Published online 22 November 2019

Nucleic Acids Research, 2020, Vol. 48, No. 2 ***e10***
doi: 10.1093/nar/gkz1088

TFregulomeR reveals transcription factors' context-specific features and functions

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

[1]Cancer Science Institute of Singapore, National University of Singapore, Singapore 117599, Singapore, [2]Computational Systems Biology Team, Institut de Biologie de l'École Normale Supérieure (IBENS), CNRS, INSERM, École Normale Supérieure, PSL Research University, Paris 75005, France, [3]Department of Biochemistry, National University of Singapore, Singapore 117596, Singapore and [4]Discipline of Genetics, Faculty of Medicine, Memorial University of Newfoundland, St. John's, NL A1B 3V6, Canada

Received August 01, 2019; Revised October 25, 2019; Editorial Decision October 31, 2019; Accepted November 01, 2019

ABSTRACT

Transcription factors (TFs) are sequence-specific DNA binding proteins, fine-tuning spatiotemporal gene expression. Since genomic occupancy of a TF is highly dynamic, it is crucial to study TF binding sites (TFBSs) in a cell-specific context. To date, thousands of ChIP-seq datasets have portrayed the genomic binding landscapes of numerous TFs in different cell types. Although these datasets can be browsed *via* several platforms, tools that can operate on that data flow are still lacking. Here, we introduce TFregulomeR (https://github.com/benoukraflab/TFregulomeR), an R-library linked to an up-to-date compendium of cistrome and methylome datasets, implemented with functionalities that facilitate integrative analyses. In particular, TFregulomeR enables the characterization of TF binding partners and cell-specific TFBSs, along with the study of TF's functions in the context of different partnerships and DNA methylation levels. We demonstrated that TFs' target gene ontologies can differ notably depending on their partners and, by re-analyzing well characterized TFs, we brought to light that numerous leucine zipper TFBSs derived from ChIP-seq experiments documented in current databases were inadequately characterized, due to the fact that their position weight matrices were assembled using a mixture of homodimer and heterodimer binding sites. Altogether, analyses of context-specific transcription regulation with TFregulomeR foster our understanding of regulatory network-dependent TF functions.

INTRODUCTION

Transcription factors (TFs) are the key components that regulate spatiotemporal gene transcription (1). Aberrant TF activities result in gene dysregulation, which is associated with several disorders such as cancer (2). A TF usually recognizes a 3–15 bp DNA sequence, *e.g.* 3 bp for c-JUN monomer (3) and 15 bp for CTCF (4), within proximal (promoter) or distal (enhancer) *cis*-regulatory elements. In the past decades, *in vitro* and *in vivo* techniques have been established to uncover TF sequence binding preferences. Protein-binding microarray (PBM) (5), bacterial one-hybrid (B1H) (6) and systematic evolution of ligands by exponential enrichment (SELEX) (7) are widely adopted *in vitro* approaches to identify TF binding sites (TFBSs) in high-throughput scales (1,8). However, although these methods have uncovered the DNA binding motifs for thousands of TFs, the characterization of TFBS on a naked DNA environment neglects key *in vivo* conditions that influence binding sites, such as chromatin structure, TF cooperativity and DNA modification, and thus may reveal only one facet of TF–DNA interaction (9). The current gold standard method to determine *in vivo* TF binding sequences is chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) (10). This technique allows to enrich chromatin fragments that interact with a TF of interest and, with the help of *in silico* procedures, compute a DNA binding motif (11,12). Nonetheless, these chromatin fragments usually cover ∼100–500 bp, while the actual TFBS size ranges from 3–15 bp, which challenges the depiction of the exact TF binding locus. To overcome this issue, Rhee and Pugh have developed a refined ChIP-protocol including a DNA trimming step to cleave DNA fragments unprotected by DNA binding proteins (13). This method, called ChIP-exo (ChIP exonuclease), has greatly enhanced the mapping resolution of TF binding sites (14,15). Nonetheless,

*To whom correspondence should be addressed. Tel: +1 709 864 6671; Fax: +1 709 864 6537; Email: tbenoukraf@mun.ca

the majority of TFs bind to DNA with other TF partners across a cluster of TFBSs, also termed *cis*-regulatory module (CRM) (16). Consequently, in spite of the great improvement made by ChIP-exo, DNA regions bound by transcriptional complexes cannot be split into individual TF binding sites by exonuclease, which still obscures the delimitation of the precise binding site for a TF of interest (17). In other words, current *in vivo* protocols may generate TF DNA binding motifs that are generally biased by the presence of co-factors. For example, Starick *et al*. observed a much lower occurrence of glucocorticoid receptor (GR) binding motif but a higher fraction of GATA-related binding sites in the peaks derived from K562 GR ChIP-exo, compared to other cell types such as IMR90 and U2OS. They speculated a cell-specific tethering recruitment of GR onto DNA sequences in K562 (14). In addition, Cohen *et al*. have recently described that variations in CEBPB motif were context-dependent (15). In their study, they described higher preservation of canonical CEBPB motif in the cell-type independent context compared to cell-type specific contexts. Hence, a straightforward analysis of DNA binding motif derived from ChIP-based methods may not be sufficient to predict the actual TF binding site. For some TFs, additional systematic analyses are required to precisely delineate the TF recognition site, taking the presence of co-factors into account. Overall, cooperative binding of multiple TFs in the CRM enables high binding specificity and fine-tuning in gene regulation (18). Hence, it is of great importance to understand TF binding behaviours collectively instead of individually.

Integrated with DNA methylation profiles, our recently launched cell-specific TF-binding profile database, MethMotif, has clearly showed that cell context influences TFBS at both sequence and epigenetic levels (19). DNA methylation information is therefore an additional layer to take into consideration for a comprehensive TF binding analysis, in order to delineate precise DNA sequence affinity and uncover TF function in a specific cell context. Nevertheless, although numerous cistrome databases and tools facilitating the prediction of CRMs and cofactors of a TF of interest are available (20–22), an environment connecting a large compendium of TF binding sites and respective DNA methylation profiles is still lacking. To fill this gap, we have established a large collection of standard ChIP-seq and Whole Genome Bisulfite Sequencing (WGBS) datasets derived from the GTRD repository (23) and MethMotif database (19). This information can be browsed, analyzed and compared to other data sources *via* TFregulomeR, an R package that contains processing and visualization functions. To our knowledge, TFregulomeR is the first platform enabling the query and the analysis of context-specific TF modules according to cell type, tissue origin and disease state. This is also the first computational tool allowing a standard integration of TF module binding sites with their respective DNA methylation profiles (Supplementary Table S1). In essence, TFregulomeR facilitates the analysis of context-specific transcription regulation and fosters our understanding of regulatory network-dependent TF functions.

MATERIALS AND METHODS

Cistrome and methylome data in TFregulomeR

TF motif position weight matrices (PWM) recorded in the TFregulomeR compendium are compiled from the MethMotif database (19) and GTRD (23). MethMotif includes a comprehensive collection of TF motifs complemented with methylation information. Briefly, relying on the Irreproducibility Discovery Rate (IDR) (24), ChIP-seq peaks were standardly called using MACS2 (25), followed by a motif *de novo* enrichment analysis with MEME-ChIP (11), focusing on regions defined by a ±100 bp window around peak summits. Subsequently, DNA methylation profiles were inferred using whole genome bisulfite sequencing (WGBS) datasets in all peak regions and predicted TFBSs.

