Published online 31 October 2018 *Nucleic Acids Research, 2019, Vol. 47, Database issue* ***D145–D154***
doi: 10.1093/nar/gky1005

MethMotif: an integrative cell specific database of transcription factor binding motifs coupled with DNA methylation profiles

Quy Xiao Xuan Lin[1], Stephanie Sian[1], Omer An[1], Denis Thieffry[2], Sudhakar Jha[1,3] and Touati Benoukraf[1,4,*]

[1]Cancer Science Institute of Singapore, National University of Singapore, Singapore, Singapore, [2]Computational Systems Biology Team, Institut de Biologie de l'École Normale Supérieure (IBENS), INSERM, École Normale Supérieure, PSL Research University, Paris, France, [3]Department of Biochemistry, National University of Singapore, Singapore, Singapore and [4]Discipline of Genetics, Faculty of Medicine, Memorial University of Newfoundland, St. John's, NL, Canada

Received August 15, 2018; Revised October 04, 2018; Editorial Decision October 09, 2018; Accepted October 10, 2018

ABSTRACT

Several recent studies have portrayed DNA methylation as a new player in the recruitment of transcription factors (TF) within chromatin, highlighting a need to connect TF binding sites (TFBS) with their respective DNA methylation profiles. However, current TFBS databases are restricted to DNA binding motif sequences. Here, we present MethMotif, a two-dimensional TFBS database that records TFBS position weight matrices along with cell type specific CpG methylation information computed from a combination of ChIP-seq and whole genome bisulfite sequencing datasets. Integrating TFBS motifs with TFBS DNA methylation better portrays the features of DNA loci recognised by TFs. In particular, we found that DNA methylation patterns within TFBS can be cell specific (e.g. MAFF). Furthermore, for a given TF, different DNA methylation profiles are associated with different DNA binding motifs (e.g. REST). To date, MethMotif database records over 500 TFBSs computed from over 2000 ChIP-seq datasets in 11 different cell types. MethMotif portal is accessible through an open source web interface (https://bioinfo-csi.nus.edu.sg/methmotif) that allows users to intuitively explore the entire dataset and perform both single, and batch queries.

INTRODUCTION

Nowadays, DNA methylation is recognised as a key component in numerous biological mechanisms such as gene regulation, RNA splicing, nucleosome positioning and transcription factors (TF) binding dynamics (1). Until recently, it has been well accepted that the writing of a methyl group into a cytosine can prevent TF binding events. Indeed, methylation within the binding site of the insulator-binding protein CTCF abrogates its DNA binding ability, resulting in imprinted expression of H19 and Igf2 (2). Similarly, AP-1 and SP1 interactions with DNA can be inhibited by methylating the CpGs adjacent to their respective DNA binding sites (3, 4). Nonetheless, this rule is not universal. Accumulating evidence during the last few years has shown that numerous TFs require methylated DNA to bind to their motifs, whereas other TFs bind to DNA regardless the methylation status (5–9). For example, Hu *et* al. characterised 41 methylated-DNA binding proteins among 1321 tested TFs using a protein microarray-based approach (8). More recently, by using methylation-sensitive systematic evolution of ligands by exponential enrichment (SELEX), Yin *et* al. segregated 519 TFs into five groups based on their binding strength to methylated DNA sequences (9). The first group includes 175 TFs whose binding activities are enhanced by methylated DNA; the second group contains 117 TFs whose binding abilities are compromised by DNA methylation; the third group is composed of 33 TFs that bind to DNA regardless the methylation status; the fourth group encompasses 25 TFs whose binding activities are differently affected by DNA methylation depending on different sub-motifs or different CpG sites in the same motif; and finally, the last group of TFs (169) does not contain any CpG site within their motifs. These systematic *in vitro* studies have unveiled the disparate mechanisms involved in TFs/DNA interactions with DNA methylation being a key leverage in this process. However, it has been shown that only a small fraction of TF binding sites (TFBS) found *in vitro* can be validated *in vivo* (10), due to a more complex and crowded nuclear environment involving the chromatin structure and non-DNA molecules such as RNAs, histones

*To whom correspondence should be addressed. Tel: +65 6516 1025; Fax: +65 6873 9664; Email: benoukraf@nus.edu.sg

and non-histone proteins (11). Therefore, further efforts are required to characterise the methylation status of TFBS in the cell context.

Most of the current TFBS databases such as JASPAR (12), HOCOMOCO (13) and GTRD (14) provide TFBS motifs derived from *in vivo* ChIP-seq datasets. Nevertheless, none of these databases have combined TFBS motifs with a precise depiction of their respective DNA methylation profiles yet. This is the reason why we have developed MethMotif (Methylation in Motif), a TFBS database that combines both position weight matrices (PWM), and DNA methylation profiles, in a cell specific manner.

MATERIALS AND METHODS

Data collection and pre-processing

Raw ChIP-seq datasets (fastq files) in different cell types were downloaded from ENCODE consortium and GEO databases (Figure 1A, Table 1). After read quality checking with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and read trimming when required, using Trimmomatic (15), short reads were aligned to the human genome assembly hg38, using STAR (16) with the spliced function blocked, for its speedy and accurate aligning performance (BioRxiv: https://doi.org/10.1101/053686). Relying upon the Irreproducibility Discovery Rate (IDR) pipeline (17), TF genomic occupancy regions with high consistency in replicates were called by MACS2 (18) using default parameters.

We took advantage of Whole Genome Bisulfite Sequencing (WGBS) datasets to measure the genome-wide DNA methylation levels in each cell type at the CpG resolution. WGBS datasets were obtained from ENCODE consortium and GEO databases (Table 2). Raw datasets underwent read quality check using FastQC, as well as quality and adapter trimming using Trim Galore. High quality short reads were aligned to hg38 using Bowtie2 with Bismark (v0.16.3) and default parameters (19,20). Methylation states in CpG context were extracted with the Bismark methylation extractor module. Whenever their correlation coefficients were confirmed over 0.8, as recommended by ENCODE standards, WGBS datasets with biological replicates were merged to increase the sequencing coverage using methylKit (21).

