4acC-IP-seq is similar to methylated DNA IP (MeDIP) and 6mA-IP-seq,15,16 applying the 4acC-specific antibody IP method to enrich DNA fragments containing 4acC modification and then constructing a high-throughput sequencing library. 4acC-IP-seq revealed that the enriched 4acC modification peaks were mostly distributed around the transcription start sites of protein-coding genes among euchromatin regions. Potential interactions of 4acC with 5mC and histones were also observed in the context of gene-expression regulation. Furthermore, the existence of 4acC in genomic DNA samples of rice, maize, mouse, and Homo sapiens was also confirmed by mass spectrometry. Therefore, precise identification of 4acC in eukaryotic DNA is crucial for exploring its biological function and its interplays with other epigenetic marks.

In practice, wet-lab experiments to detect modifiable sites, such as mass spectrometry and antibody-based sequencing method, are often time consuming with a high cost, and the specific antibodies used for IP sequencing restrict the accuracy of sequencing results.17 To date, various computational models based on DNA or RNA sequences have been proposed to serve as useful alternatives. The existing algorithms can be roughly divided into two categories: feature-based algorithms and deep-learning-based algorithms. Examples of the former include iDNA6mA-Rice,18 i6mA-Fuse,19 SDM6A,20 Methylator,21 MethCGI,22 iDNA-Methyl,23 4mCPred,24 4mCPred-SVM,25 4mCPred-EL,26 4mCPred-IFL,27 Apart from 5mC to 6mA, other DNA methylations have not been extensively detected and explored. Recently, inspired by direct analogs of chemical modifications of RNA and DNA, such as m5C/5mC, hm5C/5hmC, and m6A/6mA, as well as the widely distributed and highly conserved RNA modification N4-acetylcytosine (ac4C),12,13 Wang et al. proposed 4acC immunoprecipitation followed by sequencing (4acC-IP-seq) to explore the presence and function of 4acC in DNA of Arabidopsis thaliana.14 The protocol of 4acC-IP-seq is similar to methylated DNA IP (MeDIP) and 6mA-IP-seq,15,16 applying the 4acC-specific antibody IP method to enrich DNA fragments containing 4acC modification and then constructing a high-throughput sequencing library. 4acC-IP-seq revealed that the enriched 4acC modification peaks were mostly distributed around the transcription start sites of protein-coding genes among euchromatin regions. Potential interactions of 4acC with 5mC and histones were also observed in the context of gene-expression regulation. Furthermore, the existence of 4acC in genomic DNA samples of rice, maize, mouse, and Homo sapiens was also confirmed by mass spectrometry. Therefore, precise identification of 4acC in eukaryotic DNA is crucial for exploring its biological function and its interplays with other epigenetic marks.

In practice, wet-lab experiments to detect modifiable sites, such as mass spectrometry and antibody-based sequencing method, are often time consuming with a high cost, and the specific antibodies used for IP sequencing restrict the accuracy of sequencing results.17 To date, various computational models based on DNA or RNA sequences have been proposed to serve as useful alternatives. The existing algorithms can be roughly divided into two categories: feature-based algorithms and deep-learning-based algorithms. Examples of the former include iDNA6mA-Rice,18 i6mA-Fuse,19 SDM6A,20 Methylator,21 MethCGI,22 iDNA-Methyl,23 4mCPred,24 4mCPred-SVM,25 4mCPred-EL,26 4mCPred-IFL,27
Meta-4mCpred,28 and i4mC-ROSE, which rely on hand-crafted features to represent sequence context.29 For instance, iDNA6mA-Rice applied multiple encoding schemes, including PseKNC, single-nucleotide binary encoding, and natural vectors, and predicted 6mA using random forest on rice DNA. While the performance of feature-based methods relies heavily on choosing the best representation in each case, many deep-learning-based methods have recently been proposed to learn from raw sequences and achieve superior performance, including iDNA6mA-Rice-DL,30 DNA6mA-MINT,31 4mCPred-CNN,32 Deep6mAPred,33 BiLSTM-5mC,34 and so on. For instance, 4mCPred-CNN is the first method based on a convolutional neural network (CNN) to identify 4mC sites in the mouse genome.32

