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

**HIGHLIGHTS**
A computational tool was developed for identification of 5hmC, 6mA, and 4mC.

6mA and 4mC mark similar regions in the *C. equisetifolia* and *F. vesca* genomes.

5hmC enriches in the initial and middle of the DNA loops.

A user-friendly webserver was available at [http://lin-group.cn/server/iDNA-MS](http://lin-group.cn/server/iDNA-MS)
iDNA-MS: An Integrated Computational Tool for Detecting DNA Modification Sites in Multiple Genomes

Hao Lv,1 Fu-Ying Dao,1 Dan Zhang,1 Zheng-Xing Guan,1 Hui Yang,1 Wei Su,1 Meng-Lu Liu,1 Hui Ding,1 Wei Chen,2,3 and Hao Lin1,4,*

SUMMARY
5hmC, 6mA, and 4mC are three common DNA modifications and are involved in various of biological processes. Accurate genome-wide identification of these sites is invaluable for better understanding their biological functions. Owing to the labor-intensive and expensive nature of experimental methods, it is urgent to develop computational methods for the genome-wide detection of these sites. Keeping this in mind, the current study was devoted to construct a computational method to identify 5hmC, 6mA, and 4mC. We initially used K-tuple nucleotide component, nucleotide chemical property and nucleotide frequency, and mono-nucleotide binary encoding scheme to formulate samples. Subsequently, random forest was utilized to identify 5hmC, 6mA, and 4mC sites. Cross-validated results showed that the proposed method could produce the excellent generalization ability in the identification of the three modification sites. Based on the proposed model, a web-server called iDNA-MS was established and is freely accessible at http://lin-group.cn/server/iDNA-MS.

INTRODUCTION
The common kinds of DNA modifications, namely, 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), N6-methyladenine (6mA), and N4-methylcytosine (4mC), have been identified in diverse species (Hu et al., 2019; Ratel et al., 2006). All these modifications are essential in a number of key biological reactions (Fu and He, 2012). As an epigenetic mechanism, DNA methylation is introduced into a DNA molecule by adding methyl or hydroxymethyl groups to nucleotides (see also Figure S1), which plays an essential role for the normal development of organism, including genomic imprinting, X chromosome inactivation, repression of transposable elements, aging, and carcinogenesis (Bergman and Cedar, 2013; Smith and Meissner, 2013).

The 5hmC is generated from 5mC by Ten-eleven translocation (Tet) family proteins. Various studies indicate that 5hmC not only acts as an intermediate during 5mC demethylation, but also plays a potential active role during maintenance of pluripotency in embryo stem cells (ESCs), neural system development, and tumorigenesis (Thomson and Meehan, 2017). Moreover, 5hmC may be involved in regulating gene expression by association with different regulatory elements and processes (Szulwach et al., 2011). The 6mA is formed by transferring methyl groups to the sixth position of the adenine ring catalyzed by methyltransferases. It is a non-canonical DNA modification that is present at low levels in different eukaryotes (Greer et al., 2015; Mondo et al., 2017). 6mA possesses similar characteristics between prokaryotes and eukaryotes (Heyn and Esteller, 2015) and has a variety of roles, such as guiding the discrimination between original and newly synthesized DNA strand after replication (Wion and Casadesus, 2006), regulating gene transcription and repressing transposable elements, and reducing the stability of base pairings for opening DNA duplexes during the cell cycle (Fang et al., 2012). Strikingly, the protection from methylation is an inherited state that, however, can be modified by environmental conditions (Wion and Casadesus, 2006). The 4mC is catalyzed by the N-4 cytosine-specific DNA methyltransferase (DNMT) that specifically methylates the amino group at the fourth position of cytosine in DNA (Timinskas et al., 1995). It is a member of the restriction modification (RM) systems and can protect the host DNA against degradation by restriction enzymes (Schweizer, 2008). In prokaryotes, 4mC is primarily used for distinguishing self from foreign DNA and correcting DNA replication errors (Iyer et al., 2011; Modrich, 1991).
Recently, several studies have revealed the genome-wide distribution of 5hmC and 6mA in Homo sapiens and Mus musculus (Hu et al., 2019) (Wu et al., 2016; Xiao et al., 2018) and the 4mC in Tolypocladium sp SUPS-1 (Stamps et al., 2015), Casuarina equisetifolia (Ye et al., 2019), and so on. Although these studies testified the presence of 5hmC, 6mA, and 4mC in various genomes based on experiments and indeed achieved encouraging results, the short-read sequencing and long-read sequencing techniques still have drawbacks. For example, bisulfite sequencing with short-read techniques is widely used to call methylated cytosines by converting unmethylated cytosines to uracil (Liu et al., 2019), but its positioning efficiency is low and genome coverage is uneven, resulting in poor sequencing quality; PacBio followed by long-read sequencing can identify DNA modifications in genome-wide scale, but it lacks high signal-to-noise ratio for DNA modifications. Moreover, in the organism kingdom, the levels of 5hmC, 6mA, and 4mC were suggested to be low and to be detectable by highly sensitive technologies. Thus, computationally prediction of 5hmC, 6mA, and 4mC is a good choice to reduce the experimental costs and provide informational guidance for the experimental study on epigenetic modification research.