GTRD is a large collection of publicly available ChIP-seq experiments which serves as an important complementary source to enlarge the PWM pool in TFregulomeR. ChIP-seq peaks identified by MACS peak caller from the untreated experiments were downloaded from GTRD database. Following the same pipeline in the MethMotif, centrally enriched TF motifs were localized using MEME-ChIP with the default settings in a 200 bp range surrounding peak centres. In TFregulomeR compendium, highly and centrally enriched motifs were selected and compared with the existing TF-binding profile databases, such as JASPAR (26) and HOCOMOCO (27). In some ChIP-seq, the significantly enriched motifs did not correspond to the motifs of ChIP'ed TFs. Among 1468 PWM records in TFregulomeR, we identified 91 motifs different from those reported in TF-binding profile databases. Furthermore, 136 motif matrices were not recorded for their corresponding TFs in these databases. In order to verify that these 227 motifs do not derive from the use of a specific motif discovery algorithm, we used another motif discovery tool, HOMER (28), based on different algorithm hypergeometric enrichment, to perform an additional *de novo* motif discovery. The motifs obtained with HOMER were compared with those obtained with MEME-ChIP, and their similarities were measured by normalized Pearson correlation coefficient using the *compare-matrices* function in RSAT (29) with the formula: Ncor = cor * w / w_smaller, where cor is raw Pearson correlation coefficient, w is the alignment width of two matrices from MEME-ChIP and HOMER (the minimum value of w was set as 5), and w_smaller is the width of smaller motifs from MEME-ChIP and HOMER.

Each PWM has its own unique ID, which keeps track of the source, species, cell type and TF (*e.g.* 'MM1_HSA_K562_CEBPB' and 'GTRD-EXP010975_HSA_Ishikawa_CEBPB'). Since TF binding preferences are obviously dependent on the context, all obtained TFBS datasets have been thoroughly annotated in terms of species, organ, sample type, cell or tissue origin, disease state and experiment source.

TFregulomeR functionalities

TFregulomeR provides functionalities to ease data access, retrieval, integration and analysis. These functionalities en-

able browsing of curated datasets in TFregulomeR data compendium (*dataBrowser*), plotting motif logo (or MethMotif logo if DNA methylation is available, *plotLogo*), loading peak regions (*loadPeaks*), exporting motif PWM and DNA methylation matrix (*exportMMPFM*), obtaining context in/dependent peaks (*commonPeaks* and *exclusivePeaks*), forming a peak intersection matrix for cofactor and TF interaction studies (*intersectPeakMatrix*), generating a cofactor report automatically along with DNA methylation and read enrichment scores (*cofactorReport*), profiling TFBS distribution (*motifDistrib*), annotating peak locations and functions (*genomeAnnotate* and *greatAnnotate*), and converting the PWM in TFregulomeR into an R object compatible with TFBSTools (*toTFBSTools*) (30). All these functionalities are implemented in a public R-library package (https://github.com/benoukraflab/TFregulomeR).

S4 classes in TFregulomeR

In order to efficiently store multiple data sets, several S4 classes have been created for different purposes. *MethMotif* object is a basic class to record TFBS PWM model and, if available, DNA methylation levels. During the context independent peak analysis (*commonPeaks* function), *CommonPeaksMM* has been designed to store the percentage of common peaks, common peak regions, the DNA methylation profiles in a $\pm$100 bp window around common peak summits and the *de novo* generated *MethMotif* object representing the TFBS features in the common peak regions. Similarly, the class *ExclusivePeaksMM* has been designed for the output of context dependent peak analysis (*exclusivePeaks* function). Moreover, the *IntersectPeakMatrix* class was built to store the outputs of pair-wise peak intersection analyses (*intersectPeakMatrix* function), including the inherited *MethMotif* object, percentage of intersected peaks, as well as DNA methylation states and read enrichment scores in the overlapping peak regions.

Common and exclusive peak analysis

TFregulomeR provides functionalities to perform context in/dependent peak analysis. The peak sets can be directly derived from the TFregulomeR compendium or customized by users. The context in/dependent peak regions are returned, together with *de novo* generated Motif logos (MethMotif logos if DNA methylation is available) and DNA methylation states. Particularly, the logo-plotting function in TFregulomeR uses the package ggseqlogo (31), to generate high quality (Meth)Motif logos in vector format.

Peak intersection matrix analysis for cofactor and TF interaction study

TFregulomeR allows the users to conduct pair-wise comparison analyses between collections of peak sets. This functionality enables the profiling of TF cofactors or interactome in a cell type. Two lists of peak sets, X and Y (x and y peak sets in peak list X and Y respectively), are input to form $x \times y$ intersect matrix table, with each table cell denoting pair-wise comparison from list X and Y. The peak sets in both lists can be obtained from the TFregulomeR data compendium or self-provided. An intersection matrix denoting the percentage of the intersected peak for each pair of peak sets will be returned, and simultaneously the *de novo* generated (Meth)Motif logo as well as DNA methylation status and read enrichment scores for each set of intersected peaks will be also approachable.

Peak genomic location and functional annotations

We incorporated generic functions to annotate peak genomic locations and functions to enable TFregulomeR to serve as an all-inclusive TFBS toolbox. Location annotation follows the order starting from promoter, transcription termination site (TTS), 5′ untranslated region (UTR) exon, 3′ UTR exon, intron to intergenic region. By default, the promoter is defined in the region from 1000 bp upstream to and 100bp downstream of the transcription start site (TSS), while TTS covers from 100 bp upstream to and 1000 bp downstream of the actual TTS. These two settings can be easily modified by users. Functional annotation was achieved using rGREAT, a GREAT analysis API (32). Since GREAT server doesn't support hg38 assembly, liftOver has been implemented to automatically perform a genomic conversion from hg38 to hg19. Both annotation functions generate intuitive HTML reports.

TFregulomeR compendium maintenance

MethMotif and GTRD are the main sources for the cistrome and methylome data in TFregulomeR compendium. Since its launch, MethMotif database has been updated actively, while GTRD has been also regularly maintained. In the future, TF motif records along with DNA methylation in TFregulomeR will be updated accordingly to offer up-to-date PWM collections for the research community.

RESULTS

TFregulomeR, a comprehensive toolbox to study TF binding dynamics

The study of TF context-specific binding preferences heavily relies on comprehensive knowledge on TF cell-specific binding sites and chromatin status (1,19,33). We thus compiled this information by integrating PWM collections derived from GTRD, a database encompassing publicly available ChIP-seq data (23), with the resources from our MethMotif database, which couples TFBSs with DNA methylation status (19) (Figure 1A). To date, our compendium includes 1468 PWMs and, in contrast to current available resources, we manually curated our datasets in order to match cistrome with methylome data in a cell type-specific fashion, and organized them according to cell type, tissue of origin, disease state and experiment source (Table 1 and Supplementary Table S1).