In house WGBS data generation

We complemented the existing published datasets from ENCODE and GEO by generating a WGBS of the HCT116 cell line (human colorectal carcinoma). HCT116 cells were grown in RPMI 1640 medium with 10% FBS. The DNA was extracted using Qiagen DNeasy Blood & Tissue Kits. After sample DNA testing, negative control DNAs were added into the initial DNA pool, and then fragmented into a 200–300 bp range using Covaris S220. Next, terminal repairing, A-ligation, methylation sequencing adapters ligation were performed on DNA fragments. Then, DNA fragments underwent a bisulfite treatment using the EZ DNA Methylation Gold Kit from Zymo Research. Library concentration was firstly quantified by Qubit 2.0, before a dilution to 1 ng/ul followed by insert size verification using the Agilent Bioanalyzer 2100 and qPCR. Finally, the library was paired-end sequenced using the Illumina HiSeq 4000. Raw and processed data are available through the GEO portal under the accession number GSE118030. We will use this GEO SuperSeries ID for sharing our future in-house datasets.

Profiling the DNA methylation landscapes surrounding ENCODE ChIP-ed protein binding regions

Integrative analyses of ChIP-seq and WGBS datasets were performed to profile DNA methylation landscapes surrounding the genomic occupancy regions across all ENCODE ChIP-ed proteins. For each DNA binding protein, methylation scores (beta scores) (22) of CpGs within ±100 bp surrounding peak summits were collected using the intersectBed module of bedtools (23). For each cell type, the distributions of the collected CpG methylation scores across all ChIP-ed proteins were shown in a heatmap, and proteins were clustered into different groups based on their DNA methylation landscapes surrounding their binding regions, using hierarchical clustering with the Euclidean distance. This representation allows us to classify ENCODE ChIP-ed proteins according to the DNA methylation levels of their respective binding loci (Figure 1B).

Characterisation of DNA methylation levels within TFBS

We employed the MEME-ChIP package with the default parameters to identify TF motifs in ±100 bp surrounding TF peak summits (24). Since the actual binding sites for the ChIP-ed TFs are prone to central enrichment and maximum probability in occurrence at peak centres (25), each TF motif was determined based on its centrally enriched propensity profiled by CentriMo from the MEME-ChIP package. Then, the exact TF binding sites were localised using the FIMO module, while the methylation states of CpGs within the binding motif were assessed based on the WGBS dataset using the intersectBed module from bedtools (23).

In order to intuitively exhibit the DNA methylation levels within all TFBSs, we adopted a novel MethMotif logo combining the classical motif logo commonly used, with a cumulative bar chart describing the methylation level of each CpG present in the motif (Figure 1C, D). CpG methylation scores were segregated into three interval groups: (i) methylation scores <10% (i.e. homogenously hypomethylated); (ii) methylation scores >90% (i.e. homogenously hypermethylated) and (iii) methylation scores ranging from 10% to 90% (i.e. heterogeneously methylated). We utilised WebLogo3 (26) to generate the motif logo, while the bar chart above the motif logo describing the methylation score was generated using a custom R script. While the usage of stringent methylation score thresholds prevents mis-scoring DNA methylation in TFBS due to cell heterogeneity, a more direct genome-wide investigation by sequential ChIP-bisulfite sequencing (ChIP-BS-seq) (27) will be necessary to validate these methyl-TFBS characteristics.

Web interface

A web-based interface has been established at https://bioinfo-csi.nus.edu.sg/methmotif allowing an intuitive exploration of our datasets and facilitating queries to the

