Bioinformatics, 36(11), 2020, 3327–3335
doi: 10.1093/bioinformatics/btaa143
Advance Access Publication Date: 28 February 2020
Original Paper

Sequence analysis

DNA4mC-LIP: a linear integration method to identify N4-methylcytosine site in multiple species

Qiang Tang[1,†], Juanjuan Kang[2,†], Jiaqing Yuan[1], Hua Tang[1], Xianhai Li[1], Hao Lin [2,*], Jian Huang[2,*] and Wei Chen[1,3,*]

[1]Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China, [2]Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China and [3]Center for Genomics and Computational Biology, School of Life Sciences, North China University of Science and Technology, Tangshan 063000, China

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Anthony Mathelier

Received on January 3, 2020; revised on February 12, 2020; editorial decision on February 21, 2020; accepted on February 25, 2020

Abstract

Motivation: DNA N4-methylcytosine (4mC) is a crucial epigenetic modification. However, the knowledge about its biological functions is limited. Effective and accurate identification of 4mC sites will be helpful to reveal its biological functions and mechanisms. Since experimental methods are cost and ineffective, a number of machine learning-based approaches have been proposed to detect 4mC sites. Although these methods yielded acceptable accuracy, there is still room for the improvement of the prediction performance and the stability of existing methods in practical applications.

Results: In this work, we first systematically assessed the existing methods based on an independent dataset. And then, we proposed DNA4mC-LIP, a linear integration method by combining existing predictors to identify 4mC sites in multiple species. The results obtained from independent dataset demonstrated that DNA4mC-LIP outperformed existing methods for identifying 4mC sites. To facilitate the scientific community, a web server for DNA4mC-LIP was developed. We anticipated that DNA4mC-LIP could serve as a powerful computational technique for identifying 4mC sites and facilitate the interpretation of 4mC mechanism.

Availability and implementation: http://i.uestc.edu.cn/DNA4mC-LIP/.

Contact: hlin@uestc.edu.cn or hj@uestc.edu.cn or chenweiimu@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

DNA methylation is one of the most phylogenetically widespread epigenetic modifications (Jones, 2012; Rathi *et al.*, 2018). Without altering the underlying DNA sequence, DNA methylation vastly expands the information content and structural complexity of DNA (Jaenisch and Bird, 2003; Jurkowska *et al.*, 2011; Li *et al.*, 2019). The methylation process was catalyzed by DNA methyltransferases, in which the methyl group (CH_3) was added to a certain target base (Bart *et al.*, 2005; Chen *et al.*, 2016; Li *et al.*, 2019; Poh *et al.*, 2016; Smith and Meissner, 2013). The 6-methyladenone (6mA), 5-methylcytosine (5mC) and 4-methylcytosine (4mC) are three most common and major modification observed in both prokaryotic and eukaryotic genomes (Liang *et al.*, 2018; Ratel *et al.*, 2006; Unger and Venner, 1966; Vanyushin *et al.*, 1968, 1970). DNA 6mA is ubiquitous in prokaryotic genomes and involves in the regulation of diverse pivotal biological processes. 6mA not only protects the host DNA against degradation by restriction enzymes through distinguishing the host DNA from foreign pathogenic DNA, but also involves in bacterial DNA replication and repair, cell-cycle progression and gene regulation (Casadesus and Low, 2006; Collier *et al.*, 2007; Heyn and Esteller, 2015; Loenen *et al.*, 2014; Lu, 1994; Luo *et al.*, 2015; Messer and Noyer-Weidner, 1988; Pingoud *et al.*, 2014; Pleska *et al.*, 2016; Rao *et al.*, 2014; Wion and Casadesús, 2006). In eukaryotic organisms, 5mC plays important roles in crucial biological processes, such as the regulation of gene expression, development, normal cognitive function, the inactivation of X-chromosome and so on (Bergman *et al.*, 2003; Csankovszki *et al.*, 2001; Li *et al.*, 1993; Moore *et al.*, 2013; Scarano *et al.*, 2005; Tao *et al.*, 2011; Zhang *et al.*, 2006). In contrast, 4mC modification, which is a

member of the restriction modification systems, plays a supplementary role by correcting DNA replication errors and controlling DNA replication and cell cycle (Cheng, 1995; Modrich, 1991). However, compared with 6mA and 5mC, our knowledge about the functions of 4mC modification is still far from sufficient (Schweizer, 2008). In order to further reveal function and regulatory mechanism of 4mC, it is vital to detect its distribution in the genome.

Owing to the development of high throughout sequencing technique, two main experimental methods including single-molecule and real time (SMRT) sequencing and 4mC-Tet-assisted bisulfite-sequencing (4mC-TAB-seq) have been proposed to detect 4mC sites (Flusberg *et al.*, 2010; Yu *et al.*, 2015). SMRT sequencing was designed to directly detect DNA methylation regardless of whether an assembled genome exists or not, and it has been successfully applied for detecting 4mC modifications in several species (Flusberg *et al.*, 2010). Although this method fills a gap for detecting 4mC modifications in experimental analysis, SMRT is still not a good solution to analyze thousands of genomes already exist in the public domain (Yu *et al.*, 2015). To address this issue, 4mC-TAB-seq, a next-generation sequencing method, was proposed to accurately uncover the genome-wide locations of 4mC for bacterial. Although those approaches provide important information and facilitate the detection of 4mC, they are still time-consuming and expensive, especially in performing the genome-wide detection. Therefore, developing efficient and accurate computational tools are necessary.

To this end, several machine learning based methods have been proposed to detect 4mC sites in multiple species, including *Caenorhabditis elegans* (*C.elegans*), *Drosophila melanogaster* (*D.melanogaster*), *Arabidopsis thaliana* (*A.thaliana*), *Escherichia coli* (*E.coli*), *Geoalkalibacter subterraneus* (*G.subterraneus*) and *Geobacter pickeringii* (*G.pickeringii*). Since the first computational tool, namely iDNA4mC was proposed for identifying DNA 4mC sites (Chen *et al.*, 2017), the 4mCPred (He *et al.*, 2019), 4mCPred-SVM (Wei *et al.*, 2019), 4mCPred-IFL (Wei *et al.*, 2019) and Meta-4mCPred (Manavalan *et al.*, 2019) were proposed in succession. All of these methods were trained and validated based on the same benchmark datasets proposed by Chen *et al.* (2017). As the pioneer work, the iDNA4mC extracted both nucleotide chemical properties and nucleotide frequency from the sequences as features to build the support vector machine (SVM) model. Although iDNA4mC met quite satisfactory performance, the new predictor 4mCPred achieved much higher accuracy. In 4mCPred, the DNA sequences were firstly encoded by position-specific trinucleotide propensity (PSTNP) and electron-ion interaction potential, and then the *F*-score based feature selection method was used to select optimal features used as the input of SVM (Meng *et al.*, 2018). Later on, Zou *et al.* proposed a new predictor called 4mCPred-IFL, in which a two-step feature optimization strategy was used to filtered out the noisy features obtained based on sequence information. More recently, by using a feature combination method, Lee *et al.* developed a meta-predictor, named Meta-4mCPred. Meta-4mcPred integrated two-layer machine learning algorithms with more informative sequential features to overcome the limits of generalizability.