It is important to note that, in some ChIP-seq datasets, the highly enriched motif (labeled as binding motif of the TF of interest in TFregulomeR compendium) does not necessarily resemble the canonical motif of the ChIP'ed TF. This is presumably due to the fact that in the given cell type, the ChIP'ed TF is mostly recruited by other TFs instead of directly binding to DNA, and/or that the high presence of

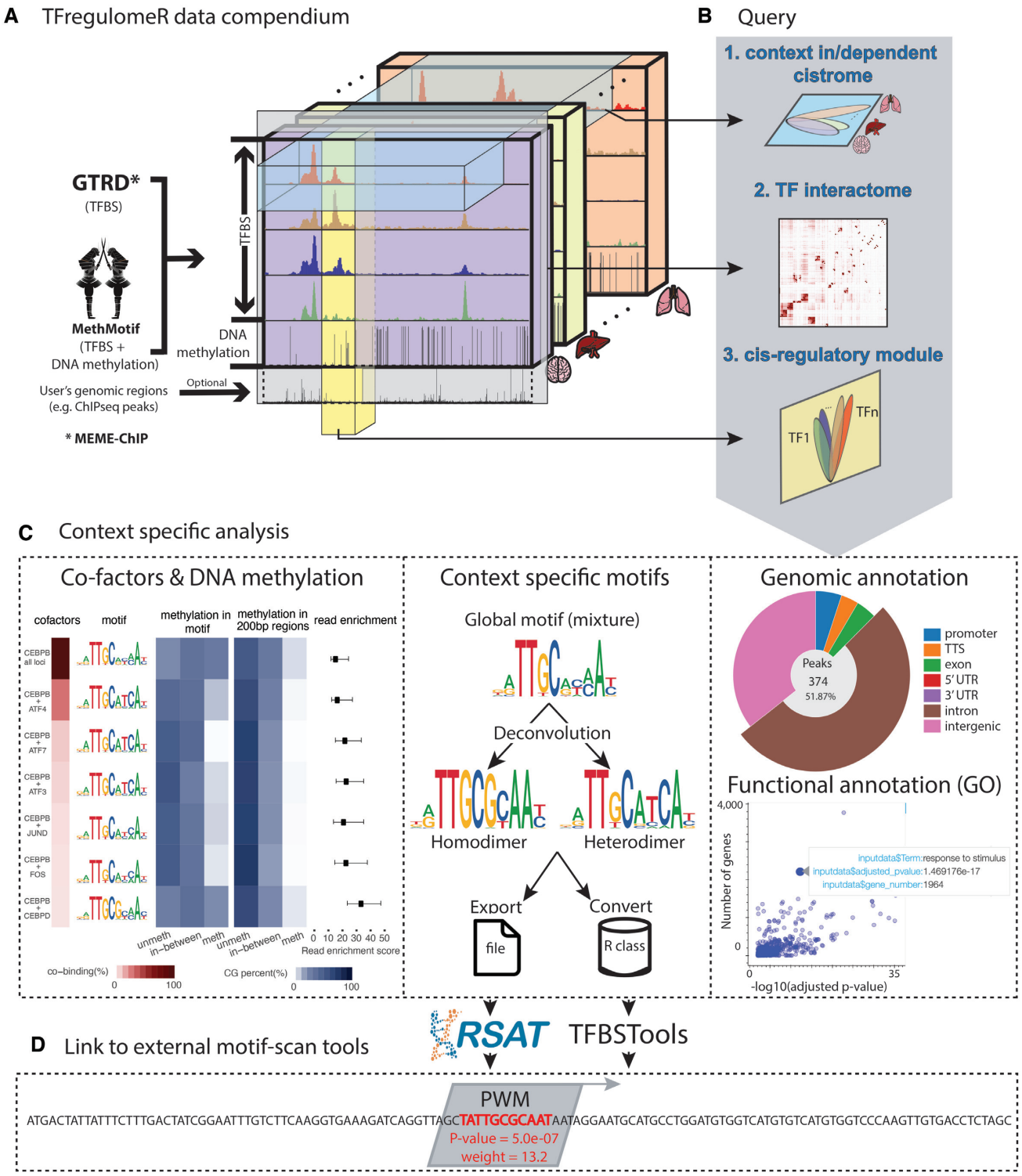

Figure 1. TFregulomeR key functionalities. (**A**) TFregulomeR consists of a toolbox linked to a large timely updated compendium of motif PWMs along with DNA methylation derived from MethMotif and GTRD. Users are also allowed to include their own genomic regions (*e.g.* ChIP-seq peaks) in TFregulomeR for peak meta-analysis. The PWM annotations recorded in TFregulomeR data compendium have been manually curated regarding species, organ, sample type, cell or tissue origin, disease state and data source. (**B**) TFregulomeR supports query of context in/dependent cistrome, TF interactome as well as *cis*-regulatory module (CRM). (**C**) Its functionalities allow (i) the study of TF co-factors along with DNA methylation and read enrichments, (ii) the characterization of context-specific binding sites and (iii) context-specific genomic and functional annotations. (**D**) Furthermore, TFregulomeR enables a direct conversion of newly generated PWM models to objects compatible with TFBSTools. These PWM models can also be exported to MEME and TRANSFAC formatted files for further downstream analyses using third-party resources such as RSAT.

Table 1. TFBSs in TFregulomeR compendium

Item	Count
PWM	1468
Unique TF	415
Species	human
Organ	stem_cell, blood_and_lymph, connective_tissue, colorectum, brain, bone, stomach, prostate, breast, pancreas, skin, kidney, lung, eye, esophagus, heart, muscle, uterus, spleen, cervix, testis, liver, adrenal_gland, neck_and_mouth, pleura, ovary, thymus, fallopian, vagina
Sample type	primary_cells, cell_line, tissue
Cell or tissue	414
Disease state	normal, tumor, Simpson_Golabi_Behmel_syndrome, progeria, metaplasia, unknown, immortalized, premetastatic
Source	GTRD, MethMotif

some non-targeted motifs repeatedly observed across ChIP-seq datasets, also known as 'zinger' motifs, mask the motif enrichment of the ChIP'ed TF (34). In TFregulomeR compendium, we identified 91 PWMs different from TF-binding profile databases, such as JASPAR (26) and HOCO MOCO (27). In order to verify that divergent motifs do not derive from the use of a specific motif discovery algorithm (in our case MEME-ChIP based on an expectation maximization approach), we used another algorithm, HOMER (based on hypergeometric enrichment), to perform an additional *de novo* motif discovery. Motif consistency details have been compiled into the TFregulomeR compendium, and divergent PWMs were flagged to allow users to interpret their results with caution.

TFregulomeR compendium is accessible *via* an effective R-package application program interface (API), called TFregulomeR. This API offers the required functions for easy data access, analysis and integration in pipelines. TFregulomeR facilitates (i) the recognition of context in/dependent TF binding locations (cistrome); (ii) the identification of context-specific TF-interacting partners (interactome); and (iii) the characterization of functionally active *cis*-regulatory modules (CRM) (Figure 1B). As demonstrated in the next sections, TFBS context-comparative analysis enables the identification of novel TFBS features, in particular the characterization of co-factors and their influence on TF functions. TFregulomeR also includes generic functionalities such as TFBS genomic annotation, which can be used to locate the genomic binding site landscapes of a TF of interest depending on cell type and/or partner combination (Figure 1C right). Interestingly, algorithms implemented in our toolbox allow the detection and segregation of multiple cofactor-derived motifs present in the PWM directly computed from whole genome context (Figure 1C, middle, *cf.* detailed hereafter). In contrast to all available tools (20–22), by integrating DNA methylome datasets, TFregulomeR uniquely emphasizes the impact of DNA modifications in TF-module recruitment and reveals how DNA methylation can affect TF functions (Figure 1C, left and Supplementary Table S1). At last, objects generated by the TFregulomeR package are compatible with third-party packages such as TFBSTools, which provides means for TFBS matrix handling and motif scanning (30), and rGREAT, for ontology analysis (32). These TFregulomeR objects can be also exported for downstream analyses using external servers like RSAT (29) (Figure 1D). In the following case studies, we focus on CEBPB, MAFF and ATF3 binding data to demonstrate the capacity of TFregulomeR to reveal co-factor partnership and perform motif deconvolution.