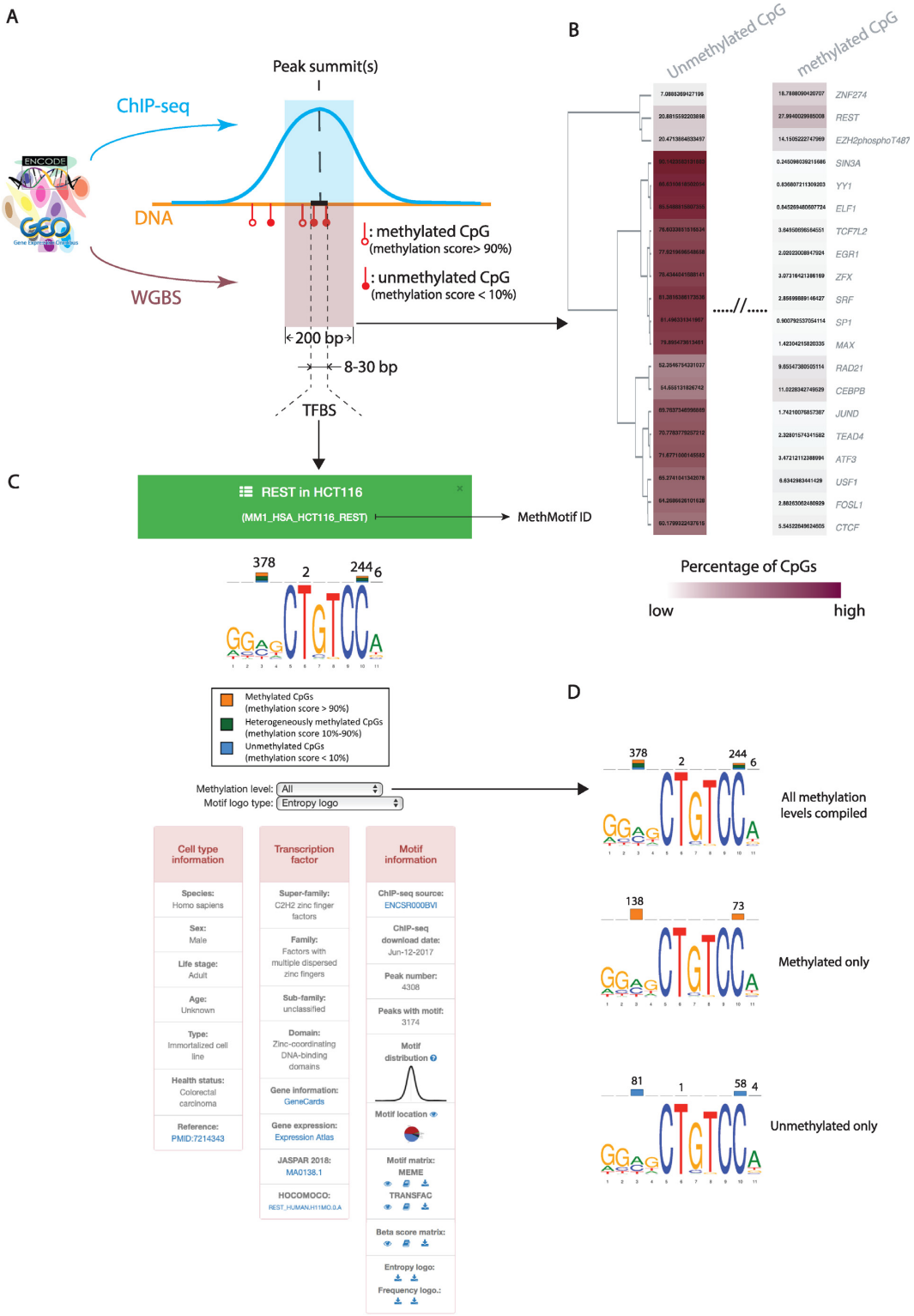

Figure 1. Construction of MethMotif database. Integrative analysis of ChIP-seq and WGBS datasets, downloaded from ENCODE and GEO databases, allows the profiling of the DNA methylation landscapes surrounding transcription factor binding regions in various cell types (**A**). Firstly, the methylation levels (beta scores or methylation scores) of CpGs located within all ChIP-ed protein peak regions are captured over the 200 bp surrounding peak summits. The distribution of the corresponding CpG methylation scores is then profiled in a heatmap for each cell type, to present the DNA methylation levels surrounding peak regions across all ChIP-ed proteins (methylation score less than 10% is defined as homogenously unmethylated, while methylation score more than 90% is regarded as homogenously hypermethylated). These heatmaps are accessible from the 'Explore' section of the MethMotif website (**B**). Finally, the direct binding motifs of sequence-specific TFs are identified. The DNA methylation level within TFBS is captured and shown in a MethMotif logo. MethMotif logos are collected in 'MethMotif database' section of the MethMotif website (**C**). The methylation levels within each logo can be displayed according to three states: (i) all methylation levels compiled, (ii) methylated only and, (iii) unmethylated only (**D**).

Table 1. ChIP-seq datasets used in MethMotif database

Cell ID	Organism	Cell type/tissue	Number of ChIP-seq experiments	Source	Download date
HeLa-S3	Human	Cervix	57	ENCODE	20 August 2017
HEK293	Human	Kidney	199	ENCODE	12 March 2018
IMR-90	Human	Lung	15	ENCODE	26 March 2018
SK-N-SH	Human	Brain	30	ENCODE	26 March 2018
A549	Human	Lung	42	ENCODE	26 March 2018
K562	Human	Blood	279	ENCODE	2 April 2018
HepG2	Human	Liver	138	137-ENCODE, 1-GEO	7 April 2018
GM12878	Human	Blood	143	ENCODE	12 April 2018
MCF-7	Human	Breast	99	ENCODE	13 April 2018
H1-hESC	Human	Stem cell	65	ENCODE	16 April 2018
HCT116	Human	Colorectum	22	ENCODE	16 April 2018

Table 2. WGBS datasets used in MethMotif database

Cell ID	Organism	Cell type/tissue	Source ID	Release date
HeLa-S3	Human	Cervix	GSM2175341	30 January 2017
HEK293	Human	Kidney	GSM1254259	11 December 2015
IMR-90	Human	Lung	ENCSR888FON	31 July 2013
SK-N-SH	Human	Brain	ENCSR145HNT	13 December 2017
A549	Human	Lung	ENCSR481JIW	4 December 2017
K562	Human	Blood	ENCSR765JPC	22 March 2016
HepG2	Human	Liver	ENCSR881XOU	13 October 2015
GM12878	Human	Blood	ENCSR890UQO	23 February 2016
MCF-7	Human	Breast	GSM1328112	3 July 2014
H1-hESC	Human	Stem cell	ENCSR617FKV	13 October 2015
HCT116	Human	Colorectum	GSM3317488	10 August 2018

database. Users can freely access MethMotif *via* these three modes:

1. '*Motif database direct query*'. Like many TFBS webservers, MethMotif includes a search engine that allows users to query any DNA binding protein present in our database using either its official gene name or its MethMotif ID (see an example of MethMotif ID in Figure 1C). Queries can be refined by selecting a given cell line or tissue of interest. Query results include the MethMotif logo, information about the cell type, ChIP-seq and WGBS assays, as well as motif distributions, motif locations and downloadable processed files (PWM and methylation matrices), along with links to JASPAR 2018 and HOCOMOCO (Figure 1C and D).
2. '*Explore*'. Users can intuitively explore all DNA binding proteins available in the released cell types in our database *via* dynamic heatmaps that classify each protein according to its CpG methylation pattern in a ±100 bp window surrounding ChIP-seq peak summits. Details about any DNA binding protein of interest can be easily obtained by hovering over the corresponding heatmap's row (Figure 1B). To date, we have made available 11 dynamic heatmaps corresponding to the 11 analysed cell types. The heatmap dynamic interface takes advantage of InCHlib, an open source interactive JavaScript library (28).
3. '*Batch query*'. MethMotif allows users to analyse the occurrences of TFBSs along with their respective methylation states across a given list of genome loci. This functionality is particularly useful to characterise binding sites for co-factors of a protein of interest, along with their respective DNA methylation levels. Here, MethMotif generates *de novo* DNA binding motifs based only on the sequences that overlap with the input genomic regions. TFBS methylation profiles are represented as beeswarm boxplots, where each dot represents the methylation level of a CpG from a given binding site in the genome (Figure 2).