Although the prediction performance assessment of these methods is acceptable as indicated by cross-validation test, the prediction versus accommodation and risk of over-fitting cannot be sufficiently assessed by cross-validation test. Obviously, one emerging critical issue for the state-of-the-art predictors is the lack of systematic assessment based on independent dataset. On the other hand, although the above-mentioned methods yielded promising results, there is still room for the improvement of prediction performance.

Notably, all these methods took advantages of distinct features to represent DNA sequences. It is likely that some of the features may be complementary to each other. Keeping this in mind, in the present work, we proposed a novel meta-predictor, called DNA4mC-LIP, by integrating the existing models with a preliminary optimal weight to improve the performance of identifying 4mC sites. To the best of our knowledge, DNA4mC-LIP is the first classifier that integrates combining approach in the prediction of 4mC sites. To demonstrate its performance, the DNA4mC-LIP was objectively evaluated on independent datasets. The performance comparisons on the independent datasets illustrated that DNA4mC-LIP outperforms existing methods for identifying 4mC sites. For the convenience of scientific community, a freely available web server was established at http://i.uestc.edu.cn/DNA4mC-LIP/.

2 Materials and methods

2.1 Collection of 4mc prediction methods

By querying the PubMed database with the keywords '(N4-methylcytosine) AND (prediction OR identify)', six predictors for computationally identifying 4mC sites were available (Table 1). Among them, five predictors were trained and validated based on the benchmark datasets derived from previous work (Chen *et al.*, 2017). This benchmark datasets includes the 4mC containing sequences from *C.elegans*, *D.melanogaster*, *A.thaliana*, *E.coli*, *G.subterraneus* and *G.pickeringii*. The remaining one called 4mCPred-EL was designed to identify the 4mC sites in mouse genome (Manavalan *et al.*, 2019). As the result, the five predictors except for 4mCPred-EL were retained for further analysis. It should be pointed out that the predictor 4mCPred contains two independent predictors 4mCPred_I and 4mCPred_II. The 4mCPred_I was constructed by using PSTNP features. To develop a more powerful model, the 4mCPred_II was constructed based on a combination of hybrid features including the optimal PSTNP and electron-ion interaction pseudopotential features obtained according to the *F-score* measurements (He *et al.*, 2019).

2.2 Independent datasets construction

In this study, we employed the independent datasets from the previous work to objectively evaluate the proposed method and its counterparts (Manavalan *et al.*, 2019). The independent datasets including the 4mC containing sequences from the above mentioned six species and were constructed with the same protocol as that of iDNA4mC (Chen *et al.*, 2017). All the positive samples obtained from the MethSMRT database were 41 bp long with the 4mC sites in the center (Ye *et al.*, 2017). On the other hand, the negative samples had the same length with the unmethylated cytosine in the center and shared the equal numbers with the positive samples for each species. Moreover, the sequence identity for both positive and negative samples are less than 70%. For a fair comparison, the samples that couldn't be predicted by any of the six predictors were removed. The final information of the independent datasets was reported in Table 2. For the convenience of the following analysis, the independent test data from the mouse genome (Manavalan *et al.*, 2019) were also included in this study.

2.3 Evaluation the prediction performance

The area under the receiver operating characteristic curve (AUC) is an objective metric for model evaluation and has been were widely used to fully measure the performance of prediction models (Kang *et al.*, 2019; Xu *et al.*, 2019). Considering that iDNA4mC has only one model for detecting 4mC sites across the different species, it was not compared with the state-of-the-art predictors in the work of Meta-4mCPred (Manavalan *et al.*, 2019). Actually, iDNA4mC was trained with the merged benchmark dataset containing six species, and it has the capability to identify 4mC sites for these species. Therefore, all the predictors were evaluated by using AUC on the independent datasets. Based on the prediction scores, we calculated the AUC by the 'pROC' package in R.

2.4 Model construction with iterative integration of predictors

Multiple predictor integration method has been used in the realm of bioinformatics, such as PhD7Faster, PrDSM and iRSpot-Pse6NC2.0 and so on (Cheng *et al.*, 2019; Huang *et al.*, 2019; Ru *et al.*, 2014; Yang *et al.*, 2019), and demonstrated its better superiority than single predictor. Therefore, in the present work, the six predictors (i.e. iDNA4mC, 4mCPred_I, 4mCPred_II, 4mcPred-SVM, 4mcPred-IFL

Table 1. Summary of the predictors for identifying 4mC sites

Predictor	Species	PubMed ID	Published date
iDNA4mC	*C.elegans*, *D.melanogaster*, *A.thaliana*, *E.coli*, *G.subterraneus* and *G.pickeringii*	28961687	November 15, 2017
4mCPred_I	As above	30052767	February 15, 2019
4mCPred_II	As above	30052767	February 15, 2019
4mcPred-SVM	As above	30239627	April 15, 2019
4mcPred-IFL	As above	31099381	May 17, 2019
Meta-4mCpred	As above	31146255	June 7, 2019
4mCpred-EL	*M.musculus*	31661923	October 28, 2019

and Meta-4mCpred) were used to construct an ensemble predictor through a linear integration strategy.

For a given species, the workflow of integrating predictors was shown in Fig. 1. The integration process was started from the predictor with the highest AUC, and iteratively combined with the other five predictors. For each round of iteration, we integrated one predictor that could maximize the value of AUC into the ensemble predictor. The predictors were combined based on the preliminary weighted summing of their prediction scores, where the preliminary weight of each single predictor was optimized and updated for every possible predictor combination at each round of iteration. Moreover, the preliminary weight of each single predictor was ranged from 0.05 to 1 with a step of 0.05. At last, the combination predictors with the highest AUC were chosen to construct our integrated predictors. Their weights were normalized when dividing individual preliminary weight by the sum of six predictors' preliminary weight to scale the prediction scores of combination predictors to 0–1, and the weight normalized formula as showed Eq. (1).

$$\varphi_i = \omega_i / \sum_{i=1}^{6} \omega_i, \tag{1}$$

where the value i ranges from 1 to 6 corresponding to ith predictor that added to the integrated predictors. So, the ω_i and the φ_i was the preliminary and normalized weight of the ith predictor that integrated into the ensemble predictor, respectively. Finally, the ensemble predictor for identifying 4mC sites in each species was defined by Eq. (2).

$$P = \sum_{i=1}^{6} \varphi_i * p_i, \tag{2}$$

where the p_i was the ith predictor that added to the integrated predictors, and φ_i was its normalized weight. If the prediction score P was larger than 0.5, the corresponding sequence was predicted as a 4mC site containing sequence.