Deconvolution of transcription factor motifs: the case of CEBPB

Using an *in vitro* HT-SELEX approach, CEBPB specific binding sequence has been characterized as a dimeric motif ATTGCGCAAT (35), where CAAT (or its reverse complement ATTG) is the well-known DNA-binding motif of all CEBP family members, and the CG dinucleotide is the spacer between two coupled binding sites (36). However, among the 16 CEBPB ChIP-seq experiments recorded in the TFregulomeR compendium, only a few shows a motif enrichment fully consistent with the *in vitro* canonical sequence. Instead, most returned motif logos display a conserved ATTG/CAAT half site along with the other degenerated half motif that looks like a combination of ATTG/CAAT and TCA/TGA, suggesting that the motif is composed by a mixture of homodimer and heterodimer binding sites (Supplementary Figure S1). Actually, this phenomenon has already been described previously for another CEBP family member. Indeed, Cai *et al.* have shown that CEBPA binds novel genomic sites when it dimerizes with AP-1 family members, which recognize the motif TCA/TGA (37). More recently, by analyzing high-throughput sequencing datasets, Cohen *et al.* have extended this observation to CEBPB. In their study, they compared the CEBPB binding sites across six cell types and showed a higher occurrence of canonical CEBPB motif in the cell-type shared regions compared to the cell-type specific loci (15). TFregulomeR enabled the systematic validation of this phenomenon in a bigger cohort encompassing 16 cell types and, more importantly, provided a possible explanation for motif variations depending on context.

After loading all binding sites present in our data compendium, we segregated the K562 CEBPB binding loci according to the number of cell types where they are enriched. This analysis shows that less cell-specific are the K562 CEBPB peaks, less is the number of peaks, but higher is the read enrichment within these loci (Supplementary Figure S2A). In line with the study by Cohen *et al.*, analysis with TFregulomeR showed an increased enrichment of the homodimer motif (ATTGCGCAAT) when the K562 CEBPB peaks were less cell-specific, while a more degenerate motif was over-represented across the K562-specific binding loci

(Figure 2A and Supplementary Figure S2B). Such observation was not unique to K562 cells but a widespread phenomenon across all other 15 cell types (Supplementary Figure S3).

This case study further illustrates how TFregulomeR can be used to provide an explanation of the observations by Cohen *et al*. in terms of partnership and DNA methylation levels. Indeed, one function implemented in TFregulomeR has been designed for the characterization of co-factors within the sub-ensembles of binding sites (*e.g.* K562 CEBPB peaks shared by different number of cell types, Figure 2B). Among all other 130 TFs profiled in K562, CEBPD and ATF4 were characterized as the two main recurrent CEBPB cofactors. Indeed, near than 37% of CEBPB binding sites were co-occupied by CEBPD, a member of CEBP family, and a third of CEBPB binding sites were co-localized with ATF4, a member of activating transcription factor (ATF) family (Figure 2C). Interestingly, the TFregulomeR context-specific analysis clearly shows that CEBPD is overrepresented in CEBPB binding loci shared across all cell types, whereas ATF4 co-binds with CEBPB mainly in K562-specific CEBPB binding sites (Figure 2B). To further confirm the connection between the TCA motif and the ATF4 co-dimer, we computed the DNA binding site motif of the intersection of ATF4 and CEBPB binding loci within the K562-specific CEBPB peaks. As expected, the resulting motif logo is a clear CEBPB-ATF4 heterodimer (Figure 2D, left). Therefore, this observation suggests that ATF4 is not merely a CEBPB co-factor, but rather dimerizes with CEBPB, thereby corroborating with a previous study by Jolma *et al*. using an *in vitro* approach called consecutive affinity-purification systematic evolution of ligands by exponential enrichment (CAP-SELEX) (38). A similar analysis was performed on the K562 shared CEBPB peaks and the prevalence of the homodimer-type motif (ATTGCGCAAT) could be explained by the fact that, within these loci, CEBPB principally dimerizes with CEBPD, which shares the same recognition DNA binding site (ATTG/CAAT) (39) (Figure 2D, right). In addition, our TFregulomeR analysis also pointed to a link between CEBPB cofactors and the dinucleotide spacer composition. Remarkably, the homodimer-type motifs (CEBP-CEBP) were clearly enriched with CG dinucleotide in the spacers, while CA dinucleotides were more present in heterodimer motifs. Consequently, CEBP-CEBP binding loci are more subjected to DNA methylation compared to heterodimer binding sites. More precisely, TFregulomeR revealed a considerable portion of hypermethylated CGs associated with CEBPB-CEBPD partnership in the cell type-independent regions, whereas the CGs around CEBPB-ATF4 co-binding regions were prone to hypomethylation (Figure 2C). In addition to connecting DNA methylation profiles to context-specific TF binding sites, the TFregulomeR compendium also includes read enrichment scores to assess the binding strength of a given TF in the context of different partners or DNA methylation profiles (Figure 2C).

The CEBPB TFBS motif is currently described as a mixture of CEBPB and other cofactors in all main TF-binding profile databases, including JASPAR (26), HOCOMOCO (27) and MethMotif (19) (Figure 2E). Here, we have illustrated how TFregulomeR can improve the accuracy of PWM through the deconvolution of TFBS motifs and the characterization of TF partners influencing DNA binding sites at the sequence and methylation levels.

Deconvolution of binding partners: the case of MAFF

TFregulomeR compendium records MAFF ChIP-seq experiments from three different cell types: K562, HeLa-S3 and HepG2, extracted from the MethMotif database (respective accession number: MM1_HSA_K562_MAFF, MM1_HSA_HeLa-S3_MAFF and MM1_HSA_HepG2_MAFF). Surprisingly, MAFF displays distinct cell-specific DNA binding preferences in terms of DNA sequences and DNA methylation levels. The inconsistence of MAFF binding motifs is also present in other TF-binding profile databases (Figure 3A). These binding sequences are composed by a common TCAGCA motif and a TGA trinucleotide, which is highly enriched in K562, moderately present in HeLa-S3, but absent in HepG2 (Figure 3A). This observation was accentuated when we focused only on MAFF cell-specific peaks (Figure 3B). The differences of motif enrichments across these three cell types were also associated with varied DNA methylation states. We observed a more hypomethylated DNA profile within the MAFF peaks in K562, while a considerable portion of CGs in HepG2 MAFF peaks were hypermethylated (Figure 3A and B).

As for our CEBPB analysis, we used TFregulomeR to search for co-factors associated with these DNA sequence variations and methylation differences across cell types. Throughout MAFF peaks specific to K562 cells, TFregulomeR highlighted several co-factors that were highly co-bound with MAFF, including Nuclear Factor Erythroid 2 (NFE2) and Erythroid 2 Like 2 (NFE2L2), which are known to bind an AP1-like binding site (TGA/TCA) and thus explain the presence of this motif in MAFF's K562 peaks (Figure 3C). In contrast, NFE2L2 was not detected as MAFF's co-factor across HeLa-S3 and HepG2 cell lines (NFE2 ChIP-seq data are not available for the two cell lines), supporting the fact that the TGA trinucleotide enrichment in K562 MAFF binding loci coincides with the presence of NFE2 and NFE2L2 (Figure 3C). Altogether, these observations suggest that the TGA sub-motif in K562 context actually correspond to NFE2 and NFE2L2 binding sites. To further validate this hypothesis, we generated the motif logo of MAFF K562-specific binding sites without the binding loci shared with NFE2 and NFE2L2, which encompass 517 regions. As expected, the resulting pattern displays a significant decrease in TGA trinucleotide enrichment (Figure 3D).