Batch query module

The functionality of batch query module implemented in MethMotif website is achieved by mapping the given list of genome loci to all TFBS genomic coordinates, together with their respective CpG methylation scores, in MethMotif database using in-house python scripts. The visible results, namely beeswarm boxplots and *de novo* generated MethMotif logo, are produced using WebLogo3 (26) and in-house R scripts with the grImport package (29).

Database maintenance

Established ChIP-seq working standards and guidelines in ENCODE consortium foster reliable ChIP-seq data (17). Since the beginning of the project, MethMotif database has been regularly updated with ChIP-seq datasets from ENCODE. We plan to further enrich the MethMotif database by expanding cell types using newly generated WGBS datasets from ENCODE, GEO and in-house experiments. Furthermore, we intend to take advantage of the Gene Transcription Regulation Database (GTRD), which is a comprehensive compendium of ChIP-seq experiments covering the existing multiple ChIP-seq databases (14), to expand the coverage of TFs in MethMotif.

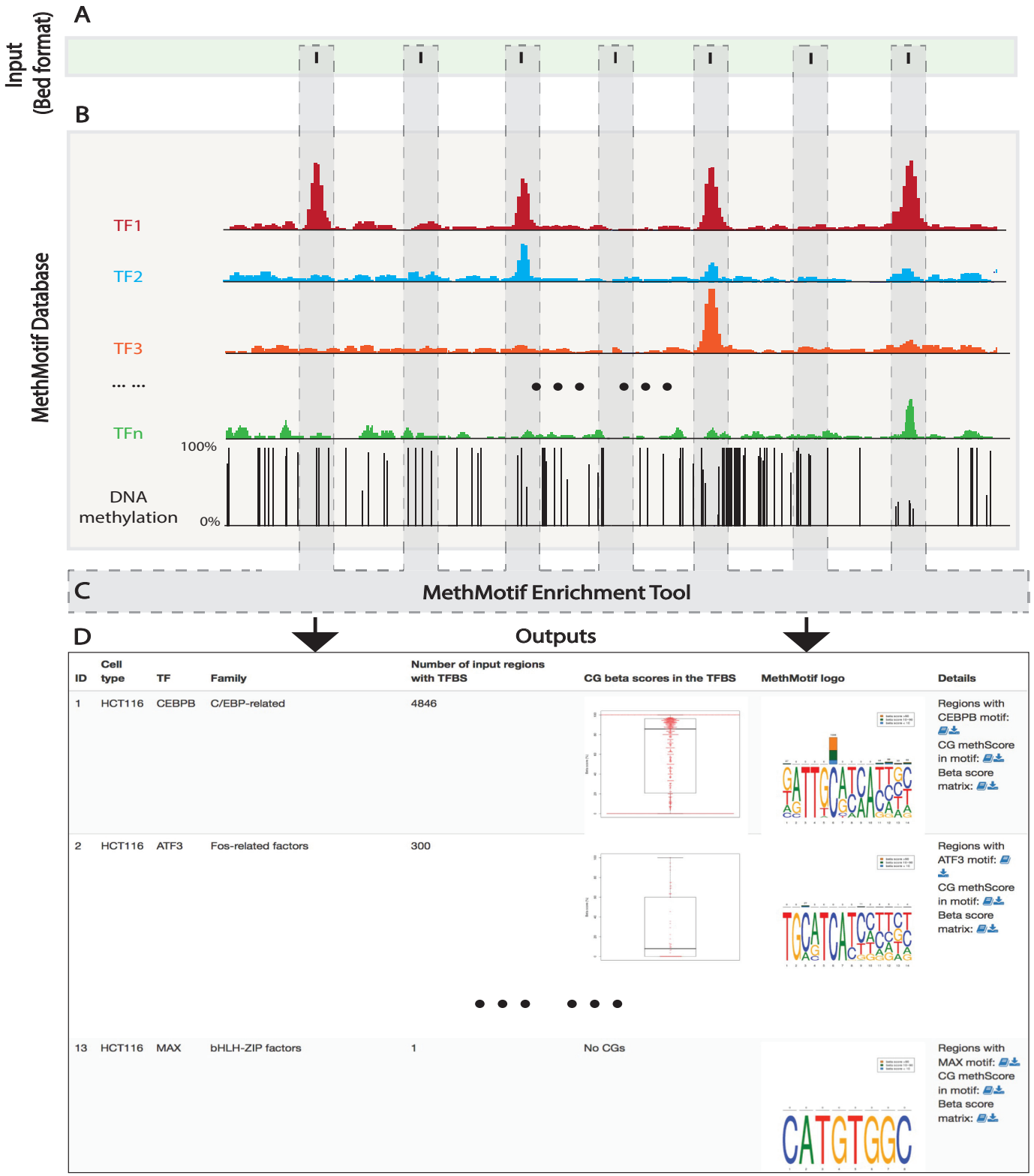

Figure 2. Workflow of MethMotif Batch Query. MethMotif Batch Query is available via the MethMotif website, which allows users to study the occurrences of TFBSs along with DNA methylation states in a given list of genomic loci. Users can upload the coordinates of the regions of their interest in BED format (**A**). MethMotif database will be queried (**B**) and the presence of TFBSs together with respective DNA methylation levels is then analysed using the MethMotif Enrichment Tool (**C**) in the given regions. If loci from the input regions overlap with any TFBS present in the MethMotif database, the Batch Query tool will output these overlapped loci along with their respective TFBS DNA methylation information *via* a beeswarm boxplot (where each dot represents the methylation level of a CpG site within the TFBS) and a MethMotif logo generated *de novo* (**D**).

