Published online 31 August 2023

Nucleic Acids Research, 2023, Vol. 51, No. 18 ***e95***
https://doi.org/10.1093/nar/gkad693

Widespread effects of DNA methylation and intra-motif dependencies revealed by novel transcription factor binding models

Jan Grau [1,*], Florian Schmidt [2,3,4,5] and Marcel H. Schulz [2,3,6,7]

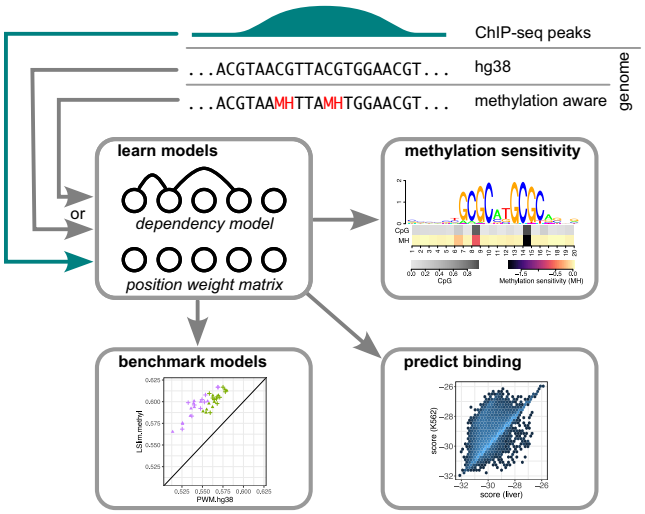
[1]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle 06120, Germany, [2]Goethe-University Frankfurt, Institute for Cardiovascular Regeneration, Theodor-Stern-Kai 7, 60590 Frankfurt, Germany, [3]Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken 66123, Germany, [4]Systems Biology and Data Analytics, Genome Institute of Singapore, Singapore 13862, Singapore, [5]ImmunoScape Pte Ltd, Singapore 228208, Singapore, [6]German Center for Cardiovascular Research, Partner site Rhein-Main, 60590 Frankfurt am Main, Germany and [7]Cardio-Pulmonary Institute, Goethe University, Frankfurt am Main, Germany

Received March 01, 2022; Revised July 20, 2023; Editorial Decision July 27, 2023; Accepted August 10, 2023

ABSTRACT

Several studies suggested that transcription factor (TF) binding to DNA may be impaired or enhanced by DNA methylation. We present MeDeMo, a toolbox for TF motif analysis that combines information about DNA methylation with models capturing intra-motif dependencies. In a large-scale study using ChIP-seq data for 335 TFs, we identify novel TFs that show a binding behaviour associated with DNA methylation. Overall, we find that the presence of CpG methylation decreases the likelihood of binding for the majority of methylation-associated TFs. For a considerable subset of TFs, we show that intra-motif dependencies are pivotal for accurately modelling the impact of DNA methylation on TF binding. We illustrate that the novel methylation-aware TF binding models allow to predict differential ChIP-seq peaks and improve the genome-wide analysis of TF binding. Our work indicates that simplistic models that neglect the effect of DNA methylation on DNA binding may lead to systematic underperformance for methylation-associated TFs.

GRAPHICAL ABSTRACT



INTRODUCTION

Transcription Factors (TFs) are essential regulatory proteins with diverse roles in transcriptional regulation, such as chromatin remodelling or the initiation of transcription (1). Hence, a key step to improve our understanding of the function of TFs is to identify the genomic location of TF binding sites (TFBS). It was shown that TFs usually bind to accessible chromatin (2) and therefore a variety of computational methods (3) has been developed to combine chromatin accessibility data (e.g. DNase1-seq, ATAC-seq, NOMe-seq) with TF motif information as encoded in Position Weight Matrices (PWMs) (4–7) to elucidate the tissue-specific binding profiles of TFs. Recently, LSlim-models, which capture intra-motif dependencies, have been successfully applied to overcome the nucleotide independence as-

*To whom correspondence should be addressed. Tel: +49 345 5524768; Fax: +49 345 5527039; Email: grau@informatik.uni-halle.de

sumption of PWMs (8). Further approaches that allow for intra-motif dependencies include improved energy models (9), transcription factor flexible models (10), parsimonious Markov models (11) and Bayesian Markov models (12).

To provide the community with a systematic comparison of the plethora of TFBS prediction approaches, the *ENCODE-DREAM in vivo Transcription Factor binding site prediction challenge* (https://www.synapse.org/#!Synapse:syn6131484/wiki/402034) was conducted in 2016. The competing methods considered, aside from epigenomics data, also DNA shape, sequence conservation, and/or sequence composition. Interestingly, the median area under the precision recall curve (AUC-PR) for one of the winning methods across all classifiers is only 0.4 (5), suggesting that important molecular signatures influencing TF binding are not incorporated yet.

One of those signatures is DNA methylation in a CpG context. The analysis of DNA methylation has been a major focus of epigenomics research and several experimental approaches have been proposed to characterize DNA methylation *in vivo* (13): while early methods used methylation sensitive restriction enzymes in PCR and gel-based approaches (14), the usage of microarrays allowed a scale-up of CpG methylation analysis (15). Array-based methods are nowadays used to characterize the methylation levels of pre-selected CpGs, e.g. for diagnostic purposes (16). With the advancements of next-generation sequencing, several sequencing based approaches to characterize DNA methylation on a genome-wide scale have been proposed (17,18). Most techniques used currently require bisulfite-treated DNA as input. Bisulfite treatment causes unmethylated cytosines to be converted to uracils, whereas methylated cytosines remain unchanged (19).

Large-scale bisulfite sequencing studies have been performed by several international consortia such as Blueprint, Roadmap and ENCODE, to generate DNA methylation data for several tissue and primary cell types.

DNA methylation in a CpG context has been reported previously to have a repressive effect on TF binding (20). Additional studies using protein binding microarrays (21), DAP-seq (22) or methylation-sensitive systematic evolution of ligands by exponential enrichment (SELEX) (23) indicated that DNA methylation can also promote TF binding.

Functionally, the addition of a methyl group to cytosines influences their steric and hydrophobic environment and renders it similar to that of a thymine (24). This is known as *thymine mimicry* (25). Specifically, CpG methylation leads to a widening of the major groove and narrows the minor groove (26,27). It also affects roll and propeller twist and results in an increase of helix stiffness (27).

As summarized in (24), there are two modes how TFs can recognize DNA methylation: i) the 5 methyl-cytosine-arginine-guanine triad detection and ii) the presence of van der Waals interactions between the methyl group of the cytosine and methyl groups of hydrophobic amino acids or methylene groups of polarized amino acids.

Methylation dependence has been studied in depth for several TFs such as KLF4 (28), P53 (29), CEBP complexes (23), NRF1 (30) and ZFP57 (31).

The MeDReaders database catalogues TF binding motifs that were learned on TF ChIP-seq peaks separated by low or high average methylation level in the peak region using MEME (32). While this constitutes a straight-forward approach, methods specifically designed to include information about DNA methylation directly into the *de novo* discovery of binding motifs are rare. The MEPIGRAM (33) software is an extension of the EPIGRAM algorithm for motif detection (34). MEPIGRAM derives motifs by constructing PWMs considering a sequence set derived from TF ChIP-seq data. Specifically, MEPIGRAM computes the most enriched k-mers within the ChIP-seq peak regions compared to a randomly shuffled set of sequences. These k-mers are treated as 'seeds' and subsequently extended both up and downstream. To incorporate DNA methylation in this process, the alphabet considered in PWM construction has been extended with a separate symbol for methylated cytosines. Viner *et al.* (35) use an alphabet with additional symbols for differently methylated cytosines and further symbols for the corresponding guanines on the opposite strand. *De novo* motif discovery is then performed by an enhanced version of the MEME suite. To analyse data generated by the *Methyl-Spec-seq* assay, Zuo *et al.* (31) use a similar extended 6-letter alphabet for PWM construction with separate symbols for methylated cytosines and guanines opposite of methylated cytosines.

Recently, the METHMOTIF database, which combines TF motifs with associated DNA methylation profiles, has been made available (36). In METHMOTIF, occurrences of known TF motifs are detected with CENTRIMO in ChIP-seq data from ENCODE. Subsequently, the genomic loci that are enriched for the tested motifs are overlayed with CpG methylation data from GEO. The found motifs and the CpG methylation signatures are visualized in so called *MethMotif* logos. A possible demerit of the approach pursued in METHMOTIF, compared with those mentioned previously, is that the methylation dependence has not been incorporated into the discovery of the TF motif. In addition, neither METHMOTIF nor MEPIGRAM provide the user with means to perform methylation-aware genome wide TFBS predictions.

Although the aforementioned methods demonstrated significant advantages in the characterization of TF binding sites by including DNA methylation, they do suffer from the simplifying *independence of nucleotide assumption* made in PWM models. Even without considering DNA methylation, several recent studies demonstrated that including intra-motif dependencies improves the accuracy of motif models. The models employed for this purpose include variable-order Bayesian networks (37), Bayesian Markov models (12), transcription factor flexible models (10), parsimonious Markov models (11,38) and sparse local inhomogeneous mixture (Slim) models (8). Considering DNA methylation, the independence assumption is obviously violated in a CpG methylation context.

Here, we present MEDEMO (Methylation and Dependencies in Motifs), a toolbox using an extension of SLIM models capturing intra-motif dependencies, which accounts for the presence of DNA methylation. The DIMONT framework for de novo motif discovery employed by MEDEMO learns PWM models or more complex motif models from input sequences, for instance, sequences under ChIP-seq peaks. The PWM models learned by DIMONT have been

benchmarked on ChIP-seq data against those generated by the alternative approaches POSMO (39), CHIPMUNK (40), MEME (41), DME (42), DREME (43) and HMS (44) previously (45), and DIMONT showed to yield the largest number of correct motifs. Likewise, SLIM/LSLIM models have been shown to perform better than other dependency models, also when learned from ChIP-seq data within the DIMONT framework (8). Here, we focus on the influence of using dependency models, considering DNA methylation in TF binding sites, and the combination of both. Since all modelling variants (PWM models and LSlim models with and without methylation information, respectively) are learned in the common DIMONT framework, we eliminate additional influence of algorithmic differences between motif discovery approaches in this analysis. However, since MEPIGRAM (34) has been developed for the same purpose as MEDEMO, namely the discovery of methylation-aware motif variants based on ChIP-seq data, we perform benchmark analyses comparing MEDEMO and MEPIGRAM.

We illustrate that the combination of methylation information and intra-motif dependencies considered by MEDEMO typically yields an improved prediction performance compared with a standard PWM-based approach. To this end, we analysed the DNA methylation dependence of hundreds of TFs in cell-lines and primary cells using DEEP and ENCODE data. MEDEMO is available as a stand-alone tool allowing both the inference of methylation-aware TF motifs and to obtain genome-wide TFBS predictions.

MATERIALS AND METHODS

Data

We downloaded whole genome bisulfite sequencing data for three cell-lines (K562 (ENCFF867JRG, ENCFF721JMB), HepG2 (ENCFF064GJQ, ENCFF369YQW), GM12878 (ENCFF79HCL, E2NCFF835NTC)) from ENCODE as well as for two replicates of primary human hepatocytes (DEEP (41_Hf01_LiHe_Ct, 41_Hf03_LiHe_Ct (available via EGA, https://ega-archive.org, EGAD00001002527)). The ENCODE data has been processed following the uniform ENCODE-Processing pipeline, the DEEP data has been processed following the DEEP *MCSv3* pipeline (https://github.molgen.mpg.de/DEEP/comp-metadata, doi:10.17617/1.2W). Furthermore, we downloaded TF-ChIP seq peak calls (IDR thresholded peaks) from ENCODE for 336 experiments in K562, 145 in HepG2, 129 in GM12878 and 25 in primary human hepatocytes (liver). Data accession IDs for TF-ChIP-seq data are provided in Supplementary Table S1.

Generation of methylation-aware genomes

To generate a methylation-aware genome sequence, where a methylated C is replaced by 'M' and a G opposite of a methylated C is replaced by 'H', we discretized the methylation calls from whole genome bisulfite data using BETAMIX (46) and the parameter *--components unimodal unimodal*, which refers to a mixture model of two unimodal distributions.

Training procedure

Motif models, i.e. PWMs and LSlim models, are learned from ChIP-seq data by the discriminative maximum supervised posterior principle within the DIMONT/SLIMDIMONT framework (8,45). To this end, we use as positive training sets genomic regions under all ChIP-seq positive peaks (optimal IDR thresholded peaks) as downloaded from the ENCODE project and extract the sequence of length 1000 bp around the peak center. Here, we do not explicitly check for overlaps between peaks, and, hence, positive sequences in the training set may be partially overlapping. In addition, we use two different sets of negative training sets. First, we randomly draw 10 000 regions uniformly from the complete genome excluding any ChIP-seq positive region of the TFs studied (random) and again extract the sequence of length 1000 bp around the center of each region. Second, we consider dinucleotide shuffled versions of each positive sequence in the training set (shuffled). To balance the influence of positive and negative sequences, we assign each negative training sequence a weight that is considered when evaluating the objective function. Specifically, if the training data contain N positive sequences and M negative sequences, each negative training sequence is assigned a weight of N/M, such that the total weight of all negative sequences is $M \cdot N/M = N$. In either case, we extract sequences from the original *hg38* genome with standard DNA nucleotides and, alternatively, sequences from the genomes including methylation calls (Section Generation of methylation-aware genomes). As the methylated genomes are cell type-specific we always use those matching the cell type of the corresponding ChIP-seq experiment. Sequences from the negative sets are also extracted from the matching genome versions. Models that are discovered *de novo* from these data sets are (i) standard position weight matrices and (ii) LSLIM models (8) with a maximum distance of 5bp between putatively dependent positions. In general, motif discovery within the DIMONT/SLIMDIMONT framework (8,45) may report multiple motifs per input data set. For the remainder of the analyses described here, we only consider the first reported motif according to the ranking by the value of the maximum supervised posterior objective function used internally in the DIMONT/SLIMDIMONT framework as proposed previously (45).