Although the MAFF-NFE2 binding complex has been already described in the JAPSAR database (accession number: MA0501.1, Figure 3A), here we showed how TFregulomeR is able to systematically discover cell type-specific interactions between MAFF and other TFs influencing DNA binding motifs. Furthermore, TFregulomeR adds an epigenetic layer to the binding site information, thereby enhancing our understanding of the role of TF complexes in gene regulation and chromatin modeling. In this particular example, we observed a correlation between the number of

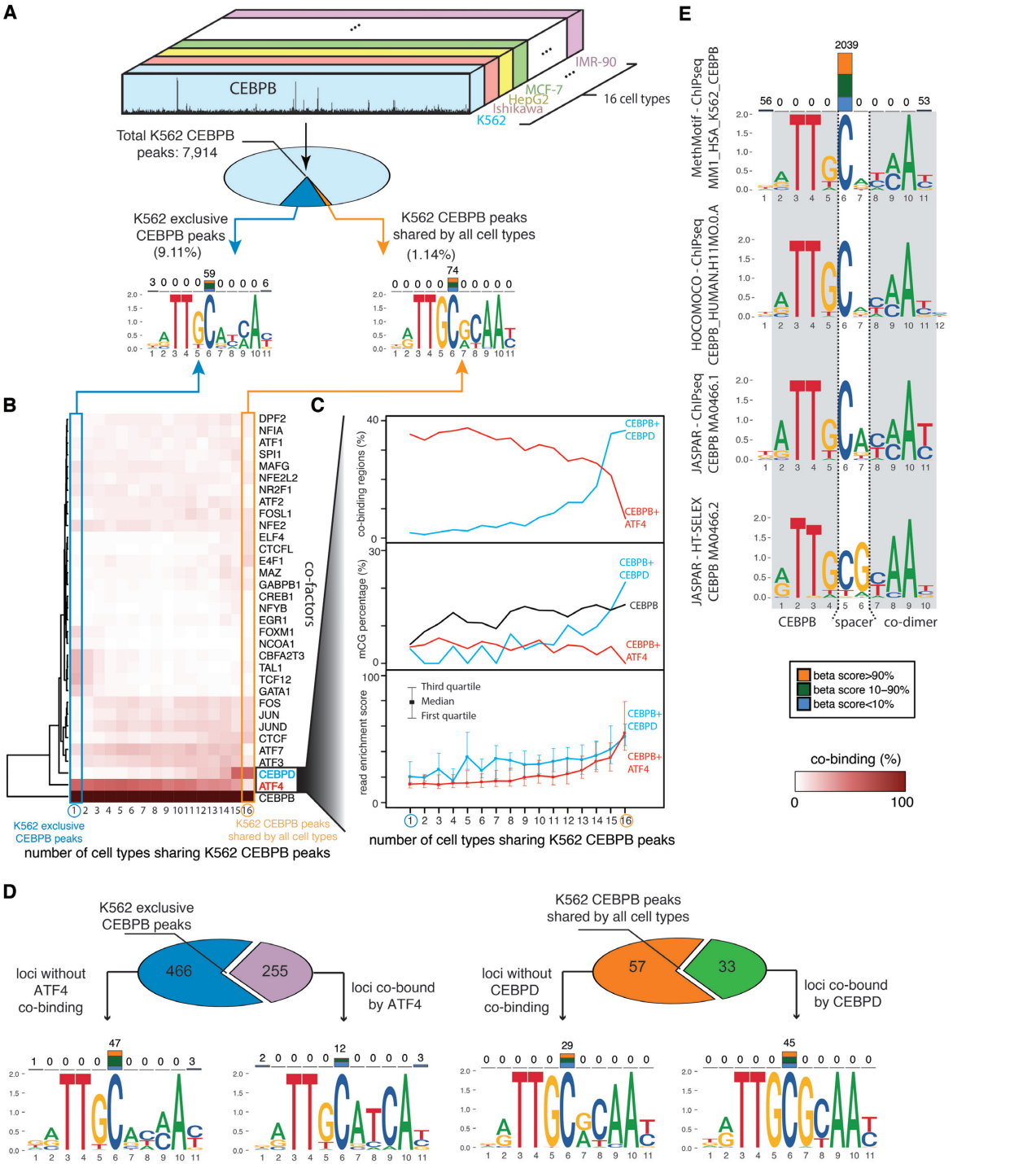

Figure 2. Analysis of binding partners and motifs in K562 context-specific CEBPB peaks. (**A**) Among 7914 CEBPB peaks in K562, 9.11% of peaks were unique to K562 while 1.14% of peaks were shared across all cell types. MethMotif logos display the motif enrichments along with DNA methylation states in K562 shared and exclusive CEBPB peaks. The blue, orange and green bars stacked above motif logo denote the numbers of CpGs homogenously unmethylated, homogenously methylated and heterogeneously methylated, respectively. (**B**) The heatmap shows the cofactor binding profiles in 16 sub-ensembles of K562 CEBPB peaks segregated according to their number of cell types where they are enriched. Each row represents a TF, each column denotes K562 CEBPB sub-ensemble peaks, and color intensity denotes a specified TF co-binding percentage within a given sub-ensemble of peaks. Here, the TF with co-binding percentages <5% in all 16 sub-ensembles were excluded, and the heatmap underwent row-wise hierarchical clustering based on Euclidean distance. (**C**) These three plots show, from the top to the bottom, the co-binding percentages, methylated CG percentages within the co-binding peaks, as well as ChIP-seq read enrichment scores in the co-binding peaks for CEBPB-CEBPD (blue) and CEBPB-ATF4 combinations (red) across 16 K562 CEBPB sub-ensembles. In the middle plot, the overall methylated CG percentages across 16 K562 CEBPB sub-ensembles are further reported in black. (**D**) The MethMotif logos display sequence preferences together with DNA methylation states enriched in the K562 shared CEBPB peaks with/without CEBPD co-binding loci, and in the K562 exclusive CEBPB peaks with/without ATF4 co-binding loci. (**E**) CEBPB motif logos extracted from different TF-binding profile databases.