3 Result and discussion

3.1 Performance assessment of predictors on independent datasets

We firstly compared the performance of the above-mentioned six predictors on the independent datasets. It is likely that the predictors may have inconsistent accuracy within specific range of probability. To check this possibility, we first sorted the prediction scores and stratified the results into the top 25% and last 25%, and further investigated the accuracy between these two groups across six species for six predictors. The results are shown in Fig. 2. Most of the predictors showed slight worse performance for predicting the last 25% group than the top 25% group in *C.elegans*, *G.pickeringii* and *G.subterraneus*. In the other three species, there was no preference in the majority of predictors between these two groups. Interestingly, we also noted that iDNA4mC had better ability to predict non-4mC sites in all species except in *C.elegans*.

As demonstrated by Lee *et al.* (Manavalan *et al.*, 2019), the Meta-4mCPred performed better than the 4mCPred in terms of accuracy for identifying 4mC sites in *A.thaliana*, *D.melanogaster*, *G.subterraneus* and *G.pickeringii*. However, from the aspect of stratified accuracy, the performance of Meta-4mCPred was better than 4mCPred only in *G.pickeringii* and *E.coli*. That is to say, the accuracy was not suitable for intuitively evaluating the overall performances.

Table 2. Summary of the updated independent datasets

Species	Original positives	Original negatives	Updated positives	Updated negatives
A.thaliana	1250	1250	1235	1226
C.elegans	750	750	747	748
D.melanogaster	1000	1000	999	1000
E.coli	134	134	132	132
G.pickeringii	300	300	198	199
G.subterraneus	350	350	350	349
M.musculus	180	180	180	180

To address this issue, we analyzed the ROC curves with their AUC of the above-mentioned six predictors on the independent datasets (Fig. 3). As shown in Fig. 3, all of the six predictors achieved AUC higher than 0.6 for identifying 4mC sites in the six species, suggesting their contributions for the prediction of 4mC sites. Besides, all the six predictors yielded the AUC higher than 0.9 for identifying 4mC sites in *C.elegans*. It was also observed that these predictors exhibit distinct performance for identifying 4mC sites in different species. For example, the performance of 4mCPred_II is the best for identifying 4mC sites in *A.thaliana*, and *C.elegans*; the Meta-4mCPred rank the first for identifying 4mC sites in *E.coli* and *G.pickeringii*; the 4mcPred_I and 4mcPred-IFL rank the first for identifying 4mC sites in *D.melanogaster* and *G.subterraneus*, respectively. Nevertheless, there is no particular predictor with the best performance for all species.

As a result, the predictors with best performance varied from species to species. That was very difficult for users to distinguish which predictor should be used to perform analysis on novel data. Therefore, a more powerful and robust method was necessary and warranted for the community to make a better choice.

3.2 Integrate predictors improve the predictive performance

Multiple classifier integration method is an important pattern classification technique and can obtain better performance than individual classifiers (Huang *et al.*, 2017; Kozlowski and Bujnicki, 2012; Schaduangrat *et al.*, 2019; Tang *et al.*, 2015). Therefore, to improve the accuracy for identifying 4mC sites, we performed the iterative combination procedure in the present work. For *A.thaliana*, the iteration started from the predictor 4mCPred_II that has the highest AUC. In the second round of iteration, 4mCPred_II was combined with the predictor Meta-4mCPred and achieved the AUC 0.9208. The AUC reached the highest value of 0.9250 in the fourth round of iteration, and the AUC no longer improved by adding the predictor iDNA4mC in fifth round or predictor 4mCPred-SVM in sixth round of iteration. After six rounds of iteration, four predictors (4mCPred_II, Meta-4mCPred, 4mCPred-IFL and 4mCPred_I) with

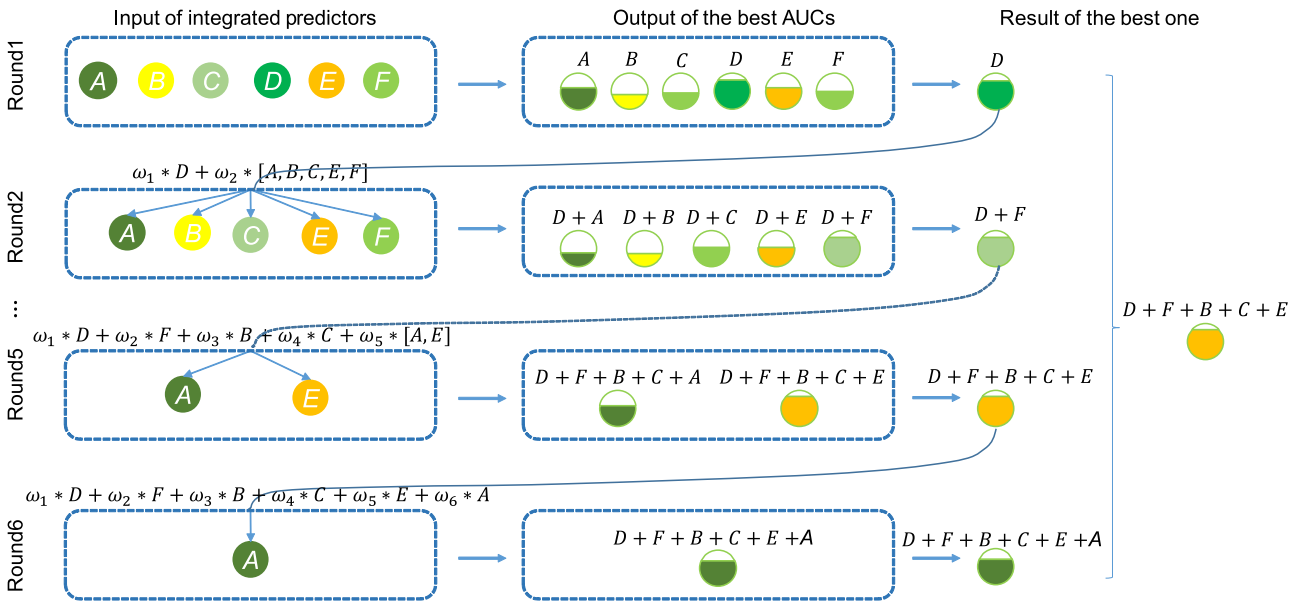

Fig. 1. The workflow of integrating predictors. A, B, C, D, E, F represent iDNA4mC, 4mCPred_I, 4mCPred_II, 4mCPred-SVM, 4mcPred-IFL and Meta-4mCPred, respectively. The ω_i was the preliminary weight of the ith predictor