In fact, several machine learning-based methods have emerged as an attractive approach for DNA methylation sites identification (Basith et al., 2019; Manavalan et al., 2019b). Recently, Ni et al. (Ni et al., 2019) extracted sequence information from Nanopore sequencing reads and proposed a deep learning-based classifier called DeepSignal to identify 6mA and 5mC sites in H. sapiens and Escherichia coli. Later, Liu et al. (2019) designed DeepMod, a bidirectional recurrent neural network with long short-term memory to identify 6mA and 5mC sites in E. coli, Chlamydomonas reinhardtii, and H. sapiens. Subsequently, Chen et al. (2019) developed a webserver named i6mA-Pred based on nucleotide chemical properties and support vector machine (SVM) to predict 6mA sites in rice genome. Similarly, Yu et al. (Yu and Dai, 2019) proposed another computational tool-based deep learning, called SNNRice6mA, for 6mA identification in rice genome. By using manually crafted DNA sequence features, Kong et al. (Kong and Zhang, 2019) built a bagging classifier to identify 6mA sites in rice genome. In addition, Basith et al. (2019) developed a novel computational predictor, called the Sequence-based DNA N6-methyladenine predictor (SDM6A), which is a two-layer ensemble approach for identifying 6mA sites in the rice genome. Based on the experimentally confirmed 4mC sites data, Chen et al. (2017) first built a novel predictor called iDNA4mC to identify 4mC sites in multiple species. Later on, based on Chen et al.’s dataset (Chen et al., 2017), Wei et al. (2019) proposed an iterative feature representation algorithm that enables one to learn informative features from several sequential models in a supervised iterative mode to identify 4mC sites. Manavalan et al. (2019a) developed Meta-4mCpred and 4mCpred-EL for the identification of 4mC sites in the mouse genome and six other species, respectively, using meta and ensemble approaches. Hasan et al. (2019) developed a tool called 4mC-ROSE for identifying 4mC sites in the genomes of Fragaria vesca and Rosa chinensis based on random forest classifier with various aspects of DNA sequence information. However, the above-mentioned predictors are trained by using species-specific data and thus may produce a low true-positive rate with a high false-positive rate when generalized to other species. Thus, it is urgent to develop novel 5hmC, 6mA, and 4mC site predictors that will be applicable for different species.

In view of the aforementioned description, this study was devoted to developing new methods for identifying 5hmC, 6mA, and 4mC sites in different species. We first collected and constructed 17 objective benchmark datasets including experimentally confirmed 5hmC/6mA/4mC sequences and non-5hmC/non-6mA/non-4mC sequences from 17 genomes. Subsequently, three kinds of sequence encoding features were proposed to formulate samples and then used as the input of the Random Forest algorithm (RF) to discriminate 5hmC/6mA/4mC from non-5hmC/non-6mA/non-4mC, respectively. Several experiments were performed to investigate the performance of the proposed methods. Finally, based on the proposed methods, a predictor called iDNA-MS was established.

RESULTS AND DISCUSSION

Sequence Analysis

In genome, some DNA consensus sequences are widespread and are conjectured to have biological functions, such as serving as transcription factor binding sites (Xiao et al., 2018). Finding the potential oligonucleotide distribution patterns of sequences around modification site will be an effective step in understanding why the site is modified and in revealing the biological functions of modifications (Smith and Meissner, 2013). To investigate the nucleotide distribution surrounding modification sites, the Two Sample Logos (Crooks et al., 2004) was plotted to observe the statistical difference of nucleotide occurrence between positive and negative samples. As shown in Figures 1 and S2, the modification sequences are significantly
different (t test, p value < 0.05) from non-modification samples in terms of nucleotide distribution. For 5hmC shown in Figure 1A, the nucleotide distribution patterns between *H. sapiens* and *M. musculus* are extremely similar. A highly conserved guanine appears at position 22 (+1 site related to 5hmC site) of the 5hmC-containing sequence; in contrast, the corresponding position in non-5hmC-containing sequence prefers to adenine, cytosine, and thymine. Moreover, we also observed the enrichment of cytosine in upstream regions of 5hmC site. These results imply that these special nucleotide distribution patterns could provide functional signal to MTases. Thus, it is reasonable to extract the positional information of the sequences to construct prediction model.

As shown in Figure 1B, we found some similar patterns of the nucleotide distributions around 6mA among seven species. First, a consensus motif, namely, [G/A]AGG, was observed in 6mA samples, which has been also reported in publications (Lv et al., 2019; Zhou et al., 2018). Second, there are high-frequency repeat A-containing segments in both upstream and downstream of 6mA sites, which is consistent with the periodic pattern of deposition in zebrafish (Liu et al., 2016). Especially, the upstream of 6mA sites have a consensus sequence of AAAAA from position 15 to 19 (from −6 to −2 upstream 6mA site). Finally, we further investigated the sequence motif of the 6mA sites in other species (See also Figure S2). We found that the nucleotide composition bias regions exist from position 22 to 23 (GG) in *F. vesca*, *R. chinensis*, and *Xoc. BLS256*.