RESULTS

MethMotif classifies DNA-interacting proteins according to DNA methylation profiles

The integration of WGBS with ChIP-seq datasets allowed us to profile the DNA methylation landscapes surrounding binding loci of DNA-bound proteins *in vivo*, at the genome scale. We employed this strategy to capture the CpG methylation patterns in a ±100 bp window surrounding the peak summits of ENCODE ChIP-ed proteins with high reproducibility in different cell types (Figure 1A). As described previously (9), DNA-binding proteins exhibit distinct DNA methylation levels around their occupancy regions and accordingly can be classified into different sub-groups (Figure 1B, Supplementary Figure S1). In myelogenous leukaemia K562 cells, we were able to classify 250 ChIP-ed proteins based on DNA methylation landscapes around binding regions into three groups (I, high; II, medium and III, low methylated DNA binding affinity) and one outlier (very high methylated binding affinity — ZBTB33) (Supplementary Figure S1). Notably, our ZBTB33 epigenetic binding profile corroborates previous *in vivo* studies performed with a limited number of loci (30,31), and generalises the ability of ZBTB33 to bind methylated DNA at the genome scale. Furthermore, group III includes TFs known to bind specifically to non-methylated DNA such as RUNX1 (32) and SP1 (4). As shown in the MethMotif website, this distinct binding propensity to methyl-CpGs among different DNA-binding proteins is not a unique observation in K562, but a widespread phenomenon across different cell types.

MethMotif introduces a new dimension to Position Weight Matrix (PWM)

A key feature of TFs is their ability to bind DNA by the recognition of a specific DNA motif. However, ChIP-seq peak regions are not necessarily associated with direct binding sites (25) and DNA methylation profiling around all aforementioned peaks may not reflect the influences of DNA methylation on the recruitment of TFs onto DNA sequences. Therefore, we inferred TFBSs by selecting genomic loci that contained a centrally enriched DNA motif surrounding ChIP-seq centres (25). TFBS was characterised using the MEME-ChIP package (24) in ±100 bp regions surrounding peak centres discovered by the IDR pipeline. This strategy allowed us to capture the DNA methylation landscape within each TFBS motif. In order to intuitively illustrate TFBS coupled with DNA methylation, we introduced a novel graphical representation combining a motif-logo (26) with cumulative bar charts displaying three different DNA methylation levels: (i) homogenously low (less than 10% sequenced CpGs of a given site are methylated); (ii) homogenously high (more than 90% of sequenced CpGs of a given site are methylated) and (iii) heterogeneously methylated (10–90% sequenced CpGs of a given site are methylated) (Figure 1C and D). The MethMotif logo represents these three DNA methylation levels as a cumulative bar chart where the blue, orange and green colours denote the proportions of CpGs falling in the group 1, 2 and 3, respectively. Additionally, we display the number of total CpGs covered by WGBS on the top of each bar. We believe that adding an epigenetic dimension to the sequence TFBS information will improve our understanding of TF binding dynamics.

Numerous TFs exhibit distinct binding site sequences and/or distinct binding site methylation profiles across different cell types

In contrast to pre-existing TFBS databases, MethMotif records TFBSs computed in a cell specific manner. Our meta-analysis shows that numerous TFs exhibit either distinct TFBSs, or distinct DNA methylation profiles, or both, depending on cell types. Notably, when comparing ChIP-seq performed in K562 and GM12878 cells, we found that the basic leucine zipper factor NFE2 has the ability to bind distinct DNA motifs (Figure 3A). Indeed, in K562, NFE2 recognises an extended AP-1-like motif, (T/C/A)GCTGA(C/G)TCA(T/C), while the motif identified in GM12878 is similar to the USF binding site (Supplementary Figure S2A). Interestingly, both motifs are supported by the literature. Specifically, NFE2 has been shown to bind an AP-1-like motif as a heterodimer with MAF family members, such as MAFF and MAFK (33); and, in another context, NFE2 was characterised to belong to a complex including USF (34). We were able to corroborate these facts by using the MethMotif batch query module with all NFE2 binding sites in K562 and GM12878 cells (Supplementary Figure S2B).

In addition, we found that the length of spacers of leucine zipper dimers, such as JUN, varies with cell types. For example, in hepatocyte carcinoma HepG2 and embryonic stem cell line H1-hESC, the main motif spacer of JUN is a CG dinucleotide, while it amounts to only a single base, either C or G, in breast cancer MCF-7, lung adenocarcinoma A549 and myelogenous leukaemia K562 cells (Figure 3B). Interestingly, the UniProbe TFBS database, which records *in vitro* characterised DNA binding sites, contains another AP-1 leucine zipper TF, JUNDM2 (Jun dimerisation protein 2), that recognises motifs similar to JUN with a spacer that can vary from 1 to 2 nucleotides (UniProbe accession number: UP00103) (35). Such variations in spacers' size can be explained by different dimerisations of JUN or JUNDM2 with other TFs (36), or by post-translational modifications and modified secondary structure.

We further noticed that some TFs have a conserved DNA motif with distinct DNA methylation profiles across cell types. This is the case for the leucine zipper MAFF. In K562, CpGs sites composing the MAFF motif are prone to hypomethylation, while in HeLa-S3 and HepG2, we observed a significant number of motifs with methylated CpGs (Figure 3C). To the best of our knowledge, MAFF was not previously described to play different roles across these cell lines. However, ChIP-seq results clearly show that MAFF targets distinct loci, and consequently regulates distinct biological functions (Supplementary Figure S3).