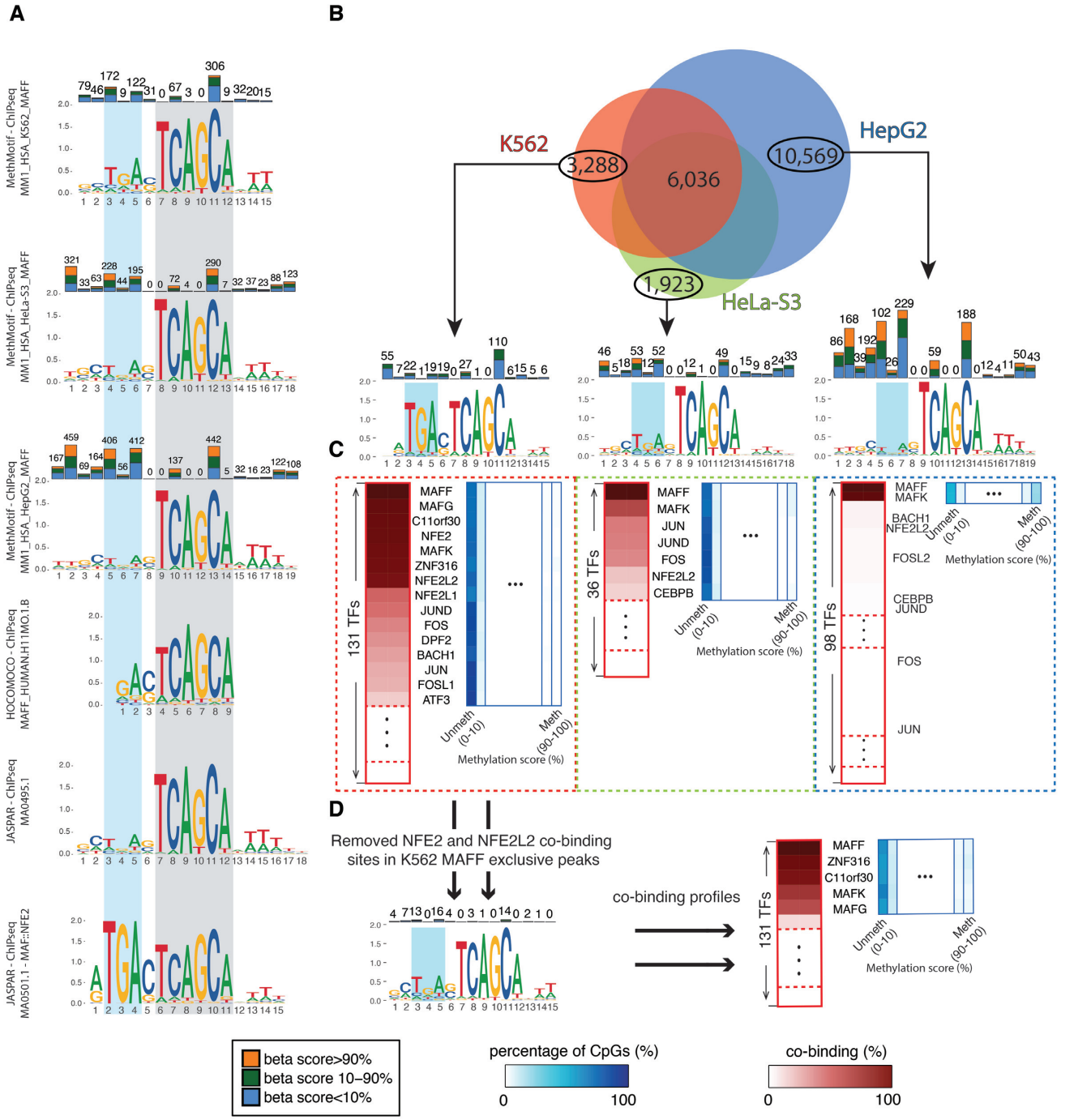

Figure 3. Analysis of MAFF binding partners and motifs in cell-specific peak regions. (**A**) MAFF motif logos from different TF-binding profile databases. (**B**) The MethMotif logos display the MAFF motif enrichments along with the DNA methylation states in cell-specific peak regions of three cell types. The blue, orange and green bars stacked above motif logo denote the numbers of CGs homogenously unmethylated, homogenously methylated and heterogeneously methylated respectively. (**C**) MAFF co-binding factors in the cell-specific regions are reported in a heatmap for each cell type (shades of red). DNA methylation states for the co-factors with more than 10% co-binding percentage are also portrayed in the regions (±100 bp around peak summits) co-bound by MAFF with those factors as the heatmaps (shades of blue). In this DNA methylation heatmap, each row represents a co-factor, while each column shows a methylation score interval. The color intensity implies the percentage of CGs with methylation scores in the given interval. (**D**) The MethMotif logo on the left displays sequence and DNA methylation preferences in the K562 MAFF cell-specific binding regions without the co-occurrences of NFE2 and NFE2L2. The right sides of this panel show the co-factor binding profiles in the regions, together with the DNA methylation for each co-factor with more than 10% co-binding percentage.

MAFF co-factors and DNA methylation states across different cell types. Indeed, in hypomethylated K562 unique MAFF loci, we found evidence for being co-bound by multiple co-factors, in contrast with HepG2 exclusive MAFF peaks, characterized by a great portion of hypermethylated CG where only MAFK emerged as a co-factor (Figure 3C). Our analysis shows that MAFF is able to bind methylated DNA only in the absence of NFE2 and NFE2L2 (Figure 3), suggesting distinct regulatory roles of MAFF conferred by the presence of NFE2 and NFE2L2.

Change of TF functions depending on partnership

As we illustrated previously (19), a TF could play distinct roles in different cell types. Moreover, the high occupancy of heterotypic dimers across cell type-specific CEBPB binding loci encouraged us to speculate that TF cell-specific regulatory roles may be partly attributed to binding partners. Indeed, we observed the distinct ontologies of targeted genes between CEBPB-CEBPD and CEBPB-ATF4 in K562 cells (Supplementary Figure S4). It is known that the collaborative interaction of TF across CRM contributes to the context-specific gene expression (40). Here, we focus on the case of ATF3 to demonstrate the capability of TFregulomeR to reveal partner-dependent roles across different cell types. In HCT116 and K562, the bound loci were highly enriched with a typical AP-1 type motif (TGA[G|C]TCA), while a different motif was extracted from GM12878, H1-hESC and HepG2 binding sites (GTCACGTG, Figure 4A). The differences at the sequence level are associated with different cofactors. JUN and FOSL1 were the predominant ATF3 co-factors in HCT116 and K562 cells, while USF proteins were found to be the main ATF3 partners in GM12878, H1-hESC and HepG2 (Figure 4B). Likewise, different partnerships lead to different binding loci. ATF3–USF complex was more inclined to occupy gene promoter regions, compared to the ATF3–JUN/FOSL1 complex (Figure 4B). Interestingly, ∼70% of the regions bound by the ATF3–USF complex were conserved with other cell types (Figure 4C), suggesting a shared function. More precisely, 704 ATF3–USF bound loci shared across all cell types were targeting genes involved in lysosome organization, intracellular transport and transferrin transport (Figure 4D). In contrast, the majority of binding sites bound by the ATF3–JUN/FOSL1 in HCT116 and K562 cells were cell-type specific (Figure 4C), and revealed a distinct set of biological functions (Figure 4D).

DISCUSSION

The dynamics of chromatin accessibility (41), epigenetic states (42) and TF cooperativity (43) across different cell types shape the genomic binding preference of a TF. Hence, it is logical to study TF genomic regulatory properties in a cell type-specific manner. In recent years, important efforts were devoted to investigate cell-specific TF binding activities (19,44–45). ENCODE-DREAM *in vivo* Transcription Factor Binding Site Prediction Challenge was established to promote the development and assessment of cell type-specific TF binding prediction tools, which were expected to replace time-consuming and expensive biological experiments such as ChIP-seq (44). However, performances of the state-of-the-art approaches proved to be limited, implying the necessity of further experiments to infer high confidence TFBS. Furthermore, a recent study has highlighted cell specificity in TFBS by introducing a modified protein binding microarray protocol, nextPBM (45). Interestingly, in this new protocol, the replacement of purified or overexpressed proteins with cell nuclear extracts enabled the consideration of cell-specific TF post-translational modifications and cooperativity. Nonetheless, nextPBM overlooks other important factors influencing TF binding, such as DNA methylation (46). Hence, an approach based on an *in vivo* and cell-specific framework is necessary to fully decipher TF genomic preferences. Furthermore, the collaborative interaction among TF is a common phenomenon across CRM at the genome scale, and such heterotypic combinations impact TF recognized DNA sequences. A systematic protein–protein interaction (PPI) analysis has provided compelling evidence for the widely spread heterodimerization across the bZIP family, which results in diversity of motif patterns (36). Therefore, it is appealing to study TF binding behaviors collectively instead of individually. In this respect, TFregulomeR enables combinatorial analyses of TF genomic binding propensities in an *in vivo* and cell-specific fashion.