As shown in Figure 1C (a) and (b), the consensus motif, namely, CCC[C/G][G/C], was observed over the range of position 17–23 (from −4 to +2 around 4mC site) in *C. equisetifolia* and *F. vesca*. For *S. cerevisiae* and Ts. SUP5-1, we found that two motifs, namely, CAA and AAC, were located at positions 15–17 (from −6 to −4 upstream 4mC site) and 25–27 (from +4 to +6 downstream 4mC site), respectively. These results indicated that the 4mCs have different nucleotide-enriched regions. Their nucleotide distribution patterns are species specific; however, within plants (*C. equisetifolia* and *F. vesca*) or fungi (*S. cerevisiae* and Ts. SUP5-1), some conserved motifs can still be observed.

**Performance Evaluation on Different Features**

In this section, we attempt to answer the following question: which features are the most important for identifying 5hmC/6mA/4mC in each species? For this, we first investigated the prediction performances of different kinds of features, namely, K-tuple nucleotide frequency component (KNFC), nucleotide chemical property and nucleotide frequency (NCPNF), and mono-nucleotide binary encoding (MNBE) and their
four combinations for identifying three types of modification in 17 genomes. The prediction results of the RF-based models were recorded in Figure 2 and Table S1. We noticed that the best prediction performances for the 5hmC recognition are obtained by NCPNF and NCPNF-MNBE in the *H. sapiens* and *M. musculus*, respectively. However, for all 6mA and 4mC identifications, the best prediction performances are always generated by MNBE. Although NCPNF is not better than MNBE, its area under the curves (AUCs) for all species are acceptable. The reason is that, although the core ideas of the two feature extraction methods are different, when they are used to transform the sequence into a feature matrix, some same features are produced, whereas MNBE can capture more position-specific information of one single nucleotide in positive and negative samples so as to obtain higher prediction performance. In fact, the KNFC is the worst descriptor among all features. This is because an active methyltransferase methylates nearly all (often >95%) target sequence motifs in a prokaryotic genome (Blow et al., 2016; Fang et al., 2012). In contrast, 6mA is a much less motif-driven modification in eukaryotes (Luo et al., 2016; Wu et al., 2016), likely owing to their involvement in functional regulation rather than RM systems. Moreover, we speculated that the 5hmC and 4mC modifications have a similar motif-driven rule as the 6mA modification. Taken together, the final models of 17 genomes were established based on their best features. For instance, the optimal features for identifying 5hmC in *H. sapiens* and *M. musculus* are NCPNF and NCPNF-MNBE, respectively. However, the MNBE is the best feature for identifying 6mA and 4mC sites in the other 15 genomes.

**Performance Evaluation on Different Algorithms**

To further testify the superiority of our proposed method, we investigated the discriminant capabilities of three classic algorithms, i.e., Naive Bayes, Bayes Net, and Decision Tree, on the benchmark dataset by using 5-fold cross-validation. All algorithms were implemented in WEKA (Frank et al., 2004). The results are shown in Figure 3 and Table S2. It was found that RF was the best one among the four compared algorithms in some cases for identifying modification sites. But in some species, Naive Bayes and Bayes Net can also produce the best predictive performances. Thus, the final model was built based on the best classification algorithm. For instance, all models that identify 5hmC sites were constructed based on Bayes Net, whereas models that recognize 6mA and 4mC sites were built almost exclusively on RF.

**Performance Evaluation on Independent Dataset**

In order to further evaluate the performance of our proposed method, the independent datasets were used to measure the stability and generality of the proposed model (Basith et al., 2020). The predictive
results thus obtained were listed in Figure 4 and Table S3. We observed that all models produced satisfactory results regardless of whether the modification type was 5hmC or 6mA, suggesting that our method is robust and reliable. However, the AUC obtained by the 4mC site recognition models was around 0.8, indicating that it is difficult to identify 4mC sites in fungi (S. cerevisiae) and plants (C. equisetifolia, Ts. SUP5-1). But these models are still the state-of-the-art tools for methylation sites analysis.

Cross-Species Validation

To explore whether knowledge transfer information can be applied to study the relationships of interacting species, it is necessary to demonstrate whether a model trained with the data from one species could recognize the modification sites in other species. To this end, we trained 17 species-specific models using the species-specific 5hmC/6mA/4mC data and validated these models on the 5hmC/6mA/4mC data from other species. The predictive accuracies thus obtained were shown in Figure 5 and Table S4.

As shown in Figure 5A, when H. sapiens and M. musculus data were regarded as independent testing data, we evaluated the H. sapiens- and M. musculus-based models and obtained satisfactory results. The accuracies for H. sapiens and M. musculus are 97.09% and 94.92%, respectively, suggesting that the knowledge based on a transfer learning method from the source domain (H. sapiens or M. musculus data) was well applied in the target domain (H. sapiens or M. musculus data). As shown in Figure 5B and Table S4, we can clearly find the poor prediction accuracies (<71%) for C. elegans, T. thermophile, and Xoc. BLS256 obtained by using other species-based models. This indicates a huge difference between these species in the 6mA modification pattern. Moreover, as shown in Figure 5C, the best accuracy is always obtained by model built based on the data from itself. However, the result based on the S. cerevisiae model is unsatisfactory with the accuracy of 74.24%. The reason is perhaps that the S. cerevisiae-based model was trained on less samples than other models. In summary, our proposed models display high accuracy, robustness, and generality for identifying the modification sites.
6mA and 4mC Mark Similar Regions in the *C. equisetifolia* and *F. vesca* Genomes