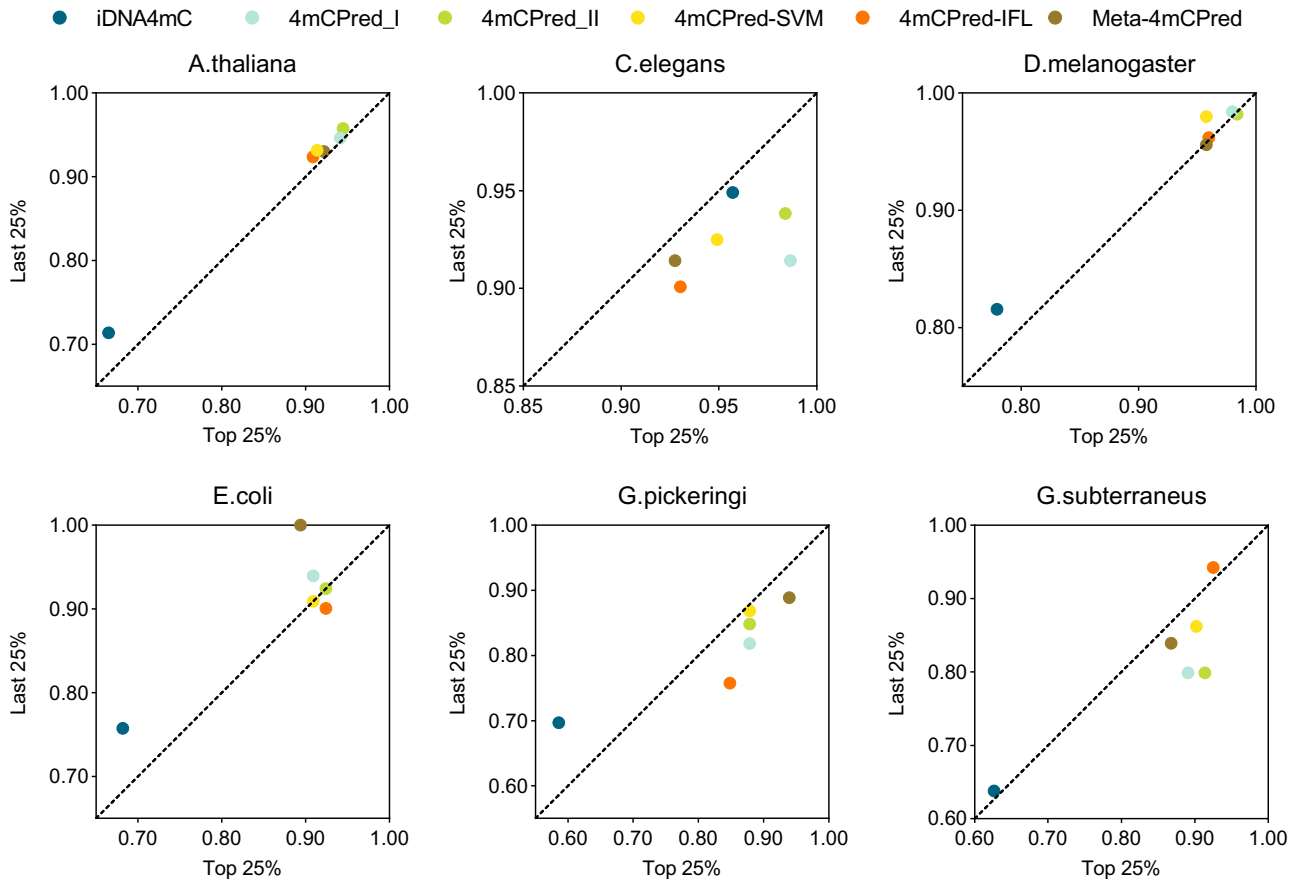

Fig. 2. The statistics of accuracy between the top 25% and last 25% of highly scored predictions for identifying 4mC sites in the independent datasets

the corresponding optimal weights (0.2903, 0.2258, 0.2742 and 0.2097) were integrated. In this case, the best AUC of 0.9250 was for predicting 4mC sites from *A.thaliana*, which is better than the reported AUC of 0.9066 by 4mCPred_II. The same iterative predictor combination process was also performed for the other five species. The normalized weights (also called contributions) of the single classifier in each species were summarized in Table 3. The normalized weight of the best predictor of each species was always the largest. For example, the normalized weight of the best predictor 4mcPred-IFL for *G.subterraneus* was 0.9524, which showed more than half of the importance for the combination predictor. These results indicate that, for each species, the combination