Finally, a group of TFs such as the zinc finger REST revealed distinct binding profiles at the DNA sequence and DNA methylation levels. Remarkably, in this specific example, changes in DNA methylation levels correlate with changes in DNA sequence. Indeed, REST is known to bind a sequence of 6 and 7 conserved nucleotides linked

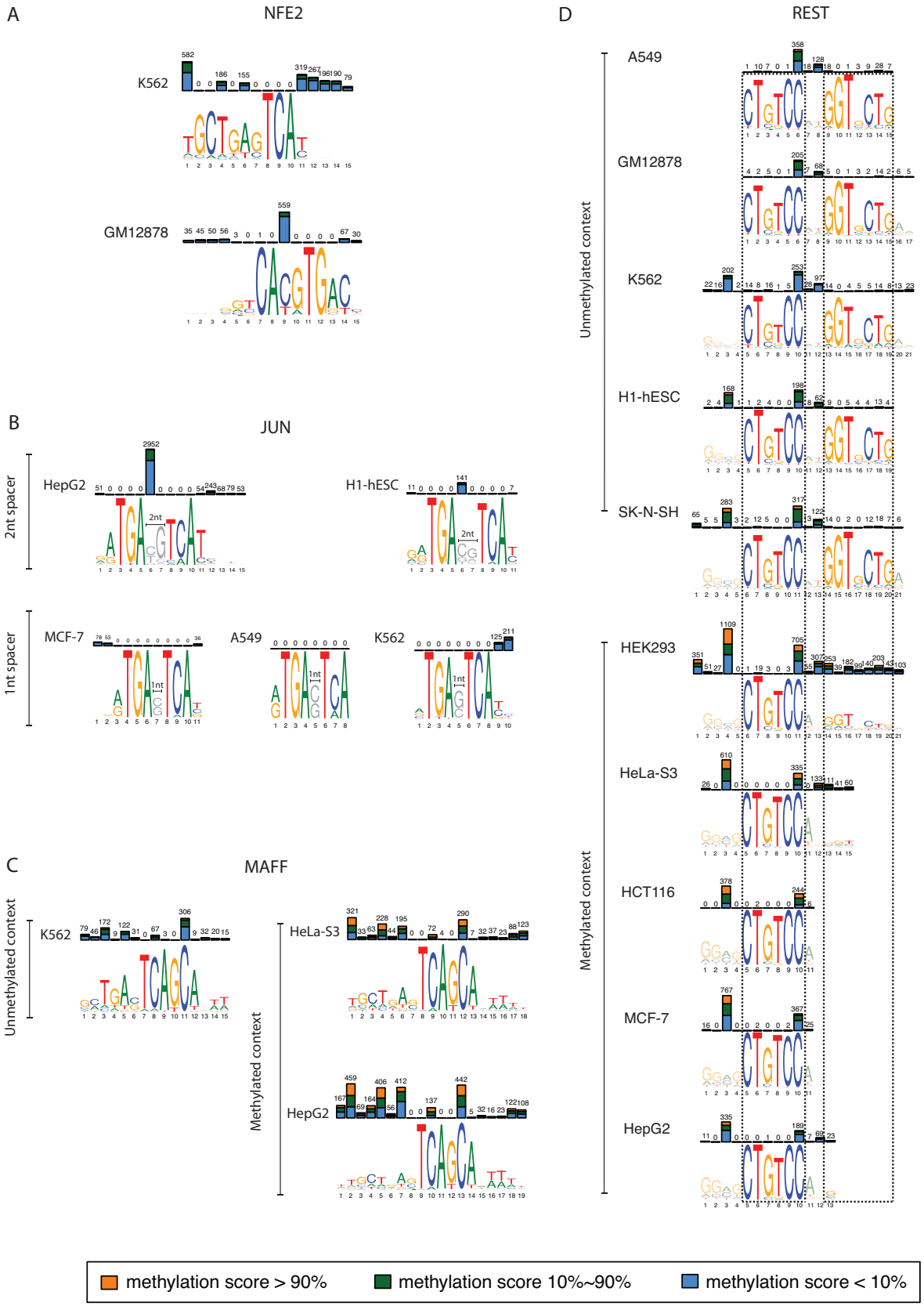

Figure 3. Distinct motifs and DNA methylation levels in motifs across cell types. Examples of MethMotif logos from the MethMotif database show that TF binding propensities can be altered in terms of motifs (**A** and **B**), DNA methylation levels inside the motifs (**C**), or even both (**D**), across different cell types.

by a 2-nucleotide spacer: CTGTCCNNGGTGCTG (cf. JASPAR accession number: MA0138.1; HOCOMOCO accession number: REST_HUMAN.H11MO.0.A). Although this motif is confirmed in A549, GM12878, K562 and H1-hESC cells, where REST binds unmethylated DNA, we observed that REST requires only the first part of its conserved motif (CTGTCC) to bind methylated DNA, as shown in HEK293, HeLa-S3, HCT116, MCF-7 and HepG2 cells (Figure 3D). It is interesting to note that, in another context, non-CpG methylation facilitates REST binding events (37).

MethMotif batch query: a case study

MethMotif portal includes a batch query tool which allows users to study the occurrences of TFBSs and their respective DNA methylation states within a given list of genomic loci. As a proof of concept, we ran MethMotif batch query using a list of CEBPB binding sites characterised by a ChIP-seq experiment in HCT116 cells (Supplementary Data 1). As expected, the most enriched TFBS found by MethMotif is CEBPB (4846 overlapped occurrences, Figure 2D and Supplementary Figure S4). Furthermore, MethMotif batch query brought to light 12 co-factors that overlap our input list from 1 to 300 co-occurrences (Supplementary Figure S4). Results include known CEBPB co-factors such as ATF3 (300 co-occurrences) (38) and JUND (74 co-occurrences) (39). Interestingly, only CEBPB binding loci harbour methylated CpG sites (Supplementary Figure S4). This observation is in line with its pioneer factor function (40). Indeed, here we can hypothesise that CEBPB is firstly recruited within methylated DNA across repressive chromatin, and then, once the DNA is demethylated and more accessible, co-factors are recruited.

DISCUSSION

In contrast to the previously widely accepted assumption, DNA methylation is not necessarily associated with the inhibition of TF binding activities (6,8,9). Interestingly, in some contexts, the recruitment of methyl-CpG specific binding TF can be abrogated by erasing DNA methylation using drugs such as 5-aza-2′-deoxycytidine (5-aza-dC) (41). Therefore, DNA methylation becomes a key event for the recruitment of certain TFs. As DNA methylation is also portrayed as an epigenetic mark of cellular memory (42), we can speculate that this memory requires the recruitment and/or interaction of specific factors that either inhibit or prime the chromatin. This is the reason why the investigation of methyl-TFBS has to be done *in vivo* and in a cell specific manner. MethMotif brought to light that, for given factors, TFBSs can vary according to the DNA methylation patterns.

