Research

New Phytologist

# A deep learning approach to automate whole-genome prediction of diverse epigenomic modifications in plants

**Yifan Wang[1]*, Pingxian Zhang[1]*, Weijun Guo[1]*, Hanqing Liu[1]*, Xiulan Li[1], Qian Zhang[1], Zhuoying Du[1], Guihua Hu[1], Xiao Han[2], Li Pu[1], Jian Tian[1] and Xiaofeng Gu[1]**

[1]Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China; [2]College of Biological Science and Engineering, Fuzhou University, Fuzhou 350108, China

Authors for correspondence:
*Li Pu*
*Email: puli@caas.cn*

*Jian Tian*
*Email: tianjian@caas.cn*

*Xiaofeng Gu*
*Email: guxiaofeng@caas.cn*

Received: *26 March 2021*
Accepted: *9 July 2021*

*New Phytologist* (2021) **232:** 880–897
**doi**: 10.1111/nph.17630

**Key words:** artificial intelligence, convolutional neural networks, deep learning, DNA methylation, histone modification, RNA methylation.

## Summary

- Epigenetic modifications function in gene transcription, RNA metabolism, and other biological processes. However, multiple factors currently limit the scientific utility of epigenomic datasets generated for plants.
- Here, using deep-learning approaches, we developed a Smart Model for Epigenetics in Plants (SMEP) to predict six types of epigenomic modifications: DNA 5-methylcytosine (5mC) and N6-methyladenosine (6mA) methylation, RNA N6-methyladenosine ($m^6A$) methylation, and three types of histone modification.
- Using the datasets from the *japonica* rice Nipponbare, SMEP achieved 95% prediction accuracy for 6mA, and also achieved around 80% for 5mC, $m^6A$, and the three types of histone modification based on the 10-fold cross-validation. Additionally, > 95% of the 6mA peaks detected after a heat-shock treatment were predicted. We also successfully applied the SMEP for examining epigenomic modifications in *indica* rice 93-11 and even the B73 maize line.
- Taken together, we show that the deep-learning-enabled SMEP can reliably mine epigenomic datasets from diverse plants to yield actionable insights about epigenomic sites. Thus, our work opens new avenues for the application of predictive tools to facilitate functional research, and will almost certainly increase the efficiency of genome engineering efforts.

## Introduction

Epigenetic modifications, including DNA methylation, RNA methylation, and histone modifications, act as major drivers for the regulation of gene transcription and RNA metabolism, as well as various biological pathways in eukaryotes. DNA methylation and histone modifications are tightly regulated by chromatin architecture, and these modifications selectively trigger the activation or repression of gene expression. Histone H3 lysine 27 trimethylation (H3K27me3), and histone acetylation have been extensively investigated and shown to exert a functional role in transcriptional repression associated with silenced regions, while H3K4me3 and histone deacetylation have been frequently associated with the activation of gene expression from active chromatin regions (Liu *et al.*, 2014; Pu & Sung, 2015). With DNA methylation, 5-methylcytosine (5mC) usually occurs at CpG islands and functions in the negative regulation of gene expression (Wu & Zhang, 2017), whereas N6-methyladenosine (6mA) is thought to exert a mostly positive regulatory impact on gene expression in eukaryotes (Liang *et al.*, 2018a). Recently, RNA modifications like N6-methyladenosine ($m^6A$) have been identified as a new layer of epigenetic regulation (Roundtree *et al.*, 2017). In plants, epigenetic marks such as 5mC and 6mA, histone modifications like H3K27me3 and H3K4me3, and RNA methylation like $m^6A$ are known to strongly impact plant development and to participate in plant responses to biotic and abiotic stresses (Lämke & Bäurle, 2017; He & Li, 2018; Shen *et al.*, 2019). It seems clear that monitoring such modifications will continue to broaden our knowledge about gene regulation and epigenetic inheritance in plants.

Numerous well-established experimental methods in plant biology are used to quantify the abundance of various epigenetic marks (even at a genome-wide scale), including DNA bisulfite sequencing (BS-seq) for 5mC, single-molecule real-time sequencing (SMRT-seq) for 6mA, methylated RNA immunoprecipitation sequencing (MeRIP-seq) for $m^6A$, and chromatin immunoprecipitation sequencing (ChIP-seq) for histone modifications (Shen *et al.*, 2019; Liang *et al.*, 2020; Zhang *et al.*, 2020b). Studies using these methods have demonstrated various mechanisms through which epigenetic and transcriptional

*These authors contributed equally to this work.

regulation interact (Li *et al.*, 2018), have provided unprecedented resolution for genome-wide distribution patterns of modifications (Liang *et al.*, 2018b; Zhang *et al.*, 2018b; Zhou *et al.*, 2018), and have also revealed epigenetic diversity at the population scale from analyses of large germplasm diversity panels (Kawakatsu *et al.*, 2016). We are now firmly in the 'Big Data' era of epigenomics in plants. However, the dynamic nature of epigenomic modifications makes them distinct from genomics studies; that is, any given experiment-based epigenomic study cannot account for all of the potential sources of variation that are known to promote and/or repress epigenomic modifications. Indeed, experiments are typically limited to a narrow set of tissues, genotypes, or environmental conditions, obviously limiting the scope of any particular study and preventing large-scale explorations of the spatial and temporal regulation of genes and pathways throughout the plant life cycle and in diverse environmental contexts. Thus, there are good reasons to explore the development of *in silico* approaches that can predict likely sites for epigenomic modifications in plants.

One promising approach for developing predictive tools for plant epigenomics is deep learning, a subfield of machine learning that uses multi-layered convolutional neural networks (CNNs) to extract novel features from input data (Lecun *et al.*, 2015). Deep learning is an artificial intelligence (AI)-based technology that has been successfully applied to image recognition and robotics, and is generally most effective when applied to large datasets (Kang *et al.*, 2019; Qiu *et al.*, 2019). There are recent demonstrations for the successful applications of deep learning methods for biomedicine, genetics, and genomics (Ainscough *et al.*, 2018; Eraslan *et al.*, 2019; Zou *et al.*, 2019; Arbab *et al.*, 2020; Kim *et al.*, 2021), as well as some examples from plant biology and agricultural science (Washburn *et al.*, 2019; Wang *et al.*, 2020; Dunker *et al.*, 2021; Warman *et al.*, 2021). Owing to the availability of big datasets generated by epigenetic and epigenomic research, it makes sense that deep learning models have shown promise for the identification of DNA methylation (Angermueller *et al.*, 2017; Holder *et al.*, 2017; Lv *et al.*, 2020; Li *et al.*, 2021), histone modifications (Xu *et al.*, 2017; Hoffman *et al.*, 2019), RNA methylation (Sun *et al.*, 2019; Wang & Wang, 2020), and chromatin interactions (Zhang *et al.*, 2018a; Yang *et al.*, 2020). However, these deep learning models are mostly implemented in nonplant species, so whether these models can be applied usefully in plant species remains unclear. In plant epigenomics, deep learning has been used to generate a DNA 5mC predictor for *Populus* (Champigny *et al.*, 2019) and there are 6mA predictors in rice (Chen *et al.*, 2019; Lv *et al.*, 2019; Pian *et al.*, 2019; Yu & Dai, 2019; Zhang *et al.*, 2020a). Notably, most of these only predict a single epigenetic mark. The general complexity of epigenetic and epi-transcriptomic regulatory networks in cells, and increasing awareness of the interrelated regulatory impacts from diverse epigenomic modification types, together highlight a need for predictive tools that encompass multiple levels of epigenomic phenomena in an integrated way.

In this study, we constructed a Smart Model for Epigenetics in Plants (SMEP), a model that uses CNNs to accurately predict the likely genome sites for six distinct types of epigenetic modifications at a genome-wide scale. Our SMEP predictor includes 5mC and 6mA, $m^6A$, H3K27me3, H3K4me3, and H3K9ac. We demonstrate its utility with training data and examples from the *japonica* rice Nipponbare. After demonstrating proof-of-concept for the SMEP with DNA, RNA, and histone modifications, we explored how it can provide insights about the dynamic 6mA states of plants grown under heat stress conditions. We also applied the SMEP for predicting these modifications in *indica* 93-11 and even achieved accurate predictions in maize (*Zea mays* cultivar B73). To further facilitate the use of the SMEP, we developed a SMEP prediction tool website (http://www.elabcaas.cn/smep/index.html). SMEP prediction offers plant science researchers a powerful tool for studying the nature and impact of dynamic epigenomic modifications and gene expression in plants.

## Materials and Methods

### Plant datasets

The training, validation and testing datasets were from a number of published studies: the datasets for DNA methylation, including 6mA using SMRT-seq and 5mC using BS-seq, and gene expression data (RNA-seq) were from a study of 3-wk-old rice seedlings grown under short day conditions (Zhang *et al.*, 2018b). The $m^6A$ dataset using $m^6A$-IP-seq from leaves of *japonica* Nipponbare seedlings (Li *et al.*, 2014), and the H3K4me3, H3K27me3, and H3K9ac datasets were for Nipponbare seedlings were published previously (He *et al.*, 2010). We also used epigenomic datasets from *indica* 93-11 to evaluate the SMEP performance. The datasets for 5mC and 6mA, H3K27me3, H3K4me3, and H3K9ac in 93-11 seedlings were obtained from previous studies (He *et al.*, 2010; Zhang *et al.*, 2018b). We used the models to predict epigenomic sites including 5mC, $m^6A$, and H3K4me3 in maize (*Z. mays*) previously reported from B73 seedlings (Li *et al.*, 2015; Perduns *et al.*, 2015; Miao *et al.*, 2020), and a 6mA dataset in *Arabidopsis* from our previous report (Liang *et al.*, 2018b). All datasets were downloaded from National Center for Biotechnology Information (NCBI).

### Data processing of 6mA and 5mC

Detection and filtering of 5mC and 6mA sites was performed using processing approaches that followed our previous report (Zhang *et al.*, 2018b) and with input data from our eRice database (http://www.elabcaas.cn/rice/index.html; Zhang *et al.*, 2020a). To construct a high-quality dataset for 5mC and 6mA sites, we filtered DNA 5mC or 6mA sequence as follows: (1) we initially chose a window of 5mC or 6mA positioned within gene loci, or upstream (1000 nt)/downstream (200 nt) of the locus from the rice genome; (2) following the previously reported i6mA-Pred (Chen *et al.*, 2019), MM-6mAPred (Pian *et al.*, 2019), and SNNRice6mA (Yu & Dai, 2019) methods, we extracted ±20 nt sequences for each of the 5mC or C sites and

each of the 6mA or A sites. Because of the large number of fragments with unmethylated C or A positions in the negative dataset, only those fragments which appeared more than once were retained in the final negative dataset. This helped to reduce the imbalance between the number of positive and negative datasets. The ratio of positive samples and negative samples was *c.* 1 : 1. Detailed information for sample size of positive and negative datasets in all training datasets is shown in Supporting Information Table S1.

## Data processing of m$^6$A and histone modifications

ChIP-seq clean tags for histone modification libraries were mapped to the rice reference genome (Zhang *et al.*, 2018b) using Bowtie2 (v.2.1.0) (Trapnell *et al.*, 2012). Macs software (Zhang *et al.*, 2008) was used to recall the ChIP-seq peaks, using default parameters (bandwidth, 300 nt; model fold, 10, 30; *P* value, 1.00e$^{-5}$) for data normalization. For the m$^6$A-IP-seq dataset, m$^6$A peaks were also mapped to the reference genome (Zhang *et al.*, 2018b), followed by previously reported processing steps (Shen *et al.*, 2016).

It should be noted that the ChIP-seq and m$^6$A-IP-seq datasets were peak sequences (not as precise as DNA 5mC or 6mA) with a range of fragment lengths. We extracted the histone modification sequences of peaks on the chromosomes between 20 nt and 800 nt as the positive dataset. Similar to histone modifications, we extracted the peak sequences from RNA m$^6$A-IP-seq datasets (Li *et al.*, 2014) between 20 nt and 800 nt as the positive dataset. Negative datasets were mined based on the same lengths for peaks, but using upstream/downstream sequences positioned near the positive-dataset peaks. The ratio of positive samples and negative samples was 1 : 2. The detailed information for sample size of positive and negative datasets in all training datasets is shown in Table S1.