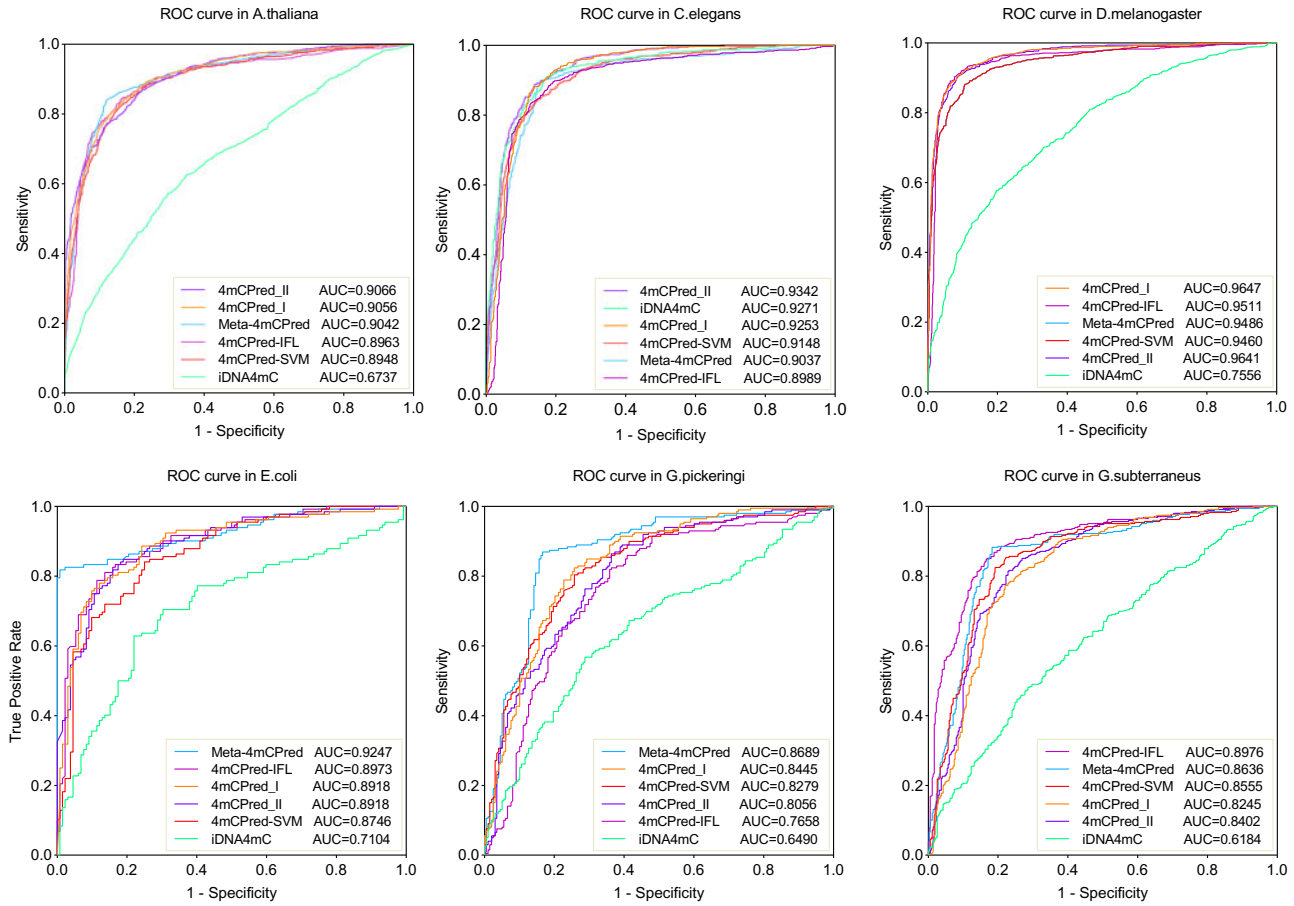

Fig. 3. Overall performance of six predictors for identifying 4mC sites in independent datasets

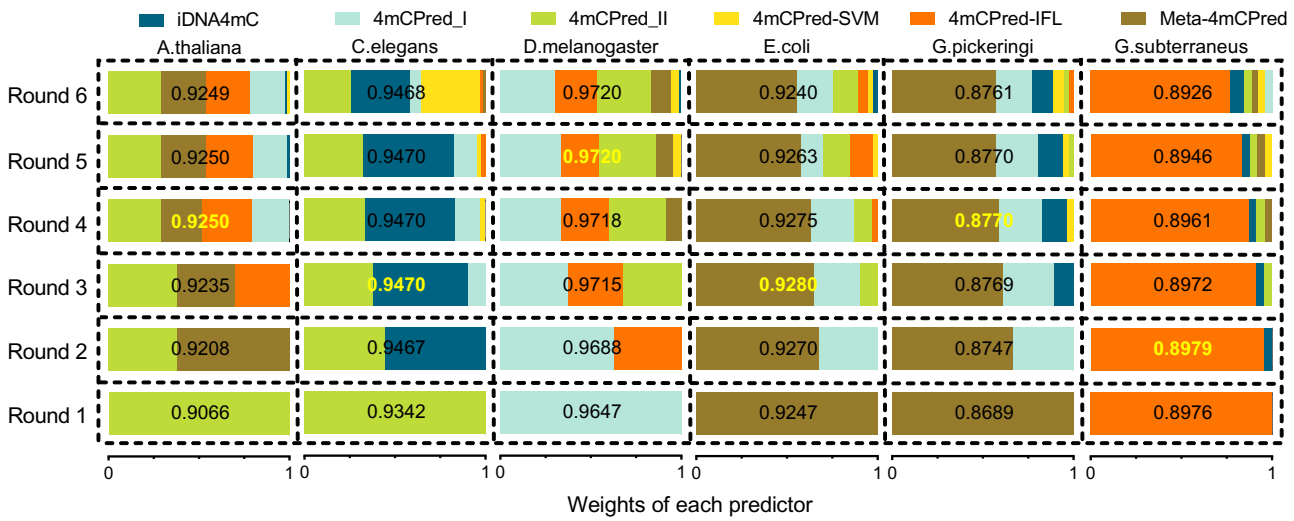

Fig. 4. AUC improvement with iterative integration of different prediction methods for six species. For each species, the predictor integrated at each round of iteration are indicated using different colors and the proportion of color is calculated based on their weights. For each round of iteration, the AUC of the combined predictor are marked on chart and the best AUC are highlighted in color of yellow. (Color version of this figure is available at *Bioinformatics* online.)

of predictors can always maximize the performance after a series of iterative processes (Fig. 4). Accordingly, for each species, the combined predictors with highest AUC was called the DNA4mC-LIP.

The predictive results of different methods for identifying 4mC sites in the independent dataset were listed in Supplementary Table S1, where the metrics used to measure the performance were defined as those mentioned in a recent work (Chen *et al.*, 2019). The AUC of DNA4mC-LIP was 25.13% and 1.84% higher for *A.thaliana*, 4.81% and 1.28% higher for *C.elegans*, 21.64% and 0.73% higher for *D.melanogaster*, 21.76% and 0.33% higher for *E.coli*, 22.8% and 0.81% higher for *G.pickeringii*, 27.95% and 0.03% higher for *G.subterraneus*, than the last-ranked and top-

Table 3. The normalized weights of the combined predictors at the highest AUC

Species	iDNA4mC	4mCPred_I	4mCPred_II	4mCPred-SVM	4mCPred-IFL	Meta-4mCPred
A.thaliana	0.0000	0.2097	0.2903	0.0000	0.2742	0.2258
C.elegans	0.5238	0.0952	0.3810	0.0000	0.0000	0.0000
D.melanogaster	0.0000	0.3333	0.3095	0.0476	0.2143	0.0952
E.coli	0.0000	0.2500	0.1000	0.0000	0.0000	0.6500
G.pickeringii	0.1379	0.2414	0.0000	0.0345	0.0000	0.5862
G.subterraneus	0.0476	0.0000	0.0000	0.0000	0.9524	0.0000

