iScience, Volume *23*

Supplemental Information

iDNA-MS: An Integrated Computational Tool for Detecting DNA Modification Sites in Multiple Genomes

Hao Lv, Fu-Ying Dao, Dan Zhang, Zheng-Xing Guan, Hui Yang, Wei Su, Meng-Lu Liu, Hui Ding, Wei Chen, and Hao Lin

Transparent Methods

Benchmark Dataset

The 5hmC site containing sequences for *H. sapiens* and *M. musculus* were collected from NCBI Gene Expression Omnibus (GEO) database (Hu et al., 2019). The 6mA site data for 11 species (*Arabidopsis thaliana* (*A. thaliana*)*, Caenorhabditis elegans* (*C. elegans*), *Casuarina equisetifolia* (*C. equisetifolia*), *Drosophila melanogaster* (*D. melanogaster*), *Fragaria vesca* (*F. vesca*), *H. sapiens*, *Rosa chinensis* (*R. chinensis*), *Saccharomyces cerevisiae* (*S. cerevisiae*), *Tolypocladium sp SUP5-1* (*Ts*. *SUP5-1*), *Tetrahymena thermophile* (*T. thermophile*) and *Xanthomonas oryzae pv. Oryzicola* (*Xoc*) *BLS256* (*Xoc*. *BLS256*)) were obtained from the MethSMRT database (Ye et al., 2017), published reference (Ye et al., 2019), MethSMRT database (Ye et al., 2017), MDR database (Liu et al., 2019), published reference (Xiao et al., 2018), MDR database (Liu et al., 2019), MethSMRT database (Ye et al., 2017), GEO database (Wang et al., 2017) and NCBI Genome database (Xiao et al., 2018), respectively. The 4mC site data for 4 species (*C. equisetifolia*, *F. vesca*, *S. cerevisiae* and *Ts*. *SUP5-1*,) were obtained from the MDR database (Liu et al., 2019) and MethSMRT database (Ye et al., 2017). Preliminary trials indicated that when the length of the segments is 41 nt with the 5hmC/6mA/4mC in the center, the highest predictive results could be obtained. Thus, the sequences of all positive samples are 41 nt. In order to construct a high quality benchmark dataset, the following two steps were performed. Firstly, for jump-seq data, to ensure the effectiveness of 5hmC calls, we selected the samples with percentage of 5hmC (5hmC calls number/sequencing depth) greater than 95%. For SMRT data, as illustrated in the Methylome Analysis Technical Note (Ye et al., 2017), when the modification QV (modQV) is set to 30 for calling a position as modified, the accuracy can reach to 99.9%. Thus, the sequences with modQV no less than 30 are left for the subsequent analysis. It should be noted that in order to obtain statistically significant results, if the raw data is too small, this step was ignored to get more samples. Secondly, to avoid redundancy and reduce homology bias , sequences with more than 80% sequence similarity were removed using the CD-HIT program (Li and Godzik, 2006). After the above two steps, the objective and strict positive datasets for above species were obtained.

The negative samples (non-5hmC/non-6mA/non-4mC site containing sequences) for the above mentioned 17 genomes were collected by satisfying the requirement that the 41 nt long sequences with Cytosine/Adenine in the center which was not proved to be methylated by experiments. By doing so, a large number of negative samples were obtained. If a model was established on an unbalanced benchmark dataset, its performance will bias. Thus, we randomly extracted negative samples with the same number of positive samples in each of the 17 genomes. Details about these data were shown in Figure S5.

In order to provide a more objective evaluation on performances of the proposed method, we randomly divided the benchmark dataset into two parts by a ratio of 1:1 (See also Table S5). One part is

used as training dataset, the other one is testing dataset. The former part was used to train the model, while the other part was used to test the performance of the corresponding model, which made sure that the training dataset and testing dataset are independent of each other.

The details of the datasets are freely available at (http://lin-group.cn/server/iDNA-MS/download.html)

Feature description

Adopting an effective feature extraction method is a key step in producing an excellent predictor (Manavalan et al., 2018b; Song et al., 2019; Stephenson et al., 2019). This study introduced three feature extraction techniques to formulate 5hmC, 6mA and 4mC samples.

K-tuple Nucleotide Frequency Component

Given an DNA sequence **D** with *L* nucleic acid residue (here *L*=41), its most straightforward expression is:

$$\mathbf{D} = \ R_1 R_2 R_3 R_4 \ \cdots R_i \ \cdots R_{L-1} \ R_L \tag{1}$$

where R_i represents the *i*-th nucleic acid residue at position *i* in the DNA sequence.