Prediction procedure

Given a trained motif model and an input set of sequences, we compute for each sequence the log-likelihood of all overlapping sub-sequences on both strands matching the motif length. We then chose as predicted value for that sequence the maximum over all these log-likelihood values. In contrast to alternative scores, like the *sum occupancy score* (47) integrating over all log-likelihood values, this procedure makes sure that the score of a sequence can be attributed to one specific sub-sequence with its methylation pattern.

Cross validation procedure

For benchmarking the different models learned from sequence with and without methylation information, we

follow a 10-fold cross validation procedure. Specifically, we partition ChIP-seq positive regions and (for the first training variant) drawn negative regions into 10 equally sized sets, where in each cross validation fold, the union of 9 of these sets is used for training and the remaining set is used for testing. The partitioning into training and test sets for the individual folds is already performed on the level of peak files. Afterwards, the corresponding partitions of peaks are considered when extracting sequences under the peaks from the methylation-aware or the original hg38 genome variant. Hence, training and test sets in the different cross validation folds are identical (aside from methylation information) between the different genome variants.

Evaluating performance

For evaluating performance of a model trained on and applied to sequences from a specific genome version, we consider a classification problem discriminating ChIP-seq positive from negative sequences. The positive set comprises all sequences extracted under ChIP-seq positive regions from the corresponding test partition. The negative set, in turn, comprises sequences from genomic regions that are again randomly drawn uniformly from the complete genome, in this case excluding all ChIP-seq positive regions for all TFs studied and also excluding the negative regions used for training. In total, this negative set contains 100 000 regions, which are again partitioned into 10 test sets to capture variability among different choices of negatives. Given a model, scores for all sequences in positive and negative sets are computed as described in section Prediction procedure. The ability of these scores to distinguish positives from negatives is then evaluated by the area under the precision recall curve (AUC-PR) as determined by the PRROCR package (48). Models trained on the training partition of one ENCODE data set for one specific TF are evaluated (i) on the test partition of the same data set, (ii) on the corresponding test partition of other data sets for the same TF and cell type and (iii) on the corresponding test partition of other data sets for the same TF in other cell types. We refer to the first two cases as *within cell type*, and to the latter case as *across cell type*. As one baseline, we consider a random classifier, i.e., a classifier that randomly assigns positive and negative labels with equal probability. The random classifier generates a true positive with probability $\frac{N}{N+M}$ and a false positive with probability $\frac{M}{N+M}$, where N and M are the number of positive and negative sequences, respectively. Hence, the AUC-PR of the random classifier can be derived analytically as $\frac{N}{N+M}$ (49).

Training and evaluating mEpigram models

In order to compare the motif discovery of MEPIGRAM and MEDEMO on common ground, we use the same methylation-aware genomes and derived sequences for both approaches. Technically, our methylation-aware M/H alphabet needs to be converted to an E/F alphabet to serve as input of the MEPIGRAM routines. For training MEPIGRAM models, we follow the procedure proposed by the authors. First, we extract training sequences based on the ChIP-seq peak files and the corresponding methylation-aware genome variants using the `bedToFasta.py` script provided with MEPIGRAM. We further generate the *k*-mer background model using the `bgModel.py` script with parameter `-k 7` from the methylation-aware genome variants. We chose 7-mers instead of 8-mers, because for the 'typeEF' variant of MEPIGRAM not limited to fully methylated CpGs, only the (required) 7-mer graph is available. We then learn motifs using the `mepigram_wrapper.py` script using the sequences extracted from the peak file, the 7-mer graph, the background file for the training genome, and parameter `typeEF`. Aside from a motif file in MEME format, this script outputs a file `enrichments.tsv`, which lists the per-motif enrichment on the positive training examples compared with shuffled negatives. For evaluating performance, we chose the motif with the largest enrichment value, since ranking by enrichment value has been suggested in the original publication (33). We then apply the selected motif to the test sequences using the provided `quickPssmScanBestMatchLiteTypeEF.jl` script reporting the best motif score for each input sequence, which matches the prediction schema applied for the MEDEMO models.

Model visualization

Since parameters of models learned by discriminative learning principles may be skewed to optimize prediction accuracy, a direct visualization of these parameters may lead to un-intuitive results. Hence, we follow the approach of (8) and visualize models based on their predicted binding sites on the training data represented by traditional sequence logos and dependency logos generated by the DEPLOGOR package (50).

Methylation sensitivity of binding models

We investigate the methylation sensitivity of a trained model again based on predicted binding sites. To this end, we consider models learned from sequences using the extended, methylation-aware alphabet and binding sites predicted from the corresponding training data set. Each of these binding sites is first converted to the standard DNA alphabet replacing occurrences of M with C and of H with G. We use this modified sequence to compute a *base score* without methylation. We then consider each CpG dinucleotide within the sequence (regardless if it was methylated in the original sequence) and change both the C to M and the G to H (MH). We compute the score of the modified sequence according to the model and determine its difference relative to the base score. If this score is larger than the base score, we consider the influence of methylation on such a nucleotide (in this sequence context) as *beneficial*, and as *detrimental* otherwise. For each binding site position, we also compute the relative abundance of CpGs in the predicted binding sites and the average of the score differences of the MH case relative to the base score. We further assess if the binding motif of a TF contains a prominent CpG in its core. Here, we define the core of the motif as all positions between the first and the last position with an information content above 0.5 and we consider a CpG is prominent if

the relative frequency of CpG di-nucleotides at a motif position is above 0.25, which is four times more frequent than could be expected by chance for a uniform distribution over the nucleotides. We also record the maximum methylation sensitivity within the motif core and outside the motif core.

Differential binding

For analyzing the association between scores of predicted binding sites and differential binding, we consider pairs of cell types, A and B, with ChIP-seq data available for the same TF. In this analysis, we distinguish common peaks that overlap between the two cell types, and unique peaks present in only one of the cell types. We further predict one binding site per ChIP-seq peak at the position yielding the maximum score as described in Section Prediction procedure.

For the common peaks, we only compare (scores of) binding sites that are predicted at exactly the same genomic location in the methylation-aware genomes of both cell types, as this allows for a direct comparison of prediction scores. This requirement is reasonable as, in principle, the position of a predicted binding site could change due to differences in the methylation states of the two cell types. Since only such common binding sites are considered, we identify common peaks by the presence of predictions at identical genomic locations within the two methylation-aware genomes. Predicted binding sites in both cell types are recorded together with the corresponding prediction scores and the peak heights (column 7 of the narrowPeak format) of the surrounding peaks.

We further identify unique peaks for cell type A using the bedtools (51) command 'bedtools intersect -v -a peaksA.bed -b peaksB.bed > onlyA.bed'. For binding sites predicted from the methylation-aware genomes of cell type A, we extract the corresponding sequence from the methylation-aware genome of cell type B, and record prediction scores for these two predicted sites. We proceed in complete analogy to identify unique peaks for cell type B.

For these analyses, we aim at using the same models that have also been considered for classification-based benchmarks. However, as these benchmarks are based on 10-fold cross validation experiments, we also obtain a set of 10 models per TF and training data set. For this reason, we perform the above-mentioned procedure for each of the 10 models, and average prediction scores per ChIP-seq peak before proceeding with statistical analysis and visualization.

As a reference, we also consider a simple baseline, which measures methylation levels of the sequences under ChIP-seq peaks. Specifically, we extract sequences of length 1000 bp and determine, on either strand of the DNA sequence, the fraction of cytosines that are methylated according to the methylation-aware genome of a cell type. We center the extracted sequences at the position of the predicted target site instead of the (cell type-specific) peak center or peak summit to ensure that methylation levels in different cell types are measured for the same genomic region.

Genome-wide predictions within the Catchitt framework

Catchitt is a framework for predicting *in vivo* TF binding sites based on motif models and cell type-specific chromatin accessibility data, and is a streamlined version of the approach winning the 'ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge' (5). Here, we use Catchitt to compare the performance of methylation-agnostic and methylation-aware motif models when used for genome-wide predictions of TF binding sites. Specifically, we obtain ATAC-seq data from ENCODE (Supplementary Table S2) as BigWig files containing the fold change over control for the four cell types considered. These serve as input of the 'access' tool of Catchitt to obtain chromatin features. In addition, we train PWM and LSlim models using TF-specific ChIP-seq data from all chromosomes except the test chromosomes chr1, chr8 and chr21 as described in Section Training procedure. Trained models are then used to compute motif-based features using the 'motif' tools of Catchitt based on the respective original or methylation-aware genome variants. Labels (bound, unbound) of 200 bp genomic regions shifted by 50 bp along each chromosome are generated from the corresponding ChIP-seq data using the 'labels' tool of Catchitt. Catchitt models using the chromatin features and motif-based features of individual motif models are then trained on the training chromosomes using the 'itrain' tool of Catchitt with default parameters and using chromosomes chr10, chr11, chr12, chr13 and chr14 for the iterative training procedure (cf. (5)). Finally, predictions for the test chromosomes chr1, chr8 and chr21 are generated using the 'predict' tool of Catchitt for the training cell type (within cell type) and for the remaining cell types (across cell type) with ChIP-seq data available for the TF of interest. AUC-PR of bound vs. unbound regions is computed using the PRROC R package (48).

Method implementation

We implement the model, training procedure and prediction procedure based on the existing implementation of the SLIMDIMONT approach (8). The basic modification compared with the version published previously is the extension of the alphabet to A, C, G, T, M and H, where M is complementary to H. This extension allows us to include information about methylation while preserving the possibility to compute reverse complements of input sequences, which is necessary because in ChIP-seq data binding sites may be located on either DNA strand. We provide this methylation-aware toolbox termed MEDEMO for motif discovery as i) stand-alone binary versions with graphical user interface and command line interface (cf. 'Data Availability').

RESULTS AND DISCUSSION

To test whether the inclusion of cell type-specific methylation information and explicitly modelling dependencies within DNA-binding sites is beneficial for a specific TF, we follow the procedure illustrated in Figure 1. We start from whole-genome bisulfite sequencing data for the cell type at hand, discretize methylation calls by the betamix (46) approach, and use these binary methylation calls to convert the original hg38 genome sequence into a methylation-aware genome version. Specifically, we convert methylated 'C' to 'M' and 'G' opposite of a methylated 'C' to 'H', yielding an extended 6-letter alphabet.

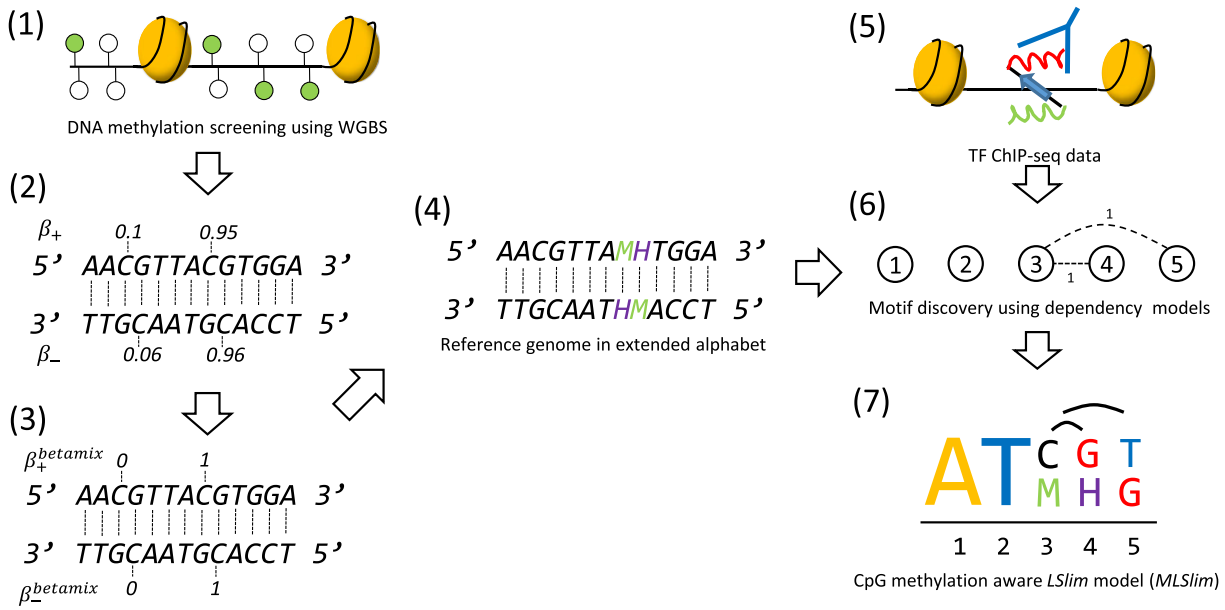


Figure 1. Overview of the MeDeMo workflow: (1) DNA methylation is assessed using whole genome bisulfite sequencing. (2) DNA methylation is quantified using β-values. (3) Methylation calls (β-values) are discretised using the BETAMIX approach resulting in a binary methylation state for each Cytosine in a CpG context. (4) A novel reference genome is generated by denoting occurrences of methylated cytosines with the letter M and occurrences of guanines opposite of a methylated cytosine with the letter H. (5) *In vivo* transcription factor binding site information are obtained using peak calls from TF -ChIP-seq data. (6) TF binding data is used for motif discovery with LSLIM models on the methylation aware reference genomes; (7) resulting in methylation aware TF motif representations.

Table 1. Overview of the combinations of genome variants and motif models considered in this study

		Genome variant	
		Original hg38	Methylation-aware
Model	PWM	*PWM.hg38*	*PWM.methyl*
	LSlim(5)	*LSlim.hg38*	*LSlim.methyl*

Based on the ChIP-seq peaks downloaded from ENCODE, we extract sequences under the peaks, which serve as input to the de novo motif discovery. As statistical binding site models, we use either Position Weight Matrix (PWM) (52,53) assuming independence of nucleotides, or LSim(5) models capturing dependencies between nucleotides over a distance of at most five nucleotides (8). Both types of models are applied to sequences under peaks extracted from the original hg38 genome, or to sequences under peaks extracted from the methylation-aware genome version for the cell type of the ChIP-seq experiment. This results in four modelling alternatives (Table 1), namely (i) PWM applied to original hg38, (ii) PWM applied to the methylation-aware genome, (iii) LSlim applied to original hg38 and (iv) LSlim applied to the methylation-aware genome. The binding site models of all four modelling alternatives are determined within the common Dimont framework. As a method published previously for the same purpose, we include PWMs learned by the MEPIGRAM approach (34) from the methylation-aware genome into the comparison.