There is evidence that 6mA and 4mC have similar functions in prokaryotes and are primarily used for distinguishing self from foreign DNA (Iyer et al., 2011). These modifications are considered to be signaling or epigenetic modifications because they are predicted not to disrupt DNA base pairing (Iyer et al., 2011). These suggest that 6mA and 4mC may have synergistic effects through co-localization during the development of the organism. Therefore, we calculated the distant distribution between 6mA sites and 4mC sites to investigate the co-localization of the two modifications in the genome. The results showed that 6mA and 4mC had significant co-localization in *C. equisetifolia* and *F. vesca*, and slight co-localization in *S. cerevisiae*, whereas 6mA generally does not co-localize with 4mC in Ts. SUP5-1 (Figure 6). This indicates that 6mA and 4mC are two similar marks in plant genome (*C. equisetifolia* and *F. vesca*) rather than fungal genome (*S. cerevisiae* and Ts. SUP5-1) (Quandt et al., 2014).

**Figure 5. The Heatmap of Cross-Species Validation**

The heatmap showing the cross-species prediction accuracies for identifying 5hmC (A), 6mA (B), and 4mC (C). Once a species-specific model was established on its own training dataset, it was tested on the data from the other species.

**6mA and 4mC Mark Similar Regions in the *C. equisetifolia* and *F. vesca* Genomes**

Distance between 6mA and 4mC was plotted to show the correlation between the two modifications in *C. equisetifolia* (A), *F. vesca* (B), *S. cerevisiae* (C), and Ts. SUP5-1 (D).
Distribution of 5hmC on DNA Loop

Genome-wide 3D chromosome organization mapping technologies have revealed important insights on genome folding that the spatial organization of genome plays a significant role in the transcriptional control of genes (Dekker and Mirny, 2016; Dixon et al., 2012). The DNA loop, as the smallest spatial structure that can be recognized by chromosome conformation capture technology, is thought to reflect the activities of transcription factors (TFs), cohesion, and CCCTC-binding factor (CTCF) (Dowen et al., 2014; Gorkin et al., 2014). Although the role of histone marks in DNA loops has been extensively explored, we do not yet know the relationship between the DNA modifications and loops (Bonev and Cavalli, 2016; Dowen et al., 2014). To address this question, we examined the distribution of 5hmC in loops from M. musculus embryonic stem cells (ESs) (Salameh et al., 2019). The results in Figure 7 showed that 5hmC modification sites enrich in the initial and middle regions of the loop. The reason might be that ESs need a large number of promoter-associated loops to maintain the active transcriptional state of the genes. These promoters locate at the 5’ region of the loop and function independently of CTCF (Bonev and Cavalli, 2016). More experiments showed that 5hmC is enriched at both extended promoter regions of Polycomb-repressed developmental regulators and gene bodies of actively transcribed genes (Wu et al., 2011).

Web-Server

More and more researches have demonstrated that the database and the web-server can provide scholars with more convenient services. Thus, for convenience of researchers, we established a user-friendly web-server called iDNA-MS to identify 5hmC, 6mA, and 4mC modification sites in 17 genomes, which can be freely accessed at http://lin-group.cn/server/iDNA-MS.

Below, we give researchers a step-by-step guideline on how to use the web-server. Users can open the homepage as shown in Figure 8 to see a short introduction about iDNA-MS. One may first click the “Web-server” button, then type or copy/paste the DNA sequence in the input box, or upload the FASTA format file. Note that the length of each sequence should be greater than 41 nt long. Subsequently, after clicking the “submit” button, the predicted results will appear on a new page. We hope that it will provide a convenient way for users concerned on DNA modifications.
Limitations of the Study

Novel feature description methods should be used. In addition, the positional distribution of epigenetic modification sites and different DNA elements can be further studied, for example, the relationship between DNA modification and nucleosome in whole genome.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

DATA AND CODE AVAILABILITY

The datasets are publicly available at http://lin-group.cn/server/iDNA-MS. All software used are published and/or in the public domain. All pipelines and Python scripts used in the study are available at http://lin-group.cn/server/iDNA-MS/download.html.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.100991.

ACKNOWLEDGMENTS

This work has been supported by the National Natural Science Foundation of China (61772119, 31771471), Sichuan Provincial Science Fund for Distinguished Young Scholars (20JCQN0262), the Natural Science Foundation for Distinguished Young Scholar of Hebei Province (No. C2017209244), and the Science Strength Promotion Programme of UESTC.
AUTHOR CONTRIBUTIONS

Conceptualization, H.D., W.C., and H. Lin.; Investigation, W.S. and M.-L.L.; Coding, D.Z.-X.G., and H.Y.; Writing - Original Draft, H. Lv. and F.-Y.D.; Writing – Review & Editing, W.C. and H. Lin.; Funding Acquisition, W.C. and H. Lin.

DECLARATION OF INTERESTS

The authors declare that they have no competing interests.

Received: February 6, 2020
Revised: February 23, 2020
Accepted: March 11, 2020
Published: April 24, 2020

REFERENCES

Basith, S., Manavalan, B., Shin, T.H., and Lee, G. (2019). SDMA: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. Mol. Ther. Nucleic Acids 18, 131–141.