TFregulomeR is linked to a large and up-to-date pool of TF cistrome derived from ChIP-seq experiments, along with the methylome landscape inferred from WGBS. TFregulomeR includes efficient and ready-to-use functionalities to ease the access, integration and analysis of large TFBS datasets. These manually curated datasets and designed functionalities together constitute a comprehensive toolbox to study CRM and TF interactions. We showed that the heterotypic dimerization of CEBPB with other bZIP family members leads to a degenerate (half-site) binding motif, which overall obscures the genome-wide motif pattern. Similarly, MAFF partners impact on the binding motifs recognized in different cell types. Interestingly, while TFregulomeR performs motif deconvolution by analyzing the TFBSs remaining in the context-specific peak regions, the significance of such TFBSs enrichment within the corresponding peak subsets was confirmed by *de novo* motif discovery for all our case studies (Supplementary Table S2). Users can therefore verify the statistical relevance of these novel TFregulomeR-generated motifs by exporting peak subsets and performing *de novo* motif discovery. Furthermore, similar to the case studies focusing on CEBPB and MAFF, in which ATF4 and NFE2 motifs were recovered in K562 CEBPB and MAFF peaks respectively, we were also able to identify CEBPB and MAFF motifs in K562 ATF4 and NFE2 peaks respectively (Supplementary Figure S5).

Even though efforts have been made to separate closely positioned TFBSs, approaches to effectively deconvolute motifs are still lacking. Although CSDeconv discriminates TFBSs separated as few as 40 bp (47), this spacer resolution apparently fails to distinguish two motifs closely positioned (*e.g.* MAFF and NFE2 motifs are only 1 bp distance in K562) or even to dissect heterodimer binding motifs (*e.g.* ATF4-CEBPB heterodimer in K562). Other motif discovery approaches, such as Maskminent (48), which enriches adjacent cofactor motifs by masking primary TF binding sites or *vice versa*, and dyad analysis (49), which

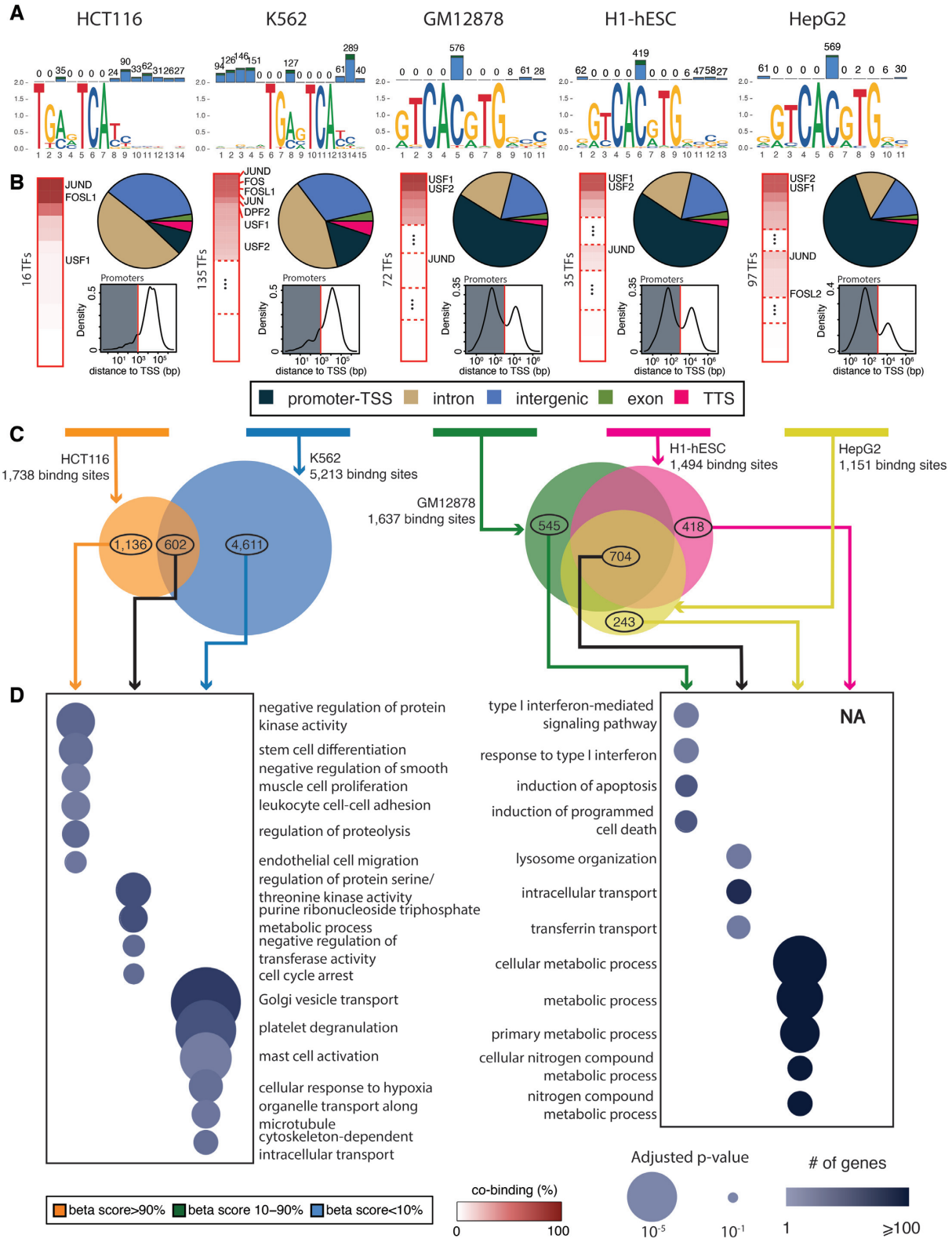

Figure 4. Analysis of ATF3 binding partners and functions. (**A**) Five MethMotif logos display the ATF3 sequence and DNA methylation preferences in five different cell types. The blue, orange and green bars stacked above motif logo denote the numbers of CGs homogenously unmethylated, homogenously methylated and heterogeneously methylated respectively. (**B**) The heatmaps represent the cofactor profiles around all ATF3 binding sites across five cell types. Furthermore, for each cell type, a pie chart illustrates the genomic locations of ATF3 binding loci, while the density plot profiles the distribution of distances between ATF3 peak summits and the nearest gene promoters. (**C**) The Venn diagrams denote the overlaps of ATF3 binding sites across different cell types. (**D**) The bubble plots show the enriched ontologies of targeted genes by different ATF3 binding subsets. The bubble size is proportional to the adjusted p-value, while the color intensity represents the number of targeted genes. Due to the long lists of gene ontology results from ATF3 binding loci in HCT116 and K562 cells, some informative terms were selected.

identifies two motifs separated by a spacer, cannot dissect a degenerate part in a motif from a mixture of homo- and hetero-dimers. This may be due to spacer conservation (for example, we found high-level of CG dinucleotide in CEBP-CEBP spacers), a feature that is ignored in current deconvolution algorithms, which expect spacers as 'non-relevant' sequences. Altogether, *in vivo* cistrome data empowers TFregulomeR as a reliable tool to dissect different TFBSs from a generally characterized PWM matrix. Several strategies could be adopted to achieve motif deconvolution using TFregulomeR. Firstly, it has been shown that main leucine zipper factors usually directly dimerize with their partners. Therefore, motif deconvolution can be done by selecting the subsets of sequence binding sites exclusive to the TFs and their partners (Supplementary Figure S5), or by removing overlapped peaks by co-factors (Figure 3D). Secondly, given the fact that TF partners can be cell specific, motif deconvolution thus can be done by selecting the subsets of sequence binding sites from cell specific peaks of a TF of interest (Figures 2 and 3B).