In the remainder of this section, we first investigate for which TFs the introduction of a methylation-aware genome and the inclusion of dependencies yield an improvement in classification performance discriminating bound from unbound sequences. We then consider specific examples of TFs that show such an improvement, discuss their binding motifs in relationship to methylation, and study general trends in sensitivity of binding models to methylation in binding sites. We finally present prototypical examples of TFs for which the combination of methylation information and modelling dependencies is pivotal to optimal performance.

Investigating the impact of DNA methylation on binding

For benchmarking MEPIGRAM and the different modelling alternatives of MEDEMO, we follow a classification-based approach. Here, motif models are tested for their capability of distinguishing bound from unbound sequences. We consider as sequences bound by a specific TF those under a ChIP-seq peak, whereas unbound sequences sampled uniformly across the genome (cf. Materials and Methods). Since for the majority of TFs, this is a highly imbalanced classification problem, we use the area under the precision-recall curve (48) as a performance measure. For each TF, we collect all data sets that are available from ENCODE for the cell types under study (GM12878, HepG2, K562, liver), which might include replicate experiments for the same combination of cell type and TF, e.g., performed in different labs.

We further follow a 10-fold cross validation strategy to be able to also assess classification performance on the data from the same experiment. For each partition of the 10-fold cross validation, we consider the motif reported on rank 1 by the SLIMDIMONT framework during training (cf. section

Training procedure) for evaluating model performance on test data.

In the following, we distinguish *within* cell type (i.e. training and test cell types match) and *across* cell type (i.e. training and test cell types are different) classification performance. For each of these sub-sets of classification problems, we collect all AUC-PR values and perform a one-sided Prentice rank sum test (54,55) (using `prentice.test` from R-package `muStat`) between each pair of modelling alternatives considering cross-validation folds as replicates of the same experiment (replicated block design) and using a significance level of $\alpha = 0.05$. In addition, we count the number of data sets, for which one alternative yielded a higher classification performance than the second one. Finally, we visualize the differences of AUC-PR values in violin plots as shown in Figure 2.

Analysis of binding models for ATF3 (across cell type setting) is presented as an example in Figure 2A. We show the corresponding results for all six pairwise comparisons of the four modelling alternatives. For instance, from the left-most panel of Figure 2A, we observe that the difference between the AUC-PR values of LSlim.methyl and LSlim.hg38 are mostly positive indicating an improved performance of LSlim models on the methylation-aware genome compared with standard hg38. This difference is statistically significant with a p-value of 5.78×10^{-13}, where for 123 cases (data sets $\times$ cross validation folds) LSlim.methyl performs better than LSlim.hg38, whereas the opposite holds for only 37 cases. Similarly, we find a significant improvement of LSlim.methyl over PWM.methyl (indicating that dependencies are beneficial), of LSlim.methyl over PWM.hg38, of LSlim.hg38 over PWM.hg38 and of PWM.methyl over PWM.hg38. For the comparison of LSlim.hg38 (only dependencies) with PWM.methyl (only methylation information), we do not observe a significant difference, which indicates that both aspects of the novel approach contribute to a similar degree to the final classification performance of LSlim.methyl. Together, these results make ATF3 a prototypical example of a TF for which the combination of methylation information and modelling dependencies is important for yielding the best classification performance among the considered classification approaches.

In Figure 2B, we present further examples of TFs for which the combination of methylation information and modelling dependencies is beneficial. These cases also illustrate the varying quantity of combinations of training and test data sets from different cell types available for different TFs (each split into 10 cross validation folds). Here, these span from 2 (USF2, one ChIP-seq data set for each of two cell types) to 18 (JUND and MAX). In all cases, the improvement of LSlim.methyl over PWM.hg38 is statistically significant, although the magnitude of the improvement in classification performance (y-axis) as well as the proportion of cases where one model performs better than the other differ among these TFs.

Finally, Figure 2C shows examples of TFs for which the improvement of PWM.methyl over PWM.hg38 is significant but the improvement of LSlim.methyl over PWM.methyl is not, i.e. TFs for which inclusion of methylation information is beneficial but modelling dependencies does not lead to further improvements.

While we refer to differences of AUC-PR values for comparing modelling alternatives in Figure 2 for a compressed representation, we show scatter plots of absolute performance of the modelling alternatives in Supplementary Figures S1–S3. In addition, we include a comparison of PWM.methyl and LSlim.methyl models, respectively, to a random classifier in Supplementary Figures S4–S6. In general, we observe that prediction performance is highly dataset dependent. Partly, this can be explained by differing numbers of peaks in different ChIP-seq data sets, which affects the class ratio between positives and negatives. This also becomes apparent in the comparisons to the random baseline classifier, which obtains AUC-PR values according to the class ratio. However, the difference in absolute performance may also be a result of cell type-specific differences if predictions are made on the same test dataset using models trained on different training cell types (e.g., HepG2 for MNT, K562 for TBL1XR1).

Comparing scatter plots against the PWM.hg38 model (Supplementary Figures S1–S3) with the scatter plots against the random baseline classifier (Supplementary Figures S4–S6), we observe that in some cases (especially prominent for ARID3A, ARNT and NONO in the across cell type setting) the PWM.methyl model even performs worse than the random baseline classifier. However, the corresponding combinations of training and test cell types are identical to those where performance is low in general (low number of peaks) and where we do not observe a clear improvement of the PWM.methyl model over the PWM.hg38 model. Hence, this observation does not impair our general comparison of the PWM.methyl/LSlim.methyl and PWM.hg38 models.

Supplementary Figures S7–S10 include a detailed comparison of the performance of MEPIGRAM with the corresponding MEDEMO models (PWM.methyl and LSlim.methyl, respectively) for all TFs considered in Figure 2. In summary, MEDEMO performs similarly to MEPIGRAM for some TFs but yields an improved prediction performance consistently across the two training methods for the majority of examples, and we compare the two methods on all datasets in the following section.

Benchmarking of modelling alternatives

We compile an overview of pairwise comparisons of modelling alternatives within MEDEMO as well as MEPIGRAM in Figure 3. Here, we apply stringent criteria for counting one modelling alternative to perform better than a second one for the TF at hand. Specifically, we require the improvement to be significant i) in the within cell type *and* across cell type settings consistently for both training variants (shuffled and randomly drawn negatives, cf. Materials and Methods). For the comparison to MEPIGRAM, we require consistent results for within and across cell type comparisons as well.

Of the 335 TFs considered in total, ChIP-seq data sets for at least two cell types are available for 144 TFs, while all remaining cannot meet the stringent criteria by definition.

Among 144 TFs, we observe a significant improvement of the methylation-aware MEDEMO modelling alternatives over the methylation-aware MEPIGRAM motifs for 66

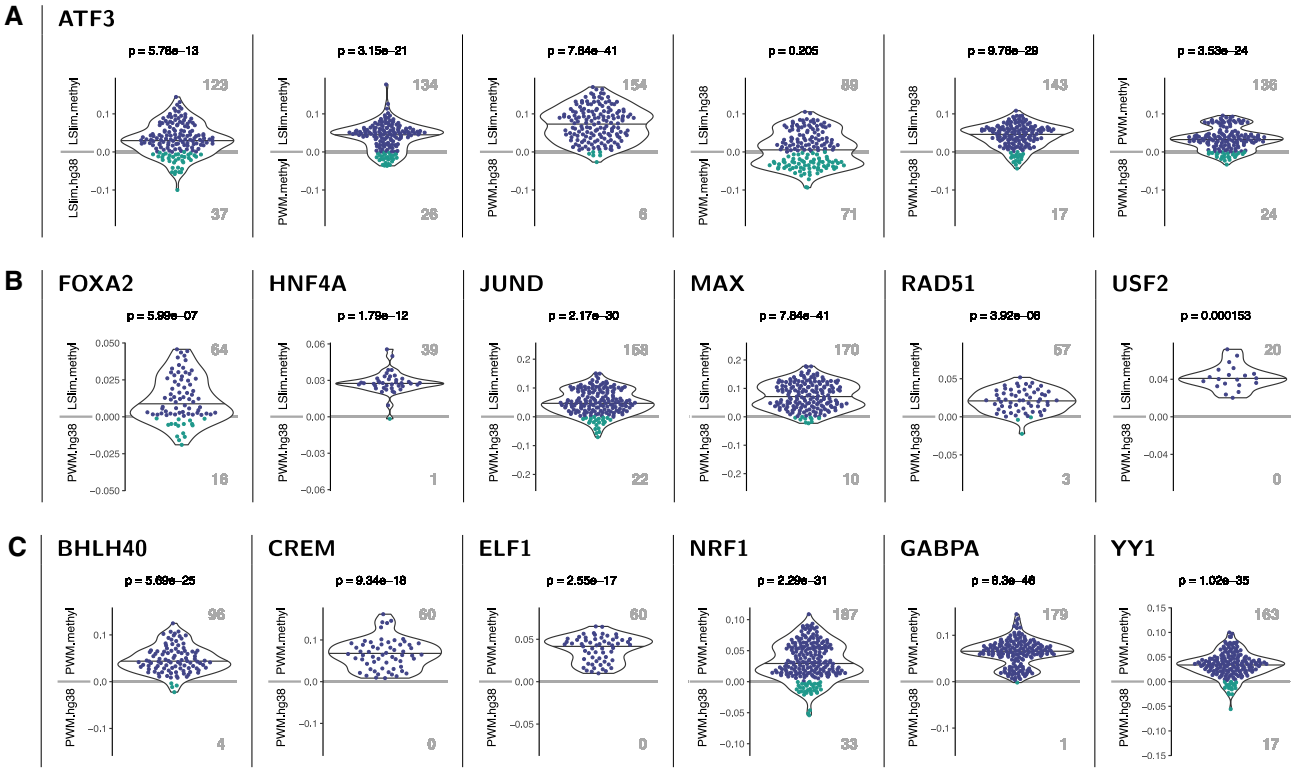


Figure 2. Examples of TFs with significantly improved classification performance (AUC-PR) in across cell type predictions using a methylation-aware genome. Each panel shows a pairwise comparison of models as indicated by the y-labels above and below the zero line. Each dot represents a case (data sets × cross validation folds) with different colours for positive (i.e., top model performs best) and negative (i.e., bottom model performs best) differences of AUC-PR values. Total number of cases where one model performs better than the other are shown as boldface, grey numbers. In addition, points are summarised by a violin plot and corrected p-values for the H_0 that both models perform identical (Prentice test) are given in the header. (**A**) Pairwise comparison of different modelling variants for ATF3. We find that all methylation-aware models perform better than their counterparts learned on the original hg38 genome and that dependency models (LSlim) perform better than PWM models on the same genome variant. For instance, LSlim.methyl performs better than LSlim.hg38 in 123 cases, whereas the opposite is true for only 37 cases, leading to a P-value of 5.78×10^{-13}. (**B**) Comparison of methylation-aware dependency models (LSlim.methyl) with PWM models using standard hg38 (PWM.hg38) for TFs with a clear advantage of the combination of methylation information and modelling dependencies. (**C**) Comparison of PWM models learned from the methylation-aware genome with those learned from the standard hg38 genome.

(PWM.methyl) and 75 (LSlim.methyl) TFs. Notably, even the training variants of MEDEMO using the original hg38 genomes often outperform MEPIGRAM. By contrast, MEPIGRAM performs significantly and consistently better than the MEDEMO models only for 3–6 TFs. These TFS are ASH2L, BRD4, TBP, MAZ, SMAD1 and ZBTB7A, where MEPIGRAM performs better than the LSlim.methyl variant of MEDEMO only for MAZ, SMAD1 and ZBTB7A. We provide a detailed comparison on the level of individual data sets and binding motifs between MEPIGRAM and MEDEMO in Supplementary Figures S11–S16. In summary, we find that MEPIGRAM often discovers GC-rich motifs, which in multiple cases resemble the known SP1 motif, whereas MEDEMO generates more diverse motifs which, however, are less capable of distinguishing positive from negative sequences and, hence, yield lower AUC-PR values.

Among PWM and LSlim models trained by MEDEMO, we observe the largest number of TFs (51) with significant and consistent improvement comparing LSlim.methyl (methylation information *and* dependencies) against PWM.hg38 (neither of the two). We also find improvements for a substantial number of TFs when considering intra-motif dependencies in addition to methylation information (i.e. LSlim.methyl compared with PWM.methyl, 27 TFs), or when considering methylation information in addition to intra-motif dependencies (i.e. LSlim.methyl compared with LSlim.hg38, 18 TFs). Modeling only dependencies (LSlim.hg38 vs. PWM.hg38) or including only methylation information (PWM.methyl vs. PWM.hg38) yields an improvement for 33 and 23 TFs, respectively. For the direct comparison of either including only dependencies (LSlim.hg38) or only using a methylation-aware genome (PWM.methyl), we find balanced numbers of TFs with an improvement in either direction (16 and 13 TFs). The opposite comparisons yield a significant improvement only for a minority of at most one TF. Considering the traditionally used PWM model using the standard hg38 genome, we find a better performance for PWM.hg38 compared with LSlim.hg38, LSlim.methyl or PWM.methyl for none of the TFs studied. We find one TF (HDAC2) for which the PWM model yields a better performance than the LSlim model on the methylation-aware genome. In this case, the PWM.methyl model significantly outperforms all other modelling alternatives and adding dependencies appears to be rather detrimental. Further, we find one TF (CTCF) for which the LSlim model works better on the

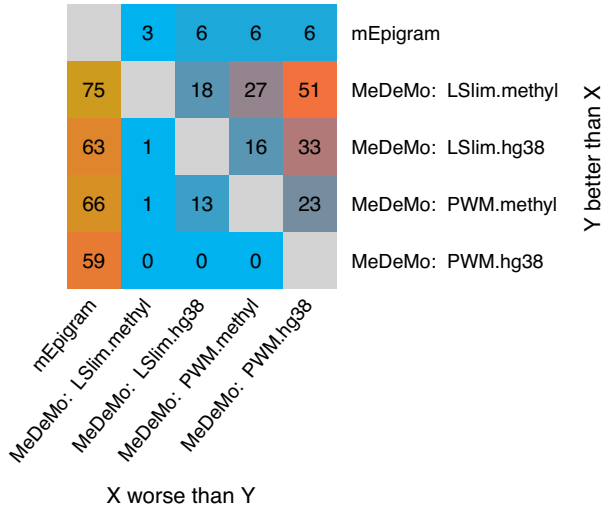


Figure 3. Pairwise comparison of different modelling variants. For MEPIGRAM, and each of the MEDEMO models (PWM, LSlim) and each genome variant (original: hg38, methylation aware: methyl), we determine the number of TFs for which the model listed in the row performs significantly better than the model listed in the column i) within and across cell types, and ii) consistently using randomly drawn and shuffled negatives.

original than on the methylation-aware genome. Converse to HDAC2, intra-motif dependencies seem to be of greater importance for CTCF than methylation information, and the LSlim.hg38 outperforms any other modelling alternative.