Some sequence alignment-based tools, such as BLAST and Bowtie could search local similarity regions between sequences. However, these methods tend to lose sample information and even do not work when processing low-similar sequences. Fortunately, machine learning methods could make up for this shortcoming. However, most of machine leaning methods can only handle vectors with same dimension. Thus, it is a big challenge in bioinformatics to transfer each sample into a fixed length of the feature vector. The *k*-tuple composition (or called *k*-mer) is a smart strategy and has been widely used in genome analysis (Yang et al., 2019). Its principle is to convert each sample into a 4^k dimension vector expressed as:

$$\mathbf{D} = \left[f_1^{k-tuple} f_2^{k-tuple} \cdots f_i^{k-tuple} \cdots f_{4^k}^{k-tuple}\right]^T \tag{2}$$

where the symbol T represents the transposition of the vector, and $f_i^{k-tuple}$ represents the frequency of the i-th k-tuple composition in the DNA sequence sample. Here, we set k=1, 2, 3, 4, which means 4+16+64+256=340 features.

Nucleotide chemical property and Nucleotide frequency

The four nucleic acids have different chemical properties. In terms of ring structures, A and G are purines containing two rings, whereas C and T are pyrimidines containing one ring. In terms of forming secondary structures, C and G form strong hydrogen bonds, whereas A and T form weak hydrogen bonds. In terms of chemical functionality, A and C can be classified into the amino group, while G and T

can be classified into the keto group (Chen et al., 2019). Therefore, three coordinates (*x, y, z*) were used to represent the chemical properties of the four nucleotides and the value of 0 and 1 was assigned to the three coordinates. If the *x* coordinate stands for the ring structure, *y* for the hydrogen bond, and *z* for the chemical functionality, a nucleotide in DNA sequence can be encoded by (x_i, y_i, z_i), where

$$x_i = \begin{cases} 1 \ if \ s_i \in \{\mathrm{A}, \mathrm{G}\} \\ 0 \ if \ s_i \in \{\mathrm{C}, \mathrm{T}\} \end{cases}, \ y_i = \begin{cases} 1 \ if \ s_i \in \{\mathrm{A}, \mathrm{T}\} \\ 0 \ if \ s_i \in \{\mathrm{C}, \mathrm{G}\} \end{cases}, \ z_i = \begin{cases} 1 \ if \ s_i \in \{\mathrm{A}, \mathrm{C}\} \\ 0 \ if \ s_i \in \{\mathrm{G}, \mathrm{T}\} \end{cases} \quad (3)$$

Accordingly, A, C, G and T can be represented by the coordinates (1, 1, 1), (0, 0, 1), (1, 0, 0) and (0, 1, 0), respectively.

For the purpose of extracting nucleotide composition surrounding the modification sites, the density d_i of any nucleotide n_j at position *L* in a sequence was defined as follows

$$d_i = \frac{1}{|N_j|} \sum_{j=1}^{L} f(n_j), \ f(n_j) = \begin{cases} 1 \ if \ n_j = q \\ 0 \ other \ cases \end{cases} \quad (4)$$

where *L* is the sequence length, $|N_j|$ is the length of the *i*-th prefix string $\{n_1, n_2, \cdots, n_i\}$ in the sequence, and $q \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$.

By integrating nucleotide chemical properties and nucleotide frequency, an *L* nt long sequence will be encoded by a (4×*L*)-dimensional vector.

Mono-nucleotide binary encoding

The third feature extraction technique is to transfer nucleotide to a binary code formulated as:

$$n = \begin{cases} (1,0,0,0), \ when \ n = A \\ (0,1,0,0), \ when \ n = C \\ (0,0,1,0), \ when \ n = G \\ (0,0,0,1), \ when \ n = T \end{cases} \quad (5)$$

In our dataset, the sequences are all 41 nt. Thus, an arbitrary DNA sequence with 41 nucleotides can be described as a vector of 164 (4 × 41) features (Wei et al., 2019).

Random Forest (RF)

The RF algorithm is a very powerful algorithm and has been widely used in many areas of computational biology (Schaduangrat et al., 2019; Win et al., 2017; Win et al., 2018). It is a flexible and practical machine learning method. It can handle thousands of input variables without variable deletion and generate an internal unbiased estimate of the generalization error. The principle of RF is to randomly generate many trees by recursive partitioning approach and then aggregate the results according to voting rules. In this study, the number of trees is set to 100 with the seed of 1. The detailed procedures of RF and its formulation have been very clearly described in the reference (Breiman, 2001).

Performance evaluation

Cross-validation is a statistical analysis method for evaluating the performance of a classifier. In order to save computational time, the five-fold cross-validation test was used to estimate the performance of the proposed method on training data in this study. Once the models were determined, the independent datasets were used to evaluate the models. We employed sensitivity (Sn), specificity (Sp), overall accuracy (Acc) and Matthew's correlation coefficient (MCC) to measure the predictive capability of the proposed model (Manavalan et al., 2018a; Song et al., 2018).

$$\begin{cases} Sn = \frac{TP}{TP+FN} \times 100\% \;\; 0 \le Sn \le 1 \\ Sp = \frac{TN}{TN+FP} \times 100\% \;\; 0 \le Sp \le 1 \\ Acc = \frac{TP+TN}{TP+FN+TN+FP} \times 100\% \;\; 0 \le Acc \le 1 \\ MCC = \frac{(TP \times TN)-(FP \times FN)}{\sqrt{(TP+FN) \times (TP+FP) \times (TN+FN) \times (TN+FP)}} \;\; -1 \le MCC \le 1 \end{cases} \quad (6)$$

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

In addition, we also calculated the AUC (area under the receiver operating characteristic curve) to objectively evaluate the proposed model. The AUC ranges from 0 to 1. A model with a higher AUC indicates a better performance.

