the partial data preprocessing. This study was supported by The Children's Hospital of Philadelphia Endowed Chair in Genomic Research to H.H. and an Institutional Development Award to the Center for Applied Genomics from The Children's Hospital of Philadelphia. This work was supported by Extreme Science and Engineering Discovery Environment (XSEDE) through allocation CIE160021 and CIE170034 (supported by National Science Foundation grant no. ACI-1548562).

Author contributions

Z.W. and H.H. conceived and supervised the project. F.T., T.T. and X.H. designed the methods and conducted the experiments with input from L.G. T.T., X.Y., B.D.G., F.M. and F.T. conducted the validation experiments. F.T., T.T., Z.W. and H.H. wrote the manuscript. All authors approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42256-020-0211-4.

Correspondence and requests for materials should be addressed to Z.W. or H.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Zhi Wei

Last updated by author(s): Jun 23, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The custom computer code is available in Github: https://github.com/tanfei2007/DeepM6A/tree/master/Code The DOI is 10.5281/zenodo.3887349
Data analysis	The custom computer code is available in Github: https://github.com/tanfei2007/DeepM6A/tree/master/Code The DOI is 10.5281/zenodo.3887349

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The SMART-seq data that support the findings of this study are available in Github:
https://github.com/tanfei2007/DeepM6A/tree/master/Data
The sequencing data of A. thaliana is available on the GEO database with the accession no. GSE149060. The data for different stages of D. melanogaster embryos is available on the GEO database with the accession no. GSE86795. The raw data of 6mA-DNA-IP-Seq of D. melanogaster is available from https://trace.ddbj.nig.ac.jp/DRASearch/study?acc=SRP055483

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	*Describe how sample size was determined, detailing any statistical methods used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data exclusions	*Describe any data exclusions. If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Replication	*Describe the measures taken to verify the reproducibility of the experimental findings. If all attempts at replication were successful, confirm this OR if there are any findings that were not replicated or cannot be reproduced, note this and describe why.*
Randomization	*Describe how samples/organisms/participants were allocated into experimental groups. If allocation was not random, describe how covariates were controlled OR if this is not relevant to your study, explain why.*
Blinding	*Describe whether the investigators were blinded to group allocation during data collection and/or analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National*

Research sample	*Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access and import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☐ Eukaryotic cell lines
☐	☐ Palaeontology
☐	☐ Animals and other organisms
☐	☐ Human research participants
☐	☐ Clinical data

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☐	☐ Flow cytometry
☐	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Antibody against N6-methyladenosine modifications of RNA and DNA (Synaptic Systems, m6A - 202 003)
Validation	See validation experiment on: https://www.sysy.com/products/m6a/facts-202003.php

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	*State the source of each cell line used.*

Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	*For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE149060
Files in database submission	input_fq_1.fq.gz; input_fq_2.fq.gz; rep1_fq_1.fq.gz; rep1_fq_2.fq.gz; rep2_fq_1.fq.gz; rep2_fq_2.fq.gz
Genome browser session (e.g. UCSC)	Not applicable

Methodology

Replicates	Two Columbia-0 (Col-0) ecotype of Arabidopsis thaliana were vertically grown at 1/2 Murashige and Skoog medium (MS) salts (Phytotech) in the chambers under long days conditions (16-hour photoperiods) under 22 °C.
Sequencing depth	Input: 268M; Rep1: 15M; Rep2: 7.2M
Antibodies	Antibody against N6-methyladenosine modifications of RNA and DNA (Synaptic Systems, m6A - 202 003)
Peak calling parameters	callpeak -c input_ara.sort.bam -t rep1_ara.sort.bam -f BAMPE --call-summits -g 1.35e8 -n rep1 -B -q 0.05 --bdg callpeak -c input_ara.sort.bam -t rep2_ara.sort.bam -f BAMPE --call-summits -g 1.35e8 -n rep1 -B -q 0.05 --bdg
Data quality	Raw fastq files were trimmed by trim-galore to remove adapters and low-quality bases with default parameters. Peaks were called by the FDR level 0.05. Rep1 has 297 peaks and rep2 has 376 peaks.
Software	trim-galore; bowtie2; macs2

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*

Behavioral performance measures

State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).

Acquisition

Imaging type(s)

Specify: functional, structural, diffusion, perfusion.

Field strength

Specify in Tesla

Sequence & imaging parameters

Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.

Area of acquisition

State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software

Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).

Normalization

If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.

Normalization template

Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

Noise and artifact removal

Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference
(See Eklund et al. 2016)

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.