The DNA binding ability regardless of DNA methylation is a reminiscence of pioneer factors, which form a specific class of TFs required and sufficient to trigger enhancer competency (43). MethMotif is able to detect this capacity for some pioneer factors such as CEBPA (44) and CEBPB (40). Moreover, numerous zinc finger proteins have been characterised with methyl-DNA recognition ability (45,46). With MethMotif, we can systematically reveal that occupancy at methylated cytosine by zinc finger proteins is not a sporadic event but a widespread phenomenon. Interestingly, the position weight matrices encompassing methylated CpGs are inclined to be specifically enriched in TpG dinucleotide at the same position, corroborating the fact that a methylated cytosine base structurally resembles a thymine (47) (Supplement Figure S5).

Identification of methyl-cytosine binding TFs by MethMotif is the initial but crucial step to study the crosstalk between TFs and DNA methylation. While we provide compelling evidence that numerous TFs bind differentially to methylated and non-methylated DNA, ChIP-BS-seq experiments (27) will be required to confirm these interactions. Nonetheless, the current version of MethMotif remains a valuable resource for guiding researchers in the usage of ChIP-BS-seq as an alternative to ChIP-seq for studying methyl-binding TFs.

In conclusion, MethMotif complements current TFBS databases by integrating DNA methylation as a new epigenetic dimension with TFBS, thereby providing a useful resource for the study of DNA methylation and TF recruitment dynamics.

DATA AVAILABILITY

MethMotif is an academic and open source database under GPL licence and available at https://bioinfo-csi.nus.edu.sg/methmotif.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Morgane Thomas-Chollier and Daniel G. Tenen for useful discussions and comments during the preparation of this manuscript, as well as Mathanapriya Naidu for her proof-reading.
Authors contributions: Q.X.X.L. and T.B. designed the platform; Q.X.X.L. programmed the website and performed all analysis; O.A. participated in the graphical user interface implementation; S.S. generated the HCT116 WGBS dataset; Q.X.X.L., D.T., S.J. and T.B. interpreted the data; Q.X.X.L. and T.B. wrote the manuscript. T.B. directed the study.

FUNDING

Work in the T.B. laboratory is supported by the National Research Foundation, the Singapore Ministry of Education under its Centres of Excellence initiative, the National Medical Research Council of Singapore [NMRC/BNIG/2035/2015]; RNA Biology Center at the Cancer Science Institute of Singapore, NUS, as part of funding under the Singapore Ministry of Education's AcRF Tier 3 grants [MOE2014-T3-1-006]. T.B. and D.T. collaboration is supported by the Institut Français à Singapour [Merlion Project grant number 6.10.14]. S.J. was supported by grants from the National Research Foundation Singapore and the Singapore Ministry of Education under

its Research Centers of Excellence initiative to the Cancer Science Institute of Singapore [R-713-006-014-271]; National Medical Research Council (NMRC) [CBRG-NIG BNIG11nov001]; Ministry of Education Academic Research Fund [MOE AcRF Tier 1 T1-2012 Oct-04 and T1-2016 Apr-01]; RNA Biology Center at CSI Singapore, NUS, from funding by the Singapore Ministry of Education's AcRF Tier 3 grants [MOE2014-T3-1-006]. Funding for open access charge: National Research Foundation, the Singapore Ministry of Education under its Centres of Excellence initiative and the Singapore Ministry of Education Academic Research Fund Tier 3 [MOE2014-T3-1-006].
Conflict of interest statement. None declared.