## Coding scheme

We transformed the sequence as a one-Hot-Coded matrix with A encoded by vector (1,0,0,0), T encoded by vector (0,1,0,0), G encoded by vector (0,0,0,1), and C encoded by vector (0,0,1,0):

$$B = (b_{n1}, b_{n2}, b_{n3}, b_{n4}, \ldots),\ b \in \begin{cases} \mathrm{A} = 1, 0, 0, 0 \\ \mathrm{T} = 0, 1, 0, 0 \\ \mathrm{C} = 0, 0, 1, 0 \\ \mathrm{G} = 0, 0, 0, 1 \end{cases},\ n \in (\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G})$$

## Convolutional neural network architecture

A typical CNN architecture comprises three parts: convolution layers, pooling layers, and fully connected layers. We trained CNN models with the constructed dataset based on a 10-fold cross-validation strategy. Briefly, the CNN model architecture we deployed was trained on the TensorFlow platform (Rampasek & Goldenberg, 2016) based on Keras 2.0.1 in a Python v.2.7 programming environment (compatible with Python v.3.5).

The final architecture comprised two convolutional layers, and each layer group was followed by a maximum pooling layer and a dropout layer, as well as the fully connected layers, each followed again by a dropout layer, and a final prediction layer. In this model, the rectified linear unit (ReLU) activation function was used (except for the final prediction layer) as follows:

$$\mathrm{ReLU}(X) = \begin{cases} 0, & \text{if } x < 0 \\ x, & \text{else} \end{cases}$$

where $x$ denotes the feature map from the convolution operation (the weighted sum of a neuron). The final prediction layer used a softmax activation function. For DNA 6mA, RNA m$^6$A, and histone modifications, we predict binary labels divided into their modification sites/peaks and nonmodification sites/peaks. But for 5mC datasets, we use multiclass labels based on its three methylation types (CG, CHG, CHH). The optimizer, adaptive moment estimation (Adam), and the categorical cross-entropy loss function, were used to optimize the parameters in the CNN-based model. We used class weight parameters for training CNN model to deal with the unbalanced dataset. We have optimized the various hyper-parameters in the CNN architecture, including the number of layers, number of kernels, kernel size, fully connected layer size, dropout rate, learning rate, batch size, activation functions, number of nodes, and optimizers using the optimization package hyperas (https://github.com/maxpumperla/hyperas). The final parameters used for the prediction models are shown in Table S2. Note that we have also optimized the various hyper-parameters for other machine learning algorithms (Methods S1; Dataset S1).

To avoid overly-optimistic performance by our SMEP, we have also implemented CD-HIT (Fu *et al.*, 2012) to cluster the training data (DNA sequence identity threshold = 0.7). An early stopping technique to detect the prediction accuracy on the test dataset in the training process was employed to avoid over-training (and hence) over-fitting of the prediction model. A 10-fold cross-validation approach was used to validate the performance of the proposed method. The training time of each epigenetic mark by SMEP was listed in Table S3. The codes for the above CNN architecture with Python v.2.7 (compatible with v.3.5) environments are available from GitHub (https://github.com/BRITian/smep).

## Evaluation of Smart Model for Epigenetics in Plants prediction performance

A 10-fold cross-validation strategy was used to train and evaluate the performance of the prediction models. The performance of the prediction models was evaluated with metrics including accuracy, recall, precision, F1-score, and the area under the receiver operating characteristic (AUROC) curve score, which were each calculated based on a 10-fold cross-validation. The data in the metrics was calculated using the Keras package as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} (0 \le \text{Recall} \le 1)$$

$$\text{Precision} = \frac{\text{TN}}{\text{TN} + \text{FP}} (0 \le \text{Precision} \le 1)$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} (0 \le \text{F1} \le 1)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{FP} + \text{FN} + \text{TN} + \text{FP}} (0 \le \text{Accuracy} \le 1)$$

where TP, TN, FP, and FN indicate true positive, true negative, false positive, and false negative, respectively.

### Predicting epigenomic sites in the Nipponbare genome

Genome sequences (in the FASTA format) were split with a window size of 41 nt and a step size of 1 nt. When the nucleotide in the middle of the sequence fragment was A or C, the fragment was coded based on our earlier description of coding scheme and then fed to the 6mA or 5mC prediction model. For the $m^6A$ and the histone modification prediction, the genome sequences were split with a window size of 800 nt and a step size of 100 nt. Fragments were coded and fed to the corresponding prediction CNN model with the various optimized hyper-parameters. Numerous epigenomic sites were yielded, and then displayed the overlap between experimental and predicted epigenomic datasets using Venn diagrams.

Data visualization was performed using Circos plots and R, with calculations for data density for the entire genome. The overall distribution density of the predicted epigenomic sites was calculated, and the predicted values were correlated with the sites as determined by the experimental datasets. The distribution density was $\log_2$ transformed, then subjected to Pearson correlation coefficient analysis between the experimental and SMEP prediction datasets. For the distribution of histone modifications, we plotted the visualized pattern using R with the CHIPSEEKER package (Yu *et al.*, 2015). For plotting the $m^6A$ distribution pattern, we used R with calculations for probability distribution of 5′-untranslated regions (5′UTRs), coding sequences (CDSs), and 3′UTRs of messenger RNAs (mRNAs) following a previous report (Shen *et al.*, 2016).

### Validation of the Smart Model for Epigenetics in Plants predictions in Nipponbare seedlings under heat shock treatment

Three-week-old Nipponbare seedlings (short-day photo-regime) were exposed to a heat shock (HS) treatment for 36 h at 45°C as reported in our previous study (Zhang *et al.*, 2018b). Briefly, genomic DNA was independently isolated from aerial tissues of Nipponbare seedlings, followed by sonication to generate *c.* 250 nt fragments. Sonicated DNA was further incubated for 2 h at 4°C with a specific anti-6mA antibody (202003; Synaptic Systems, Göttingen, Germany), followed by immunoprecipitation by incubation with Protein A/G Plus-Agarose (Santa Cruz, Dallas, TX, USA). After extensive washing, the target DNA was eluted from the beads, followed by sequencing libraries for input and immunoprecipitated DNA on the Illumina HiSeq 4000 platform (Illumina, San Diego, CA, USA). Raw data were trimmed to remove adapters, and quality control steps were implemented with a previous report (Zhang *et al.*, 2018b). Reads with more than 15% nitrogen (N), more than 50% low-quality bases, or shorter than 18 nt were filtered out. After alignment, 6mA-enriched regions were called using a cut-off false discovery rate (FDR) < 0.05 and fold change > 2. Further data processing with SMEP followed the procedures described earlier.

### Validation of selected 6mA, $m^6A$ and H3K4me3 modification sites under heat shock treatment

To confirm selected 6mA, $m^6A$, and H3K4me3 modification sites predicted by SMEP, we performed 6mA-IP-qPCR, $m^6A$-IP-qPCR, and ChIP-qPCR assays as previously reported (Shen *et al.*, 2016; Zhang *et al.*, 2018b, 2021), using specific anti-$m^6A$ antibody (202003; Synaptic Systems) and anti-trimethyl H3K4 (05-745R; Millipore, Bedford, MA, USA) antibodies. Quantitative polymerase chain reaction (qPCR) was conducted to analyze the amounts of *HsfA1*, *AREB1*, and *DREB2A* fragments on a Roche LightCycler480 II System using SYBR Green PCR master mix. The primers for specific 6mA, $m^6A$ and H3K4me3 modification sites are listed in Table S4.

### Application of the Smart Model for Epigenetics in Plants with the *indica* cultivar 93-11, *Arabidopsis thaliana*, and maize

To further assess the utility of SMEP, we also applied it with epigenomic datasets (6mA, 5mC, H3K4me3, H3K27me3 and H3K9ac) from *indica* cultivar 93-11, and even for other plants, including 5mC, $m^6A$ and H3K4me3 in maize and 6mA in *A. thaliana*. The epigenomic data processing and training of SMEP were the same for the maize epigenomic datasets as for the SMEP trained on rice data.

### DEEPLIFT

DEEPLIFT (Shrikumar *et al.*, 2017) is a feature attribution method for computing the contribution of each base (feature) in an input sequence to a specific scalar output prediction from a CNN model. DEEPLIFT decomposes the difference between the output prediction from an input sequence vs that of a neutral reference input sequence as an additive combination of contribution scores of all bases in the input sequence. Here, the sequence fragments containing modified epigenetic sites were used as the input sequence. We also found that the method REVEALCANCEL-CONV-RS reduced noise relative to pure REVEALCANCEL, which was selected to interpret the models and analyze the motifs. The detailed calculational codes and steps were obtained from website of DEEPLIFT (https://github.com/kundajelab/deeplift).

## Results

### The Smart Model for Epigenetics in Plants

Our SMEP predicts likely epigenetic sites at a genome-wide scale based on epigenetic experimental data including 5mC and 6mA methylation, $m^6A$ methylation and three distinct types of histone modifications (H3K4me3, H3K27me3 and H3K9ac) that is transformed using a deep learning task model (Fig. 1a). Standard epigenomic analyses typically rely on a normalization step when processing epigenetic signals (Fig. 1b). When training the SMEP for 5mC and 6mA data, we extracted 41 nt sequences as the positive dataset (i.e. the methylated 5mC or 6mA site plus 20 nt upstream and downstream). For the negative dataset, unmethylated A or C sites located within genes or in upstream (1000 nt) or downstream (200 nt) regions were extracted (again with 20 adjacent nt on both sides). The ratio of positive and negative samples was *c.* 1 : 1 for 5mC and 6mA (Table S1). For training $m^6A$ methylation and histone modifications, we isolated the peak sequences ranging from 20 to 800 nt for the positive datasets, and the negative datasets were mined based on the same lengths for peaks but using sequences positioned near the peaks. The ratio of positive and negative samples was 1 : 2 for $m^6A$ and histone modifications (Table S1).

Our deep learning process was based on a CNN approach, which explicitly models the nucleotide sequence context to train a predictive model (Figs 1c, S1; Table S2). *In silico* prediction of SMEP produces a large database with nucleotide-level scores (Fig. 1d,e). In theory, our SMEP tool should allow researchers to apply an intelligent, multi-modification-type epigenomic prediction tool to improve basic studies and applied strategies of epigenomic modifications in rice and other plants (Fig. 1f).