Most existing frameworks are based on strong supervision, which requires the precise location of the modified bases. However, such data are currently unavailable for 4acC. The only high-throughput-sequencing-based technology, 4acC-IP-seq, only allows the detection of 4acC-carrying DNA fragments of at least 200–400 bases in length.14 Because there usually exist multiple cytosines in each fragment, it is unclear which ones are modified. Such data do not allow the model to learn modification-specific sequence contexts from fixed-length sequences centered on the target cytosine as strongly supervised learning does. Instead, the only label information that can be used for training is associated with DNA regions of different lengths, i.e., whether the region contains at least one 4acC site. To address the challenge of learning from these coarse-grained labels, here we consider a multiple-instance learning (MIL) framework, one of the weakly supervised learning algorithms.

Weakly supervised learning aims to construct predictive models by learning from noisy, limited, or imprecise sources. In genomics, weakly supervised learning, especially MIL, is widely considered in protein-DNA interaction prediction,47–50 where the bound DNA sequence may contain multiple binding sites, and the exact location is unknown. WSCNN and its updated version, WSCNNLSTM, combine MIL with deep neural networks and have achieved superior results in in vivo and in vitro transcription factor binding site prediction.49,50 In addition to DNA modeling, MIL has been applied to automate the annotation of protein functions,51 protein splice variants,52 specific functional binding sites in microRNA targets,53 and proteome-wide interactions.54 More recently, to predict RNA modifications from only low-resolution epitranscriptome data, WeakRM combined MIL and attention mechanisms and showed promising performance on three RNA modifications, including ac4C.55 WeakRM divided the RNA sequence into multiple fixed-length subsequences. For peak regions called by bioinformatics tools, at least one subsequence contains a target modification and should cover specific sequence patterns. Thus, the integrated representation of all subsequences can be linked to a positive label. Whereas for regions from the same transcript but not detected as peaks, modification-specific sequence patterns should not be included in any subsequences, and their integration remains negative. Inspired by the WeakRM framework and considering the direct analog of ac4C and 4acC, we consider the 4acC prediction from 4acC-IP-seq data naturally and inherently an MIL task.

We propose 4acCPred, the first prediction framework for high-accuracy identification of 4acC-carrying regions from Arabidopsis genomic DNA sequences. Under the MIL framework, we combine CNN and bidirectional long short-term memory (LSTM) to exploit their advantages in local motif extraction and long-term interaction learning, respectively. A simplified graphic framework of 4acCPred is illustrated in Figure 1. Evaluation of all four conditions provided by 4acC-IP-seq (wild type, NH2OH treatment, met1 mutant, and

![Figure 1. A simplified graphic illustration of the proposed 4acCPred framework](image_url)
ros1dm12dml3 [rdd] mutant) demonstrated the general effectiveness of our approach in predicting 4acC from DNA regions (average areas under the receiver operating characteristic curve [AUCs] in a 10-fold cross-validation test were 0.9877 and 0.9899 in independent testing). Furthermore, model interpretation showed that 4acCPred also captured motifs consistent with existing knowledge. The 4acCPred web server, accessible via http://www.rnamd.org/4accpred, is designed to help users predict 4acC modifications and visualize captured motif patterns on Arabidopsis genomic DNA. All data used in this study (the peak information called by MACS2 with annotation) and trained 4acCPred models have also been uploaded to the web server for user convenience. We anticipate that our newly proposed model and the web server can take full advantage of limited experimental data and facilitate the study of DNA 4acC modification by providing alternative computational prediction approaches.