The examples previously shown in Figure 2A and B are in the intersection of all three sets for which LSlim.methyl performs better than any of the other three modelling alternatives within the MEDEMO framework (second row of Figure 3), whereas those shown in Figure 2C are from the union of LSlim.methyl vs. LSlim.hg38, LSlim.methyl vs. PWM.hg38 and PWM.methyl vs. PWM.hg38, excluding TFs where one model on the original hg38 genome performs better than its methylation-aware counterpart.

We present a list of those TFS for which methylation information was beneficial for prediction performance in Table 2. Here, we exclude TFs without direct and sequence-specific DNA binding (as discussed for BRCA1 below), while we provide a complete list of TFs in Supplementary Table S3. For all motifs listed in Table 2, we also provide a comparison of the motifs discovered using the PWM.methyl model with those learned using the PWM.hg38 and corresponding motifs from the databases Jaspar (56), Hocomoco (57), Factorbook (58) and CIS-BP (59) in Supplementary Figures S17–S21.

For 10 TFs, we find that the combination of methylation information and modelling intra-motif dependencies (i.e. LSlim.methyl) yields a better prediction performance than either using methylation information (PWM.methyl) or modelling dependencies alone. For three of these (NONO, RAD51, USF2) we find that modelling dependencies alone does not improve prediction performance over the standard PWM model on the hg38 genome, but only in combination with methylation information. For the remaining 7 TFs, our results are compatible with independent contributions from including methylation information and modelling intra-motif dependencies.

The motifs of 18 out of 28 TFs (64%) listed in Table 2 show a prominent CpG in their core motif, whereas this is true for only 53% (76 out of 144) of the motifs of all TFs in the collection. We further examine the remaining $76 - 18 = 58$ TFs that contain a prominent CpG in their core motif and find inconclusive results (improvement only for one training variant) for 32 TFs. For these, our stringent filtering aiming at a low number of false positives does not allow for giving a final assessment of the influence of DNA methylation. The complementary set of 26 TFs contains many non sequence-specific binders according to Factorbook (58). However, there are 8 TFs (CEBPB, CTCF, EGR1, MXI1, NR2C1, ZBTB33, ZNF143 and ZNF592) that specifically bind DNA and contain a prominent CpG in their core motif but still, prediction performance does not profit from including methylation information in our evaluation.

For four of these (MXI1, NR2C1, ZNF143 and ZNF592), our data contain only one experiment in each of two cell types, which complicates statistical assessment of the improvement, while for ZBTB33, we obtain discordant results for the within cell type and the across cell type setting (Supplementary Figure S22). For CEBPB (Supplementary Figure S23), the improvement is only significant in the within cell type setting, although the discovered motif is highly similar to the Jaspar (56) motif and we find the corresponding PWM model to be methylation sensitive. For CTCF, we find an improvement of prediction performance in neither setting, but also the CTCF motif resembles the Jaspar motif and two positions of the model are methylation sensitive (Supplementary Figure S24). Finally, we find an improvement for EGR1 only in the across cell type setting, although the discovered motif is similar to the Jaspar motif and two positions of the model are methylation sensitive (Supplementary Figure S25).

In the literature, CEBPB (25,60–62) and CTCF (30,63–65) have been found to be methylation sensitive TFs. Hence, these TFs may represent potential false negatives of our classification-based search strategy for methylation-associated TFs.

For all TFs listed in Table 2, we further compare the performance of PWM models learned by MEPIGRAM with the different models learned by MEDEMO in Supplementary Figures S7–S10, which further illustrate our findings summarized in Figure 3. Of the 10 TFs that show an improvement for the combination of methylation information and modelling dependencies, MEDEMO using LSlim models performs better than PWMs from MEPIGRAM for 8 TFs, while for MAX we only find an improvement when the LSlim model is learned using randomly sampled negatives, and for SP1, the MEPIGRAM PWM works better than the LSlim model learned using shuffled negatives.

Methylation sensitivity of binding models

Having established a set of TFs for which the inclusion of methylation information leads to an improvement in the benchmark study, we further investigate binding preferences of TFs in the context of their binding motifs. To

Table 2. Summary of TFs that profit from considering DNA methylation in the motif models. For each TF, we list the availability of ChIP-seq data sets for the four cell types studied. Columns 'Methylation' and 'Methyl. & Deps.' indicate a significant and consistent improvement (y: yes, n: no) by including information about methylation in general and/or in combination with modelling intra-motif dependencies, respectively. We also note if a binding motif contains a prominent CpG in its core (CpG) and if methylation sensitivity in the core is larger than outside the core (Core). In the last column, we note references to the literature for TFs that have already been reported to be methylation sensitive, where '–', '+' and 's' indicate negative or positive influence of methylation or general methylation sensitivity according to the referenced publications, respectively

TF	GM12878	HepG2	K562	liver	Methylation	Methyl. & Deps.	CpG	Core	Literature
ARID3A	x	x	x		y	n	n	y	new
ARNT	x	x	x		y	n	y	y	- (30)
ATF3		x	x	x	y	y	y	y	- (23)
ATF7	x	x	x		y	n	y	y	- (23)
BHLHE40	x	x	x		y	n	y	y	- (23)
CREM	x	x	x		y	n	y	y	- (23)
ELF1	x	x	x		y	n	y	y	- (23,60)
FOXA1		x	x	x	y	n	n	n	- (85)
FOXA2		x		x	y	y	n	n	new
FOXK2	x	x	x		y	n	n	y	new
GABPA	x	x	x	x	y	n	y	y	- (23)
HNF4A		x		x	y	y	n	y	new
HNF4G		x		x	y	y	n	y	new
JUND	x	x	x	x	y	y	n	y	- (23)
MAX	x	x	x	x	y	y	y	y	s/- (23,30)
MNT		x	x		y	n	y	y	s/- (30)
NFATC3	x		x		y	n	n	y	+ (23)
NONO		x	x		y	y	y	y	- (86)
NR2C2	x		x		y	n	y	y	new
NRF1	x	x	x		y	n	y	y	- (30)
PKNOX1	x		x		y	n	n	y	new
RAD51	x	x	x		y	y	y	y	new
SIX5	x		x		y	n	y	y	new
SP1		x	x	x	y	y	y	y	+/– (23,70,71–71)
TBL1XR1	x	x	x		y	n	n	y	new
USF2	x		x		y	y	y	y	- (23)
YY1	x	x	x	x	y	n	y	y	different motif (68)
ZBTB40	x		x		y	n	y	y	new

this end, we compute a position-specific profile of methylation sensitivity by altering CpG dinucleotides within putative binding sites to their fully methylated variant MpH and recording the resulting differences in the corresponding binding scores according to the motif model. By this means, we may decode the information about methylation preference captured by the motif model. If the difference of binding scores is positive, this corresponds to MpH dinucleotides (i.e. methylated DNA) being preferred over CpG dinucleotides by the model at a given position, and vice versa. By referring to the level of predicted binding sites, this measure of methylation sensitivity is easily transferred to LSlim models, where methylation sensitivity may depend on the sequence context.

In Figure 4, we present six examples of such profiles of methylation sensitivity according to the corresponding PWM models, plotted below the sequence logo of their predicted binding sites. As might be expected, all these examples have in common that their motifs contain prominent CpG dinucleotides, although with different frequencies and in different contexts. For ELF1, CREM and MAX, we observe one prominent CpG dinucleotide as part of their motifs, where CpG content varies between 0.57 (ELF1) and 0.85 (CREM). In all three cases, methylation of this CpG dinucleotide according to the model leads to a decrease in the prediction score, indicating that methylation is detrimental for binding affinity or that TF binding has a negative influence on DNA methylation. Similar patterns also occur for YY1 with one prominent and several less frequent CpG positions, and for BRCA1 and NRF1 exhibiting two prominent CpG dinucleotides each.

For the NRF1 model, it appears as if methylation affects one of the CpGs (position 8/9) to a lesser degree than the other (position 14/15). However, ChIP-seq does not provide strand information and the strand model encapsulating the PWM allows for switching the strand orientation of the binding site. For these reasons, and because the motif of NRF1 is clearly palindromic, this phenomenon needs to be interpreted with care. An alternative explanation might be that once one of the CpGs present in NRF1 binding sites is methylated, additional methylation of the other CpG does not lead to a substantial further effect. Notably, the binding motif discovered for BRCA1 does not match the canonical motif present in HOCOMOCO (57). BRCA1 has been reported to bind DNA directly but without sequence specificity (66). The ZBTB33-like motif discovered by our approach could possibly be due to indirect binding, and a similar motif has been reported for BRCA1 before (67).

Strikingly, the influence of methylation on the prediction score at high-CpG positions is negative in all examples presented in Figure 4, suggesting that DNA methylation may lead to reduced binding affinity or binding negatively influences DNA methylation for many TFs. In order to investigate if this observation constitutes a general tendency among the studied TFs, we consider all TFs with a significant and consistent improvement in prediction performance

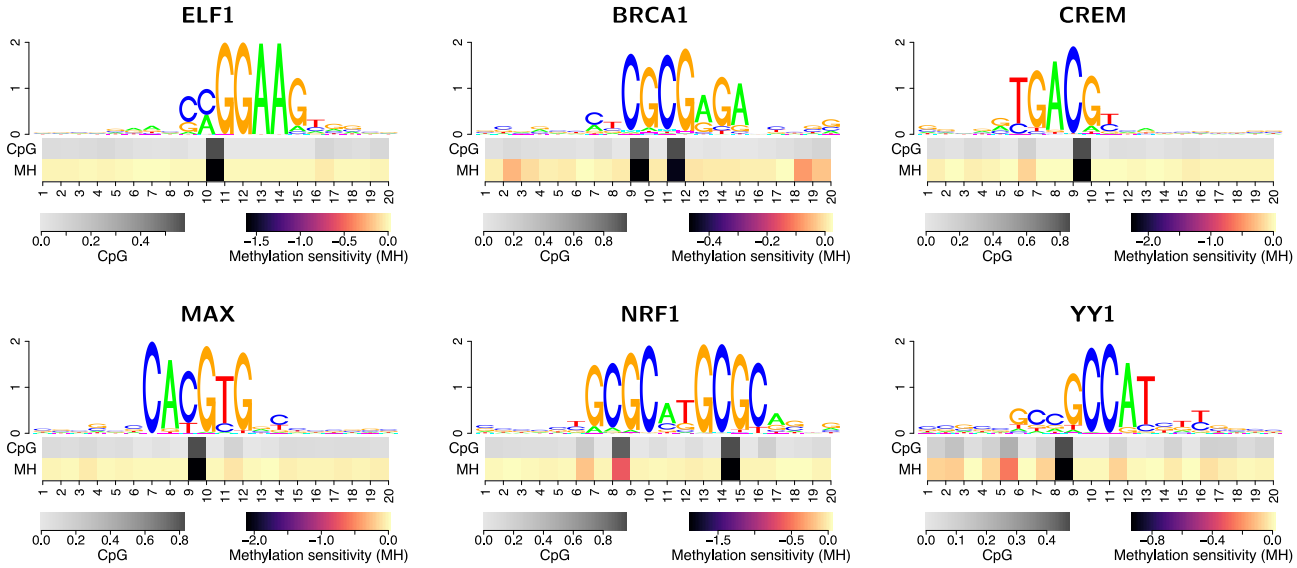


Figure 4. Methylation sensitivity of binding models with improved performance using a methylation-aware genome. In each panel, the top part show a sequence logo of the discovered motif using the extended alphabet. However, since the model learned to penalize methylated DNA in all six cases, additional symbols are only visible in case of BRCA1. In the bottom part of the plot, we visualize position-specific CpG content (top row with grey scale) and methylation sensitivity (bottom row with colour scale) within predicted binding sites. Positive values of methylation sensitivity indicate preferred binding of methylated DNA, whereas negative values indicate methylated DNA being disfavored. For all six TFs, we observe a detrimental effect of DNA methylation at frequent CpG positions.

when including methylation information (cf. Table 2). For each of these TFs, we compute the corresponding profiles of methylation sensitivity of their binding model per data set and record the range of values (i.e. minimum value to maximum value) present in the profile. Strong deviations from 0 of the maximum or minimum value indicate a clear preference for methylated or unmethylated DNA according to the model, respectively. From Figure 5, we observe that the maximum value is only slightly above 0 for the wide majority of TFs, whereas for many TFs, the minimum value is clearly below 0. This indicates that for most TFs, the profiles of methylation sensitivity indeed are similar to those presented in Figure 4. There are a few examples of TFs (FOXA1, FOXA2, HNF4A, HNF4G, RAD21), for which neither the maximum nor the minimum of methylation sensitivity shows a strong amplitude. These TFs do not have a prominent CpG in their binding motifs. Nonetheless, inclusion of methylation information leads to an improvement in prediction performance. We discuss possible explanations of this observation for two examples below (FOXA1 and FOXA2, Figure 6).