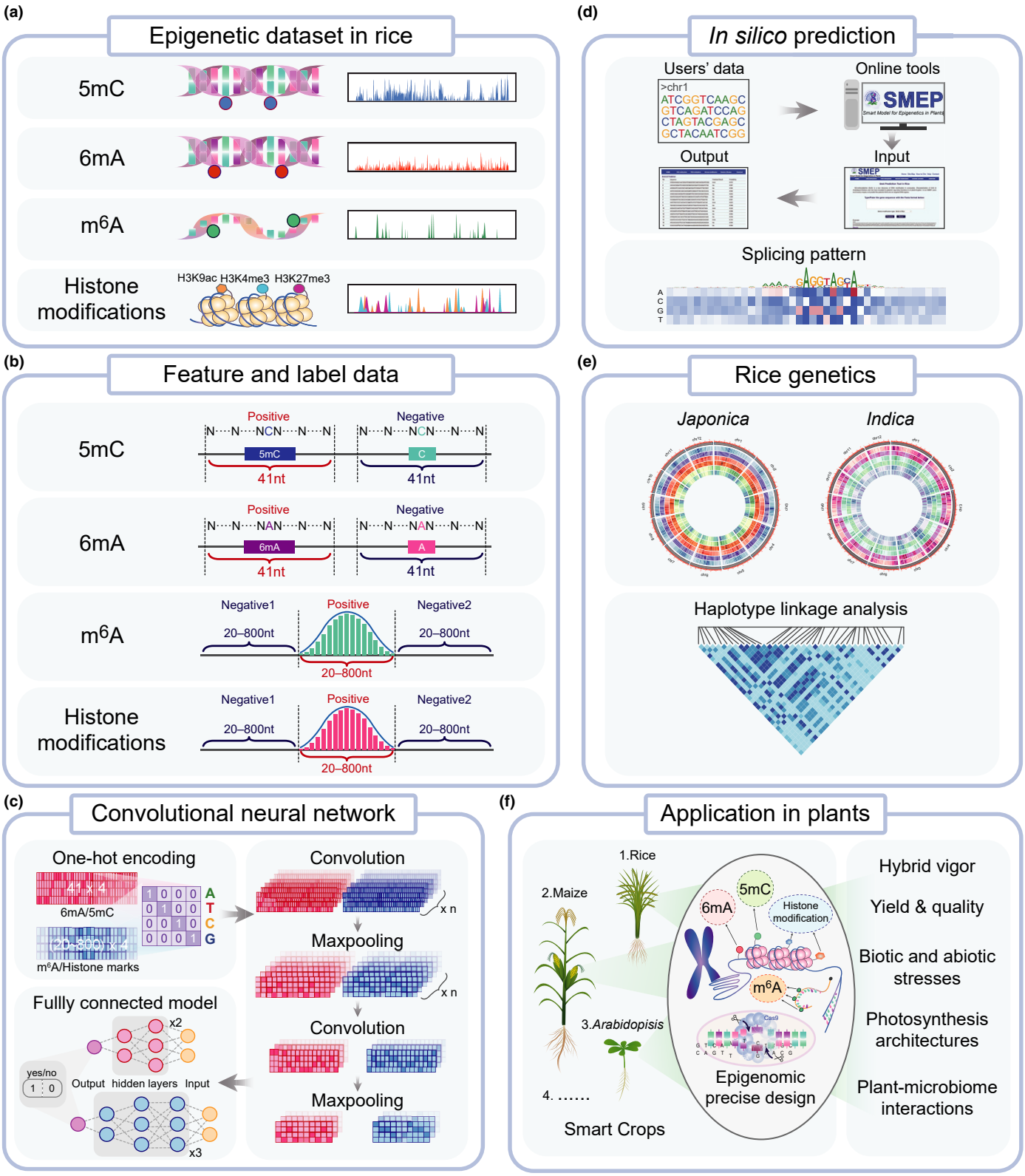
### Performance of Smart Model for Epigenetics in Plants for predicting 5mC and 6mA sites

First, we assessed the ability of SMEP to predict 5mC and 6mA states from aprevious study in Nipponbare seedlings (Zhang *et al.*, 2018b). Specifically, our baseline approach was to use receiver operating characteristic (ROC) curve and Precision-Recall curve (PRC) analyses to evaluate the prediction performance for the following types of sites: three types of methylated C (CG, CHG or CHH, where H represents A, C or T) and methylated A (6mA), unmethylated C, and unmethylated A (Fig. 2a,b). Moreover, both the AUROC and the area under PRC (AUPRC) scores were relatively high for 5mC and 6mA, respectively, suggesting high accuracy for SMEP performance in predicting 5mC and 6mA sites (Fig. 2a,b). We also considered a range of alternative metrics, including the prediction accuracy, the F1 score, the precision score, and the recall score; all of these supported the accuracy of SMEP predictions (Fig. S2). Using the SMEP 5mC and 6mA prediction models, we assessed a total of 53 135 897 potential 5mC and 7474 922 potential 6mA sites (Fig. 2c), which covered almost all genes in the Nipponbare genome (Table S5; Dataset S2). More than 99% of the 5mC and 95% of the 6mA sites identified in a previous SMRT-seq study were among the sites predicted by SMEP (Fig. 2c).

Note that in addition to the CNN algorithm we eventually selected, we also explored the use of four other machine-learning algorithms (Naive Bayes, Logistic Regression, Decision Tree, and Random Forest) for predicting 5mC and 6mA sites. Compared to the CNN-based iteration of SMEP, the AUROC scores for the iterations with the algorithms were all lower for 5mC and for 6mA (Fig. S3a,b), likely implying that the CNN algorithm for our SMEP model was more efficient than the common machine learning algorithms. Next, we performed DeepLIFT analysis to examine different genomic contexts of the 5mC (with CG, CHG, and CHH types) and 6mA modifications. These comparisons of motifs from site context data for the experimental and SMEP-predicted 5mC and 6mA sites supported strong conservation of motifs likely to promote 5mC and 6mA modifications across the rice genome (Figs 2d, S4a). Similar to previous reports (Zhang *et al.*, 2018b), by DeepLIFT analysis, we also extracted the GAGG motif for 6mA, which is highly conserved in eukaryotes (Liang *et al.*, 2020). To investigate the extent of false positives among the predicted modification sites, we have investigated the correlation between probability value in prediction data and fraction value experimental data (by SMRT-seq). We found that prediction sites with higher probabilities of 5mC and 6mA are the most like to be modified (Fig. 2e), these should be selected for further downstream biological research where suitable. The users can easily select those sites with high predicted probability ($\geq$ 0.7) from our easy-to-use SMEP web server.

We employed the CNN-enabled SMEP to preliminarily evaluate the correlation between gene transcription and 5mC or 6mA site. Compared to the performance of the predicting models trained from the experimental 5mC and 6mA datasets, the *r* values of 5mC and 6mA predicted by SMEP were significantly increased (Fig. S4b). Additionally, the *r* value evaluated by the models trained from the combined 5mC and 6mA datasets predicted by SMEP was also significantly increased (Fig. S4b).

**Fig. 1** Computational workflow for Smart Model for Epigenetics in Plants (SMEP) based on deep learning task models. (a) Epigenomic datasets (including BS-seq for 5mC, SMRT-seq for 6mA, MeRIP-seq for $m^6A$, and ChIP-seq for histone modifications) were used as input for training the SMEP model to predict epigenomic modifications in rice at a genome-wide scale. (b) Feature and label data for DNA sequences from rice genomic regions. (c) The convolutional neural network (CNN) algorithm at the core of SMEP. The nucleotide sequences are coded as the one-hot matrix. A neural network thoroughly convolutes the data in fully connected layers and uses pooling to predict epigenetic modifications from DNA sequences, doing so at a genome-wide scale. The schematic CNN models and optimized hyper-parameters were shown in Supporting Information Fig. S1 and Table S2, respectively. (d) An SMEP website was constructed to enable easy access for online prediction based on user data. For each epigenomic modification, users can submit input data with the Fasta format, and then obtain the prediction results in the output page. (e, f) SMEP offers a convenient tool for rice genetic studies (e) and downstream applications in plant research and biotechnology (f). SMEP lays a foundation for applying multi-modification-type epigenomic prediction tools to improve basic studies and applied strategies for epigenomic modification in rice and other plants.




## Prediction performance for transcriptome-wide m$^6$A sites

For training RNA m$^6$A, we isolated peak sequences ranging from 20 to 800 nt. As with our performance validation for 6mA and 5mC sites in DNA, we again used ROC curve and PRC to evaluate the prediction performance of SMEP for the sites of methylated A marks in mRNA (m$^6$A) and unmethylated A (Fig. 3a). As with the DNA modifications, we evaluated the accuracy value, the AUROC value, the F1 score, the precision score, and the recall score for our trained

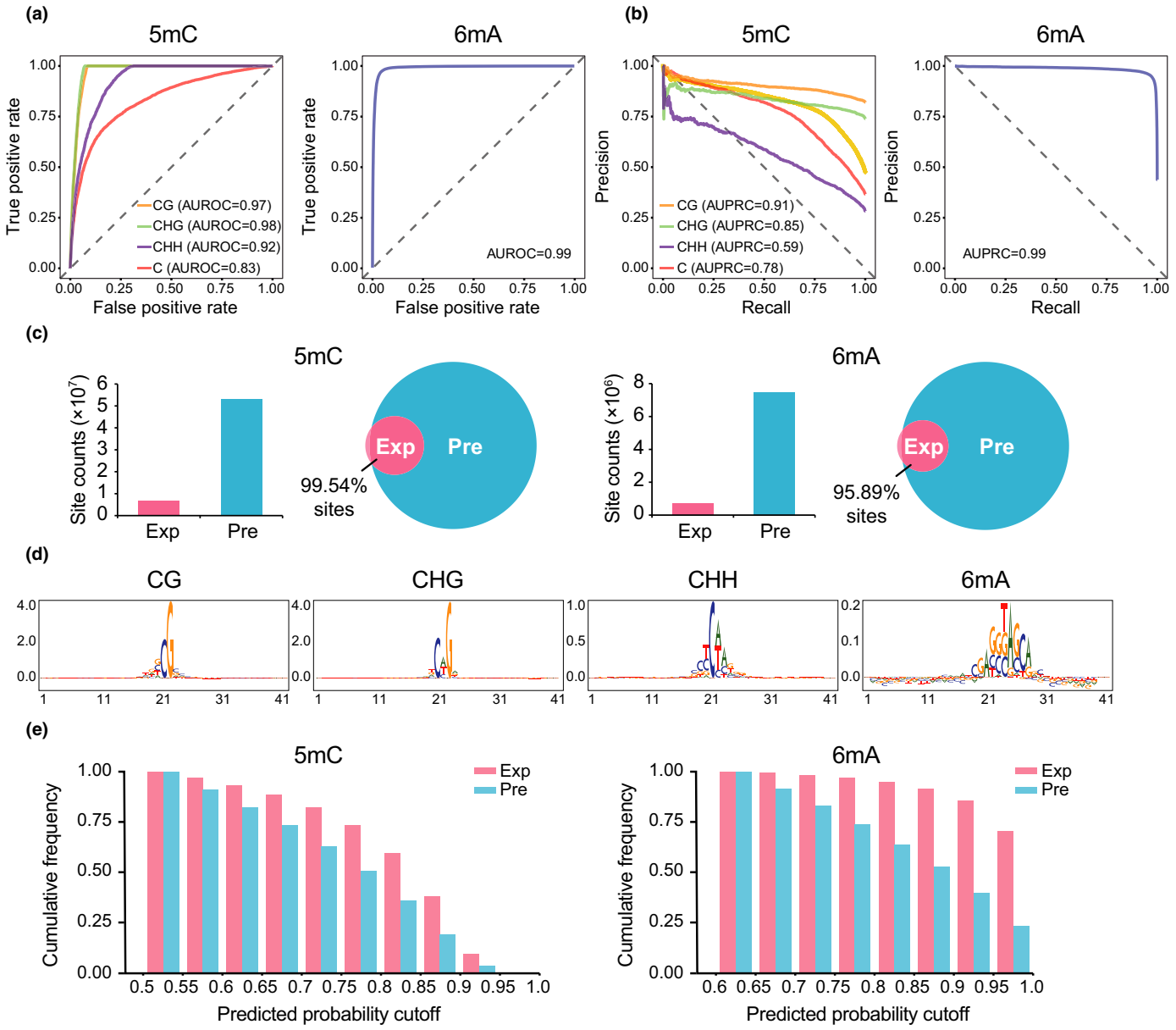



**Fig. 2** Smart Model for Epigenetics in Plants (SMEP) accurately predicts DNA 5mC and 6mA states in the *japonica* rice cultivar Nipponbare. (a, b) Receiver operating characteristic (ROC) curve (a) and Precision-Recall curve (PRC) (b) of SMEP output for predicting 5mC and 6mA sites in Nipponbare rice. ROC and PRC were calculated from classifying preferential site usage on 5mC and 6mA based on the results of 10-fold cross-validation. (c) Site counts and Venn diagrams of 5mC and 6mA sites in Nipponbare rice. We assessed the 5mC and 6mA state sites between our SMEP predictions (Pre) and data from previously reported experiments (Exp); the Venn diagram reveals the overlap between the prediction and experimental datasets. (d) Motif analysis for 5mC and 6mA prediction datasets in Nipponbare rice. The sequence logo representations of the consensus motifs were calculated by DeepLIFT. (e) Cumulative histogram on the probability cut-off of all experimentally determined and predicted sites in Nipponbare rice.