RESULTS

We developed the first DNA 4acC modification predictor based on MIL. To demonstrate the model’s stability, we formed four benchmark datasets based on four conditions provided in 4acC-IP-seq (GEO: GSE168538) and evaluated the model performance separately. The initial learning rate, decayed learning rate, instance length, and instance stride are set as 5 × 10⁻⁴, 1 × 10⁻⁵, 40, and 5, respectively. Details can be found in the materials and methods. For each dataset, one-third of the data was selected as an independent test dataset. The proposed method was found to be robust on both cross-validation and independent testing.

The whole structure of the model contains a four-layer encoding module (see Figure 1): after the first convolutional layer, a max-pooling layer withdraws weak features in datasets to expand the receptive field. There also involves a dropout layer to prevent overfitting in training the model. Finally, a bidirectional LSTM layer captures the hidden long-term dependencies between sequential patterns. Each instance passes through the same encoding module (weights are shared) and outputs instance-level features. The network learns weights for each instance and sums all instance features as features for the entire input sequence for final classification.

Comparison with existing frameworks developed for protein-DNA binding prediction

We compared our algorithms 4acCPred with WSCNN and its updated version WSCNNLSTM using the same datasets, which were originally developed for transcription factor binding site prediction. We split each dataset equally into ten partitions to perform 10-fold cross-validation. Table 1 shows the average performances in terms of the AUC, average precision (AP), Matthews correction coefficient (MCC), and accuracy of different algorithms under cross-validation. In the wild-type case, the AUC of 4acCPred is about 0.1 higher than that of WSCNN and WSCNNLSTM. 4acCPred also achieved an improvement of at least 0.07 AUC under the other three conditions. Under all four conditions, 4acCPred had an AUC of at least 0.97 and an AP of at least 0.96, indicating its promising performance in predicting 4acC-carrying regions.

Performance evaluation on independent test datasets

To further test the performance and robustness of our newly proposed predictor in finding 4acC-carrying regions, we apply the ten models obtained from the cross-validation for each condition to held-out independent test datasets and show their average results. As shown in Table 2, with accuracy, AUC, MCC, and AP as evaluation metrics, the results of our model are consistent with those in cross-validation. A significant improvement from the baseline model to 4acCPred can also be observed. The average performances in terms of AUC, AP, MCC, and accuracy of different algorithms under the ensemble of these ten models are shown in Table S1. It is pleasing that all metrics of 4acCPred are higher than WSCNN and WSCNNLSTM.

Table 1. Performance of 4acCPred under 10-fold cross-validation with standard deviations

| Model     | Group | Accuracy | AUC      | AP       | MCC       |
|-----------|-------|----------|----------|----------|-----------|
| WSCNN     | WT    | 0.7089   | 0.8052   | 0.8041   | 0.4474    |
|           | NH2OH | 0.6860   | 0.8306   | 0.8271   | 0.4265    |
|           | met1  | 0.6788   | 0.8705   | 0.8695   | 0.4337    |
|           | Rdd   | 0.6668   | 0.8558   | 0.8584   | 0.4121    |
| WSCNNLSTM | WT    | 0.8064   | 0.8801   | 0.8453   | 0.6175    |
|           | NH2OH | 0.8124   | 0.8881   | 0.8578   | 0.6263    |
|           | met1  | 0.8511   | 0.9226   | 0.9015   | 0.7027    |
|           | Rdd   | 0.8433   | 0.9216   | 0.8985   | 0.6884    |
| 4acCPred  | WT    | 0.9485   | 0.9794   | 0.9695   | 0.8981    |
|           | NH2OH | 0.9506   | 0.9857   | 0.9800   | 0.9019    |
|           | met1  | 0.9679   | 0.9928   | 0.9975   | 0.9360    |
|           | Rdd   | 0.9706   | 0.9928   | 0.9890   | 0.9414    |