Supplemental References

Breiman, L. (2001). Random Forests. Machine Learning *45*, 5-32.

Chen, W., Lv, H., Nie, F., and Lin, H. (2019). i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. Bioinformatics *35*, 2796-2800.

Hu, L., Liu, Y., Han, S., Yang, L., Cui, X., Gao, Y., Dai, Q., Lu, X., Kou, X., Zhao, Y.*, et al.* (2019). Jump-seq: Genome-Wide Capture and Amplification of 5-Hydroxymethylcytosine Sites. J Am Chem Soc *141*, 8694-8697.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658-1659.

Liu, Z.Y., Xing, J.F., Chen, W., Luan, M.W., Xie, R., Huang, J., Xie, S.Q., and Xiao, C.L. (2019). MDR: an integrative DNA N6-methyladenine and N4-methylcytosine modification database for Rosaceae. Hortic Res *6*, 78.

Manavalan, B., Govindaraj, R.G., Shin, T.H., Kim, M.O., and Lee, G. (2018a). iBCE-EL: A New Ensemble Learning Framework for Improved Linear B-Cell Epitope Prediction. Front Immunol *9*, 1695.

Manavalan, B., Shin, T.H., and Lee, G. (2018b). PVP-SVM: Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine. Front Microbiol *9*, 476.

Schaduangrat, N., Nantasenamat, C., Prachayasittikul, V., and Shoombuatong, W. (2019). ACPred: a computational tool for the prediction and analysis of anticancer peptides. Molecules *24*, 1973.

Song, J., Li, F., Leier, A., Marquez-Lago, T.T., Akutsu, T., Haffari, G., Chou, K.C., Webb, G.I., Pike, R.N., and Hancock, J. (2018). PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. Bioinformatics *34*, 684-687.
Song, J., Wang, Y., Li, F., Akutsu, T., Rawlings, N.D., Webb, G.I., and Chou, K.C. (2019). iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. Brief Bioinform *20*, 638-658.
Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., Zhang, J., Chan, L., and Cao, R. (2019). Survey of Machine Learning Techniques in Drug Discovery. Curr Drug Metab *20*, 185-193.
Wang, Y., Chen, X., Sheng, Y., Liu, Y., and Gao, S. (2017). N6-adenine DNA methylation is associated with the linker DNA of H2A.Z-containing well-positioned nucleosomes in Pol II-transcribed genes in Tetrahymena. Nucleic Acids Res *45*, 11594-11606.
Wei, L., Luan, S., Nagai, L.A.E., Su, R., and Zou, Q. (2019). Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. Bioinformatics *35*, 1326-1333.
Win, T.S., Malik, A.A., Prachayasittikul, V., JE, S.W., Nantasenamat, C., and Shoombuatong, W. (2017). HemoPred: a web server for predicting the hemolytic activity of peptides. Future Med Chem *9*, 275-291.
Win, T.S., Schaduangrat, N., Prachayasittikul, V., Nantasenamat, C., and Shoombuatong, W. (2018). PAAP: a web server for predicting antihypertensive activity of peptides. Future medicinal chemistry *10*, 1749-1767.
Xiao, C.L., Zhu, S., He, M., Chen, Zhang, Q., Chen, Y., Yu, G., Liu, J., Xie, S.Q., Luo, F*., et al.* (2018). N(6)-Methyladenine DNA Modification in the Human Genome. Mol Cell *71*, 306-318 e307.
Yang, H., Yang, W., Dao, F.Y., Lv, H., Ding, H., Chen, W., and Lin, H. (2019). A comparison and assessment of computational method for identifying recombination hotspots in Saccharomyces cerevisiae. Brief Bioinform.
Ye, G., Zhang, H., Chen, B., Nie, S., Liu, H., Gao, W., Wang, H., Gao, Y., and Gu, L. (2019). De novo genome assembly of the stress tolerant forest species Casuarina equisetifolia provides insight into secondary growth. Plant J *97*, 779-794.
Ye, P., Luan, Y., Chen, K., Liu, Y., Xiao, C., and Xie, Z. (2017). MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. Nucleic Acids Res *45*, D85-D89.