SMEP prediction for $m^6A$ sites (Fig. S2). We further analyzed the accuracy of $m^6A$ prediction tool in SMEP at the genome-wide scale. Approximately 78% of the $m^6A$ peaks in MeRIP-seq $m^6A$ dataset were detected in the SMEP-predicted sites throughout the genome of Nipponbare rice. For $m^6A$ prediction, we screened 708 647 potential $m^6A$ sites (Fig. 3b), positioned within 37 328 genes (Table S5; Dataset S2).

A previous study in rice using MeRIP-seq for transcriptome-wide $m^6A$ profiling revealed that $m^6A$ marks are mainly enriched across the CDS of mRNA transcripts, with two major peaks located after the start codon and before the stop codon (Li *et al.*, 2014). We first compared $m^6A$ peak counts in the mRNA between MeRIP-seq experiments and SMEP predictions. The results showed that $m^6A$ is preferentially located in the CDS region rather than untranslated regions (5′UTR and 3′UTR), both in experimental data and the SMEP predictions (Fig. 3c–e). In addition, we found that both the experimental data and SMEP predictions featured enrichment in modifications at particular $m^6A$ motifs, for example the GAVGA motif (Fig. 3f). Together,

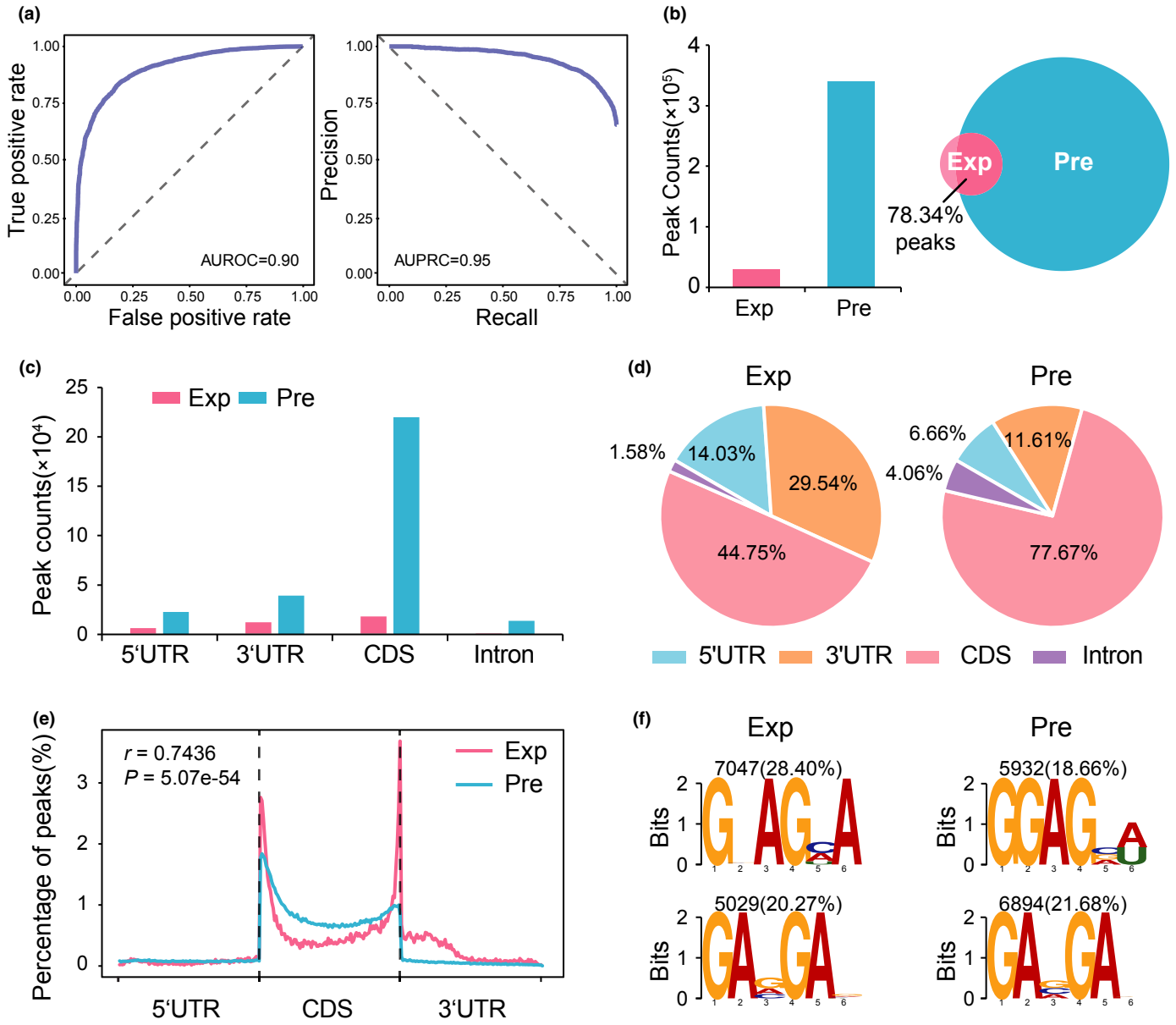



**Fig. 3** Smart Model for Epigenetics in Plants (SMEP) accurately predicts RNA m$^6$A methylation states in Nipponbare rice. (a) Receiver operating characteristic (ROC) curve and Precision-Recall curve (PRC) for RNA m$^6$A. ROC curve was calculated from classifying preferential site usage for RNA m$^6$A data based on the 10-fold cross-validation in Nipponbare rice. (b) Comparison of peak counts and Venn diagram analysis of global RNA m$^6$A states in the SMEP prediction (Pre) and experimental (Exp) datasets. We assessed the RNA m$^6$A state site data between our SMEP predictions and data from previous experiments; the Venn diagram reveals the overlap between the prediction and experimental datasets. (c,d) Peak counts (c) and distribution (d) of m$^6$A marks in messenger RNAs (mRNAs). The m$^6$A states were mapped in mRNA to 5′UTR, 3′UTR and CDS regions. CDS, coding sequence; UTR, untranslated region. (e) Distribution pattern of m$^6$A mapped to mRNA in Nipponbare rice. The m$^6$A marks were enriched in CDS regions in both the SMEP prediction and experimental datasets. (f) Motif analysis for m$^6$A between the experimental and prediction datasets. The sequence logo representations of the consensus motifs were identified by MEME with parameter (-zoops -n 10 -minw 5 -maxw 9, others were default). The number and percentage of occurrences of each motif are shown relative to total number of m$^6$A-containing motifs.

these results support the utility of SMEP for predicting RNA m$^6$A modifications.

## Prediction performance for histone modifications

After we assessed the ability of SMEP to predict DNA 5mC and 6mA, and RNA m$^6$A, we next trained the SMEP for histone modification marks. For training H3K4me3, H3K27me3 and H3K9ac, we isolated peak sequences ranging from 20 to 800 nt. SMEP yielded accurate predictions of histone modification marks (Figs 4a,b, S2) when we compared the SMEP predictions and ChIP-seq experimental datasets (Fig. 4c). We also evaluated the accuracy of histone modification prediction tool in SMEP at the genome-wide scale. More than 80% of peaks in the experimental datasets overlapped with the SMEP-predicted sites throughout the genome of Nipponbare rice (Fig. 4d). Moreover, we analyzed the location of

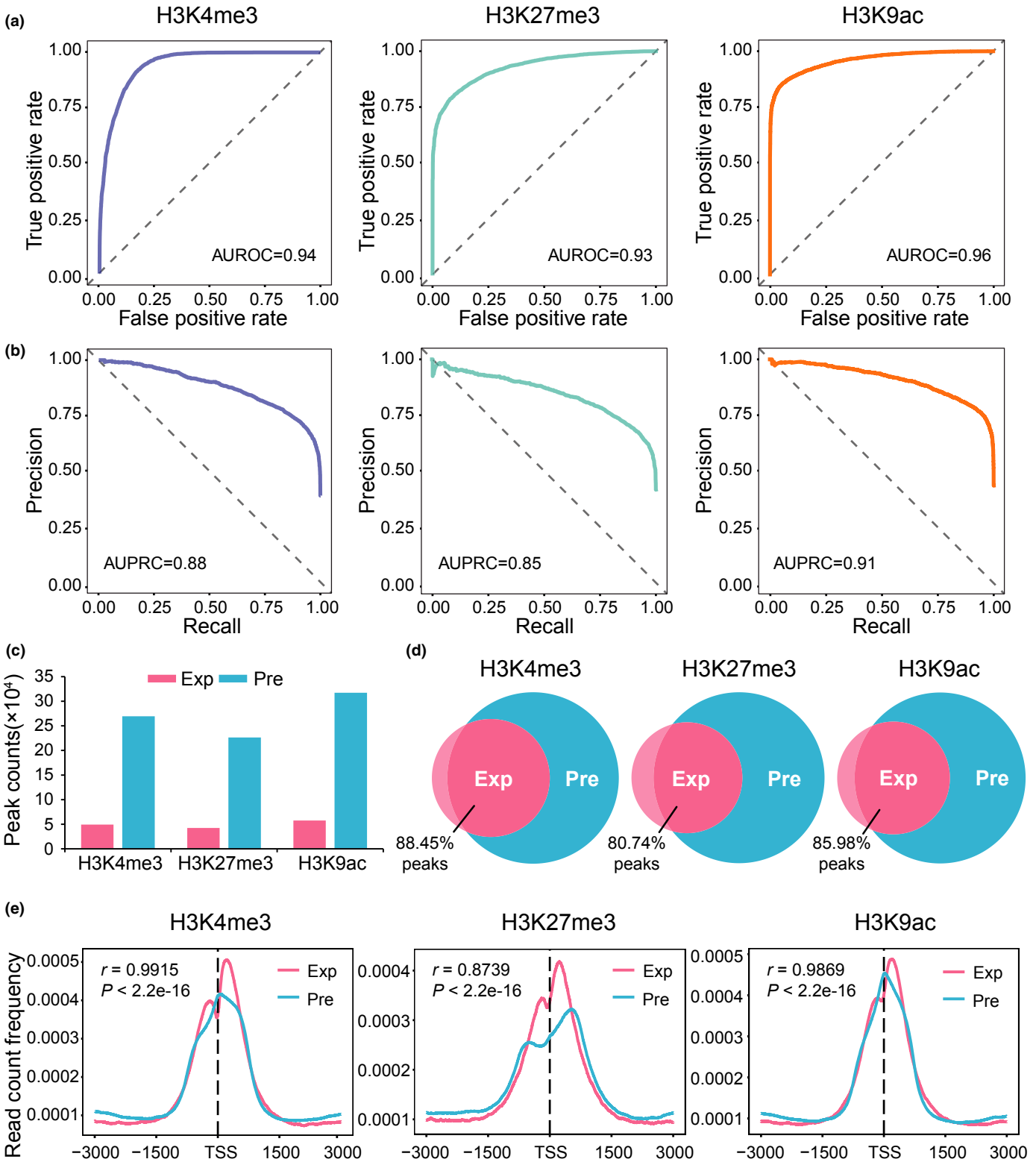



**Fig. 4** Smart Model for Epigenetics in Plants (SMEP) accurately predicts H3K4me3, H3K27me3 and H3K9ac states in Nipponbare rice. (a, b) Receiver operating characteristic (ROC) curve (a) and Precision-Recall curve (PRC) (b) of H3K4me3, H3K27me3, and H3K9ac. ROC and PRC were calculated from classifying preferential site usage on H3K4me3, H3K27me3 and H3K9ac data based on the 10-fold cross-validation in Nipponbare rice. (c, d) Peak counts (c) and Venn plot (d) analysis of H3K4me3, H3K27me3, H3K9ac. We gathered statistics of H3K4me3, H3K27me3 and H3K9ac sites between our SMEP predictions (Pre) and previous experiments (Exp). Venn diagrams were used to plot the overlap between the SMEP predictions and the experimental datasets. (e) Distribution pattern of H3K4me3, H3K27me3, and H3K9ac in genic regions. The H3K4me3, H3K27me3 and H3K9ac modifications were mapped to rice genic regions 3000 nt upstream and downstream of transcription start site (TSS). We compared the epigenomic features of H3K4me3, H3K27me3 and H3K9ac in the SMEP prediction and experimental datasets.