All methods were evaluated using the same datasets with a positive-to-negative ratio of 1:1. The performance is given as average ± standard deviation. The threshold for accuracy is 0.5. *WT represents the condition of wild type.
same BED results from MACS2, 58,59 unsurprisingly, the motif
Compared with the known motifs corresponding motifs on two strands, respectively.
’Figure 2, our proposed model tends to assign high weights to CT-
training data, it is natural to obtain two motifs separately from
of supporting seqlets. The top 1 result from wild-type models is
quence, we selected the consensus motif with the highest number
representative motifs. After pruning through the overall letter fre-
cant importance scores, then clusters and aligns all seqlets to obtain
some evidence that our model learns true biological signals rather than technical biases
and that attention weights have the potential to be used to predict
acetylation at higher resolutions.
Motifs identified by 4acCPred are consistent with existing knowledge
In 4acCPred, we quantified attribution scores for each input feature
using the integrated gradient (IG) method. 56 Then, we extracted
consensus motifs from instances by using TF-MoDISco. 57 TF-
MoDISco first identifies subsequences (called seqlets) with significant importances, then clusters and aligns all seqlets to obtain representative seqlets. After pruning through the overall letter frequency, we selected the consensus motif with the highest number of supporting seqlets. The top 1 result from wild-type models is shown in Figure 2. Due to the lack of strand information in the training data, it is natural to obtain two motifs separately from the two strands, and they should complement each other. According to Figure 2, our proposed model tends to assign high weights to CT-enriched or GA-enriched (complement) subsequences. Since 4acC is a kind of DNA modification, and we input sequence on both '+' and '-' strands to train the model, 4acCPred provides two corresponding motifs on two strands, respectively.

Compared with the known motifs identified by HOMER with the same BED results from MACS2, 58,59 unsurprisingly, the motif learned by 4acCPred with the strictest score (a total of 229 seqlets were found) is consistent with the top 1 motif found by HOMER. To numerically measure the similarity between these two motifs, we applied the motif comparison tool MEME-Tomtom, 60 resulting in a significant p value of 0.0019. More results from the other three conditions can be found in Figures S1–S3, and the top 5 motifs found by 4acCPred for each condition are shown in Figures S4–S7. Together, we show that the motifs revealed by our model are consistent with existing knowledge, providing some evidence that our model learns true biological signals rather than technical biases and that attention weights have the potential to be used to predict acetylation at higher resolutions.

Web implementation
To facilitate the use of our model and assess of the used data, a web server has been developed using Hyper Text Markup Language (HTML), Cascading Style Sheets (CSSs), and JavaScript (JS) and is accessible at http://www.rnamd.org/4accpred (see Figure 3). The web server allows users to upload DNA sequences in FASTA format and provides predicted probability scores under user-specified conditions (wild type [WT], NH2OH, rdd, and met1 mutant). All the results could be downloaded in a CSV format file. The display of motifs is available upon request. In addition, all processed data originating from 4acC-IP-seq and analyzed using MACS2 can be freely downloaded from the server.

DISCUSSION
4acC is a recently discovered abundant DNA chemical modification that is involved in various gene-expression functions. Precise prediction of 4acC modification-containing regions is vital for scientific investigations to understand its role in biological regulation. In this study, we designed the first 4acC predictor, 4acCPred, on Arabidopsis based on weakly supervised learning and LSTM. We collected data from all four conditions provided by 4acC-IP-seq and divided them into cross-validation and independent test datasets. 4acCPred achieved average AUCs of 0.9877 and 0.9899 in the two datasets, respectively. Additionally, the motif discovered by 4acCPred with the strictest score is consistent with motifs found in existing knowledge. Together, these results demonstrate the robustness of our model as a useful alternative to detect DNA 4acC acetylation. To facilitate

Since an essential question in deep learning is to what extent the system learns peculiarities of a particular experimental setup (rather than the underlying biology), we designed an experiment for performance comparisons where training and evaluation data come from a different experiment to deal with this. 4acCPred also makes the best prediction (see Table S2).