Supplementary Figures

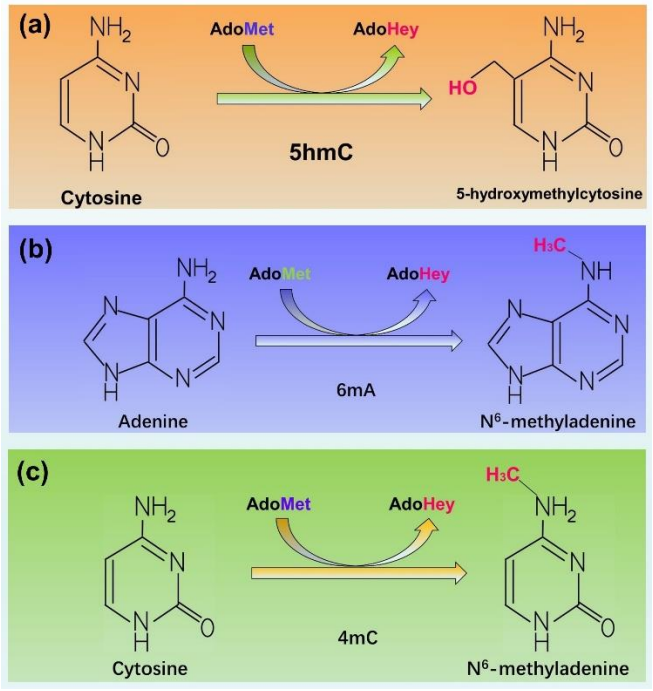

Figure S1. A schematic drawing to show the three types of modifications (5hmC, 6mA, 4mC). These processes are catalyzed by adenine- or cytosine-specific DNA methyltransferases (MTases) that transfer a methyl group from the donor S-adenosyl-L-methionine (AdoMet) to the substrate and generate methylated DNA and S-adenosyl-L-homocysteine (AdoHcy). Related to the Figure 1.

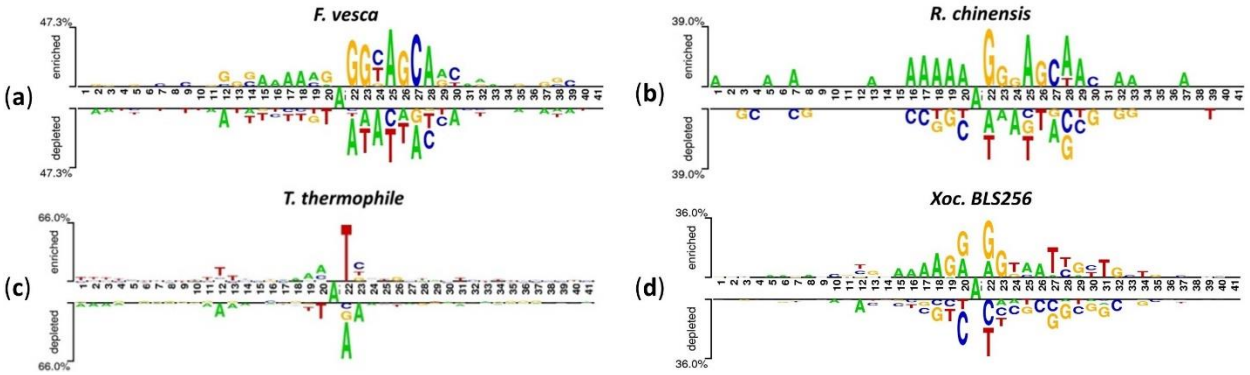

Figure S2. The nucleotide distribution around 6mA and non-6mA sites in (a) *F. vesca*, (b) *R. chinensis*, (c) *T. thermophile* and (d) *Xoc. BLS256*. In each figure, the top panel of the x-axis is for 6mA site containing sequences, while the down panel of the x-axis is for non-6mA site containing sequences. Related to Figure 1.