REFERENCES

1. Tirado-Magallanes,R., Rebbani,K., Lim,R., Pradhan,S. and Benoukraf,T. (2017) Whole genome DNA methylation: beyond genes silencing. *Oncotarget*, **8**, 5629–5637.
2. Bell,A.C. and Felsenfeld,G. (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, **405**, 482–485.
3. Fujimoto,M., Kitazawa,R., Maeda,S. and Kitazawa,S. (2005) Methylation adjacent to negatively regulating AP-1 site reactivates TrkA gene expression during cancer progression. *Oncogene*, **24**, 5108–5118.
4. Zhu,W.-G., Srinivasan,K., Dai,Z., Duan,W., Druhan,L.J., Ding,H., Yee,L., Villalona-Calero,M.A., Plass,C. and Otterson,G.A. (2003) Methylation of adjacent CpG sites affects Sp1/Sp3 binding and activity in the p21(Cip1) promoter. *Mol. Cell. Biol.*, **23**, 4056–4065.
5. 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.
6. Jin,J., Lian,T., Gu,C., Yu,K., Gao,Y.Q. and Su,X.-D. (2016) The effects of cytosine methylation on general transcription factors. *Sci. Rep.*, **6**, 29119.
7. Pozner,A., Terooatea,T.W. and Buck-Koehntop,B.A. (2016) Cell-specific kaiso (ZBTB33) regulation of cell cycle through cyclin D1 and cyclin E1. *J. Biol. Chem.*, **291**, 24538–24550.
8. 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.
9. 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.
10. Slattery,M., Zhou,T., Yang,L., Dantas Machado,A.C., Gordân,R. and Rohs,R. (2014) Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.*, **39**, 381–399.
11. Allfrey,V.G., Mirsky,A.E. and Stern,H. (2006) *The Chemistry of the Cell Nucleus*. Wiley-Blackwell, pp. 411–500.
12. Khan,A., Fornes,O., Stigliani,A., Gheorghe,M., Castro-Mondragon,J.A., van der Lee,R., Bessy,A., Chèneby,J., Kulkarni,S.R., Tan,G. *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D1284–D1284.
13. 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.* (2018) 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.
14. Yevshin,I., Sharipov,R., Valeev,T., Kel,A. and Kolpakov,F. (2017) GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res.*, **45**, D61–D67.
15. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
16. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
17. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
18. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
19. Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572.
20. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
21. Akalin,A., Kormaksson,M., Li,S., Garrett-Bakelman,F.E., Figueroa,M.E., Melnick,A. and Mason,C.E. (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.*, **13**, R87.
22. Adusumalli,S., Mohd Omar,M.F., Soong,R. and Benoukraf,T. (2015) Methodological aspects of whole-genome bisulfite sequencing analysis. *Brief. Bioinform.*, **16**, 369–379.
23. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
24. Machanick,P. and Bailey,T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **27**, 1696–1697.
25. Bailey,T.L. and Machanick,P. (2012) Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.*, **40**, e128.
26. Crooks,G.E., Hon,G., Chandonia,J.-M. and Brenner,S.E.WebLogo: A sequence logo generator. *Genome Res.*, **14**, 1188–1190.
27. Brinkman,A.B., Gu,H., Bartels,S.J.J., Zhang,Y., Matarese,F., Simmer,F., Marks,H., Bock,C., Gnirke,A., Meissner,A. *et al.* (2012) Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Res.*, **22**, 1128–1138.
28. Škuta,C., Bartůněk,P. and Svozil,D. (2014) InCHlib – interactive cluster heatmap for web applications. *J. Cheminform.*, **6**, 44.
29. Murrell,P. (2009) Importing vector graphics: the grImport package for R. *J. Stat. Softw.*, **30**, 1–37.
30. Pozner,A., Terooatea,T.W. and Buck-Koehntop,B.A. (2016) Cell-specific Kaiso (ZBTB33) Regulation of cell cycle through cyclin D1 and cyclin E1. *J. Biol. Chem.*, **291**, 24538–24550.
31. Donaldson,N.S., Pierre,C.C., Anstey,M.I., Robinson,S.C., Weerawardane,S.M. and Daniel,J.M. (2012) Kaiso represses the cell cycle gene cyclin D1 via Sequence-specific and Methyl-CpG-Dependent mechanisms. *PLoS One*, **7**, e50398.
32. Suzuki,T., Shimizu,Y., Furuhata,E., Maeda,S., Kishima,M., Nishimura,H., Enomoto,S., Hayashizaki,Y. and Suzuki,H. (2017) RUNX1 regulates site specificity of DNA demethylation by recruitment of DNA demethylation machineries in hematopoietic cells. *Blood Adv.*, **1**, 1699–1711.
33. Igarashi,K., Kataokat,K., Itoh,K., Hayashi,N., Nishizawa,M. and Yamamoto,M. (1994) Regulation of transcription by dimerization of erythroid factor NF-E2 p45 with small Maf proteins. *Nature*, **367**, 568–572.
34. Zhou,Z., Li,X., Deng,C., Ney,P.A., Huang,S. and Bungert,J. (2010) USF and NF-E2 cooperate to regulate the recruitment and activity of RNA polymerase II in the beta-globin gene locus. *J. Biol. Chem.*, **285**, 15894–15905.
35. 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.
36. Hai,T. and Curran,T. (1991) Cross-family dimerization of transcription factors Fos/Jun and ATF/CREB alters DNA binding specificity. *Proc. Natl. Acad. Sci. U.S.A.*, **88**, 3720–3724.
37. Zhang,D., Wu,B., Wang,P., Wang,Y., Lu,P., Nechiporuk,T., Floss,T., Greally,J.M., Zheng,D. and Zhou,B. (2017) Non-CpG methylation by DNMT3B facilitates REST binding and gene silencing in developing mouse hearts. *Nucleic Acids Res.*, **45**, 3102–3115.
38. Pan,Y.-X., Chen,H., Thiaville,M.M. and Kilberg,M.S. (2007) Activation of the ATF3 gene through a co-ordinated amino acid-sensing response programme that controls transcriptional

regulation of responsive genes following amino acid limitation. *Biochem. J.*, **401**, 299–307.
39. Fries,F., Nazarenko,I., Hess,J., Claas,A., Angel,P. and Zöller,M. (2007) CEBPbeta, JunD and c-Jun contribute to the transcriptional activation of the metastasis-associated C4.4A gene. *Int. J. Cancer*, **120**, 2135–2147.
40. Plachetka,A., Chayka,O., Wilczek,C., Melnik,S., Bonifer,C. and Klempnauer,K.-H. (2008) C/EBPbeta induces chromatin opening at a cell-type-specific enhancer. *Mol. Cell. Biol.*, **28**, 2102–2112.
41. Wang,H., Liu,W., Black,S., Turner,O., Daniel,J.M., Dean-Colomb,W., He,Q.P., Davis,M. and Yates,C. (2016) Kaiso, a transcriptional repressor, promotes cell migration and invasion of prostate cancer cells through regulation of miR-31 expression. *Oncotarget*, **7**, 5677–5689.
42. Kim,M. and Costello,J. (2017) DNA methylation: an epigenetic mark of cellular memory. *Exp. Mol. Med.*, **49**, e322.
43. Mayran,A., Khetchoumian,K., Hariri,F., Pastinen,T., Gauthier,Y., Balsalobre,A. and Drouin,J. (2018) Pioneer factor Pax7 deploys a stable enhancer repertoire for specification of cell fate. *Nat. Genet.*, **50**, 259–269.
44. Di Stefano,B., Sardina,J.L., van Oevelen,C., Collombet,S., Kallin,E.M., Vicent,G.P., Lu,J., Thieffry,D., Beato,M. and Graf,T. (2014) C/EBPα poises B cells for rapid reprogramming into induced pluripotent stem cells. *Nature*, **506**, 235–239.
45. Sasai,N., Nakao,M. and Defossez,P.-A. (2010) Sequence-specific recognition of methylated DNA by human zinc-finger proteins. *Nucleic Acids Res.*, **38**, 5015–5022.
46. Buck-Koehntop,B.A. and Defossez,P.-A. (2013) On how mammalian transcription factors recognize methylated DNA. *Epigenetics*, **8**, 131–137.
47. Spruijt,C.G. and Vermeulen,M. (2014) DNA methylation: old dog, new tricks? *Nat. Struct. Mol. Biol.*, **21**, 949–954.