| Table 2. Performance of 4acCPred on independent test datasets with standard deviations |
|-----------------|-----------|-------|---|---|
| Model          | Group     | Accuracy | AUC  | AP  | MCC |
| WSCNN          | WT        | 0.7101 (±0.0216) | 0.8051 (±0.0483) | 0.8044 (±0.0260) | 0.4590 (±0.0251) |
|                | NH2OH     | 0.6831 (±0.0452) | 0.8293 (±0.0652) | 0.8250 (±0.0372) | 0.4209 (±0.0601) |
|                | met1      | 0.6766 (±0.0402) | 0.8691 (±0.0740) | 0.8694 (±0.0424) | 0.4301 (±0.0524) |
|                | Rdd       | 0.6728 (±0.0590) | 0.8643 (±0.0744) | 0.8647 (±0.0426) | 0.4213 (±0.0898) |
| WSCNNLSTM      | WT        | 0.8119 (±0.0032) | 0.8866 (±0.014)  | 0.8551 (±0.0020) | 0.6288 (±0.0054) |
|                | NH2OH     | 0.8154 (±0.0033) | 0.8893 (±0.0018) | 0.8607 (±0.0035) | 0.6324 (±0.0068) |
|                | met1      | 0.8526 (±0.0036) | 0.9241 (±0.0020) | 0.9018 (±0.0032) | 0.7061 (±0.0074) |
|                | Rdd       | 0.8594 (±0.0053) | 0.9284 (±0.0011) | 0.9056 (±0.0020) | 0.7200 (±0.0107) |
| 4acCPred       | WT        | 0.9512 (±0.0015) | 0.9855 (±0.0006) | 0.9796 (±0.0007) | 0.9031 (±0.0031) |
|                | NH2OH     | 0.9530 (±0.0017) | 0.9859 (±0.0005) | 0.9779 (±0.0009) | 0.9066 (±0.0033) |
|                | met1      | 0.9728 (±0.0033) | 0.9950 (±0.0004) | 0.9926 (±0.0006) | 0.9458 (±0.0065) |
|                | Rdd       | 0.9726 (±0.0016) | 0.9932 (±0.0002) | 0.9891 (±0.0006) | 0.9453 (±0.0031) |

All methods were evaluated using the same datasets with a positive-to-negative ratio of 1:1. The performance is given as average ± standard deviation. The threshold for accuracy is 0.5. *WT represents the condition of wild type.
the use of 4acCPred, we have also built a user-friendly web server for prediction and motif visualization, available at http://www.rnamd.org/4accpred. It is worth noting that 4acCPred is currently only constructed for Arabidopsis, and its performance in other species needs further experimental verification.

MATERIALS AND METHODS

Benchmark datasets
The high-throughput 4acC-IP-seq samples were collected from the recently published article.14 All data were downloaded from NCBI Gene Expression Omnibus (GEO) under GEO: GSE168538 (see Table 3). Confident 4acC peaks were obtained through peak calling of MACS2 (https://github.com/macs3-project/MACS).59 The sequencing results of four conditions were used to form four datasets separately: WT, NH2OH treatment, met1 mutant, and rdd mutant. Since there were two biological replicates for each condition, to improve data quality, we used the intersection of the two replicates as positive data. Considering that a single DNA fragment is 200-400 nt in length, we randomly supplemented regions that were too short after the intersection to 200-nt to ensure a consistent peak width distribution. Such supplemented sequences still contain the peak information and treat the weighted sum of all instance features as instances. Specifically, a sliding window of length c moves along the bag sequence with stride s. This means that each time a subsequence of length c is extracted as an instance, the next instance starts s-bases downstream of where the previous instance started. If the length of a specific bag sequence is L, there will be \( \frac{L-c}{s} +1 \) instances in total. Window length c and stride s are two hyper-parameters. In our study, we chose a sliding window of length 50 and a stride of 10 for ac4C RNA modification in 4acCPred. The available label information is associated with the entire bag, not with each instance. The underlying logic of the MIL framework used in 4acCPred is that the network should highlight instances in positive data that contain target acetylation and capture their sequential patterns. Conversely, for negative data, the model should treat all instances as negative.