For several of the TFs shown in Figure 5, a negative influence of methylation on their binding has been reported before. This includes ARNT (30), ATF3/7 (23), CREM (23), ELF1 (23,60), GABPA (23), JUND (23), MAX (23,30), MNT (30), NRF1 (30), USF2 (23) and YY1 (68). For NFATC3, a previous study based on HT-SELEX experiments (23) found preferred binding of NFATC3 to methylated DNA, whereas our results suggest a negative association with DNA methylation. Notably, the motif detected by MEDEMO is highly similar to the motif reported in factorbook (cf. Supplementary Figure S19) but considerably different to the motif reported by Yin *et al.* (23). One reason for this observation might be the difference between the *in vitro* setting considered by Yin *et al.* and the *in vivo* ChIP-seq data considered in this study, for instance due to effects of co-binding with other TFs that are not present in the *in vitro* setting. SP1 shows a generally negative association with methylation of its binding sites in our data, although with cell type-specific strength. Previous results for SP1 have been contradictory, as some studies suggested a positive influence of binding site methylation (23), whereas others indicated no decisive influence (69), negative effects (70), or the prevention of methylation by SP1 binding (71). In general, preference for de-methylated DNA may be observed either due to the direct binding preference of the TF at hand, or due to a de-methylation of the bound region as an effect of TF binding. Based on our data, these two cases could not be distinguished.

The reasons for the mostly detrimental influence of methylation for the TFs in our study could be manifold. First, this could be a bias introduced by the specific selection of TFs under study, although no such bias has been introduced intentionally, since we consider all TFs with ENCODE data sets in at least two of the selected cell types. Specifically, CEBPB (25,61,62), SMAD5 (23) and ZBTB33 (23,72,73) have been reported to prefer methylated DNA, but we did not observe a significant and consistent improvement of prediction performance in our study. For GATA1/2/4 (60,62), IRF2 (23), KLF16 (23), NFATC1 (23), STAT1/5A (60) and ZNF274 (23), we had only data for one of the cell types studied, which prevented us from studying performance across cell types. Second, this result might be an artifact of our method. While we cannot rule out this possibility in general, we do observe clearly positive methylation sensitivity values for a few TFs. Examples (ZBTB33 with inconsistent results across cell types, and NFATC1 and ZNF274 with ChIP-seq data available only

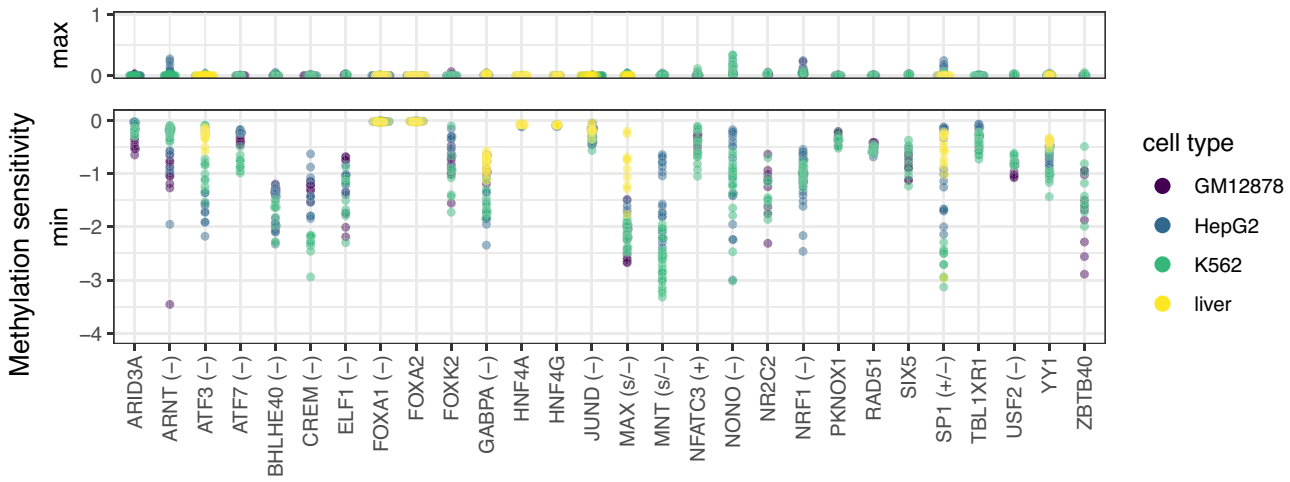


Figure 5. Methylation within binding motifs is mostly detrimental in models with significantly and consistently improved prediction performance (cf. Figures 2 and 3). For each TF and each data set, we record the profiles of methylation sensitivity as shown in Figure 4. We aggregate this profile to two values per data set by computing the minimum and maximum value of methylation sensitivity, which captures the range of values observed in the profile. Here, we plot these maximum and minimum values of methylation sensitivity across all training data sets. We observe a large amplitude of negative values for the minimum (i.e., methylated DNA being disfavored) but only slightly positive values for the maximum, indicating that—according to the models—DNA methylation is detrimental for the majority of TFs. Methylation sensitivity of TFs according to the literature is given in parentheses (if present), and the corresponding references are given in Table 2.

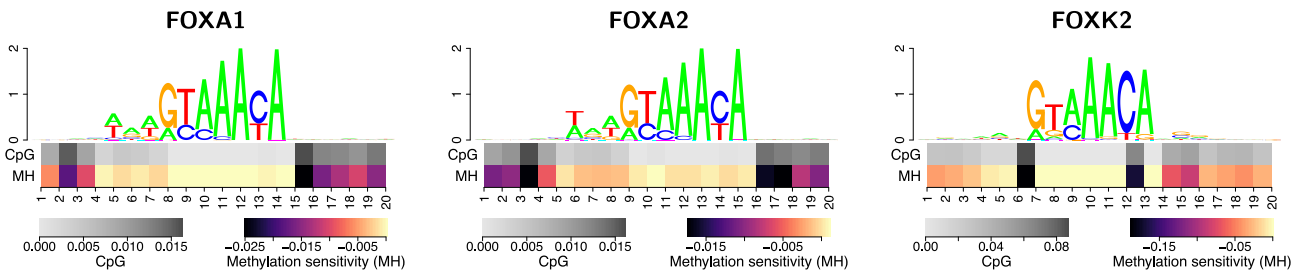


Figure 6. Methylation sensitivity may differ between members of a TF family. While methylation sensitivity of the binding models for FOXA1 and FOXA2 is highly similar in HepG2 cells, that of FOXK2 is noticeably different, although all three motifs appear to be highly similar. This behaviour is consistent between different cell types (Supplementary Figure S27).

for one cell type) are provided in Supplementary Figure S26. Hence, we may at least conclude that our method is capable of capturing such patterns in general. Third, there might also be a bias of methylation on the ChIP-seq experiment that constitute the basis of our approach, although we did not find this to be reported before. For instance, methylation might influence the amplification step in the ChIP-seq protocol, which could lead to an under-representation of reads from methylated peak regions.

Methylation sensitivity may vary within a TF family

As we had ChIP-seq data from TFs with the same binding domain (family) and similar consensus sites we wondered, whether there could be differences in the sensitivity to DNA methylation for individual family members. For example, the models for FOXA1 and FOXA2 showed a low amplitude in methylation sensitivity in Figure 5, whereas FOXK2 binding appears to be more strongly associated with DNA methylation. Although all three TFs are members of the forkhead box family, they play different roles related to development and disease (74,75). In Figure 6, we present the binding motifs and profiles of methylation sensitivity discovered by our approach for FOXA1, FOXA2 and FOXK2 in HepG2 cells. In general, all three motifs follow the consensus GTAAAYA with slight deviations. The major difference between FOXA1/FOXA2 and FOXK2 motifs is an additional A/T-rich stretch directly preceding this canonical motif. With regard to methylation sensitivity, we find more prominent difference between the three TFs. Specifically, the models for FOXA1 and FOXA2 exhibit a mildly negative effect of methylation at positions bordering their core motif. While the influence on the binding score of any of these positions individually is rather low, the combination of multiple methylated CpGs at bordering positions might still have an effect on binding site prediction. By contrast, FOXK2 shows two, still rather infrequent, CpG dinucleotides at positions 6/7 and 12/13 of the core motif, which are not present in the FOXA1/FOXA2 motifs. Both of these positions show a stronger sensitivity to methylation than any position of FOXA1/FOXA2. This general picture is consistently observed in other cell types (Supplementary Figure S27). Biologically, this observation might be linked to the mechanism of FOXA1 and FOXA2 acting

as pioneering factors (74,76), although pioneering activity has been shown for FOXK2 as well (75).

DNA methylation sensitivity depends on sequence context

In this study, we identified a substantial number of TFs, for which the combination of methylation information and modelling intra-motif dependencies yields an improvement in classification performance compared with the base model (PWM on original hg38) but also relative to the individual contributions of methylation information and/or modelling dependencies (cf. Figures 2B and 3). Here, we discuss three of those TFs in more detail that illustrate the breadth of the binding landscapes observed and how these are linked to specific profiles of methylation sensitivity. In Figure 7, we present dependency logos (8,50) of the predicted binding sites of JUND (K562 cells), USF2 (K562) and ATF3 (HepG2), which are enriched with partition-specific profiles of methylation sensitivity.

JUND binds DNA as a dimer with a variable 1–2 bp spacer (77) which may be captured by dependency models like the LSlim model employed in this study (8) and more specialized models like TFFMs (10), but not (adequately) by standard PWM models. In the dependency logo, this variable spacer is visible as two distinct blocks, the upper block starting with consensus TGA at positions 3–5 and the lower, smaller block starting with the same consensus (TGA) but already at positions 2-4. Both variants share the consensus TCA at positions 7–9. For the short-spacer variant (upper block), only a small subset of binding sites deviating from the standard consensus (TGYGTCA, 4th partition from top) has a substantial fraction of CpG dinucleotides at positions 5/6, which are moderately methylation sensitive. By contrast, about a quarter of the long-spacer variant (lower block, 6th partition from top) with consensus TGACGTCA exhibits a CpG dinucleotide at positions 5/6, which are strongly affected by methylation. Both, the variable spacer and the specific profiles of methylation sensitivity within both variants, explain why the combination of methylation information and modelling intra-motif dependencies yields a particular advantage for JUND binding sites. Notably, the JUND motif for K562 present in the MethMotif database (36) only represents the short-spacer variant and no specific methylation profile within the core motif, which is likely an effect of the database's limitation to PWM models. By contrast, our results suggest that both spacer variants and the associated patterns of methylation sensitivity are present across cell types (Supplementary Figure S28).

For USF2, we observe a canonical E-box motif with consensus CACGTG for the majority of binding sites, and consensus CACATG for a minority of binding sites displayed as the bottom partition of the dependency logo. Intra-motif dependencies are especially prominent between positions 6 and 10, but also several positions flanking the core motif. The dependency between positions 6 and 10 can be attributed to the consensus CACATG always being preceded by a T at position 6, whereas the canonical E-box motif may also be preceded by C or G. Only those binding sites following the consensus CAYGTG frequently (approx. 80%) exhibit a CpG at positions 9/10, which is then moderately (1st and 2nd partition from top) or strongly (3rd partition from top) affected by methylation. For the partition with consensus CACATG, we find an almost flat profile of methylation sensitivity. Again, this dependency structure and associated varying methylation sensitivity may adequately be captured by dependency models but not by standard PWM models.

Finally, we observe substantial heterogeneity among the binding sites of ATF3, which have been reported before (8). Starting from the top of the dependency logo, we find a partition with consensus TTTACGRC (positions 5–12), followed by a large partition with consensus YCACRTG (positions 6–12), a small partition with consensus TRACGYR (positions 6–12), a partition with consensus TGACGBCA (positions 6–13) and finally a partition with consensus TGAYGYAA (positions 6-13). The diversity of the predicted ATF3 binding sites manifests as strong intra-motif dependencies between positions 7 and 12, 7 and 11, 5 and 7, and 11 and 12. However, all partitions show a considerable fraction of CpG dinucleotides at positions 9/10, which are methylation sensitive to different degrees. Partition 3 (counted from top) exhibits an additional CpG at positions 11/12 with moderate frequency and methylation sensitivity. While each of these partitions could be modelled decently by its individual (methylation-aware) PWM model, only dependency models as proposed in this study are capable of capturing such highly heterogeneous binding landscapes without prior knowledge about their specific structure.

Dependency logos of the remaining seven TFs for which we observed an improvement by the combination of methylation information and modelling intra-motif dependencies are given in Supplementary Figure S29. As described in the previous section, we find methylation sensitive model positions flanking the core motif for FoxA2, which show mutual dependencies. For HNF4A and HNG4G, we observe methylation sensitivity for two motif positions with widely independent contributions in different partitions of binding sites, whereas dependencies are present only between directly adjacent binding site positions. For MAX, the central CpG of the CACGTG core motif shows the strongest signal of methylation sensitivity, but one specific partition of binding sites with pattern C[AG]C[AG]TGCG in addition shows methylation at two additional model positions. For NONO, we find different GC-rich sub-motifs with variable patterns of methylation sensitivity, which might explain why for NONO, the combination of methylation information and modelling dependencies is of special utility. For RAD51, we find two main sub-types of motifs with consensus CACGTGA and CATGTGA, of which only the former shows methylation sensitivity. For SP1, we find dependencies within the canonical motif with a clear signal of methylation sensitivity at the central CpG di-nucleotide. In addition, we find a sub-motif that, according to TomTom (78) is similar to a ZBTB33 motif from Jaspar (MA0527.1) and might, hence, rather represent the motif of an SP1 co-binding TF.