Basith, S., Manavalan, B., Shin, T.H., and Lee, G. (2020). Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. Med. Res. Rev. https://doi.org/10.1002/med.21658.

Bergman, Y., and Cedar, H. (2013). DNA methylation dynamics in health and disease. Nat. Struct. Mol. Biol. 20, 274–281.

Blow, M.J., Clark, T.A., Daum, C.G., Deutschbauer, A.M., Fomenkov, A., Fries, R., Froula, J., Kang, D.D., Malmstrom, R.R., Morgan, R.D., et al. (2016). The epigenomic landscape of prokaryotes. PLoS Genet. 12, e1005854.

Boney, B., and Cavalli, G. (2016). Organization and function of the 3D genome. Nat. Rev. Genet. 17, 661–678.

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

Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics 33, 3518–3523.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190.

Dekker, J., and Mirny, L. (2016). The 3D genome as mediator of chromosomal communication. Cell 164, 1110–1121.

Dix, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380.

Down, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuierers, J., Lee, T.I., Zhao, K., et al. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell 159, 374–387.

Fang, G., Munera, D., Friedman, D.I., Mandlik, A., Chao, M.C., Banerjee, O., Feng, Z., Lexic, B., Mahajan, M.C., Jabado, O.J., et al. (2012). Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. Nat. Biotechnol. 30, 1232–1239.

Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I.H.J.B. (2004). Data mining in bioinformatics using Weka. Bioinformatics 20, 2479–2481.

Fu, Y., and He, C. (2012). Nucleic acid modifications with epigenetic significance. Curr. Opin. Chem. Biol. 16, 516–524.

Gorkin, D.U., Leung, D., and Ren, B. (2014). The 3D genome in transcriptional regulation and pluripotency. Cell Stem Cell 14, 762–775.

Greer, E.L., Blanco, M.A., Gu, L., Sendinc, E., Liu, J., Aristizabal-Conreales, D., Hsu, C.H., Aravind, L., He, C., and Shi, Y. (2015). DNA methylation on N6-adenine in C. elegans. Cell 161, 868–878.

Hasan, M.M., Manavalan, B., Khatun, M.S., and Kurata, H. (2019). i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome. Int. J. Biol. Macromol. https://doi.org/10.1016/j.ijbiomac.2019.12.009.

Heyn, H., and Esteller, M. (2015). An adenosine code for DNA: a second life for N6-methyladenine. Cell 161, 710–713.

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.

Iyer, L.M., Abhiman, S., and Aravind, L. (2011). Natural history of eukaryotic DNA methylation systems. Prog. Mol. Biol. Transl. Sci. 101, 25–104.

Kong, L., and Zhang, L. (2019). iDNA-DCNP: computational identification of DNA N6-methyladenine sites in the rice genome using optimized nucleotide-based features. Genes (Basel) 10, https://doi.org/10.3390/genes10100828.

Liu, J., Zhu, Y., Luo, G.Z., Wang, X., Yue, Y., Wang, X., Zong, X., Chen, K., Yin, H., Fu, Y., et al. (2016). Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig. Nat. Commun. 7, 13582.

Liu, Q., Fang, L., Yu, G., Wang, D., Xiao, C.L., and Wang, K. (2019). Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. Nat. Commun. 10, 2449.

Luo, G.Z., Wang, F., Weng, X., Chen, K., Hao, Z., Yu, M., Deng, X., Liu, J., and He, C. (2016). Characterization of eukaryotic DNA N6-methyladenine by a highly sensitive expression enzyme-assisted sequencing. Nat. Commun. 7, 11301.

Lv, H., Dao, F.Y., Guan, Z.X., Zhang, D., Tan, J.X., Zhang, Y., Chen, W., and Lin, H. (2019). iDNA6mA-Rice: a computational tool for detecting N6-methyladenine sites in rice. Front. Genet. 10, 793.

Manavalan, B., Basith, S., Shin, T.H., Lee, D.Y., Wei, L., and Lee, G. (2019a). 4mCpred-EL: an ensemble learning framework for identification of DNA N6-methylcytosine sites in the mouse genome. Cells 8, https://doi.org/10.3390/cells8111332.

Manavalan, B., Basith, S., Shin, T.H., Wei, L., and Lee, G. (2019b). Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. Mol. Ther. Nucleic Acids 16, 733–744.

Modrich, P. (1991). Mechanisms and biological effects of mismatch repair. Annu. Rev. Genet. 25, 229–253.

Mondo, S.J., Dannebaum, R.O., Kuo, R.C., Louie, K.B., Bewick, A.J., LaButti, K., Haridas, S., Kuo, A., Salamov, A., Ahrendt, S.R., et al. (2017). Widespread adenine N6-methylation of active genes in fungi. Nat. Genet. 49, 964–968.

Ni, P., Huang, N., Zhang, Z., Wang, D.P., Liang, F., Mao, Y., Xiao, C.L., Luo, F., and Wang, J. (2019). DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. Bioinformatics 35, 4586–4595.

Quandt, C.A., Kepler, R.M., Gams, W., Araujo, J.P., Ban, S., Evans, H.C., Hughes, D., Humber, R., Hywel-Jones, N., Li, Z., et al. (2014). Phylogenetic-based nomenclatural proposals for Ophiocordycipitaceae (Hypocreales) with new
combinations in Tolypocladium. IMA Fungus 5, 121–134.