The network structure used to extract sequence features consists of one convolutional layer and one bidirectional LSTM layer. One of the greatest strengths of CNNs in genomics is that it naturally captures sequence motifs for a given target through its local receptive fields. However, it inevitably overlooks hidden long-term dependencies between sequential patterns, which can be addressed by using LSTMs. Between CNN and LSTM, we add a max-pooling layer to filter weak features and expand the receptive field and a dropout layer to prevent overfitting in model training. It is worth noting that the network uses shared weights to extract features for each instance.

A key step of the MIL framework is to merge the instance-level features to obtain the bag-level probability (i.e., the predicted value that a DNA sequence contains at least one 4acC acetylation). Unlike WSCNN and WSCNNLSTM,49,50 which first let the network output a score for each instance and then use functions such as mean, max, and Noisy-and to aggregate scores into one value for the entire bag,61,62 we assign weights to each instance using an attention mechanism and treat the weighted sum of all instance features as the final bag representation. Specifically, we use gated attention,
which consists of three fully connected layers. The attention weight \( a_k \) is calculated as follows:

\[
a_k = \frac{\exp \left( w^T \left( \tanh (V b^i_j) \odot \text{sigm}(U b^i_j) \right) \right)}{\sum_{j=1}^{K} \exp \left( w^T \left( \tanh (V b^i_j) \odot \text{sigm}(U b^i_j) \right) \right)}
\]

where \( K \) represents the number of instances in a bag; \( b^i_j \) is the hidden representation of instance \( i \); \( w, U, \) and \( V \) are weights of three neural network layers; \( \tanh \) and \( \text{sigm} \) are tanh and sigmoid activation functions; and \( T \) means transposition. The weights are normalized to guarantee that all weights sum to 1 so that the network can handle the input of any number of instances.

The advantage of this feature-merging approach is that the network can learn to assign high weights to instances with preferred sequence patterns. The attention weight can also be considered as an indicator to infer the subsequences most likely to contain acetylation. Conversely, score merging using a fixed function tends to lose information. For example, using the maximum value will force the network to make decisions based on only one instance, ignoring sequence context from upstream and downstream. Using the mean function is limited by the fact that acetylation is sparsely distributed across many instances. WSCNN and WSCNNLSTM demonstrated that Noisy-and is better than max and mean functions. However, although Noisy-and allows a learnable threshold, it is still constructed based on the average score of the instances, which may suffer from the
same drawbacks of mean function, i.e., the model is insensitive to positive instances due to a large number of negative instances.

Motif mining
In this study, we use the IG method to perform model interpretation and modification motif mining. The IG method was developed based on backpropagation, a key design of neural networks. It uses the gradient value from the final output to each input multiplied by the input value itself as the contribution of that input (known as attribution score). To solve the problem of gradient saturation, the IG method first selects a reference for each input, performs linear interpolation from the reference to the input, calculates the attribute score of each interpolation point, and then averages all scores to get final values. Its mathematical formula is given as

\[ IG_i(x) = \frac{1}{m} \sum_{k=1}^{m} \frac{\partial F(x') + \frac{1}{2} \frac{\partial}{\partial x} (x - x')}{\partial x} \times \frac{1}{m} \]

where \( x \) is the input to be interpreted, \( x' \) is a selected reference of the same shape as \( x \), and \( m \) is the number of steps in linear interpolation.

Two kinds of reference selection methods were explored in this study: fixed reference for all inputs such as zero matrices and dinucleotide-shuffled sequence for each specific input. The former is computationally efficient and can provide a cleaner view when interpreting individual sequences using one-hot encodings because, in this case, only one value remains for each base, directly corresponding to the contribution of that base. The dinucleotide-shuffled sequence refers to shuffling the sequence but maintaining the frequency of dinucleotides. Compared with the zero matrices, it takes more time but is more biologically interpretable and thus was used for motif discovery in our study. Both references can be selected in the web server analysis.