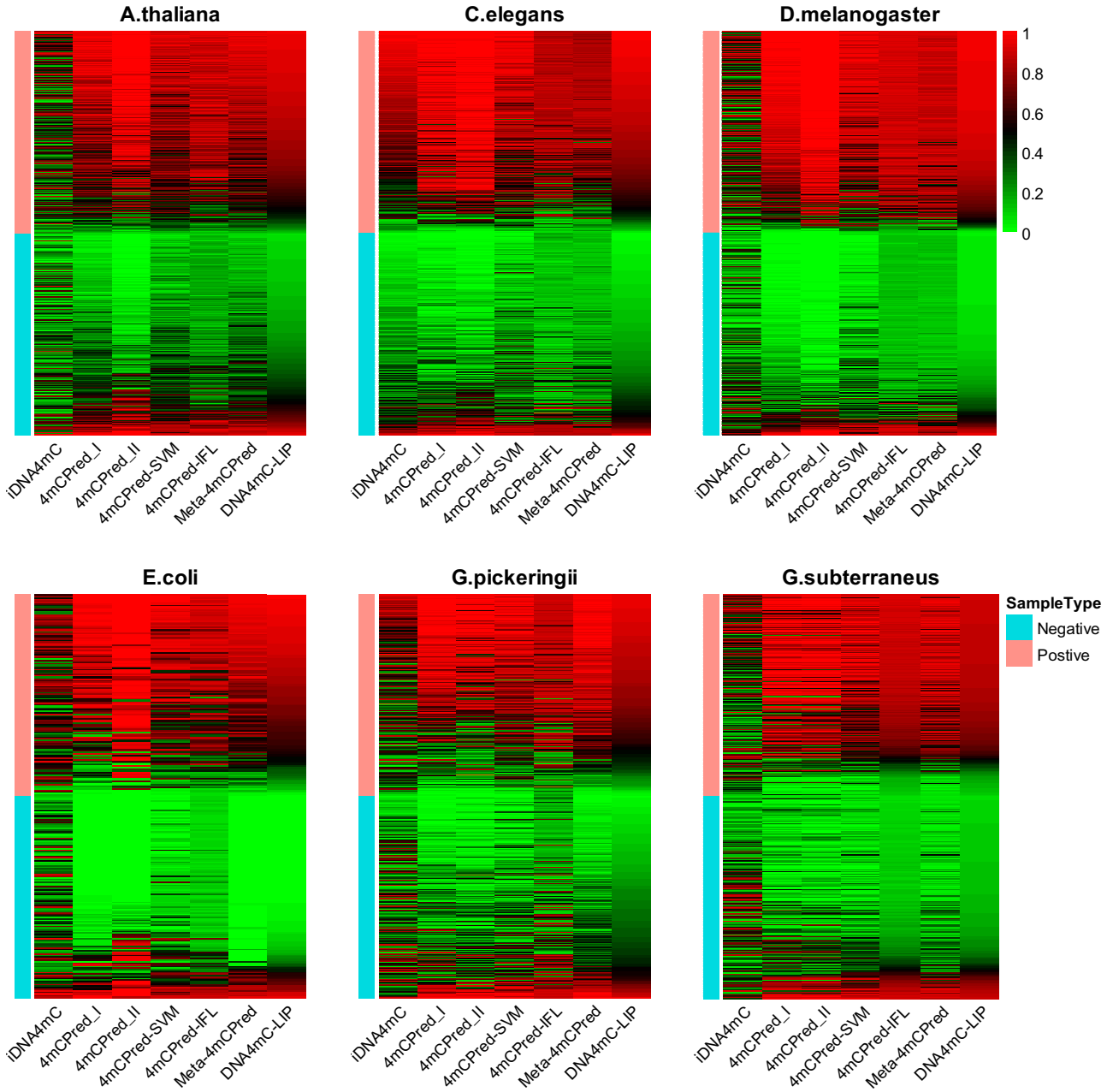

Fig. 5. The heat map showing the prediction scores of different methods based on the independent dataset. The pink and blue bar represent positive and negative samples in the independent datasets, respectively. The red color indicates a higher prediction score, and the green color indicates a lower prediction score. The higher the prediction score is, the more likely the sample is a 4mC site containing sequences. (Color version of this figure is available at *Bioinformatics* online.)

ranked predictors, respectively. We could conclude that the performance of DNA4mC-LIP is better than that of existing methods for four species (*A.thaliana*, *C.elegans*, *D.melanogaster* and *E.coli*) and is comparable with that of the top-ranked predictor Meta-4mCpred for *G.pickeringii* and *G.subterraneus*. For more clarity, the prediction scores of different methods based on the independent dataset were shown in Fig. 5. Overall, the performance of the DNA4mC-LIP outperformed the best single predictor across the

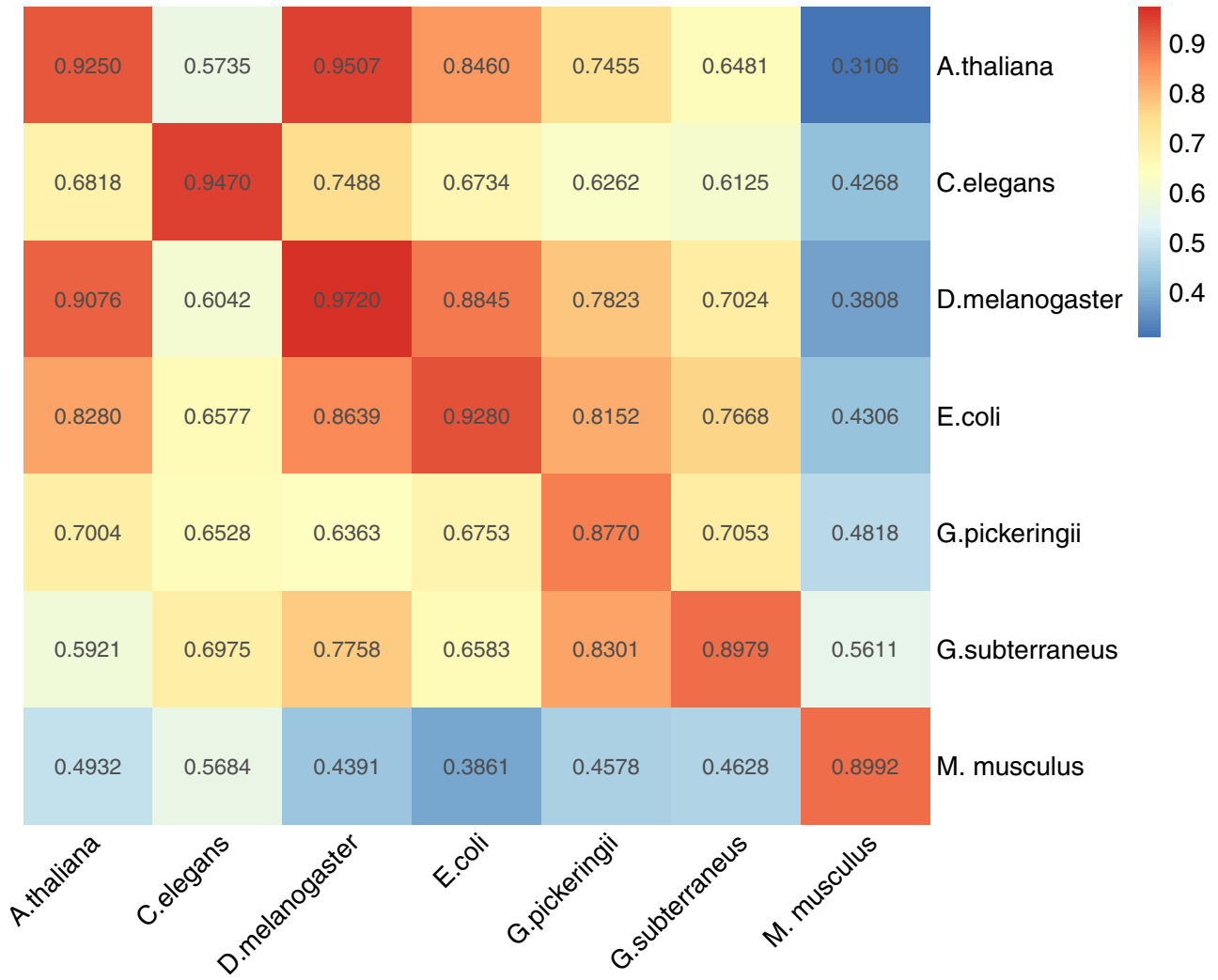

Fig. 6. The heat map showing the cross-species performance of DNA4mC-LIP and 4mCpred-EL for identifying 4mC sites. The performance is measured in terms of AUC. The top six rows are for DNA4mC-LIP and the last row is for 4mCpred-EL.