Ratel, D., Ravanat, J.L., Berger, F., and Wion, D. (2006). N6-methyladenine: the other methylated base of DNA. BioEssays 28, 309–315.

Salameh, T.J., Wang, X., Song, F., Zhang, B., Wright, S.M., Khunriraksakul, C., and Yue, F.J.b. (2019). A supervised learning framework for chromatin loop detection in genome-wide contact maps. bioRxiv, 739698, https://doi.org/10.1101/739698.

Schweizer, H. (2008). Bacterial genetics: past achievements, present state of the field, and future challenges. Biotechniques 44, 633–634, 636–641.

Smith, Z.D., and Meissner, A. (2013). DNA methylation: roles in mammalian development. Nat. Rev. Genet. 14, 204–220.

Stamps, B.W., Du, L., Mitchell, C.A., Cichecwicz, R.H., and Stevenson, B.S. (2015). Draft genomes of two sordariomycete fungi that produce novel secondary metabolites. Genome Announc. 3, https://doi.org/10.1128/genomeA.00291-15.

Szulwach, K.E., Li, X., Li, Y., Song, C.X., Wu, H., Dai, Q., Irier, H., Upadhyay, A.K., Gearing, M., Levey, A.I., et al. (2011). 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. Nat. Neurosci. 14, 1607–1616.

Thomson, J.P., and Meehan, R.R. (2017). The application of genome-wide 5-hydroxymethylcytosine studies in cancer research. Epigenomics 9, 77–91.

Timinskas, A., Butkus, V., and Janulaitis, A. (1995). Sequence motifs characteristic for DNA [cytosine-N4] and DNA [adenine-N6] methyltransferases. classification of all DNA methyltransferases. Gene 157, 3–11.

Wei, L., Su, R., Luan, S., Liao, Z., Manavalan, B., Zou, Q., and Shi, X. (2019). Iterative feature representations improve N4-methylcytosine site prediction. Bioinformatics 35, 4930–4937.

Wion, D., and Cassadesus, J. (2006). N6-methyl-adenine: an epigenetic signal for DNA-protein interactions. Nat. Rev. Microbiol. 4, 183–192.

Wu, H., D’Alessio, A.C., Ito, S., Wang, Z., Cui, K., Zhao, K., Sun, Y.E., and Zhang, Y. (2011). Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. Genes Dev. 25, 679–684.

Wu, T.P., Wang, T., Seetin, M.G., Lai, Y., Zhu, S., Lin, K., Liu, Y., Byrum, S.D., Mackintosh, S.G., Zhong, M., et al. (2016). DNA methylation on N(6)-adenine in mammalian embryonic stem cells. Nature 532, 329–333.

Xiao, C.L., Zhu, S., He, M., Chen, Zhang, Q., Chen, Y., Yu, G., Liu, J., Xin, S.Q., Luo, F., et al. (2018). N(6)-methyladenine DNA modification in the human genome. Mol. Cell 71, 306–318 e307.

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.

Yu, H., and Dai, Z. (2019). SNNRice6mA: a deep learning method for predicting DNA N6-methyladenine sites in rice genome. Front. Genet. 10, 1071.

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. Nat. Plants 4, 554–563.
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:

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

(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:

$$D = \left[ f_1^{k\text{-tuple}} \, f_2^{k\text{-tuple}} \, \cdots \, f_i^{k\text{-tuple}} \, \cdots \, f_4^{k\text{-tuple}} \right]^T$$

(2)

where the symbol $T$ represents the transposition of the vector, and $f_i^{k\text{-tuple}}$ 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

\[
x_i = \begin{cases} 
1 & \text{if } s_i \in \{A,G\} \\
0 & \text{if } s_i \in \{C,T\}
\end{cases}, \quad y_i = \begin{cases} 
1 & \text{if } s_i \in \{A,T\} \\
0 & \text{if } s_i \in \{C,G\}
\end{cases}, \quad z_i = \begin{cases} 
1 & \text{if } s_i \in \{A,C\} \\
0 & \text{if } s_i \in \{G,T\}
\end{cases}
\]

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

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

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

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

### Mono-nucleotide binary encoding

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

\[
n = \begin{cases} 
(1,0,0,0), \text{ when } n = A \\
(0,1,0,0), \text{ when } n = C \\
(0,0,1,0), \text{ when } n = G \\
(0,0,0,1), \text{ when } n = T
\end{cases}
\]

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 \((4 \times 41)\) 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 (Schaduongrat 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).

\[
\begin{align*}
Sn &= \frac{TP}{TP+FN} \times 100\% \quad 0 \leq Sn \leq 1 \\
Sp &= \frac{TN}{TN+FP} \times 100\% \quad 0 \leq Sp \leq 1 \\
Acc &= \frac{TP+TN}{TP+FP+FN+TN} \times 100\% \quad 0 \leq Acc \leq 1 \\
MCC &= \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}} - 1 \leq MCC \leq 1
\end{align*}
\]