Methylation-aware models may explain differential binding

Having established that incorporating methylation-aware genomes and/or intra-motif dependencies is often

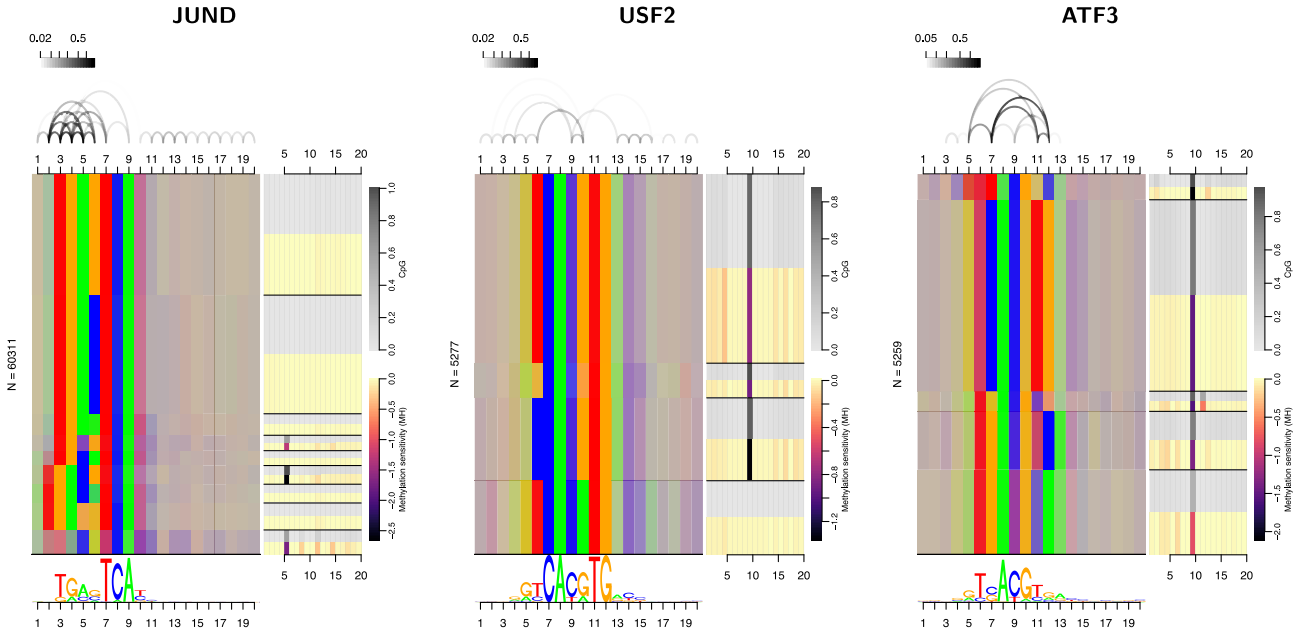


Figure 7. For JUND, USF2 and ATF3, the advantage of combining dependency models with a methylation-aware genome can be attributed to specific properties of the corresponding binding landscapes. For each TF, we visualize predicted binding sites by Dependency Logos that partition binding sites by nucleotides at the most inter-dependent positions and represent each partition using the same colors that are also used in sequence logos (provided below; A=green, C=blue, G=orange, T=red). If a partition contains a mixture of nucleotides at a certain position, colors are mixed as well. For JUND, we find the known variable spacer between the two 3 bp half motif (TGA, TCA), where only the longer spacer frequently contains a CpG. For USF2, the prevalent CpG at positions 9 and 10 shows dependencies to other binding site positions, and is not present in one specific subset (TCACATG) of binding sites. For ATF3, we find broad heterogeneity, where each sub-motif contains CpG at positions 9 and 10 in different proportions.

beneficial for modeling TF binding sites, we further investigate to which extent these models are capable of explaining differential binding across cell types as outlined in Figure 8A.

We consider TFs for which ChIP-seq data are available for two cell types. The idea is to compare whether differences in peak occurrence or ChIP-seq signal in both cell lines can be related to a change in binding scores according to our models. In addition to our models, we consider a simple baseline model, which considers average methylation levels of larger genomic regions (cf. Materials and Methods) instead of scores of individual binding sites. To associate binding scores with ChIP-seq peaks, we consider the binding sites under ChIP-seq peaks as predicted by the same model, which may have been trained on data from one of the cell types considered or from another cell type. We partition the peaks into 'common peaks', i.e. peaks that are overlapping between the two cell types, and 'unique peaks', i.e. peaks that are present only in one of the cell types.

For the common peaks, and associated binding sites and prediction scores, we separate peaks into those without differential methylation in the binding site and, accordingly, identical prediction scores ('equal'), those with a greater score in cell type A than in cell type B ('greater') and vice versa ('less'). In addition, we compute the difference in log peak height ('signal') for each pair of overlapping peaks. If the model could explain differential binding, we would expect these differences to be lower than 0 for the 'less' group, around 0 for the 'equal' group and above 0 for the 'greater' group, and we test pairwise differences in the distribution of log signal values accordingly by a one-sided Wilcoxon rank sum test.

Boxplots representing this analysis for TF CREM in K562 and GM12878 cell types using a PWM model trained from K562 data (cf. Supplementary Table S4) are shown in the left panel of Figure 8B. We find significant differences in log signal between all pairs of groups. The difference between the median values for the 'less' and 'greater' group is 0.6436, which corresponds to a 1.9-fold increase in the ratio of the cell type-specific signal values. Hence, the model appears to be capable of predicting if a peak is larger in cell type A than in cell type B, although the large number of confounding factors, including chromatin accessibility, leads to pronounced variation within each of the groups.

In addition, we plot the differences in log signal against the differences in associated prediction scores and compute the Pearson correlation coefficient between both quantities as shown in the middle panel of Figure 8B. Here, we exclude peaks without differential methylation in the binding site, since these would obtain a fixed score difference of 0. In case of CREM, we find a substantial correlation between both quantities, although only a small subset of common CREM peaks (473 peaks) participates in the analysis. This may indicate that the model is not only capable of predicting the direction of the change in peak height, but that the difference in prediction scores is associated with the magnitude of this change.

For the unique peaks present only in cell type A, we complement the predicted binding site in the methylation-aware genome of cell type A with the corresponding site in the methylation-aware genome of cell type B, and compute the model scores for both site variants. If DNA methylation as captured by the model could explain the presence and

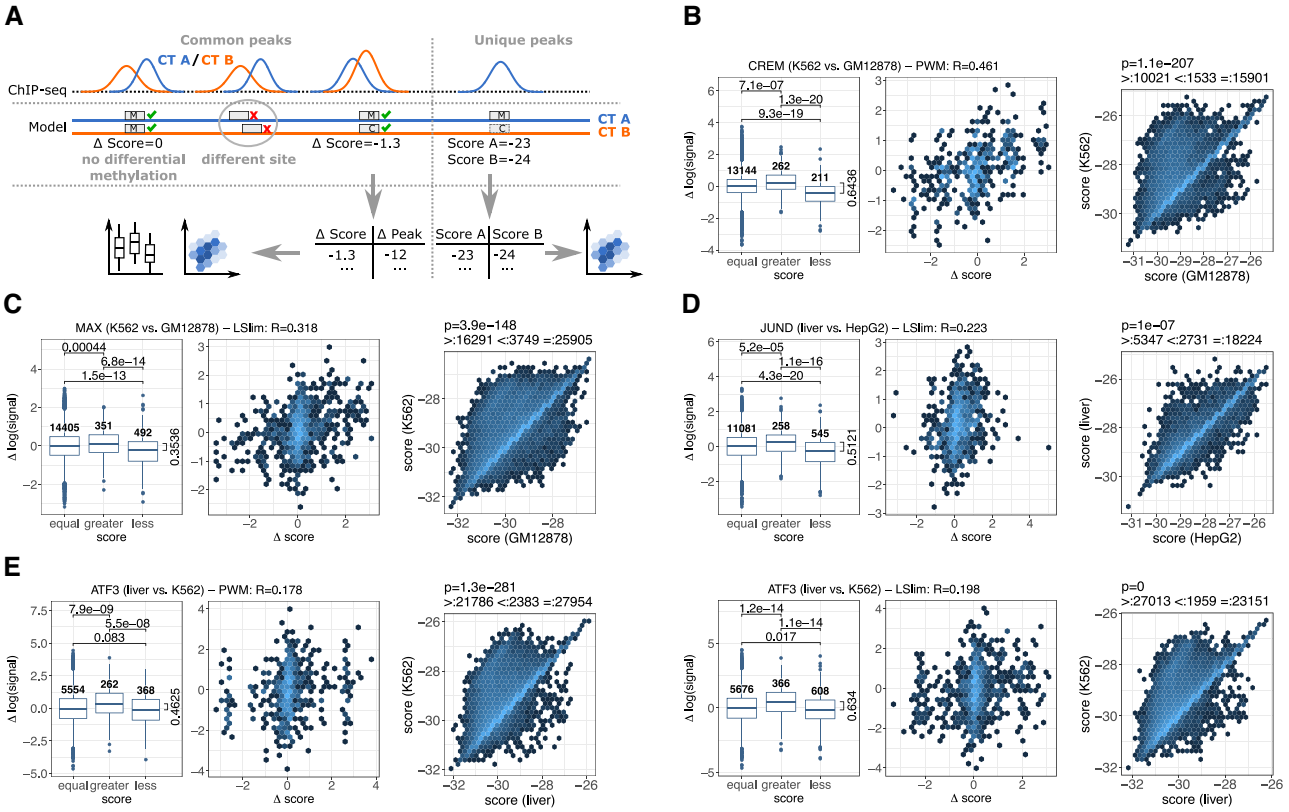


Figure 8. Association of differential model scores and differential binding according to ChIP-seq data. (**A**) Evaluation schema. For common peaks of two cell types, we consider predicted binding sites at the same location that may show differential methylation and, consequently, different model scores in the methylation-aware genomes. For the peaks containing such binding sites, we record the difference of model scores and the difference in peak height (signal). For unique peaks present in only one of the cell types, we record the scores of the binding sites predicted in the two methylation-aware genomes. (**B**) Evaluation of cell type-specific binding for CREM in K562 and GM12878 cell types. Left: Comparison of the difference in log signal for binding sites with an equal score in the methylation-aware genomes of K562 and GM12878, with a larger score in K562 than in GM12878, and vice versa. Number of peaks in each group are given above the boxplots, p-values from a one-sided Wilcoxon rank sum test are above the boxplots, and the difference of median values between the 'greater' and 'less' group is indicated. Middle: For those sites with a prediction score differing between K562 and GM12878, we find a correlation of 0.461 between the difference of the log signals and the difference of the prediction scores in those two cell types. Right: Hexbin representation of the scatter plot of scores determined from binding sites in the two methylation-aware genomes for peaks that are present only in K562. Hexbin colours in log scale. (**C**) Same as (B), but for MAX in K562 and GM12878 cell types. (**D**) Same as (B), but for JUND in liver and HepG2 cell types. (**E**) Same as (B), but for ATF3 in liver and K562 cell types using a PWM model (left group) or an LSlim model (right group).

absence of a peak, respectively, we would expect the score for cell type A to be larger than for cell type B. In the right panel of Figure 8B, we show a hexbin representation of the scatter plot of such pairs of prediction scores for CREM in K562 and GM12878 cell types. Indeed, we find a larger score for K562 than GM12878 for 10 021 sites, whereas the opposite is true only for 1 533 sites. For the majority of 15 901 sites, prediction scores in the methylation-aware genomes of both cell types are identical. Still, the pairwise difference in scores is significantly different from 0 in a Wilcoxon signed rank test ($P = 1.1 \times 10^{-207}$).

In complete analogy, we present results for TF MAX in K562 and GM12878 cell types in Figure 8C. Here, we consider an LSlim model trained on data for cell type HepG2, i.e. in this case the training cell type is different from the two cell types considered in this analysis. Again, we find significant differences between the three groups of peaks divided by the difference in prediction scores. However, the difference of median values between the 'less' and 'greater' groups is only 0.3536 in this case. Here, the correlation analysis shows a slightly lower Pearson correlation than for CREM as well with a visible enrichment of score differences around 0. Considering unique peaks, we find approximately 4-fold as many peaks with larger prediction scores in K562 than in GM12878 for peaks that are present only in K562.

Similar tendencies may be observed for JUND in liver and HepG2 cell types using a LSlim model trained from K562 data (Figure 8D). However, the results for the unique peaks are less pronounced in this case with only 2-fold difference in the number of peaks with greater and lower scores in liver than in HepG2, respectively.

Finally, we illustrate the impact of modelling intra-motif dependencies, i.e., the comparison of PWM and LSlim models, for ATF3 binding sites in liver and K562 cell types in Figure 8E. While we observe a clear advantage of the LSlim model over the PWM model for all three analyses, this advantage is less pronounced than it had been for the classification-based benchmarks in previous sections.

In Supplementary Figures S30–S39, we provide results for these and further TFs, and compare these against the

baseline model that considers average methylation levels in the regions under the ChIP-seq peaks. It is well known that methylation levels in broader regions, especially in enhancers, are highly informative of TF binding (79). In addition, binding models consider only 20 bp of DNA, which makes the presence of differential methylation less likely than for the simple model. Hence, we expect this to be a strong baseline model. For the common peaks, we indeed find that the differences between the 'equal', 'greater' and 'less' groups often obtain lower p-values for the baseline than for the methylation-aware binding models, partly due to the larger number of regions with differences in methylation levels. Notably, the binding models often surpass the baseline models for the correlation analysis. Regarding unique peaks, binding models often more clearly show an enrichment of larger scores for the cell type with a peak being present.

In summary, our results indicate that models of TF binding sites learned from methylation-aware genomes and incorporating intra-motif dependencies may indeed be indicative of presence or absence of a ChIP-seq peak and its peak height, despite the many confounding factors that are not related to DNA methylation but strongly influence TF binding.

Methylation-aware models capturing intra-motif dependencies improve genome-wide binding site predictions

The analyses in previous sections showed that the combination of information about DNA methylation and modelling intra-motif dependencies has the potential to improve classification performance on ChIP-seq data in a cross-validation setting. However, performance of genome-wide predictions of binding sites may be considered more relevant for practical applications. To investigate prediction performance in genome-wide predictions for the 10 TFs that showed a consistent improvement of LSlim.methyl models previously (Table 2), we adopt the general setting of the 'ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge' combining motif-based features with chromatin accessibility data within the Catchitt framework (5). The Catchitt framework combines assays of chromatin accessibility (DNase-seq, ATAC-seq) with genome-wide predictions of arbitrary motif models represented by aggregated score profiles over fixed-size genomic windows. Hence, all types of MEDEMO models can be directly used within the Catchitt framework and their prediction performance compared.