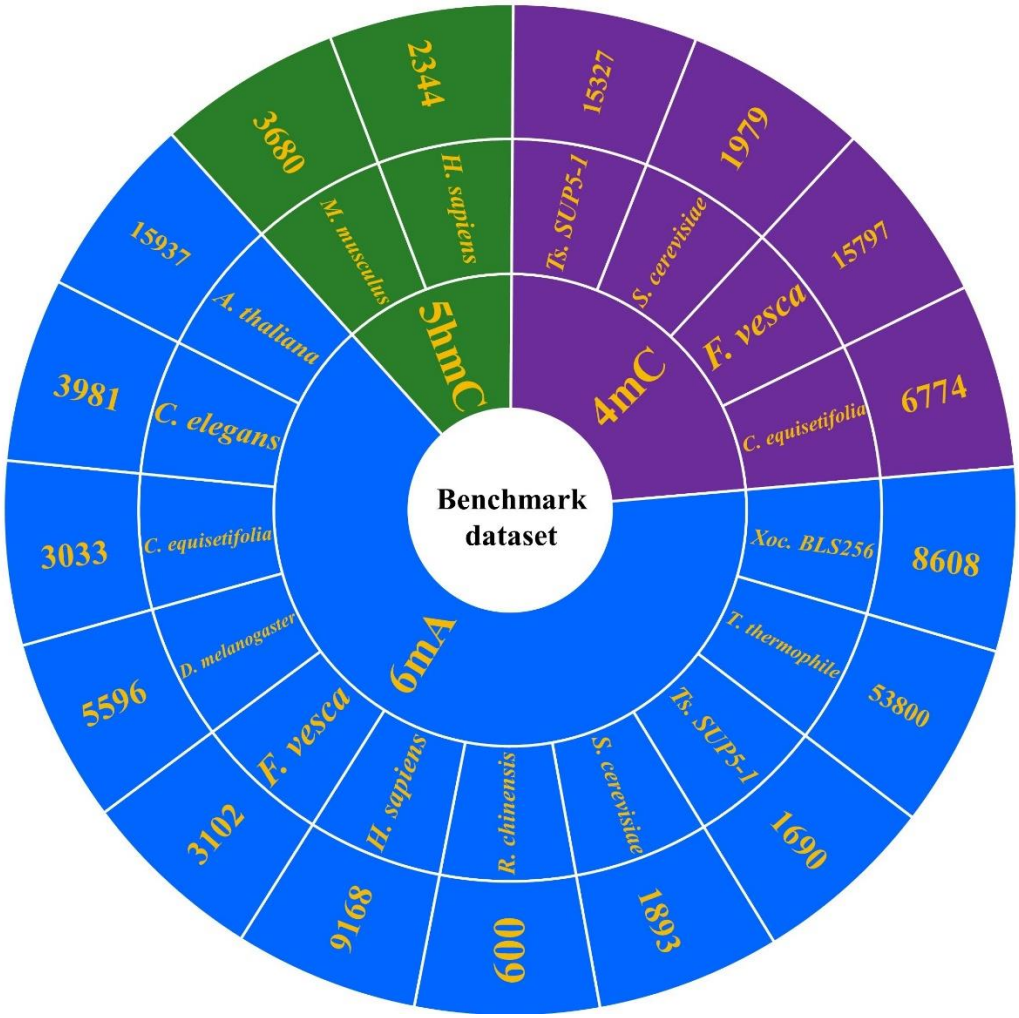

Figure S3. A diagram showing the benchmark datasets. Related to Figures 1-5.

Supplementary Tables

Table S1. Comparison of different features for identifying modification sites in 17 genomes. Related to Figure 2.

Modification type	Genome	Performance	KNFC	NCPNF	MNBE	KNFC-NCPNF	KNFC-MNBE	NCPNF-MNBE	KNFC-NCPNF-MNBE
5hmC	*H. sapiens*	Sn(%)	90.19	**97.35**	97.35	97.10	97.44	97.44	97.44
		Sp(%)	60.67	**92.83**	92.92	90.61	90.78	92.92	92.92
		Acc(%)	75.43	**95.09**	95.14	93.86	94.11	95.18	95.18
		MCC	0.532	**0.903**	0.904	0.903	0.884	0.905	0.905
		AUC	0.821	**0.962**	0.960	0.954	0.956	0.957	0.956
	M. musculus	Sn(%)	97.45	96.25	96.25	97.01	97.12	**96.25**	96.68
		Sp(%)	81.74	97.66	97.83	91.47	91.58	**97.88**	97.58
		Acc(%)	89.59	96.96	97.04	94.24	94.35	**97.07**	96.63
		MCC	0.802	0.939	0.941	0.886	0.888	**0.941**	0.933
		AUC	0.931	0.984	0.984	0.981	0.982	**0.984**	0.983
6mA	*A. thaliana*	Sn(%)	67.57	80.72	**82.02**	78.82	79.99	81.31	80.80
		Sp(%)	67.38	83.32	**84.34**	80.93	81.62	84.75	83.54
		Acc(%)	67.48	82.02	**83.18**	79.87	80.81	82.69	82.17
		MCC	0.350	0.641	**0.664**	0.598	0.616	0.654	0.644
		AUC	0.736	0.896	**0.906**	0.878	0.886	0.903	0.899
	C. elegans	Sn(%)	68.78	82.92	**85.28**	81.79	83.14	83.55	84.15
		Sp(%)	63.78	82.87	**83.95**	79.98	80.86	83.35	82.69
		Acc(%)	66.28	82.89	**84.61**	80.88	82.00	83.45	83.42
		MCC	0.326	0.658	**0.692**	0.618	0.640	0.669	0.668
		AUC	0.723	0.904	**0.922**	0.888	0.902	0.913	0.913
	C. equisetifolia	Sn(%)	59.97	69.83	**70.79**	67.49	68.88	70.89	69.83
		Sp(%)	54.50	71.25	**72.07**	66.47	66.67	70.89	69.73
		Acc(%)	57.24	70.54	**71.43**	66.98	67.77	70.89	69.78
		MCC	0.145	0.411	**0.429**	0.340	0.356	0.418	0.396
		AUC	0.591	0.775	**0.786**	0.734	0.748	0.779	0.763