six species, indicating that the classifier integration method is indeed helpful to improve the performance of identifying 4mC sites.

3.3 Cross-species validation

Its interesting to see whether the methods could accurately recognize the 4mC sites in different species. To this end, we compared the cross-species performance of DNA4mC-LIP and 4mCpred-EL (Manavalan *et al.*, 2019) for identifying 4mC sites. The performances measured by using AUC were shown in Fig. 6, where the top six rows are for DNA4mC-LIP and the last row is for 4mCpred-EL. It was found that the species-specific DNA4mC-LIP model obtained the AUC greater than 0.7 for identifying 4mC sites in other species except for in *C.elegans*, and obtained the AUC smaller than 0.5 for identifying 4mC sites in mouse. It was also noticed that the performance of 4mCPred-EL trained based on data from mouse was unsatisfactory for identifying 4mC sites in the other species, and the obtained AUC are all smaller than 0.5. These results indicate that there might be species-specific signals surrounding 4mC sites. It is necessary to extract such features and develop new methods to improve the performance of cross-specific 4mC sites prediction.

3.4 Webserver implementation

For the convenience of researchers, an easy-to-use webserver was established to implement our predictor, which can be freely accessed via http://i.uestc.edu.cn/DNA4mC-LIP/. The step-by-step guideline on how to use the webserver is as following. Firstly, the users can input FASTA format sequences into the input box or upload a file containing FASTA format sequences by clicking the upload button. The example of FASTA format sequences can be shown by clicking on the 'Example' button. Secondly, select the desired species. To get the anticipated prediction accuracy, the selected species must be consistent with the source of query sequences. Finally, clicking the 'Submit' button to get the predicted results. Moreover, the prediction results of the other methods as references were shown in the web server.

4 Conclusion

Increasing evidence has suggested a significant role for DNA methylation in human diseases (Issa *et al.*, 1993; Kulis and Esteller, 2010; Lai *et al.*, 2019; Lyko and Brown, 2005; Nakagawa *et al.*, 2017; Su *et al.*, 2018; Wilson *et al.*, 2007), which has further propelled the emergence of DNA methylation sites prediction such as 4mC, 5mC and 6mA being an active direction in the field of bioinformatics. Compared with the other kind of DNA methylation, the knowledge of 4mC is relatively scarcity due to insufficient of experimental methods. Indeed, given that the large proportion of potential 4mC sites remains unexplored, the computational approaches are excellent complements to the experimental assays. In this study, we first systematically evaluated the existing predictors for identifying 4mC sites on independent datasets. The obtained results based on the

independent datasets demonstrated that the performance of the best predictor isn't stable and varies from species to species.

To address this issue, by integrating existing predictors, an iterative integration method, called DNA4mC-LIP, was proposed to identify 4mC sites in multiple species. The optimal weight of each predictor was found by using linear iteration strategy. DNA4mC-LIP improved the AUC to 0.9250, 0.9470, 0.9720, 0.9280, 0.8770, 0.8979 for identifying 4mC sites in *A.thaliana*, *C.elegans*, *D.melanogaster*, *E.coli*, G.*pickeringii* and *G.subterraneus*, respectively, which is better than those of existing methods for the same task. This result indicates that DNA4mC-LIP holds a high potentional to be a useful tool for identifying 4mC sites. In addition, for the convenience of scientific community, a user-friendly web server for DNA4mC-LIP was provided at http://i.uestc.edu.cn/DNA4mC-LIP/. We anticipated that DNA4mC-LIP could serve as a powerful computational technique for identifying 4mC sites and facilitate the discovery of novel 4mC sites.

Acknowledgements

The authors would like to thank the anonymous reviewers for their constructive comments.

Funding

This work was supported by the National Nature Scientific Foundation of China [Nos 31771471 and 61772119]; the Natural Science Foundation for Distinguished Young Scholar of Hebei Province [No. C2017209244]; and the Youth Teacher Innovation Foundation of Xinglin Scholar of Chengdu University of Traditional Chinese Medicine [No. ZRQN2019015].

Conflict of Interest: none declared.

References

Bart,A. *et al.* (2005) Direct detection of methylation in genomic DNA. *Nucleic Acids Res.*, **33**, e124.

Bergman,Y. *et al.* (2003) Epigenetic mechanisms that regulate antigen receptor gene expression. *Curr. Opin. Immunol.*, **15**, 176–181.

Casadesus,J., and Low,D. (2006) Epigenetic gene regulation in the bacterial world, *Microbiol. Mol. Biol. Rev.*, **70**, 830–856.

Chen,K. *et al.* (2016) Nucleic acid modifications in regulation of gene expression. *Cell Chem. Biol.*, **23**, 74–85.

Chen,W. *et al.* (2019) iRNA-m7G: identifying N(7)-methylguanosine sites by fusing multiple features. *Mol. Ther. Nucleic Acids*, **18**, 269–274.

Chen,W. *et al.* (2017) iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics*, **33**, 3518–3523.

Cheng,N. *et al.* (2019) Comparison and integration of computational methods for deleterious synonymous mutation prediction, *Brief. Bioinform.*, doi: 10.1093/bib/bbz047.

Cheng,X. (1995) DNA modification by methyltransferases. *Curr. Opin. Struct. Biol.*, **5**, 4–10.

Collier,J. *et al.* (2007) A DNA methylation ratchet governs progression through a bacterial cell cycle. *Proc. Natl. Acad. Sci. USA*, **104**, 17111–17116.

Csankovszki,G. *et al.* (2001) Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. *J. Cell Biol.*, **153**, 773–784.

Flusberg,B.A. *et al.* (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods*, **7**, 461–465.

He,W. *et al.* (2019) 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. *Bioinformatics*, **35**, 593–601.

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

Huang,M.W. *et al.* (2017) SVM and SVM ensembles in breast cancer prediction. *PLoS One*, **12**, e0161501.