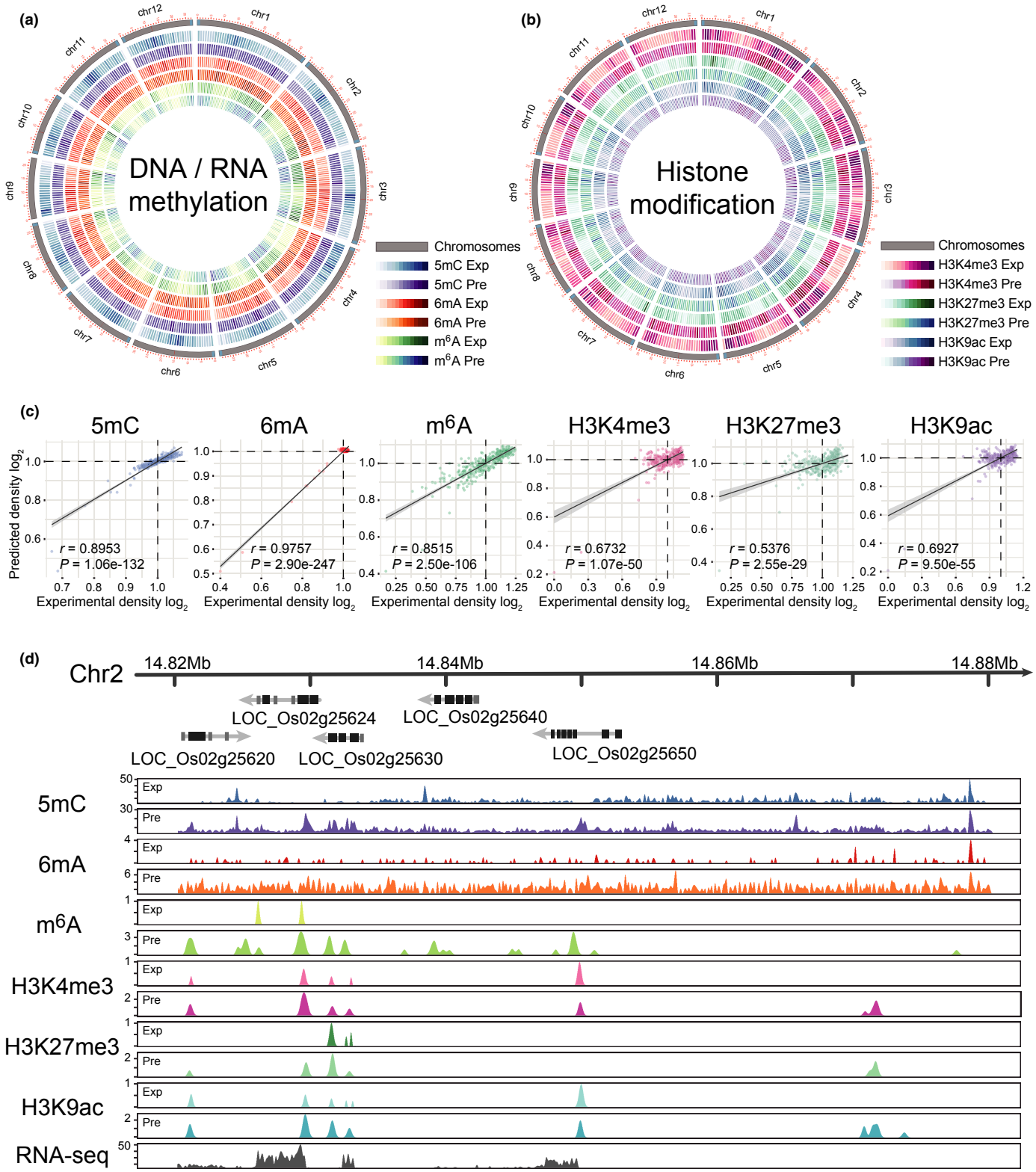



**Fig. 5** Genome-wide comparison of epigenetic modifications in the experimental and Smart Model for Epigenetics in Plants (SMEP) prediction datasets for Nipponbare rice. (a, b) Circos plot analysis of the DNA 6mA, 5mC, and RNA $m^6A$ (a), as well as histone modifications (b) in Nipponbare rice. (c) Pearson correlation analysis to compare the experimental and SMEP prediction datasets; data were plotted as the $log_2$ predictions vs $log_2$ experimental coverage. (d) Zoomed-in analysis of one genomic region of chromosome 2. Gene exons with transcriptional directions are easily identified; for each dataset of epigenetic marks and RNA-seq, the peak sites represent the enrichment coverage. For each track of epigenetic marks, the experimental version is at the top, the bottom represents the SMEP predictions.

the predicted H3K4me3, H3K27me3 and H3K9ac peaks, which positioned within 29 491, 25 806 and 31 557 genes, respectively (Table S5; Dataset S2). We then compared epigenomic features in regions 3000 nt upstream and downstream of transcription start sites (TSSs) between the ChIP-seq data and the SMEP predictions, which revealed similar distribution patterns (Fig. 4e). Taken together, these results support that SMEP can accurately predict histone modifications.

## Epigenome landscape scale comparison of Smart Model for Epigenetics in Plants predictions and experimental data

We next widened the scope of our analysis and compared features between the SMEP predictions and experimental data at a genome-wide scale. A Circos plot analysis supported strong overlap of the rice genomic landscapes for the predicted and experimentally confirmed 5mC and 6mA, and $m^6A$ modifications (Fig. 5a). Similarly, strong overlap was evident in a Circos plot comparing predicted and experimentally confirmed histone marks (Fig. 5b). We have calculated the positive rate of epigenomic modifications at a genome-wide scale (Fig. S5). We conducted a Pearson correlation analysis among the predicted and experimentally confirmed data for each modification type, which highlighted strong correlations, especially for the nonhistone modifications: the correlation coefficient ranged from a low of 0.5376 for H3K27me3 to a high of 0.9757 for 6mA (Fig. 5c).

Although there was clear overlap and strong correlations, these genome-wide comparisons also indicated that there were some genome regions where the predictions did not match the experimental data. We therefore zoomed-in on one specific region of chromosome 2 (from 14.82 to 14.88 Mb) to further examine discrepancies between the SMEP predictions and experimental data (Fig. 5d). For all types of epigenomic and epi-transcriptomic modifications predicted by SMEP, we plotted the experimental data, SMEP predictions, and RNA-seq data for this region (Fig. 5d). There were many more predicted modifications in this region than experimentally confirmed modifications. Moreover, we found that there were some positions for which the specific type of epigenetic modification did not match between the SMEP predictions (i.e. the SMEP prediction of RNA $m^6A$ and H3K27me3 was more enriched in the gene *LOC_02g25630*).

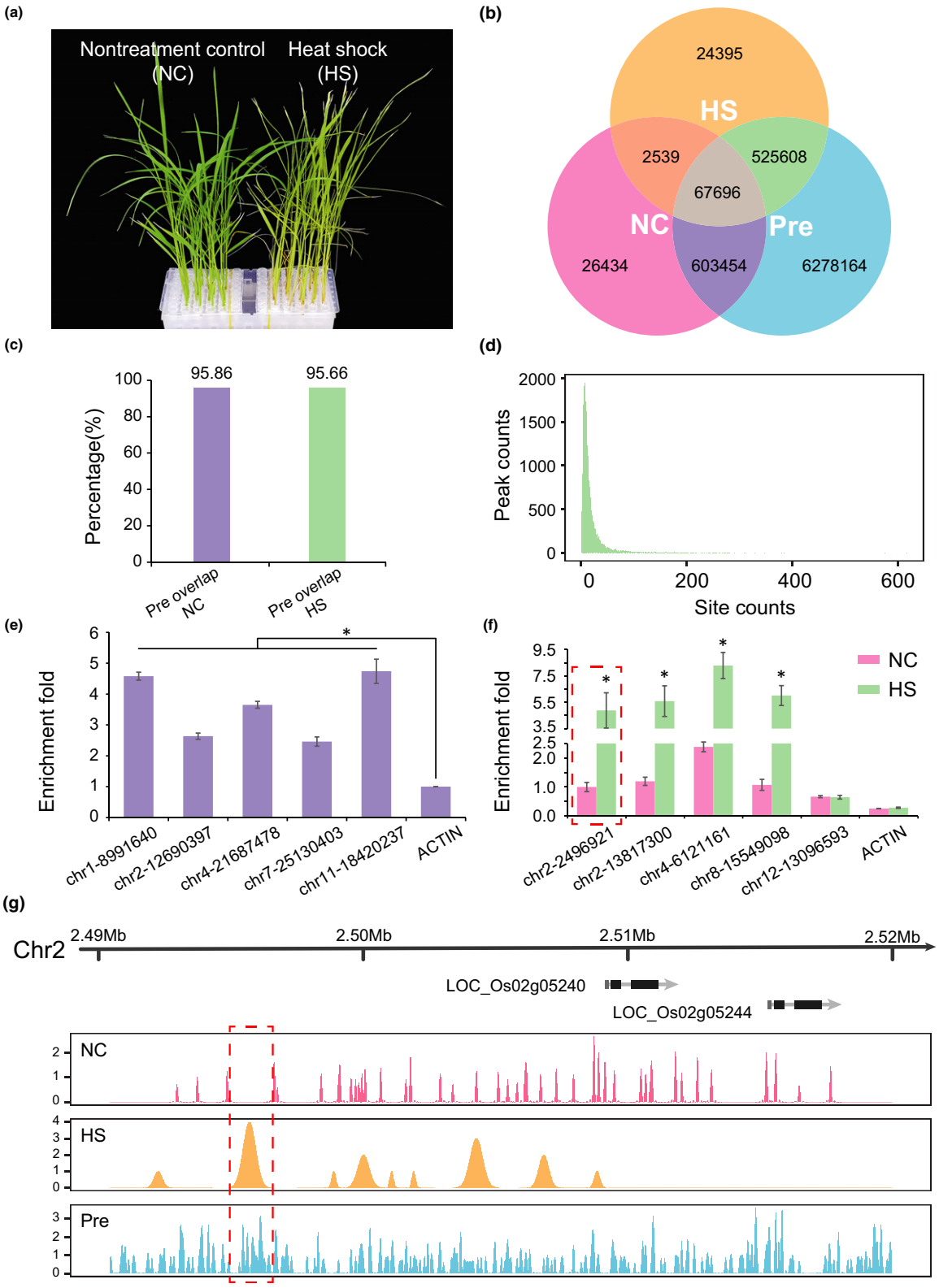
There are at least two plausible factors that impact these differences. For the first factor, unlike genomes, epigenomes are dynamic on short timescales, and the set of modifications present at the moment of a given analysis is going to differ massively depending on the various contexts that the sample plant is experiencing (e.g. growth stage, time of day, light intensity, nutrient status, biotic and abiotic stressors, etc.). On the one hand, perhaps the samples do not have these modifications because their experimental context does not induce such modifications. On the other hand, it is possible that the experimental context of the data we used to train the SMEP models could have been biased in some way that is reflected in its predictions. The second factor impacting the discrepancies is the potential for excessively numerous predictions from SMEP. That is, with the training we conducted and with the settings we deployed for our predictions of the various epigenetic modification types, SMEP may simply output a very large number of predicted modification sites.