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.
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. e. for. | 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           |
|                | Sn(%) | Sp(%) | Acc(%) | MCC  | AUC   | AUC   |
|----------------|-------|-------|--------|------|-------|-------|
| **D. melanogaster** |       |       |        |      |       |       |
| Sn(%)          | 70.07 | 88.74 | 90.60  | 84.44| 86.51 | 89.67 |
| Sp(%)          | 68.92 | 89.42 | 89.94  | 86.97| 86.99 | 90.19 |
| Acc(%)         | 69.50 | 89.08 | 90.27  | 85.70| 86.75 | 89.93 |
| MCC            | 0.390 | 0.782 | 0.805  | 0.714| 0.735 | 0.799 |
| AUC            | 0.763 | 0.955 | 0.962  | 0.930| 0.940 | 0.959 |
| **F. vesca**   |       |       |        |      |       |       |
| Sn(%)          | 71.31 | 93.23 | 94.26  | 90.39| 91.62 | 93.75 |
| Sp(%)          | 72.73 | 92.71 | 92.52  | 90.59| 90.97 | 92.33 |
| Acc(%)         | 72.02 | 92.97 | 93.39  | 90.49| 91.30 | 93.04 |
| MCC            | 0.440 | 0.859 | 0.868  | 0.810| 0.826 | 0.861 |
| AUC            | 0.802 | 0.975 | 0.977  | 0.964| 0.969 | 0.976 |
| **H. sapiens** |       |       |        |      |       |       |
| Sn(%)          | 75.29 | 84.35 | 85.23  | 83.38| 84.46 | 84.94 |
| Sp(%)          | 73.70 | 88.17 | 89.51  | 85.79| 87.01 | 89.20 |
| Acc(%)         | 74.50 | 86.26 | 87.37  | 85.38| 85.73 | 87.07 |
| MCC            | 0.490 | 0.726 | 0.748  | 0.692| 0.715 | 0.742 |
| AUC            | 0.825 | 0.935 | 0.944  | 0.925| 0.932 | 0.941 |
| **R. chinensis** |       |       |        |      |       |       |
| Sn(%)          | 74.00 | 82.00 | 84.00  | 77.00| 83.00 | 84.33 |
| Sp(%)          | 69.67 | 76.33 | 79.33  | 73.67| 74.00 | 77.67 |
| Acc(%)         | 71.83 | 79.17 | 81.67  | 75.33| 78.50 | 81.00 |
| MCC            | 0.437 | 0.584 | 0.634  | 0.507| 0.572 | 0.621 |
| AUC            | 0.774 | 0.867 | 0.902  | 0.844| 0.859 | 0.880 |
| **S. cerevisiae** |      |       |        |      |       |       |
| Sn(%)          | 65.45 | 75.86 | 77.18  | 73.80| 75.22 | 75.59 |
| Sp(%)          | 68.09 | 82.99 | 82.30  | 80.67| 80.19 | 84.31 |
| Acc(%)         | 66.77 | 79.42 | 79.74  | 77.23| 77.71 | 79.95 |
| MCC            | 0.336 | 0.590 | 0.596  | 0.546| 0.558 | 0.601 |
| AUC            | 0.725 | 0.875 | 0.883  | 0.848| 0.855 | 0.878 |
| **Ts. SUP5-1** |       |       |        |      |       |       |
| Sn(%)          | 62.37 | 70.06 | 73.02  | 69.35| 71.78 | 72.90 |
| Sp(%)          | 55.15 | 71.60 | 73.55  | 65.92| 68.58 | 71.54 |
| Acc(%)         | 58.76 | 70.83 | 73.28  | 67.63| 70.18 | 72.22 |
| MCC            | 0.176 | 0.417 | 0.466  | 0.353| 0.404 | 0.444 |
| AUC            | 0.617 | 0.777 | 0.798  | 0.744| 0.767 | 0.797 |
|          | Sn(%)  | Sp(%)  | Acc(%) | MCC   | AUC    | AUC    | AUC    | AUC    | AUC    |
|----------|--------|--------|--------|-------|--------|--------|--------|--------|--------|
| T. thermophile | 67.78  | 95.53  | 95.97  | 92.11 | 92.33  | 95.75  | 94.73  | 92.33  | 92.73  |
|           | 68.04  | 75.72  | 75.79  | 77.80 | 78.03  | 75.94  | 76.32  | 74.82  | 75.21  |
|           | 67.91  | 85.63  | 85.88  | 84.95 | 85.18  | 85.85  | 85.53  | 85.18  | 85.63  |
|           | 0.358  | 0.727  | 0.733  | 0.706 | 0.711  | 0.731  | 0.723  | 0.71  | 0.723  |
| Xac. BLS256 | 62.21  | 83.76  | 85.99  | 78.53 | 80.40  | 85.10  | 82.25  | 80.40  | 82.25  |
|           | 71.10  | 85.40  | 86.12  | 83.16 | 84.14  | 86.41  | 85.40  | 83.16  | 85.40  |
|           | 66.65  | 84.58  | 86.05  | 80.84 | 82.31  | 85.75  | 82.82  | 80.84  | 82.82  |
|           | 0.334  | 0.692  | 0.721  | 0.647 | 0.647  | 0.715  | 0.677  | 0.647  | 0.677  |
|           | 0.730  | 0.916  | 0.932  | 0.900 | 0.900  | 0.929  | 0.913  | 0.900  | 0.913  |
| 4mC      |        |        |        |       |        |        |        |        |        |
| C. equisetifolia | 59.20  | 70.94  | 72.75  | 70.28 | 70.90  | 72.08  | 71.59  | 70.90  | 71.59  |
|           | 55.83  | 68.93  | 72.48  | 65.46 | 66.58  | 69.55  | 68.48  | 66.58  | 69.55  |
|           | 57.51  | 69.94  | 72.62  | 67.87 | 68.74  | 70.81  | 70.04  | 67.87  | 70.04  |
|           | 0.150  | 0.399  | 0.452  | 0.358 | 0.375  | 0.416  | 0.401  | 0.358  | 0.401  |
|           | 0.612  | 0.768  | 0.796  | 0.739 | 0.755  | 0.781  | 0.771  | 0.739  | 0.771  |
| F. vesca |         |        |        |        |        |        |        |        |        |
|           | 68.92  | 82.71  | 84.58  | 76.97 | 79.01  | 83.57  | 80.71  | 79.01  | 80.71  |
|           | 68.05  | 78.66  | 80.78  | 75.10 | 76.83  | 80.40  | 78.47  | 75.10  | 78.47  |
|           | 68.48  | 80.68  | 82.68  | 76.03 | 77.92  | 81.99  | 79.59  | 77.92  | 81.99  |
|           | 0.370  | 0.614  | 0.654  | 0.521 | 0.559  | 0.640  | 0.592  | 0.521  | 0.592  |
|           | 0.749  | 0.884  | 0.905  | 0.839 | 0.854  | 0.898  | 0.875  | 0.839  | 0.875  |
| S. cerevisiae | 61.11  | 66.16  | 70.30  | 67.27 | 66.97  | 69.39  | 69.09  | 67.27  | 69.09  |
|           | 58.18  | 67.98  | 72.83  | 68.18 | 65.76  | 70.61  | 69.29  | 68.18  | 69.29  |
|           | 59.65  | 67.07  | 71.57  | 67.73 | 66.36  | 70.00  | 69.19  | 67.73  | 69.19  |
|           | 0.193  | 0.341  | 0.431  | 0.355 | 0.327  | 0.400  | 0.384  | 0.355  | 0.384  |
|           | 0.631  | 0.735  | 0.783  | 0.718 | 0.723  | 0.764  | 0.758  | 0.718  | 0.758  |
| Ta. SUP5-1 | 57.72  | 70.56  | 72.56  | 68.95 | 70.84  | 71.26  | 70.85  | 68.95  | 70.85  |
|           | 54.83  | 69.25  | 71.14  | 66.18 | 67.46  | 69.39  | 68.85  | 66.18  | 68.85  |
|           | 56.28  | 69.90  | 71.85  | 67.56 | 69.15  | 70.32  | 69.85  | 67.56  | 69.85  |
|           | 0.126  | 0.398  | 0.437  | 0.351 | 0.383  | 0.407  | 0.397  | 0.351  | 0.397  |
|           | 0.592  | 0.768  | 0.788  | 0.734 | 0.753  | 0.776  | 0.766  | 0.734  | 0.776  |
Table S2. Comparison of different methods for identifying modification sites in 17 genomes. Related to Figure 3.