Interestingly, it was shown that enzymes such as DNase1 and the Tn5 transposase, the two most often used enzymes for the measurement of open-chromatin, show differences in DNA cutting or insertion with respect to CpG methylation (7,80). Thus in genome-wide analysis of such data, neglecting the status of DNA methylation may be harmful in two ways. Binding may be impaired due to TFs that show reduced binding of methylation and abundance of open-chromatin reads may also be affected.

Here, we obtain ATAC-seq data for GM12878, HepG2, K562 and liver from ENCODE, and combine chromatin-based features with motif-based features of individual motif models when training Catchitt models on training chromosomes for a specific TF and cell type. These models are then used for predicting binding regions of TFs on test chromosomes in the training cell type and all remaining cell types with ChIP-seq data available for the TF at hand, and compute respective AUC-PR values.

In Figure 9, we compare the prediction performance of LSlim.methyl and PWM.hg38 models for all 10 TFs in genome-wide predictions on the test chromosomes chr1, chr8 and chr21. It has been observed before that the influence of chromatin accessibility data on the final prediction performance is substantially greater than the influence of the specific choice of motif models (5). Nonetheless, we find an improved prediction performance achieved by methylation-aware models capturing intra-motif dependencies for the majority of TF-cell type combinations in the within cell type setting (Figure 9A). Similar improvements can also be observed in the across cell type setting (Figure 9B). Turning to the performance for individual TFs in Figure 9C, we find a few notable and/or systematic cases, where PWM models considering the original hg38 genome sequence (PWM.hg38) perform better than LSlim.methyl models, namely FOXA2 trained on liver and tested on HepG2, JUND trained on GM12878 on all test data sets except liver, and MAX trained and tested on liver. In contrast to the remaining cell types, liver is a primary cell type, and the methylation data and chromatin accessibility data have not been obtained from identical donors in the available data sets, which might partly explain the special behaviour of liver in this and the following comparisons.

For several TFs, namely ATF3, HNF4A, HNF4G, RAD51 and USF2, we observe a consistent improvement of LSlim.methyl models over PWM.hg38 models across the different cell types. Comparing PWM.methyl versus PWM.hg38 (Supplementary Figure S40), we observe a similar picture for FOXA2 and MAX, but also a more balanced performance between both modelling alternatives for other TFs (JUND, NONO). Finding an improvement of methylation-aware compared with methylation-agnostic PWM models for the majority of TFs indicates that methylation information contributes to the improved prediction performance and is not fully redundant to chromatin accessibility, which was avaiable to both models within the Catchitt framework. For LSlim.methyl versus LSlim.hg38 (Supplementary Figure S41), a few additional cases occur with better performance of the methylation-agnostic models, especially for LSlim models trained for JUND on liver, whereas we find an improvement of LSlim.methyl compared with PWM.methyl (Supplementary Figure S42) for the majority of TFs.

In summary, we find that the combination of methylation information and models capturing intra-motif dependencies yields an improved prediction performance compared with traditional PWM models trained on the original hg38 genome for the TFs considered. As these TFs have been selected based on the previous benchmark in cross validation experiments, our results indicate that the observed improvement in cross validation can be largely transferred to more practical applications like genome-wide binding predictions. Based on further model comparisons, we assume that the contribution of intra-motif dependencies is larger than the contribution of methylation information in this scenario.

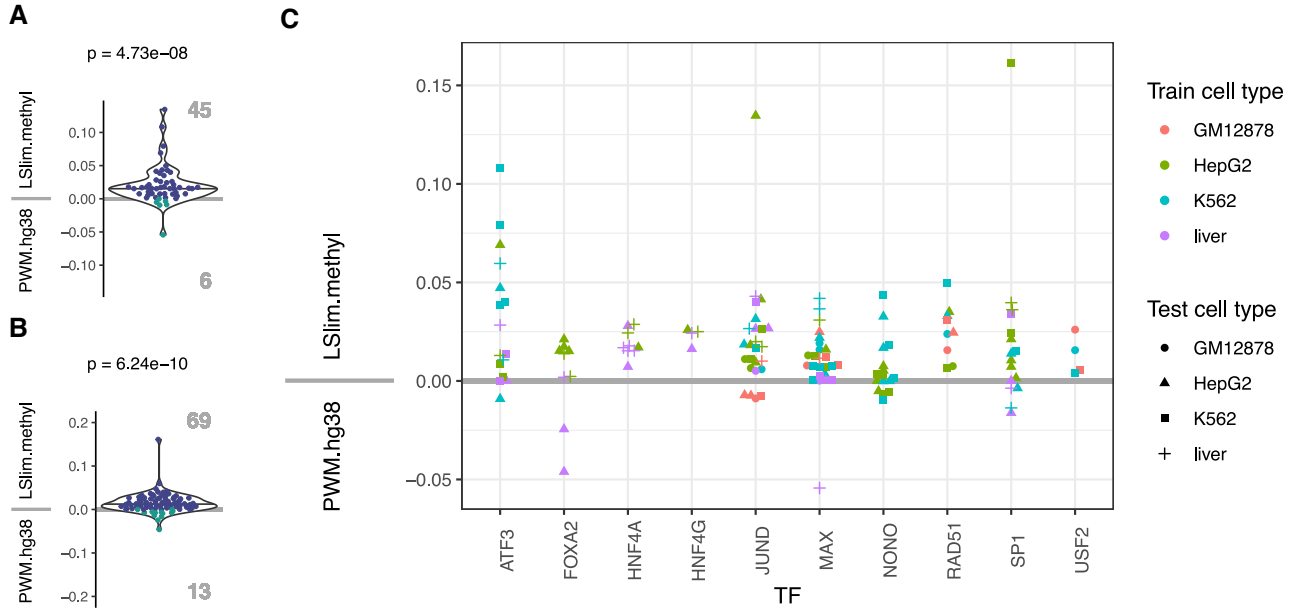


Figure 9. Comparison of the genome-wide prediction performance of PWM models learned on the original hg38 genome (PWM.hg38) and methylation-aware LSlim models (LSlim.methyl) within the Catchitt framework. (**A**) Pairwise comparison of models across all 10 TFs in the within cell type setting. LSlim.methyl performs better than PWM.hg38 for 45 test data sets, while PWM.hg38 performs better than LSlim.methyl for 6 test data sets. (**B**) Pairwise comparison of models across all 10 TFs in the across cell type setting. LSlim.methyl performs better than PWM.hg38 for 69 test data sets, while PWM.hg38 performs better than LSlim.methyl for 13 test data sets. (**C**) Comparison of performance per TF, where the training cell type is encoded by colour and the test cell type is encoded by shape. If multiple data sets are present per TF and cell type, all combinations of data sets are considered.

Conclusions

In this paper, we present MEDEMO, a novel framework for TF motif discovery and TFBS prediction that combines information about DNA methylation with models capturing intra-motif dependencies. Similar to previous approaches (31,35), MEDEMO uses an extended 6-letter alphabet with separate symbols for methylated cytosines and the corresponding guanosines on the opposite strand. In contrast to the MEPIGRAM pipeline (33), MEDEMO does not use a beta value cut-off of 0.5 to obtain a discrete methylation value. Instead, we model the distribution of all beta values using the BETAMIX (46) software to select the cut-off in an informed way. More research is necessary to study the effects of different discretization schemes for methylation-aware models with customized alphabets.

The previous approach of MEPIGRAM is PWM-based, neglecting intra-motif dependencies. Therefore, MEDEMO using PWM models can be seen as an improved instantiation of MEPIGRAM, as we find that models learned by MEDEMO typically outperform PWMs learned by MEPIGRAM. The MEME suite was also extended to predict PWMs in a similar way (35), but we do not expect a direct comparison to offer any additional value as there is no conceptual improvement over the MEPIGRAM approach or the PWM models learned within MEDEMO.

In addition, MEDEMO allows for including intra-motif dependencies when applying LSlim models to methylation-aware input data. Here, we find that the combination of methylation information and intra-motif dependencies improves the performance of binding site predictions for several methylation-associated TFs in cross validation experiments but also in genome-wide predictions. Model visualization provided by MEDEMO further facilitates the interpretation of methylation patterns in putative TF binding sites. In general, there is a smooth transition from perceived dependencies to perceived heterogeneity (8) of binding landscapes, and the latter could alternatively be modelled by (mixture models of) multiple PWM models. As both can be represented well by LSlim models, we consider the modelling approach pursued in this study a useful generalization of previous PWM-based approaches.

Further, MEDEMO allows the research community to leverage the vast amounts of TF ChIP-seq and DNA methylation datasets available to elucidate the methylation dependence of hundreds of TFs *in vivo*, without the need of performing additional experiments such as Methyl-Spec-seq (31). However, as these analyses are based on *in vivo* ChIP-seq data, the effects of DNA methylation on direct and co-/indirect binding may be harder to distinguish than in *in vitro* settings.

Apart from improving TF binding predictions, MEDEMO could also improve the interpretation of methylation QTLs (meQTLs). Methylation QTLs have been reported before to be associated to changes in TF binding, histone modification and gene expression (81). Using MEDEMO, those associations could be understood at more detail, and our analyses regarding differential binding might be a first step towards this goal. Similarly, our tool could provide valuable additional insights into the vast amount of epigenome-wide association studies (EWAS) (82).

Especially in light of upcoming single cell applications as single-cell methylation (83) and single-cell chromatin acces-

sibility assays become available (84), the need of methylation aware TFBS prediction approaches will rise even further in the near future. MEDEMO will help to fulfill these data analysis needs.

DATA AVAILABILITY

All ChIP-seq data sets analyzed in this study are available from ENCODE. ENCODE IDs of the corresponding narrowPeak files are listed in Supplementary Table S1 and can be accessed via the URL schema https://www.encodeproject.org/search/?searchTerm=<ID>. The methylation-aware genome variants and the models generated during the current study are available as a Zenodo archive from https://doi.org/10.5281/zenodo.3723984. The MEDEMO software is available as stand-alone binary versions with graphical user interface and command line interface at http://www.jstacs.de/index.php/MeDeMo. The source code of the software is available from github at https://github.com/Jstacs/Jstacs (permanent DOI: https://doi.org/10.5281/zenodo.8210619). Classes specifically implemented for this project are provided in package `projects.methyl`.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Christopher Schröder for help to run BETAMIX. We thank Ekaterina Shelest for valuable discussions. We thank Yang Gao for help and discussions in an initial test study.

FUNDING

DZHK (German Centre for Cardiovascular Research) [81Z0200101]; Cardio-Pulmonary Institute (CPI) [EXC 2026, Project ID: 390649896]. Funding for open access charge: Goethe University funds.
Conflict of interest statement. None declared.

REFERENCES

1. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
2. Natarajan,A., Yardımcı,G.G., Sheffield,N.C., Crawford,G.E. and Ohler,U. (2012) Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res.*, **22**, 1711–1722.
3. Jayaram,N., Usvyat,D. and Martin,R.A.C. (2016) Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics*, **17**, 547.
4. Pique-Regi,R., Degner,J.F., Pai,A.A., Gaffney,D.J., Gilad,Y. and Pritchard,J.K. (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, **21**, 447–455.
5. Keilwagen,J., Posch,S. and Grau,J. (2019) Accurate prediction of cell type-specific transcription factor binding. *Genome Biol.*, **20**, 9.
6. Schmidt,F., Kern,F., Ebert,P., Baumgarten,N. and Schulz,M.H. (2018) TEPIC 2—an extended framework for transcription factor binding prediction and integrative epigenomic analysis. *Bioinformatics*, **35**, 1608–1609.
7. Nordström,K.J.V., Schmidt,F., Gasparoni,N., Salhab,A., Gasparoni,G., Kattler,K., Müller,F., Ebert,P., Costa,I.G., consortium,D. *et al.* (2019) Unique and assay specific features of NOMe-, ATAC- and DNase I-seq data. *Nucleic Acids Res.*, **47**, 10580–10596.
8. Keilwagen,J. and Grau,J. (2015) Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Res.*, **43**, e119.
9. Zhao,Y., Ruan,S., Pandey,M. and Stormo,G.D. (2012) Improved Models for Transcription Factor Binding Site Identification Using Nonindependent Interactions. *Genetics*, **191**, 781–790.
10. Mathelier,A. and Wasserman,W.W. (2013) The Next Generation of Transcription Factor Binding Site Prediction. *PLoS Comput. Biol.*, **9**, e1003214.
11. Eggeling,R., Roos,T., Myllymäki,P. and Grosse,I. (2015) Inferring intra-motif dependencies of DNA binding sites from ChIP-seq data. *BMC Bioinformatics*, **16**, 375.
12. Siebert,M. and Söding,J. (2016) Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.*, **44**, 6055–6069.
13. Fouse,S.D., Nagarajan,R.O. and Costello,J.F. (2010) Genome-scale DNA methylation analysis. *Epigenomics*, **2**, 105–117.
14. Gonzalgo,M.L., Liang,G., Spruck,C.H., Zingg,J.M., Rideout,W.M. and Jones,P.A. (1997) Identification and characterization of differentially methylated regions of genomic DNA by methylation-sensitive arbitrarily primed PCR. *Cancer Res.*, **57**, 594–599.
15. Huang,T.H., Perry,M.R. and Laux,D.E. (1999) Methylation profiling of CpG islands in human breast cancer cells. *Hum. Mol. Genet.*, **8**, 459–470.
16. Bibikova,M. and Fan,J.B. (2009) GoldenGate assay for DNA methylation profiling. *Methods Mol. Biol.*, **507**, 149–163.
17. Lister,R., Pelizzola,M., Dowen,R.H., Hawkins,R.D., Hon,G., Tonti-Filippini,J., Nery,J.R., Lee,L., Ye,Z., Ngo,Q.-M. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.
18. Cokus,S.J., Feng,S., Zhang,X., Chen,Z., Merriman,B., Haudenschild,C.D., Pradhan,S., Nelson,S.F., Pellegrini,M. and Jacobsen,S.E. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, **452**, 215–219.
19. Frommer,M., McDonald,L.E., Millar,D.S., Collis,C.M., Watt,F., Grigg,G.W., Molloy,P.L. and Paul,C.L. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 1827–1831.
20. Smith,Z.D. and Meissner,A. (2013) DNA methylation: roles in mammalian development. *Nat. Rev. Genet.*, **14**, 204–220.
21. Hu,S., Wan,J., Su,Y., Song,Q., Zeng,Y., Nguyen,H.N., Shin,J., Cox,E., Rho,H.S., Woodard,C. *et al.* (2013) DNA methylation presents distinct binding sites for human transcription factors. *Elife*, **2**, e00726.
22. O'Malley,R.C., Huang,S.C., Song,L., Lewsey,M.G., Bartlett,A., Nery,J.R., Galli,M., Gallavotti,A. and Ecker,J.R. (2016) Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*, **166**, 1598.
23. Yin,Y., Morgunova,E., Jolma,A., Kaasinen,E., Sahu,B., Khund-Sayeed,S., Das,P.K., Kivioja,T., Dave,K., Zhong,F. *et al.* (2017) Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, **356**, eaaj2239.
24. Kribelbauer,J.F., Lu,X.-J., Rohs,R., Mann,R.S. and Bussemaker,H.J. (2020) Toward a Mechanistic Understanding of DNA Methylation Readout by Transcription Factors. *J. Mol. Biol.*, **432**, 1801–1815.
25. Kribelbauer,J.F., Laptenko,O., Chen,S., Martini,G.D., Freed-Pastor,W.A., Prives,C., Mann,R.S. and Bussemaker,H.J. (2017) Quantitative analysis of the DNA methylation sensitivity of transcription factor complexes. *Cell Rep.*, **19**, 2383–2395.
26. Dantas Machado,A.C., Zhou,T., Rao,S., Goel,P., Rastogi,C., Lazarovici,A., Bussemaker,H.J. and Rohs,R. (2015) Evolving insights on how cytosine methylation affects protein-DNA binding. *Brief Funct. Genomics*, **14**, 61–73.
27. Rao,S., Chiu,T.-P., Kribelbauer,J.F., Mann,R.S., Bussemaker,H.J. and Rohs,R. (2018) Systematic prediction of DNA shape changes due to CpG methylation explains epigenetic effects on protein-DNA binding. *Epigenet. Chromatin*, **11**, 6.