D. melanogaster	Sn(%)	70.07	88.74	**90.60**	84.44	86.51	89.67	87.97
	Sp(%)	68.92	89.42	**89.94**	86.97	86.99	90.19	89.24
	Acc(%)	69.50	89.08	**90.27**	85.70	86.75	89.93	88.61
	MCC	0.390	0.782	**0.805**	0.714	0.735	0.799	0.772
	AUC	0.763	0.955	**0.962**	0.930	0.940	0.959	0.951
F. vesca	Sn(%)	71.31	93.23	**94.26**	90.39	91.62	93.75	93.62
	Sp(%)	72.73	92.71	**92.52**	90.59	90.97	92.33	92.39
	Acc(%)	72.02	92.97	**93.39**	90.49	91.30	93.04	93.00
	MCC	0.440	0.859	**0.868**	0.810	0.826	0.861	0.860
	AUC	0.802	0.975	**0.977**	0.964	0.969	0.976	0.976
H. sapiens	Sn(%)	75.29	84.35	**85.23**	83.38	84.46	84.94	84.62
	Sp(%)	73.70	88.17	**89.51**	85.79	87.01	89.20	89.10
	Acc(%)	74.50	86.26	**87.37**	84.58	85.73	87.07	86.86
	MCC	0.490	0.726	**0.748**	0.692	0.715	0.742	0.738
	AUC	0.825	0.935	**0.944**	0.925	0.932	0.941	0.942
R. chinensis	Sn(%)	74.00	82.00	**84.00**	77.00	83.00	84.33	81.33
	Sp(%)	69.67	76.33	**79.33**	73.67	74.00	77.67	79.67
	Acc(%)	71.83	79.17	**81.67**	75.33	78.50	81.00	80.50
	MCC	0.437	0.584	**0.634**	0.507	0.572	0.621	0.610
	AUC	0.774	0.867	**0.902**	0.844	0.859	0.880	0.877
S. cerevisiae	Sn(%)	65.45	75.86	**77.18**	73.80	75.22	75.59	76.70
	Sp(%)	68.09	82.99	**82.30**	80.67	80.19	84.31	83.04
	Acc(%)	66.77	79.42	**79.74**	77.23	77.71	79.95	79.87
	MCC	0.336	0.590	**0.596**	0.546	0.555	0.601	0.599
	AUC	0.725	0.875	**0.883**	0.848	0.855	0.878	0.871
Ts. SUP5-1	Sn(%)	62.37	70.06	**73.02**	69.35	71.78	72.90	71.72
	Sp(%)	55.15	71.60	**73.55**	65.92	68.58	71.54	68.88
	Acc(%)	58.76	70.83	**73.28**	67.63	70.18	72.22	70.30
	MCC	0.176	0.417	**0.466**	0.353	0.404	0.444	0.406
	AUC	0.617	0.777	**0.798**	0.744	0.767	0.797	0.778

	T. thermophile	$Sn(\%)$	67.78	95.53	**95.97**	92.11	92.33	95.75	94.73
		$Sp(\%)$	68.04	75.72	**75.79**	77.80	78.03	75.94	76.32
		$Acc(\%)$	67.91	85.63	**85.88**	84.95	85.18	85.85	85.53
		MCC	0.358	0.727	**0.733**	0.706	0.711	0.731	0.723
		AUC	0.747	0.923	**0.925**	0.912	0.915	0.926	0.921
	Xoc.. BLS256	$Sn(\%)$	62.21	83.76	**85.99**	78.53	80.48	85.10	82.25
		$Sp(\%)$	71.10	85.40	**86.12**	83.16	84.14	86.41	85.40
		$Acc(\%)$	66.65	84.58	**86.05**	80.84	82.31	85.75	82.82
		MCC	0.334	0.692	**0.721**	0.647	0.647	0.715	0.677
		AUC	0.730	0.916	**0.932**	0.900	0.900	0.929	0.913
4mC	*C. equisetifolia*	$Sn(\%)$	59.20	70.94	**72.75**	70.28	70.90	72.08	71.59
		$Sp(\%)$	55.83	68.93	**72.48**	65.46	66.58	69.55	68.48
		$Acc(\%)$	57.51	69.94	**72.62**	67.87	68.74	70.81	70.04
		MCC	0.150	0.399	**0.452**	0.358	0.375	0.416	0.401
		AUC	0.612	0.768	**0.796**	0.739	0.755	0.781	0.771
	F. vesca	$Sn(\%)$	68.92	82.71	**84.58**	76.97	79.01	83.57	80.71
		$Sp(\%)$	68.05	78.66	**80.78**	75.10	76.83	80.40	78.47
		$Acc(\%)$	68.48	80.68	**82.68**	76.03	77.92	81.99	79.59
		MCC	0.370	0.614	**0.654**	0.521	0.559	0.640	0.592
		AUC	0.749	0.884	**0.905**	0.839	0.854	0.898	0.875
	S. cerevisiae	$Sn(\%)$	61.11	66.16	**70.30**	67.27	66.97	69.39	69.09
		$Sp(\%)$	58.18	67.98	**72.83**	68.18	65.76	70.61	69.29
		$Acc(\%)$	59.65	67.07	**71.57**	67.73	66.36	70.00	69.19
		MCC	0.193	0.341	**0.431**	0.355	0.327	0.400	0.384
		AUC	0.631	0.735	**0.783**	0.718	0.723	0.764	0.758
	Ts. SUP5-1	$Sn(\%)$	57.72	70.56	**72.56**	68.95	70.848	71.26	70.85
		$Sp(\%)$	54.83	69.25	**71.14**	66.18	67.46	69.39	68.85
		$Acc(\%)$	56.28	69.90	**71.85**	67.56	69.15	70.32	69.85
		MCC	0.126	0.398	**0.437**	0.351	0.383	0.407	0.397
		AUC	0.592	0.768	**0.788**	0.734	0.753	0.776	0.766