| Modification type | Genome          | Sn(%) | Sp(%) | Acc(%) | MCC | AUC  | Sn(%) | Sp(%) | Acc(%) | MCC | AUC  | Sn(%) | Sp(%) | Acc(%) | MCC | AUC  |
|-------------------|-----------------|-------|-------|--------|-----|------|-------|-------|--------|-----|------|-------|-------|--------|-----|------|
|                   |                 |       |       |        |     |      |       |       |        |     |      |       |       |        |     |      |
| 5mC               | S. cerevisiae   | 97.35 | 92.83 | 95.09  | 0.903 | 0.962 | 97.10 | 92.83 | 94.97  | 0.900 | 0.954 | 97.44 | 92.52 | 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. aquasaitida | 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. chenensis   | 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. thermophi   | 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. aquasaitida | 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|
|                   | *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. equeisetifolia* | 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|
| 6mA               | *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. equeisetifolia* | 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. aequileefolia | D. melanogaster | F. vesca | H. sapiens | R. chinensis | S. cerevisiae | T5. SUPS-1 | F. thermophile | Xoc. BLS256 |
|-----------------|-------------|------------|------------------|-----------------|---------|------------|-------------|--------------|------------|---------------|-------------|
| A. thaliana     | 100         | 67.89      | 71.76            | 87.91           | 90.01   | 83.89      | 80.17       | 77.58        | 70         | 55.07         | 61.18       |
| C. elegans      | 70.11       | 100        | 75.59            | 70.11           | 61.51   | 73.52      | 64.18       | 51.81        | 55.03      | 51.82         | 55.86       |
| C. aequileefolia| 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      | 85.97            | 78.71           | 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   | 79.33      | 75.62       | 67.96        | 54.81      | 54.40         |             |
| R. chinensis    | 70.80       | 57.01      | 85.91            | 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       | 68.52        | 51.32      | 63.14         |             |
| T5. SUPS-1      | 71.92       | 65.86      | 68.57            | 78.85           | 75.98   | 80.5       | 74.76       | 73.27        | 74.94      | 56.99         | 63.46       |
| F. 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 Training data | 5hmC Testing data | 6mA Training data | 6mA Testing data | 4mC Training data | 4mC 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             | -                 | -                |