After obtaining attribution maps for each input, we used TF-ModISCO to obtain the consensus motif for DNA 4acC acetylation. It firstly identifies input segments of user-specific length with high contribution scores, then clusters these segments based on continuous Jaccard similarity calculation, and finally aligns the segments in each cluster to form consensus motifs.

Web interface implementation
The web interface has been established by HTML, CSSs, and JS. MySQL database management systems were used to store our data. Datatables, a table plugin, was applied to show the data.

DATA AVAILABILITY STATEMENT
The raw data used in this study are publicly available in the GEO database under GEO: GSE168538. All processed data can be downloaded from the 4acCPred web server at http://www.rnamd.org/4accpred.

SUPPLEMENTAL INFORMATION
Supplemental information can be found online at https://doi.org/10.1016/j.omtn.2022.10.004.

ACKNOWLEDGMENTS
This work was supported by the National Natural Science Foundation of China (32100519 and 31671373) and the XJTLU Key Program Special Fund (KSF-E-51 and KSF-P-02).

AUTHOR CONTRIBUTIONS
D.H., J.M., and Z.W. contributed to the conception of the study; J.Z., D.H., and J.M. contributed significantly to the analysis and manuscript preparation; J.Z. and D.H. performed the experiment; J.Z. and D.H. performed the data analyses and wrote the manuscript; X.W. and J.Z. designed and conducted the web server; D.H., J.M., and Z.W. helped perform the analysis with constructive discussions.

DECLARATION OF INTERESTS
The authors declare no competing interests.

REFERENCES
1. Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. 13, 484–492. https://doi.org/10.1038/nrg3230.

Table 3. Positive peaks collected in 4acCPred

| Modification | Technology      | Condition             | Size | Species | Sample                        |
|--------------|----------------|-----------------------|------|---------|-------------------------------|
| 4acC         | 4acC-IP-seq    | Wild type (IP)        | 19,849 | Tair10  | GSM5145699<sup>a</sup>        |
| 4acC         | 4acC-IP-seq    | NH₂OH (WT IP)         | 16,682 | Tair10  | GSM5145692<sup>a</sup>        |
| 4acC         | 4acC-IP-seq    | met1 (mutate IP)      | 10,989 | Tair10  | GSM5145698<sup>a</sup>        |
| 4acC         | 4acC-IP-seq    | rdd (mutate IP)       | 9,844  | Tair10  | GSM5369300<sup>a</sup>        |
| 4acC         | 4acC-IP-seq    | Wild type (input)     | –     | Tair10  | GSM5145694<sup>b</sup>        |

The ratio of positive and negative labels is 1:1.

<sup>a</sup>Size denotes the total number of sequences.

<sup>b</sup>Wild-type input sample GSM5145694 is used to call peaks by MACS2.
39. Song, Z., Huang, D., Song, B., Chen, K., Song, Y., Liu, G., Su, J., Magalhaes, J.P.d., Rigden, D.J., and Meng, J. (2021). Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications. Nat. Commun. 12, 4011. https://doi.org/10.1038/s41467-021-24313-3.

40. Ao, C., Zhou, Q., and Yu, L. (2022). MedRF: identification of multispecies RNA 2′-O-methylation modification sites from RNA sequences. Brief. Bioinform. 23, bbab480. https://doi.org/10.1093/bib/bbab480.

41. Ma, J., Zhang, L., Li, S., and Liu, H. (2021). BRCPA: bounded robust principal component analysis to incorporate similarity network for N7-methylguanosine(m7G) site-disease association prediction. IEEE/ACM Trans. Comput. Biol. Bioinf. In preparation. https://doi.org/10.1109/TCBB.2021.3109055.

42. Zhang, L., Li, G., Li, X., Wang, H., Chen, S., and Liu, H. (2021). ELDA:m6APred: ensemble deep learning approach for mRNA m6A site prediction. BMC Bioinf. 22, 288. https://doi.org/10.1186/s12859-021-04206-4.