Table S2. Comparison of different methods for identifying modification sites in 17 genomes. Related to Figure 3.

Modification type	Genome	Random Forest					Naïve Bayes					Bayes Net					Decision Tree				
		Sn(%)	Sp(%)	Acc(%)	MCC	AUC	Sn(%)	Sp(%)	Acc(%)	MCC	AUC	Sn(%)	Sp(%)	Acc(%)	MCC	AUC	Sn(%)	Sp(%)	Acc(%)	MCC	AUC
5hmC	*H .sapiens*	97.35	92.83	95.09	0.903	0.962	97.10	92.83	94.97	0.900	0.954	**97.44**	**92.92**	**95.18**	**0.905**	**0.966**	94.03	92.24	93.13	0.863	0.919
	M. musculus	96.25	97.88	97.07	0.941	0.984	95.38	97.88	96.63	0.933	0.985	**96.30**	**97.83**	**97.07**	**0.941**	**0.987**	96.58	96.85	96.71	0.934	0.959
6mA	*A. thaliana*	**82.02**	**84.34**	**83.18**	**0.664**	**0.906**	81.14	80.17	80.66	0.613	0.883	80.32	80.92	80.62	0.612	0.881	75.20	75.88	75.54	0.511	0.735
	C. elegans	**85.28**	**83.95**	**84.61**	**0.692**	**0.922**	82.49	79.35	80.92	0.619	0.887	78.90	81.01	79.95	0.599	0.879	74.13	74.30	74.22	0.484	0.729
	C. equisetifolia	**70.79**	**72.07**	**71.43**	**0.429**	**0.786**	70.89	71.05	70.97	0.419	0.776	69.47	71.81	70.64	0.413	0.771	62.74	64.39	63.57	0.271	0.625
	D. melanogaster	**90.60**	**89.94**	**90.27**	**0.805**	**0.962**	89.64	84.67	87.15	0.744	0.942	85.35	87.62	86.48	0.730	0.938	83.92	83.74	83.79	0.676	0.819
	F. vesca	**94.26**	**92.52**	**93.39**	**0.868**	**0.977**	91.88	93.04	92.46	0.849	0.975	92.07	92.13	92.10	0.842	0.973	88.85	88.59	88.72	0.774	0.872
	H. sapiens	**85.23**	**89.51**	**87.37**	**0.748**	**0.944**	85.91	82.60	84.26	0.685	0.917	83.70	84.57	84.13	0.683	0.915	79.37	79.74	79.56	0.776	0.707
	R. chinensis	**84.00**	**79.33**	**81.67**	**0.634**	**0.902**	85.00	76.00	80.50	0.612	0.900	81.00	79.33	80.17	0.603	0.885	75.67	69.33	72.50	0.451	0.719
	S. cerevisiae	**77.18**	**82.30**	**79.74**	**0.596**	**0.883**	79.87	79.98	79.93	0.599	0.876	75.70	80.08	77.89	0.558	0.864	72.11	71.90	72.00	0.440	0.695
	Ts. SUP5-1	73.02	73.55	73.28	0.466	0.798	**74.91**	**72.66**	**73.79**	**0.476**	**0.803**	72.54	73.66	73.11	0.462	0.800	63.20	63.43	63.31	0.266	0.613
	T. thermophile	**95.97**	**75.79**	**85.88**	**0.733**	**0.925**	93.65	75.64	84.46	0.701	0.907	93.44	75.51	84.47	0.701	0.907	84.26	81.27	82.76	0.656	0.809
	Xoc. BLS256	**85.99**	**86.12**	**86.05**	**0.721**	**0.932**	80.59	76.88	78.73	0.575	0.861	79.58	77.67	78.62	0.573	0.863	82.75	82.30	82.52	0.650	0.812
4mC	*C. equisetifolia*	**72.75**	**72.48**	**72.62**	**0.452**	**0.790**	69.21	74.76	71.98	0.440	0.789	71.60	73.69	72.65	0.453	0.786	64.36	64.87	64.61	0.292	0.636
	F. vesca	**84.58**	**80.78**	**82.68**	**0.654**	**0.905**	78.43	80.14	79.28	0.586	0.871	79.34	79.87	79.61	0.592	0.876	75.45	76.03	75.74	0.515	0.739
	S. cerevisiae	70.30	72.83	71.57	0.431	0.783	**70.51**	**73.23**	**71.87**	**0.438**	**0.791**	68.84	69.39	66.62	0.333	0.736	60.51	64.14	62.32	0.247	0.626
	Ts. SUP5-1	**72.56**	**71.14**	**71.85**	**0.437**	**0.788**	74.06	67.00	70.53	0.412	0.778	70.67	70.66	70.66	0.413	0.772	64.01	63.03	64.02	0.280	0.632