Huang,Z. *et al.* (2019) Benchmark of computational methods for predicting microRNA-disease associations. *Genome Biol.*, **20**, 202.

Issa,J.P. *et al.* (1993) Increased cytosine DNA-methyltransferase activity during colon cancer progression. *J. Natl. Cancer Inst.*, **85**, 1235–1240.

Jaenisch,R., and Bird,A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.*, **33 Suppl**, 245–254.

Jones,P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, **13**, 484–492.

Jurkowska,R.Z. *et al.* (2011) Structure and function of mammalian DNA methyltransferases. *ChemBioChem*, **12**, 206–222.

Kang,J. *et al.* (2019) NeuroPP: a tool for the prediction of neuropeptide precursors based on optimal sequence composition. *Interdiscip. Sci.*, **11**, 108–114.

Kozlowski,L.P. and Bujnicki,J.M. (2012) MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinformatics*, **13**, 111.

Kulis,M., and Esteller,M. (2010) DNA methylation and cancer. *Adv. Genet.*, **70**, 27–56.

Lai,H.Y. *et al.* (2019) iProEP: a computational predictor for predicting promoter. *Mol. Ther. Nucleic Acids*, **17**, 337–346.

Li,E. *et al.* (1993) Role for DNA methylation in genomic imprinting. *Nature*, **366**, 362–365.

Li,S. *et al.* (2019) N (4)-cytosine DNA methylation is involved in the maintenance of genomic stability in *Deinococcus radiodurans*. *Front. Microbiol.*, **10**, 1905.

Liang,Z. *et al.* (2018) DNA N(6)-adenine methylation in *Arabidopsis thaliana*, *Dev. Cell*, **45**, 406–416 e403.

Loenen,W.A. *et al.* (2014) Type I restriction enzymes and their relatives. *Nucleic Acids Res.*, **42**, 20–44.

Lu,M. (1994) SeqA: a negative modulator of replication initiation in *E. coli*. *Cell*, **77**, 413–426.

Luo,G.Z. *et al.* (2015) DNA N(6)-methyladenine: a new epigenetic mark in eukaryotes? *Nat. Rev. Mol. Cell Biol.*, **16**, 705–710.

Lyko,F. and Brown,R. (2005) DNA methyltransferase inhibitors and the development of epigenetic cancer therapies. *J. Natl. Cancer Inst.*, **97**, 1498–1506.

Manavalan,B. *et al.* (2019) 4mCpred-EL: an ensemble learning framework for identification of DNA N(4)-methylcytosine sites in the mouse genome. *Cells*, **8**,

Manavalan,B. *et al.* (2019) Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol. Ther. Nucleic Acids*, **16**, 733–744.

Meng,L. *et al.* (2018) Cost-sensitive feature selection by optimizing F-measures. *IEEE Trans. Image Process.*, **27**, 1323–1335.

Messer,W. and Noyer-Weidner,M. (1988) Timing and targeting: the biological functions of Dam methylation in *E. coli*. *Cell*, **54**, 735–737.

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

Moore,L.D. *et al.* (2013) DNA methylation and its basic function. *Neuropsychopharmacology*, **38**, 23–38.

Nakagawa,T. *et al.* (2017) Frequent promoter hypermethylation associated with human papillomavirus infection in pharyngeal cancer. *Cancer Lett.*, **407**, 21–31.

Pingoud,A. *et al.* (2014) Type II restriction endonucleases—a historical perspective and more. *Nucleic Acids Res.*, **42**, 7489–7527.

Pleska,M. *et al.* (2016) Bacterial autoimmunity due to a restriction-modification system. *Curr. Biol.*, **26**, 404–409.

Poh,W.J. *et al.* (2016) DNA methyltransferase activity assays: advances and challenges. *Theranostics*, **6**, 369–391.

Rao,D.N. *et al.* (2014) Type III restriction-modification enzymes: a historical perspective. *Nucleic Acids Res.*, **42**, 45–55.

Ratel,D. *et al.* (2006) N6-methyladenine: the other methylated base of DNA. *Bioessays*, **28**, 309–315.

Rathi,P. *et al.* (2018) Selective recognition of N4-methylcytosine in DNA by engineered transcription-activator-like effectors. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **373, 1748**.

Ru,B. *et al.* (2014) PhD7Faster: predicting clones propagating faster from the Ph.D.-7 phage display peptide library. *J. Bioinform. Comput. Biol.*, **12**, 1450005.

Scarano,M.I. *et al.* (2005) DNA methylation 40 years later: its role in human health and disease. *J. Cell. Physiol.*, **204**, 21–35.

Schaduangrat,N. *et al.* (2019) Meta-iAVP: a sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation. *Int. J. Mol. Sci.* **20, 5743**.

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.

Su,J. *et al.* (2018) Homeobox oncogene activation by pan-cancer DNA hypermethylation. *Genome Biol.*, **19**, 108.

Tang,Q. *et al.* (2015) NIEluter: predicting peptides eluted from HLA class I molecules. *J. Immunol. Methods*, **422**, 22–27.

Tao,Y. *et al*. (2011) Lsh, chromatin remodeling family member, modulates genome-wide cytosine methylation patterns at nonrepeat sequences. *Proc. Natl. Acad. Sci. USA*, **108**, 5626–5631.

Unger,G., and Venner,H. (1966) [Remarks on minor bases in spermatic desoxyribonucleic acid.] *Hoppe Seylers Z. Physiol. Chem.*, **344**, 280–283.

Vanyushin,B.F. *et al*. (1968) 5-Methylcytosine and 6-methylamino-purine in bacterial DNA. *Nature*, **218**, 1066–1067.

Vanyushin,B.F. *et al*. (1970) Rare bases in animal DNA. *Nature*, **225**, 948–949.

Wei,L. *et al*. (2019) Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics*, **35**, 1326–1333.

Wei,L. *et al*. (2019) Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics, 35, 4930–4937*.

Wilson,A.S. *et al*. (2007) DNA hypomethylation and human diseases. *Biochim. Biophys. Acta*, **1775**, 138–162.

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

Xu,Z.-C. *et al*. (2019) iRNAD: a computational tool for identifying D modification sites in RNA sequence. *Bioinformatics*, **35**, 4922–4929.

Yang,H. *et al*. (2019) A comparison and assessment of computational method for identifying recombination hotspots in *Saccharomyces cerevisiae*. *Brief. Bioinform.*, doi: 10.1093/bib/bbz123.

Ye,P. *et al*. (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.

Yu,M. *et al*. (2015) Base-resolution detection of N4-methylcytosine in genomic DNA using 4mC-Tet-assisted-bisulfite-sequencing. *Nucleic Acids Res.*, **43**, e148.

Zhang,X. *et al*. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell*, **126**, 1189–1201.