28. Wan,J., Su,Y., Song,Q., Tung,B., Oyinlade,O., Liu,S., Ying,M., Ming,G.-L., Song,H., Qian,J. *et al.* (2017) Methylated cis-regulatory elements mediate KLF4-dependent gene transactivation and cell migration. *Elife*, **6**, e20068.
29. Hashimoto,H., Zhang,X., Vertino,P.M. and Cheng,X. (2015) The mechanisms of generation, recognition, and erasure of DNA 5-Methylcytosine and thymine oxidations. *J. Biol. Chem.*, **290**, 20723–20733.
30. Domcke,S., Bardet,A.F., Adrian Ginno,P., Hartl,D., Burger,L. and Schübeler,D. (2015) Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature*, **528**, 575–579.
31. Zuo,Z., Roy,B., Chang,Y.K., Granas,D. and Stormo,G.D. (2017) Measuring quantitative effects of methylation on transcription factor-DNA binding affinity. *Sci. Adv.*, **3**, eaao1799.
32. Wang,G., Luo,X., Wang,J., Wan,J., Xia,S., Zhu,H., Qian,J. and Wang,Y. (2017) MeDReaders: a database for transcription factors that bind to methylated DNA. *Nucleic Acids Res.*, **46**, D146–D151.
33. Ngo,V., Wang,M. and Wang,W. (2019) Finding de novo methylated DNA motifs. *Bioinformatics*, **35**, 3287–3293.
34. Whitaker,J.W., Chen,Z. and Wang,W. (2015) Predicting the human epigenome from DNA motifs. *Nat. Methods*, **12**, 265–272.
35. Viner,C., Ishak,C.A., Johnson,J., Walker,N.J., Shi,H., Sjöberg-Herrera,M.K., Shen,S.Y., Lardo,S.M., Adams,D.J., Ferguson-Smith,A.C. *et al.* (2023) Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet. bioRxiv doi: https://doi.org/10.1101/043794, 29 July 2022, preprint: not peer reviewed.
36. Xuan Lin,Q.X., Sian,S., An,O., Thieffry,D., Jha,S. and Benoukraf,T. (2019) MethMotif: an integrative cell specific database of transcription factor binding motifs coupled with DNA methylation profiles. *Nucleic Acids Res.*, **47**, D145–D154.
37. Ben-Gal,I., Shani,A., Gohr,A., Grau,J., Arviv,S., Shmilovici,A., Posch,S. and Grosse,I. (2005) Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics*, **21**, 2657–2666.
38. Eggeling,R., Grosse,I. and Grau,J. (2017) InMoDe: tools for learning and visualizing intra-motif dependencies of DNA binding sites. *Bioinformatics*, **33**, 580.
39. Ma,X., Kulkarni,A., Zhang,Z., Xuan,Z., Serfling,R. and Zhang,M.Q. (2011) A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information. *Nucleic Acids Res.*, **40**, e50–e50.
40. Kulakovskiy,I.V., Boeva,V.A., Favorov,A.V. and Makeev,V.J. (2010) Deep and wide digging for binding motifs in ChIP-seq data. *Bioinformatics*, **26**, 2622–2623.
41. Bailey,T.L., Williams,N., Misleh,C. and Li,W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.
42. Redhead,E. and Bailey,T.L. (2007) Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. *BMC Bioinformatics*, **8**, 385.
43. Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659.
44. Hu,M., Yu,J., Taylor,J. M.G., Chinnaiyan,A.M. and Qin,Z.S. (2010) On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res.*, **38**, 2154–2167.
45. Grau,J., Posch,S., Grosse,I. and Keilwagen,J. (2013) A general approach for discriminative de novo motif discovery from high-throughput data. *Nucleic Acids Res.*, **41**, e197.
46. Schröder,C. and Rahmann,S. (2017) A hybrid parameter estimation algorithm for beta mixtures and applications to methylation state classification. *Algorithm. Mol. Biol.*, **12**, 21.
47. Orenstein,Y. and Shamir,R. (2014) A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Res.*, **42**, e63–e63.
48. Grau,J., Grosse,I. and Keilwagen,J. (2015) PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics*, **31**, 2595–2597.
49. Keilwagen,J., Grosse,I. and Grau,J. (2014) Area under precision-recall curves for weighted and unweighted data. *PLoS One*, **9**, e92209.
50. Grau,J., Nettling,M. and Keilwagen,J. (2019) DepLogo: visualizing sequence dependencies in R. *Bioinformatics*, **35**, 4812–4814.
51. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
52. Stormo,G.D., Schneider,T.D., Gold,L. and Ehrenfeucht,A. (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res.*, **10**, 2997–3011.
53. Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–743.
54. Wittkowski,K.M. and Song,T. (2012) muStat: Prentice Rank Sum Test and McNemar Test. R package version 1.7.0.
55. Wittkowski,K.M. (1988) Friedman-type statistics and consistent multiple comparisons for unbalanced designs with missing data. *J. Am. Stat. Assoc.*, **83**, 1163–1170.
56. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C.-Y., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *et al.* (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–D115.
57. Kulakovskiy,I.V., Vorontsov,I.E., Yevshin,I.S., Sharipov,R.N., Fedorova,A.D., Rumynskiy,E.I., Medvedeva,Y.A., Magana-Mora,A., Bajic,V.B., Papatsenko,D.A. *et al.* (2017) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.*, **46**, D252–D259.
58. Pratt,H.E., Andrews,G.R., Phalke,N., Huey,J.D., Purcaro,M.J., van der Velde,A., Moore,J.E. and Weng,Z. (2021) Factorbook: an updated catalog of transcription factor motifs and candidate regulatory motif sites. *Nucleic Acids Res.*, **50**, D141–D149.
59. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K. *et al.* (2014) Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell*, **158**, 1431–1443.
60. Lea,A.J., Vockley,C.M., Johnston,R.A., Del Carpio,C.A., Barreiro,L.B., Reddy,T.E. and Tung,J. (2018) Genome-wide quantification of the effects of DNA methylation on human gene regulation. *eLife*, **7**, e37513.
61. Mann,I.K., Chatterjee,R., Zhao,J., He,X., Weirauch,M.T., Hughes,T.R. and Vinson,C. (2013) CG methylated microarrays identify a novel methylated sequence bound by the CEBPB—ATF4 heterodimer that is active in vivo. *Genome Res.*, **23**, 988–997.
62. Suzuki,T., Maeda,S., Furuhata,E., Shimizu,Y., Nishimura,H., Kishima,M. and Suzuki,H. (2017) A screening system to identify transcription factors that induce binding site-directed DNA demethylation. *Epigenet. Chromatin*, **10**, 60.
63. Stadler,M.B., Murr,R., Burger,L., Ivanek,R., Lienert,F., Schöler,A., Nimwegen,E.V., Wirbelauer,C., Oakeley,E.J., Gaidatzis,D. *et al.* (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, **480**, 490–495.
64. Feldmann,A., Ivanek,R., Murr,R., Gaidatzis,D., Burger,L. and Schübeler,D. (2013) Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. *PLoS Genet.*, **9**, e1003994.
65. Maurano,M.T., Wang,H., John,S., Shafer,A., Canfield,T., Lee,K. and Stamatoyannopoulos,J.A. (2015) Role of DNA methylation in modulating transcription factor occupancy. *Cell Rep.*, **12**, 1184–1195.
66. Paull,T.T., Cortez,D., Bowers,B., Elledge,S.J. and Gellert,M. (2001) Direct DNA binding by Brca1. *Proc. Natl. Acad. Sci.*, **98**, 6086–6091.
67. Wang,J., Zhuang,J., Iyer,S., Lin,X., Whitfield,T.W., Greven,M.C., Pierce,B.G., Dong,X., Kundaje,A., Cheng,Y. *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798–1812.
68. Kim,J.D., Hinz,A.K., Choo,J.H., Stubbs,L. and Kim,J. (2007) YY1 as a controlling factor for the Peg3 and Gnas imprinted domains. *Genomics*, **89**, 262–269.
69. Harrington,M.A., Jones,P.A., Imagawa,M. and Karin,M. (1988) Cytosine methylation does not affect binding of transcription factor Sp1. *Proc. Natl. Acad. Sci.*, **85**, 2066–2070.
70. Tian,H.-P., Lun,S.-M., Huang,H.-J., He,R., Kong,P.-Z., Wang,Q.-S., Li,X.-Q. and Feng,Y.-M. (2015) DNA Methylation Affects the SP1-regulated Transcription of FOXF2 in Breast Cancer Cells. *J. Biol. Chem.*, **290**, 19173–19183.

71. Höller,M., Westin,G., Jiricny,J. and Schaffner,W. (1988) Sp1 transcription factor binds DNA and activates transcription even when the binding site is CpG methylated. *Gene. Dev.*, **2**, 1127–1135.
72. Prokhortchouk,A., Hendrich,B., Jørgensen,H., Ruzov,A., Wilm,M., Georgiev,G., Bird,A. and Prokhortchouk,E. (2001) The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. *Gene. Dev.*, **15**, 1613–1618.
73. Daniel,J.M., Spring,C.M., Crawford,H.C., Reynolds,A.B. and Baig,A. (2002) The p120 ctn-binding partner Kaiso is a bi-modal DNA-binding protein that recognizes both a sequence-specific consensus and methylated CpG dinucleotides. *Nucleic Acids Res.*, **30**, 2911–2919.
74. Hannenhalli,S. and Kaestner,K.H. (2009) The evolution of Fox genes and their role in development and disease. *Nat. Rev. Genet.*, **10**, 233–240.
75. Ji,Z., Donaldson,I.J., Liu,J., Hayes,A., Zeef,L. A.H. and Sharrocks,A.D. (2012) The forkhead transcription factor FOXK2 promotes AP-1-mediated transcriptional regulation. *Mol. Cell. Biol.*, **32**, 385–398.
76. Cernilogar,F.M., Hasenöder,S., Wang,Z., Scheibner,K., Burtscher,I., Sterr,M., Smialowski,P., Groh,S., Evenroed,I.M., Gilfillan,G.D. *et al.* (2019) Pre-marked chromatin and transcription factor co-binding shape the pioneering activity of Foxa2. *Nucleic Acids Res.*, **47**, 9069–9086.
77. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
78. Gupta,S., Stamatoyannopoulos,J.A., Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.
79. Héberlé,É. and Bardet,A.F. (2019) Sensitivity of transcription factors to DNA methylation. *Essays Biochem.*, **63**, 727–741.
80. Lazarovici,A., Zhou,T., Shafer,A., Dantas Machado,A.C., Riley,T.R., Sandstrom,R., Sabo,P.J., Lu,Y., Rohs,R., Stamatoyannopoulos,J.A. *et al.* (2013) Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 6376–6381.
81. Banovich,N.E., Lan,X., McVicker,G., van de Geijn,B., Degner,J.F., Blischak,J.D., Roux,J., Pritchard,J.K. and Gilad,Y. (2014) Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet.*, **10**, e1004663.
82. Xiong,Z., Li,M., Yang,F., Ma,Y., Sang,J., Li,R., Li,Z., Zhang,Z. and Bao,Y. (2020) EWAS Data Hub: a resource of DNA methylation array data and metadata. *Nucleic Acids Res.*, **48**, D890–D895.
83. Karemaker,I.D. and Vermeulen,M. (2018) Single-Cell DNA methylation profiling: technologies and biological applications. *Trends Biotechnol.*, **36**, 952–965.
84. Buenrostro,J.D., Wu,B., Litzenburger,U.M., Ruff,D., Gonzales,M.L., Snyder,M.P., Chang,H.Y. and Greenleaf,W.J. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, **523**, 486–490.
85. Sérandour,A.A., Avner,S., Percevault,F., Demay,F., Bizot,M., Lucchetti-Miganeh,C., Barloy-Hubler,F., Brown,M., Lupien,M., Métivier,R. *et al.* (2011) Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers. *Genome Res.*, **21**, 555–565.
86. Park,Y., Lee,J.M., Hwang,M.Y., Son,G.H. and Geum,D. (2013) NonO binds to the CpG island of oct4 promoter and functions as a transcriptional activator of oct4 gene expression. *Mol. Cells*, **35**, 61–69.