The aforementioned results for SMEP were all for data from the *japonica* rice variety Nipponbare. However, there are multiple previous studies that have experimentally evaluated epigenetic modifications – including 5mC and 6mA, and three histone modifications (H3K4me3, H3K27me3 and H3K9ac) – occurring in the *indica* rice variety 93-11 (He *et al.*, 2010; Zhang *et al.*, 2018). We were therefore able to explore the cross-validation of our SMEP models using the 93-11 datasets. Similar to the results we obtained with SMEP prediction for the *japonica* Nipponbare, it was clear that SMEP trained from Nipponbare data can predict epigenetic modifications for *indica* rice (Table S6; Figs S6, S7). Together, these findings support that SMEP can accurately and intelligently predict the likely occurrence of epigenetic sites across the rice genome.

## A heat shock experiment to facilitate comparisons of Smart Model for Epigenetics in Plants-predicted and experimentally confirmed epigenetic modifications

We subjected rice plants to a HS stress treatment to facilitate comparisons of SMEP-predicted and experimentally confirmed epigenetic modifications for samples reflecting different environmental conditions than the sample data used for initial SMEP training. Specifically, we performed a 6mA-IP-seq analysis of Nipponbare seedlings exposed to HS conditions and normal growth conditions (nontreatment control, NC) to enable direct

**Fig. 6** Validation of 6mA Smart Model for Epigenetics in Plants (SMEP) in Nipponbare seedlings given a heat shock (HS) treatment. (a) Nipponbare seedlings grown under nontreatment control (NC), and HS treatment at 45°C for 36 h. (b) Venn diagram of 6mA-IP-seq for plants given a HS treatment, NC and SMEP-predicted datasets in Nipponbare rice. Comparison of SMEP predictions and experimental datasets for Nipponbare seedlings grown under normal conditions or with HS stress. Venn diagrams were used to plot the overlap between the SMEP predictions and experiments datasets. (c) Overlap of 6mA experiments under NC or HS treatment with SMEP predictions. (d) Peak counts of 6mA experiments under HS treatment overlapped with SMEP predictions. (e) Validation of 6mA-SMEP-prediction sites overlapped with 6mA sites under normal conditions (nontreatment control, NC) by 6mA-IP-qPCR. 6mA-IP-qPCR was conducted using 3-wk-old Nipponbare seedlings grown normal conditions. The five positive loci examined were randomly selected for 6mA-IP-qPCR analysis. The *ACTIN* serves as a negative control. Error bars, mean SD. *, significant at $P < 0.05$. (f) Validation of 6mA-SMEP-prediction sites overlapped with 6mA sites under HS treatment by 6mA-IP-qPCR. 6mA-IP-qPCR was conducted using 3-wk-old Nipponbare seedlings grown under NC or HS treatment. The five positive loci examined were randomly selected for 6mA-IP-qPCR analysis. The dotted box indicated the specific genomic region where SMEP-predicted 6mA sites were only overlapped with 6mA under HS treatment, as in the zoomed-in genomic region in (g). *ACTIN* served as a negative control. Error bars, mean SD. *, significant at $P < 0.05$. (g) Zoomed-in analysis of one genomic region of chromosome 2 (chr 2). Gene exons with transcriptional directions are easily identified. For each epigenetic mark, the peak sites represent the enrichment. For each track of epigenetic marks, the experimental version is indicated on top; the SMEP predictions are below. The dotted box indicates the specific genomic region where SMEP-predicted 6mA sites only overlapped with 6mA under HS treatment, as confirmed by 6mA-IP-qPCR.




comparisons of the SMEP 6mA predictions from the model initially trained on the 6mA-SMRT-seq data (Fig. 6a,b). As expected, we found more than 95% of the detected 6mA peaks were predicted by SMEP, but many of the detected 6mA peaks did not overlap with 6mA sites in normal growth conditions (Fig. 6b,c). Among these detected 6mA peaks under HS treatment, more than 80% of peaks shared 1–20 SMEP-predicted methylated sites (Fig. 6d). Next, we performed 6mA-IP-qPCR assays to confirm many SMEP-predicted 6mA sites under normal or HS conditions. By 6mA-IP-qPCR, we first validated 6mA sites

enriched under NC (Fig. 6e). Under HS treatment, most 6mA sites were significantly enriched compared to that in NC (Fig. 6f), indicating SMEP could accurately predict the potential epigenomic sites for diverse environmental conditions. Moreover, by zoomed-in analysis, we also noted as expected that some peaks from the 6mA-IP-seq data from the heat shocked samples were not detected in the control condition samples (Fig. 6g), and found that SMEP had predicted more than 95% of the 6mA modifications uniquely present in the control samples.

Previously, 6mA marks have been demonstrated to affect gene expression in response to diverse environmental signals in plants, especially in HS response signaling, which is associated with both dynamic changes of 6mA levels and altered expression of HS-related genes (Liang *et al.*, 2018a, 2020; Zhang *et al.*, 2018b). SMEP predicted many 6mA sites on these HS-related genes (Fig. S8a). The 6mA-IP-qPCR analysis showed that increased 6mA levels were induced by HS treatment at three HS transcriptional regulator genes, *Heat shock transcription factor A1* (*HsfA1; LOC_Os03g63750*) (Zhang *et al.*, 2018) and *ABA-responsive element binding protein 1* (*AREB1; LOC_Os06g10880*) (Yang *et al.*, 2011), *Dehydration-responsive element binding transcription factor 2A* (*DREB2A*; *LOC_Os01g07120*) (Fig. S8b–d; Table S7). Notably, we have also experimentally confirmed that the 66.7% sites (14 confirmed in 21 SMEP-predicted $m^6A$ and H3K4me3 modifications) were validated in *HsfA1*, *AREB1*, and *DREB2A* loci at the HS condition (Tables S8, S9). Therefore, our experimental-enabled case study supports that the predicted modifications are not substantially enriched with false positives in terms of responses to the different growth conditions or stages. It is thus clear that SMEP predictions are useful in facilitating the interpretation of experimental data for samples from diverse growth conditions or treatments.

### Application of Smart Model for Epigenetics in Plants in maize

We next tested the utility of SMEP with other plants, specifically focusing on the maize line B73, which is widely used in plant genetics research. Similar to the training procedure in rice, we evaluated the DNA methylation (5mC), RNA methylation ($m^6A$) and histone modification (H3K4me3) from previous studies (Li *et al.*, 2015; Perduns *et al.*, 2015; Miao *et al.*, 2020).

Using the SMEP trained from datasets of 5mC, $m^6A$ and H3K4me3, we found that the AUROC scores and other metrics for these marks were similarly high as for rice (Figs 7a, S9a,b). We also generated a Circos plot of these three modification types at a genome-wide scale for maize and found similar genome-wide patterns between experimental datasets and SMEP predictions (Fig. 7b). That is, the SMEP for maize predicted a total of 321 027 821 5mC, 579 706 $m^6A$ and 2131 992 H3K4me3 for maize, which respectively covered 98.07%, 80.05% and 83.36% of the experimental data available for maize (Fig. 7c). Additionally, we further analyzed the location of the predicted 5mC, $m^6A$ and H3K4me3 peaks, which positioned within 39 591, 4653 and 4647, respectively (Table S5; Dataset S2). We then evaluated motifs and distribution patterns for $m^6A$ and H3K4me3 modifications and found similar trends for experimental and SMEP-predicted datasets in maize (Figs 7d, S9c,d). Moreover, Pearson correlation analysis revealed strong positive correlation values between the maize predictions and maize experimental data (Fig. 7e), again supporting that SMEP can accurately predict DNA, RNA and histone marks in maize.

### The Smart Model for Epigenetics in Plants website tool

To facilitate the widespread use of our SMEP prediction tool in plant science, we have implemented SMEP as an online tool (http://www.elabcaas.cn/smep/index.html) (Fig. S10). This website contains experimental and SMEP-predicted datasets from 5mC and 6mA, RNA $m^6A$, and three histone modifications (H3K4me3, H3K27me3 and H3K9ac), and RNA-seq for rice, as well as the 5mC, RNA $m^6A$, and H3K4me3 datasets for maize. The input format for query sequences is FASTA, and users select which model is suitable for their sequence of interest (e.g. rice model trained from the Nipponbare SMRT-seq dataset); the output file from the SMEP website contains sequence names, scores, and predicted modification status across the submitted sequence (Fig. S11). Alternatively, users can simply run a genomic-data-visualization with the included JBROWSE tool (Buels *et al.*, 2016), to integrate data for epigenomic states; tracks can be chosen to obtain experimental and SMEP-predicted epigenomic sites at different regions at the genome-wide scale (Fig. S12). Detailed instructions for use of the prediction tools are available at the help page of the SMEP website.

## Discussion

Herein, we developed a SMEP tool, a deep-learning-based epigenomic predictor for multiple types of epigenetic modifications that is based on a CNN algorithm. Our SMEP predicts epigenomic modifications including 5mC and 6mA, RNA $m^6A$, and three types of histone modifications (H3K4me3, H3K27me3, and H3K9ac) in rice. We also extended its utility in the *indica* group cultivar 93-11 at a genome-wide scale, and verified our SMEP model in the maize 5mC, $m^6A$, and H3K4me3 datasets as well as in the *Arabidopsis* 6mA dataset (Fig. S13), again finding that SMEP results in accurate predictions.

It should be noted that during our initial explorations of the SMEP, we conducted a series of experiments to analyze the efficiency and to avoid over-optimistic performance. Initially, we have compared prediction accuracy of the SMEP model based on diverse input sequence lengths with the 6mA training datasets (ranging from 11 nt to 61 nt, at 10 nt intervals), as well as different ratios of positive and negative data points. Similar to previous reports about how sequence length impacts model training for 6mA (Chen *et al.*, 2019; Lv *et al.*, 2019; Pian *et al.*, 2019), we found that 41 nt sequences for the training dataset and a 1 : 1 positive : negative ratio yield the highest scores (Figs S14, S15). Using the same training datasets as previously reported (Chen *et al.*, 2019; Pian *et al.*, 2019; Yu & Dai, 2019), SMEP indicated