Table S3. Performance evaluation on independent dataset for identifying modification sites in 17 genomes. Related to Figure 4.

Modification type	Genome	Sn(%)	Sp(%)	Acc(%)	MCC	AUC
5hmC	*H. sapiens*	97.70	91.81	94.75	0.897	0.960
	M. musculus	96.85	96.68	96.79	0.936	0.984
6mA	*A. thaliana*	82.44	85.11	83.77	0.676	0.911
	C. elegans	86.76	84.37	85.57	0.712	0.935
	C. equisetifolia	71.81	70.46	71.13	0.423	0.779
	D. melanogaster	88.97	90.26	89.62	0.792	0.956
	F. vesca	93.94	90.59	92.26	0.846	0.977
	H. sapiens	86.31	90.52	88.42	0.769	0.950
	R. chinensis	87.96	82.94	85.45	0.710	0.924
	S. cerevisiae	75.38	81.72	78.55	0.572	0.868
	Ts. SUP5-1	74.25	72.59	73.42	0.468	0.813
	T. thermophile	95.79	75.48	85.63	0.728	0.922
	Xoc. BLS256	82.50	86.52	84.51	0.691	0.921
4mC	*C. equisetifolia*	71.69	70.49	71.09	0.422	0.780
	F. vesca	82.97	81.81	82.39	0.648	0.900
	S. cerevisiae	70.17	70.68	70.42	0.408	0.771
	Ts. SUP5-1	71.59	70.76	71.15	0.423	0.780

Table S4. The results of cross species prediction accuracies in 11 6mA contained genomes. Related to Figure 5.

Specie	*A. thaliana*	*C. elegans*	*C. equisetifolia*	*D. melanogaster*	*F. vesca*	*H. sapiens*	*R. chinensis*	*S. cerevisiae*	*TS. SUP5-1*	*T. thermophile*	*Xoc. BLS256*
A. thaliana	100	67.89	71.76	87.91	90.01	83.89	80.17	77.58	70	55.07	61.38
C. elegans	70.11	100	75.59	70.11	61.51	73.52	64.18	51.81	55.03	51.82	55.86
C. equisetifolia	78.08	70.13	100	81.27	84.53	77.94	76.17	74.54	69.29	57.00	64.20
D. melanogaster	77.37	65.08	67.80	100	85.33	79.98	77.00	77.26	69.44	50.49	64.78
F. vesca	73.76	54.85	65.97	78.73	99.97	78.56	78.83	66.51	67.78	54.11	55.42
H. sapiens	81.70	68.29	70.90	86.57	87.17	100	79.33	75.62	67.96	54.81	54.40
R. chinensis	70.80	57.01	65.91	79.32	83.43	72.66	100	70.29	66.66	53.23	56.12
S. cerevisiae	74.86	70.22	67.51	83.99	81.43	75.32	72.33	100	68.52	51.32	63.14
Ts. SUP5-1	71.92	65.86	66.57	78.85	75.98	80.5	74.76	73.27	74.94	56.99	63.46
T. thermophile	49.40	51.90	50.74	45.58	59.90	48.15	47	46.46	48.25	99.97	41.18
Xoc. BLS256	60.92	56.87	59.38	73.03	63.12	63.65	69.67	65.35	62.37	39.08	99.62

Table S5. Training and testing data from 17 genomes used in this study. Related to Figure S3.

Genome	5hmC		6mA		4mC	
	Training data	Testing data	Training data	Testing data	Training data	Testing data
A. thaliana	-	-	15937	15936	-	-
C. elegans	-	-	3981	3980	-	-
C. equisetifolia	-	-	3033	3033	3387	3387
D. melanogaster	-	-	5596	5595	-	-
F. vesca	-	-	1551	1551	7899	7898
H. sapiens	1172	1172	9168	9167	-	-
M. musculus	1840	1839	-	-	-	-
R. chinensis	-	-	300	300	-	-
S. cerevisiae	-	-	1893	1893	990	989
Ts. SUP5-1	-	-	1690	1689	7664	7663
T. thermophile	-	-	53800	53800	-	-
Xoc. BLS256	-	-	8608	8607	-	-