43. Zhang, L., Chen, S., Ma, J., Liu, Z., and Liu, H. (2021). REW-ISA V2: a biclustering method fusing homologous information for analyzing and mining Epitranscriptome data. Front. Genet. 12, 654820. https://doi.org/10.3389/fgene.2021.654820.

44. Ma, J., Zhang, L., Chen, J., Song, B., Zang, C., and Liu, H. (2021). m7GDisAI: N7-methylguanosine (m7G) sites and diseases associations inference based on heterogeneous network. BMC Bioinf. 22, 152. https://doi.org/10.1186/s12859-021-04007-9.

45. Zhang, L., Chen, J., Ma, J., and Liu, H. (2021). HN-CNN: a heterogeneous network based on convolutional neural network for m7G site disease association prediction. Front. Genet. 12, 653284. https://doi.org/10.3389/fgene.2021.653284.

46. Wu, X., Wei, Z., Chen, K., Zhang, Q., Su, J., Liu, H., Zhang, L., and Meng, J. (2019). m6Acomet: large-scale functional prediction of individual m6A RNA methylation sites from an RNA co-methylation network. BMC Bioinf. 20, 223. https://doi.org/10.1186/s12859-019-28642-0.

47. Gao, Z., and Ruan, J. (2015). A structure-based Multiple-Instance Learning approach through multi-instance multi-label learning. IEEE/ACM Trans. Comput. Biol. Bioinf. 11, 891–902. https://doi.org/10.1109/TCBB.2014.2323058.

48. Panwar, B., Menon, R., Eksi, R., Li, H.D., Omenn, G.S., and Guan, Y. (2016). Genome-wide functional annotation of human protein-coding splice variants using multiple instance learning. J. Proteome Res. 15, 1747–1753. https://doi.org/10.1021/acs.jproteome.5b00883.

49. Bandyopadhyay, S., Ghosh, D., Mitra, R., and Zhao, Z. (2015). MBSTAR: multiple instance learning for predicting specific functional binding sites in microRNA targets. Sci. Rep. 5, 8004. https://doi.org/10.1038/srep08004.

50. Mei, S., and Zhu, H. (2014). AdaBoost based multi-instance transfer learning for predicting proteome-wide interactions between Salmonella and human proteins. PLoS One 9, e110488. https://doi.org/10.1371/journal.pone.0110488.

51. Huang, D., Song, B., Wei, J., Su, J., Coenen, F., and Meng, J. (2021). Weakly supervised learning of RNA modifications from low-resolution epitranscriptome data. Bioinformatics 37, i222–i230. https://doi.org/10.1093/bioinformatics/btab278.

52. Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1703.01365.

53. Shulkumar, A., Tian, K., Avsec, Z., Scherberina, A., Banerjee, A., Sharmin, M., Nair, S., and Kundaje, A. (2018). Technical note on transcription factor motif discovery from importance scores (TF-MoDISco). Preprint at arXiv. https://doi.org/10.48550/arXiv.1811.00416.

54. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.G., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004.

55. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-seq (MACS). Genome Biol. 9, R137. https://doi.org/10.1186/gb-2008-9-9-r137.

56. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biol. 8, R24. https://doi.org/10.1186/gb-2007-8-2-r24.

57. Kraus, O.Z., Ba, J.L., and Frey, B.J. (2016). Classifying and segmenting microscopy images with deep multiple instance learning. Bioinformatics 32, i52–i59. https://doi.org/10.1093/bioinformatics/btw252.

58. Dauphin, Y.N., Pan, A., Auli, M., and Grangier, D. (2017). Language modeling with gated convolutional networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1612.08083.

59. Sotoudeh, M., and Thakur, A.V. (2019). Computing linear restrictions of neural network. Preprint at arXiv. https://doi.org/10.48550/arXiv.1908.06214.

60. Shulkumar, A., Greenside, P., and Kundaje, A. (2018). Technical note on transcription factor motif discovery from importance scores (TF-MoDISco). Preprint at arXiv. https://doi.org/10.48550/arXiv.1908.06214.