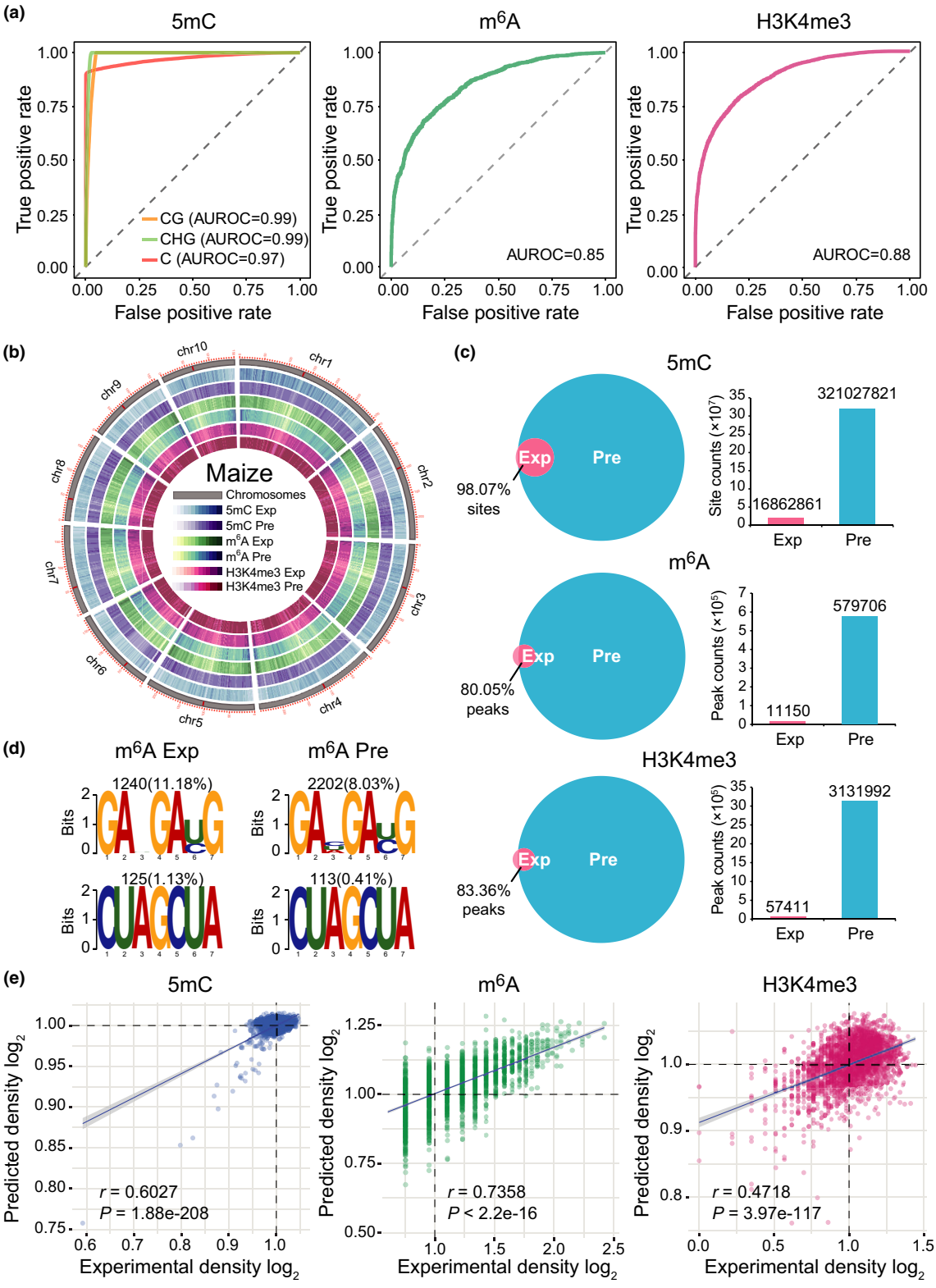



**Fig. 7** Validation of the Smart Model for Epigenetics in Plants (SMEP) using epigenomic datasets in maize. (a) Receiver operating characteristic (ROC) curve of DNA 5mC, RNA m$^6$A, and H3K4me3 in maize. ROC was calculated from classifying preferential site usage on 5mC, m$^6$A and H3K4me3 datasets based on the 10-fold cross-validation. (b) Circos plot analysis of the experimental and predicted DNA 5mC and RNA m$^6$A and H3K4me3 datasets in maize. (c) Venn plot and counts of 5mC, m$^6$A, and H3K4me3 in maize. We gathered statistics of 5mC, m$^6$A, and H3K4me3 sites or peaks between our SMEP predictions and previous experiments. Venn diagrams were used to plot the overlap between the SMEP predictions and experimental datasets. (d) Motif analysis for m$^6$A between the experimental and prediction datasets in maize. The sequence logo representations of the consensus motifs were identified by MEME with parameter (-zoops -n 10 -minw 5 -maxw 9, others were default). (e) Pearson correlation analysis to compare the experimental and SMEP prediction datasets of DNA 5mC and RNA m$^6$A and H3K4me3 in maize; data were plotted as the $\log_2$ predictions vs $\log_2$ experimental coverage.

higher performance metrics (Table S10). Moreover, upon training the SNNRice-6mA model with our SMRT-seq dataset (Yu & Dai, 2019), SMEP again outperformed the SNNRice-6mA method (Table S11).

To explore whether overfitting is a factor for our SMEP prediction, we have implemented CD-HIT software (Fu *et al.*, 2012) to cluster the training data (DNA sequence identity threshold = 0.7) to remove sequence redundancy. Compared to the prediction accuracy from our previous strategy without considering the sequence redundancy, we obtained a very slightly lower accuracy when using this training and validation split strategy (Fig. S16). Moreover, using 6mA datasets in Nipponbare and 93-11, we have separately divided the training, validation and independent testing datasets (Nipponbare and 93-11) to cross-validate the accuracy (Fig. S17a). Note that the sequence identity of independent testing datasets was less than 70% by the CD-HIT split strategy. In two models for 6mA datasets from Nipponbare and 93-11, we separately obtained high scores of prediction accuracies (Fig. S17b), and again gain high prediction probabilities (Fig. S17c). Moreover, as our previous prediction accuracy for the SMEP was conducted based on 10-fold cross-validation, now we have also conducted experiments to evaluate the efficiency of the SMEP using 50% and 10% independent testing datasets; all showed high prediction accuracy scores (Accuracy, AUROC, F1, Precision, Recall) (Fig. S18).

In addition, we have experimentally confirmed SMEP-predicted 6mA, $m^6A$ and H3K4me3 modifications that were not included in the training dataset at three HS related genes of *HsfA1*, *AREB1*, and *DREB2A* (Tables S7–S9), revealing that the predicted modifications are not substantially enriched with false positives in terms of responses to the different growth conditions or stages. Thus, it is likely that over-prediction is not inherently problematic when the predictions are deployed in an experimental design scenario. Finally, it will be possible to 'dial-in' the number of SMEP predicted modification sites to an appropriate level for a given scientific need by selecting the predicted probability (Fig. S19). For example, the users can easily select those sites with high predicted probability ($\geq 0.7$) from our easy-to-use SMEP web server.

Deep learning is now being deployed in plant epigenetics and epigenomic research (Champigny *et al.*, 2019; Chen *et al.*, 2019; Lv *et al.*, 2019; Pian *et al.*, 2019; Zhang *et al.*, 2020a), and our work illustrates how deep learning for *de novo* prediction of modifications in plant epigenomes can be achieved to generate actionable data for better designing and interpreting experiments. For crop species in particular, we anticipate that deep learning has great potential for the design of breeding strategies in a 'Breeding 4.0 era' (i.e. a post 'Big Data' era) for crop improvement (Wang *et al.*, 2020). Fundamentally, our study underscores the utility of CNN models for successfully predicting epigenetic modifications in diverse plant samples, representing a powerful new *in silico* strategy for genetic improvement research in rice and other plants. In the future, it seems very likely that deep-learning-based epigenetic predictions will be applied variously in studies of the complex, multi-layer regulatory pathways that control plant development and environmental responses. Because of the training step of CNN models, and provided that we can access larger and context-appropriate epigenomic and epi-transcriptomic experimental datasets, we fully anticipate that ever-more-accurate SMEP can be developed. As mentioned, this SMEP-enabled epigenomic prediction can be integrated with transcriptome-wide RNA-seq datasets for different developmental stages and diverse environmental stress conditions, and the models can be deployed by breeders and researchers to both drive biological discoveries and to improve the efficiency of genome engineering.

Despite recent efforts, experiment-based tools for plant epigenomics remain insufficient for accurately characterizing the entire complement of epigenomic modifications occurring at a genome-wide scale. Thus, a major challenge in epigenomic studies is the need to develop computational framework(s) that can manage for example single-cell epi-transcriptome data to accessibly present the three-dimensional (3D) landscape of plant epigenomes. Our work shows that deep learning methods can be harnessed to mine additional value from available experimental datasets in a way that increases the comprehensiveness of our understanding of the epigenome landscape in plants. Thus, alongside the growing appreciation of the major functional impacts of epigenetic modifications on economically important agronomic traits, it is clear that deep learning tools like SMEP for studying epigenetic states will facilitate large advances in plant science, plant breeding, and agriculture.

## Acknowledgements

This work was supported by the Chinese National Transgenic Major Program (2019ZX08010-002), the Central Public-interest Scientific Institution Basal Research Fund (Y2020YJ09), and the National Natural Science Foundation of China (31871606). The authors thank Dr Liwen Yang for performing RNA $m^6A$-IP assay, and Dr Yue Teng and Dr Jaie Woodard for their careful revising of the manuscript and many useful comments. The authors declare no competing interests.

## Author contributions

XG, JT and LP conceived of the study and designed the project; YW, WG and HL conducted computational experiments; PZ conducted biological experiments; HL and JT developed the deep-learning SMEP framework; JT, YW and PZ constructed the SMEP website; PZ, YW and WG performed analyses; GH, XL, QZ and ZD contributed to experiments and the SMEP website; XH and LP contributed to writing and conception; PZ, JT and XG wrote the article. YW, PZ, WG and HL contributed equally to this work.

## ORCID

Xiaofeng Gu https://orcid.org/0000-0003-1127-4261
Weijun Guo https://orcid.org/0000-0002-6001-4856
Hanqing Liu https://orcid.org/0000-0001-7085-4122
Li Pu https://orcid.org/0000-0002-3658-7662

Jian Tian https://orcid.org/0000-0002-9997-6518
Yifan Wang https://orcid.org/0000-0001-9721-2816
Pingxian Zhang https://orcid.org/0000-0001-6305-1006

## Data availability

The data that support the findings of this study are available in Gene Expression Omnibus at https://www.ncbi.nlm.nih.gov/geo/, reference no. (GSE108784, GSE19602, GSE67551), or in Sequence Read Archive at https://www.ncbi.nlm.nih.gov/sra/, reference no. (SRX1073669, SRP153627). These data were derived from the following resources available in the public domain: all of the source code for the SMEP model to GitHub (https://github.com/BRITian/smep) and processed epigenomic training datasets to the SMEP website (http://www.elabcaas.cn/smep/downloads.html).

## References

**Ainscough BJ, Barnell EK, Ronning P, Campbell KM, Wagner AH, Fehniger TA, Dunn GP, Uppaluri R, Govindan R, Rohan TE *et al.* 2018.** A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. *Nature Genetics* **50**: 1735–1743.

**Angermueller C, Lee HJ, Reik W, Stegle O. 2017.** DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology* **18**: 67.

**Arbab M, Shen MW, Mok B, Wilson C, Matuszek Ż, Cassa CA, Liu DR. 2020.** Determinants of base editing outcomes from target library analysis and machine learning. *Cell* **182**: 463–480.

**Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, Goodstein DM, Elsik CG, Lewis SE, Stein L *et al.* 2016.** JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biology* **17**: 66.

**Champigny MJ, Unda F, Skyba O, Soolanayakanahally RY, Mansfield SD, Campbell MM. 2019.** Learning from methylomes: epigenomic correlates of *Populus balsamifera* traits based on deep learning models of natural DNA methylation. *Plant Biotechnology Journal* **18**: 1361–1375.

**Chen W, Lv H, Nie F, Lin H, Hancock J. 2019.** I6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics* **35**: 2796–2800.

**Dunker S, Motivans E, Rakosy D, Boho D, Mäder P, Hornick T, Knight TM. 2021.** Pollen analysis using multispectral imaging flow cytometry and deep learning. *New Phytologist* **229**: 593–606.

**Eraslan G, Avsec Ž, Gagneur J, Theis FJ. 2019.** Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics* **20**: 389–403.

**Fu L, Niu B, Zhu Z, Wu S, Li W. 2012.** CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**: 3150–3152.

**He G, Zhu X, Elling AA, Chen L, Wang X, Guo L, Liang M, He H, Zhang H, Chen F *et al.* 2010.** Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. *Plant Cell* **22**: 17–33.

**He Y, Li Z. 2018.** Epigenetic environmental memories in plants: establishment, maintenance, and reprogramming. *Trends in Genetics* **34**: 856–866.

**Hoffman GE, Bendl J, Girdhar K, Schadt EE, Roussos P. 2019.** Functional interpretation of genetic variants using deep learning predicts impact on chromatin accessibility and histone modification. *Nucleic Acids Research* **47**: 10597–10611.

**Holder LB, Haque MM, Skinner MK. 2017.** Machine learning for epigenetics and future medical applications. *Epigenetics* **12**: 505–514.

**Kang D-H, Kim J-H, Oh S, Park H-Y, Dugasani SR, Kang B-S, Choi C, Choi R, Lee S, Park SH *et al.* 2019.** A neuromorphic device implemented on a salmon-DNA electrolyte and its application to artificial neural networks. *Advanced Science* **6**: 1901265.

**Kawakatsu T, Huang S-S, Jupe F, Sasaki E, Schmitz RJ, Urich MA, Castanon R, Nery JR, Barragan C, He Y *et al.* 2016.** Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. *Cell* **166**: 492–505.

**Kim HK, Yu G, Park J, Min S, Lee S, Yoon S, Kim HH. 2021.** Predicting the efficiency of prime editing guide RNAs in human cells. *Nature Biotechnology* **39**: 198–206.

**Lämke J, Bäurle I. 2017.** Epigenetic and chromatin-based mechanisms in environmental stress adaptation and stress memory in plants. *Genome Biology* **18**: 124.

**Lecun Y, Bengio Y, Hinton G. 2015.** Deep learning. *Nature* **521**: 436–444.

**Li Q, Gent JI, Zynda G, Song J, Makarevitch I, Hirsch CD, Hirsch CN, Dawe RK, Madzima TF, McGinnis KM *et al.* 2015.** RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. *Proceedings of the National Academy of Sciences, USA* **112**: 14728–14733.

**Li Y, Wang X, Li C, Hu S, Yu J, Song S. 2014.** Transcriptome-wide N6-methyladenosine profiling of rice callus and leaf reveals the presence of tissue-specific competitors involved in selective mRNA modification. *RNA Biology* **11**: 1180–1188.

**Li Z, Fu X, Wang Y, Liu R, He Y. 2018.** Polycomb-mediated gene silencing by the BAH–EMF1 complex in plants. *Nature Genetics* **50**: 1254–1261.

**Li Z, Jiang H, Kong L, Chen Y, Lang K, Fan X, Zhang L, Pian C. 2021.** Deep6mA: a deep learning framework for exploring similar patterns in DNA N6-methyladenine sites across different species. *PLoS Computional Biology* **17**: e1008767.

**Liang Z, Geng Y, Gu X. 2018a.** Adenine methylation: new epigenetic marker of DNA and mRNA. *Molecular Plant* **11**: 1219–1221.

**Liang Z, Riaz A, Chachar S, Ding Y, Du H, Gu X. 2020b.** Epigenetic modifications of mRNA and DNA in plants. *Molecular Plant* **13**: 14–30.

**Liang Z, Shen L, Cui X, Bao S, Geng Y, Yu G, Liang F, Xie S, Lu T, Gu X *et al.* 2018b.** DNA N6-adenine methylation in *Arabidopsis thaliana*. *Developmental Cell* **45**: 406–416.

**Liu X, Yang S, Zhao M, Luo M, Yu CW, Chen CY, Tai R, Wu K. 2014.** Transcriptional repression by histone deacetylases in plants. *Molecular Plant* **7**: 764–772.

**Lv H, Dao FY, Guan ZX, Zhang D, Tan JX, Zhang Y, Chen W, Lin H. 2019.** iDNA6mA-Rice: a computational tool for detecting N6-methyladenine sites in rice. *Frontiers in Genetics* **10**: 793.

**Lv H, Dao FY, Zhang D, Guan ZX, Yang H, Su W, Liu ML, Ding H, Chen W, Lin H. 2020.** iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes. *iScience* **23**: 100991.

**Miao Z, Zhang T, Qi Y, Song J, Han Z, Ma C. 2020.** Evolution of the RNA N6-methyladenosine methylome mediated by genomic duplication. *Plant Physiology* **182**: 345–360.

**Perduns R, Horst-Niessen I, Peterhansel C. 2015.** Photosynthetic genes and genes associated with the C4 trait in maize are characterized by a unique class of highly regulated histone acetylation peaks on upstream promoters. *Plant Physiology* **168**: 1378–1388.

**Pian C, Zhang G, Li F, Fan X, Hancock J. 2019.** MM-6mAPred: identifying DNA N6-methyladenine sites based on Markov model. *Bioinformatics* **36**: 388–392.

**Pu L, Sung ZR. 2015.** PcG and trxG in plants – friends or foes. *Trends in Genetics* **31**: 252–262.

**Qiu T, Shi X, Wang J, Li Y, Qu S, Cheng Q, Cui T, Sui S. 2019.** Deep learning: a rapid and efficient route to automatic metasurface design. *Advanced Science* **6**: 1900128.

**Rampasek L, Goldenberg A. 2016.** TensorFlow: biology's gateway to deep learning? *Cell Systems* **2**: 12–14.

**Roundtree IA, Evans ME, Pan T, He C. 2017.** Dynamic RNA modifications in gene expression regulation. *Cell* **169**: 1187–1200.

**Shen L, Liang Z, Gu X, Chen Y, Teo ZWN, Hou X, Cai WM, Dedon PC, Liu L, Yu H. 2016.** N6-methyladenosine RNA modification regulates shoot stem cell fate in *Arabidopsis*. *Developmental Cell* **38**: 186–200.

**Shen L, Liang Z, Wong CE, Yu H. 2019.** Messenger RNA modifications in plants. *Trends in Plant Science* **24**: 328–341.

**Shrikumar A, Greenside P, Kundaje A. 2017.** Learning important features through propagating activation differences. In: Precup D, Teh YW, eds. *Proc. 34th International Conference on Machine Learning*. Sydney, New South Wales, Australia: International Convention Centre PMLR, 3145–3153.

**Sun P, Chen Y, Liu B, Gao Y, Han Y, He F, Ji J. 2019.** DeepMRMP: a new predictor for multiple types of RNA modification sites using deep learning. *Mathematical Biosciences and Engineering* **16**: 6231–6241.

**Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012.** Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7**: 562–578.

**Wang H, Cimen E, Singh N, Buckler E. 2020.** Deep learning for plant genomics and crop improvement. *Current Opinion in Plant Biology* **54**: 34–41.

**Wang J, Wang L. 2020.** Deep analysis of RNA N6-adenosine methylation ($m^6A$) patterns in human cells. *NAR Genomics and Bioinformatics* **2**: lqaa007.

**Warman C, Sullivan CM, Preece J, Buchanan ME, Vejlupkova Z, Jaiswal P, Fowler JE. 2021.** A cost-effective maize ear phenotyping platform enables rapid categorization and quantification of kernels. *The Plant Journal* **106**: 566–579.

**Washburn JD, Mejia-Guerra MK, Ramstein G, Kremling KA, Valluru R, Buckler ES, Wang H. 2019.** Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. *Proceedings of the National Academy of Sciences, USA* **116**: 5542–5549.

**Wu X, Zhang Y. 2017.** TET-mediated active DNA demethylation: mechanism, function and beyond. *Nature Reviews Genetics* **18**: 517–534.

**Xu Y, Wang Y, Luo J, Zhao W, Zhou X. 2017.** Deep learning of the splicing (epi)genetic code reveals a novel candidate mechanism linking histone modifications to ESC fate decision. *Nucleic Acids Research* **45**: 12100–12112.

**Yang H, Wei Q, Li D, Wang Z. 2020.** Cancer classification based on chromatin accessibility profiles with deep adversarial learning model. *PLoS Computional Biology* **16**: e1008405.

**Yang X, Yang YN, Xue LJ, Zou MJ, Liu JY, Chen F, Xue HW. 2011.** Rice ABI5-like1 regulates abscisic acid and auxin responses by affecting the expression of ABRE-containing genes. *Plant Physiology* **156**: 1397–1409.

**Yu G, Wang LG, He QY. 2015.** ChIP seeker: an R/bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**: 2382–2383.

**Yu H, Dai Z. 2019.** SNNRice6mA: a deep learning method for predicting DNA N6-methyladenine sites in rice genome. *Frontiers in Genetics* **10**: 1071.

**Zhang P, Wang Y, Chachar S, Tian J, Gu X. 2020a.** eRice: a refined epigenomic platform for *japonica* and *indica* rice. *Plant Biotechnology Journal* **18**: 1642–1644.

**Zhang P, Wang Y, Gu X. 2020b.** RNA 5-methylcytosine controls plant development and environmental adaptation. *Trends in Plant Science* **25**: 954–958.

**Zhang P, Zhu C, Geng Y, Wang Y, Yang Y, Liu Q, Guo W, Chachar S, Riaz A, Yan S *et al.* 2021.** Rice and *Arabidopsis* homologs of yeast CHROMOSOME TRANSMISSION FIDELITY PROTEIN 4 commonly interact with Polycomb complexes but exert divergent regulatory functions. *Plant Cell* **33**: 1417–1429.

**Zhang Q, Liang Z, Cui X, Ji C, Li Y, Zhang P, Liu J, Riaz A, Yao Pu, Liu M *et al.* 2018b.** N6-methyladenine DNA methylation in *japonica* and *indica* rice genomes and its association with gene expression, plant development, and stress responses. *Molecular Plant* **11**: 1492–1508.

**Zhang Y, An L, Xu J, Zhang B, Zheng WJ, Hu M, Tang J, Yue F. 2018a.** Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. *Nature Communications* **9**: 750.

**Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W *et al.* 2008.** Model-based analysis of ChIP-Seq (MACS). *Genome Biology* **9**: R137.

**Zhou C, Wang C, Liu H, Zhou Q, Liu Q, Guo Y, Peng T, Song J, Zhang J, Chen L *et al.* 2018.** Identification and analysis of adenine N6-methylation sites in the rice genome. *Nature Plants* **4**: 554–563.

**Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. 2019.** A primer on deep learning in genomics. *Nature Genetics* **51**: 12–18.

## Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

**Dataset S1** List of hyper-parameters for machine learning algorithms.

**Dataset S2** Associated genes containing the experimental or predicted epigenetic sites of 5mC, 6mA, $m^6A$, H3K4me3, H3K27me3 and H3K9ac in Nipponbare genome, and 5mC, $m^6A$, and H3K4me3 in B73 genome.

**Fig. S1** Schematic CNN models used in this study.

**Fig. S2** Prediction performance of 5mC, 6mA, $m^6A$ and histone modifications in Nipponbare rice by SMEP.

**Fig. S3** Computational analysis for 5mC and 6mA based on CNN and machine learning algorithms.

**Fig. S4** Motif analysis and Pearson correlation analysis.

**Fig. S5** Genome-wide positive rate of epigenetic marks using CNN models.

**Fig. S6** Application of the SMEP using 5mC and 6mA datasets for the *indica* 93-11.

**Fig. S7** Application of the SMEP using H3K4me3, H3K27me3, and H3K9ac datasets for the *indica* 93-11.

**Fig. S8** Validation of the SMEP-predicted 6mA sites in HS-related genes.

**Fig. S9** SMEP can accurately predict of DNA 6mA, RNA $m^6A$ and H3K4me3 in maize.

**Fig. S10** The home page of SMEP website.

**Fig. S11** The prediction page of SMEP website.

**Fig. S12** The JBrowse page of SMEP website.

**Fig. S13** ROC curve and prediction performance by SMEP for predicting *Arabidopsis* 6mA dataset.

**Fig. S14** Performance of the different sequence lengths of 6mA training datasets in Nipponbare rice with the SMEP.

**Fig. S15** Performance obtained when using different ratios of negative and positive datasets, for the 6mA training datasets of Nipponbare rice with the SMEP.

**Fig. S16** Comparison of the performance with random and CD-HIT split using CNN model.

**Fig. S17** Strategies for deep learning to predict DNA 6mA modifications.

**Fig. S18** Comparison of the performance with different split ratio of training and independent testing datasets of 6mA using CNN model.

**Fig. S19** Cumulative histogram on the probability cut-off of all experimentally determined and predicted sites in Nipponbare rice.

**Methods S1** Supporting methods performed in this study.

**Table S1** Summary of datasets trained by SMEP.

**Table S2** List of hyper-parameters for our CNN model.

**Table S3** The training times of SMEP.

**Table S4** Summary of primers by 6mA-IP-qPCR, ChIP-qPCR.

**Table S5** Gene numbers of training datasets and prediction datasets.

**Table S6** Prediction performance of 5mC, 6mA, and histone modifications in 93-11 rice by SMEP.

**Table S7** Validation of the SMEP-predicted 6mA sites in HS-related genes.

**Table S8** Validation of the SMEP-predicted $m^6A$ peaks in HS-related genes.

**Table S9** Validation of the SMEP-predicted H3K4me3 peaks in HS-related genes.

**Table S10** Performance comparison between SMEP and several previous methods on 6mA-Rice-Chen dataset.

**Table S11** Performance comparison between SMEP and SNNRice6mA on different datasets.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.