In addition, the methylome datasets encompassed by TFregulomeR revealed distinct DNA methylation patterns in different deconvoluted motifs, suggesting an interesting correlation between binding partners and DNA methylation states for both CEBPB and MAFF (Figures 2 and 3). The inclusion of generic annotation functions completes TFregulomeR toolbox and thereby ease the study of TF functions depending on partner combinations. Indeed, using TFregulomeR, we could characterize the cooperation of ATF3 with different TF partners depending on cell types, as well as the corresponding functions of targeted genes (Figure 4). Observed differences between target genes of a TF depending on its partner across cell types could be due to the distinct chromatin accessibility patterns.

In conclusion, combining a valuable compendium of cistrome and methylome data with effective computational tools, the current TFregulomeR release constitutes a comprehensive resource to characterize TF context-specific binding partners and enriched motifs, uniquely taking into account DNA methylation.

DATA AVAILABILITY

The TFregulomeR package (encoded in R) and a user manual are available in GitHub, together with all the scripts used in the reported analyses (https://github.com/benoukraflab/TFregulomeR).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Morgane Thomas-Chollier, Thomas J. Belbin and Daniel G. Tenen for useful discussions and comments during the preparation of this manuscript, as well as Mathanapriya Naidu for her proof-reading.
Author contributions': Q.X.X.L. and T.B. designed the package. Q.X.X.L. programmed the R-library and performed all analyses. 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 project.

FUNDING

National Research Foundation, the Singapore Ministry of Education under its Centres of Excellence initiative (to T.B, S.J.); the National Medical Research Council of Singapore [NMRC/BNIG/2035/2015 to T.B.]; Canada Research Chairs Program (to T.B.); Institut Français à Singapour [Merlion Project] [6.10.14 to T.B., D.T.]; 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 to T.B., S.J.]; Cancer Science Institute of Singapore Doctoral Scholarship (to Q.X.X.L); Ministry of Education Academic Research Fund [MOE AcRF Tier 1 T1-20 12 Oct-04, T1-2016 Apr-01 to S.J.]. Funding for open access charge: National Research Foundation, the Singapore Ministry of Education under its Centres of Excellence initiative.
Conflict of interest statement. None declared.

REFERENCES

1. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The human transcription factors. *Cell*, **172**, 650–665.
2. Bradner,J.E., Hnisz,D. and Young,R.A. (2017) Transcriptional addiction in cancer. *Cell*, **168**, 629–643.
3. Vogt,P.K. and Bos,T.J. (1990) jun:Oncogene and transcription factor. *Adv. Cancer Res.*, **55**, 1–35.
4. Hashimoto,H., Wang,D., Horton,J.R., Zhang,X., Corces,V.G. and Cheng,X. (2017) Structural basis for the versatile and methylation-dependent binding of CTCF to DNA. *Mol. Cell*, **66**, 711–720.
5. Bulyk,M.L., Huang,X., Choo,Y. and Church,G.M. (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 7158–7163.
6. Meng,X., Brodsky,M.H. and Wolfe,S.A. (2005) A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat. Biotechnol.*, **23**, 988–994.
7. Oliphant,A.R., Brandl,C.J. and Struhl,K. (1989) Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Mol. Cell Biol.*, **9**, 2944–2949.
8. Jolma,A. and Taipale,J. (2011) Methods for analysis of transcription factor DNA-Binding specificity in vitro. *Subcell Biochem.*, **52**, 155–173.
9. 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.
10. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-Wide mapping of in vivo Protein-DNA interactions. *Science*, **316**, 1497–1502.
11. Machanick,P. and Bailey,T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **27**, 1696–1697.
12. Thomas-Chollier,M., Herrmann,C., Defrance,M., Sand,O., Thieffry,D. and van Helden,J. (2012) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res.*, **40**, e31.
13. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408–1419.
14. Starick,S.R., Ibn-Salem,J., Jurk,M., Hernandez,C., Love,M.I., Chung,H.R., Vingron,M., Thomas-Chollier,M. and Meijsing,S.H. (2015) ChIP-exo signal associated with DNA-binding motifs provides insight into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. *Genome Res.*, **25**, 825–835.
15. Cohen,D.M., Lim,H.-W., Won,K.-J. and Steger,D.J. (2018) Shared nucleotide flanks confer transcriptional competency to bZip core motifs. *Nucleic Acids Res.*, **46**, 8371–8384.

16. Arnone,M.I. and Davidson,E.H. (1997) The hardwiring of development: organization and function of genomic regulatory systems. *Development*, **124**, 1851–1864.
17. Inukai,S., Kock,K.H. and Bulyk,M.L. (2017) Transcription factor–DNA binding: beyond binding site motifs. *Curr. Opin. Genet. Dev.*, **43**, 110–119.
18. Reiter,F., Wienerroither,S. and Stark,A. (2017) Combinatorial function of transcription factors and cofactors. *Curr. Opin. Genet. Dev.*, **43**, 73–81.
19. Lin,Q.X.X., Sian,S., An,O., Thieffry,D., Jha,S. and Benoukraf,T. (2018) MethMotif: an integrative cell specific database of transcription factor binding motifs coupled with DNA methylation profiles. *Nucleic Acids Res.*, **47**, D145–D154.
20. Zheng,R., Wan,C., Mei,S., Qin,Q., Wu,Q., Sun,H., Chen,C.-H., Brown,M., Zhang,X., Meyer,C.A. *et al.* (2019) Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res.*, **47**, D729–D735.
21. Chèneby,J., Gheorghe,M., Artufel,M., Mathelier,A. and Ballester,B. (2018) ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.*, **46**, D267–D275.
22. Oki,S., Ohta,T., Shioi,G., Hatanaka,H., Ogasawara,O., Okuda,Y., Kawaji,H., Nakaki,R., Sese,J. and Meno,C. (2018) ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep.*, **19**, e46255.
23. Yevshin,I., Sharipov,R., Kolmykov,S., Kondrakhin,Y. and Kolpakov,F. (2019) GTRD: a database on gene transcription regulation—2019 update. *Nucleic Acids Res.*, **47**, D100–D105.
24. 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.
25. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nussbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
26. 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**, D260–D266.
27. 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.
28. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
29. Nguyen,N.T.T., Contreras-Moreira,B., Castro-Mondragon,J.A., Santana-Garcia,W., Ossio,R., Robles-Espinoza,C.D., Bahin,M., Collombet,S., Vincens,P., Thieffry,D. *et al.* (2018) RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Res.*, **46**, W209–W214.
30. Tan,G. and Lenhard,B. (2016) TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics*, **32**, 1555–1556.
31. Wagih,O. (2017) ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics*, **33**, 3645–3647.
32. McLean,C.Y., Bristor,D., Hiller,M., Clarke,S.L., Schaar,B.T., Lowe,C.B., Wenger,A.M. and Bejerano,G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.
33. Lee,T.I. and Young,R.A. (2013) Transcriptional regulation and its misregulation in disease. *Cell*, **152**, 1237–1251.
34. Worsley Hunt,R. and Wasserman,W.W. (2014) Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. *Genome Biol.*, **15**, 412.
35. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-Binding specificities of human transcription factors. *Cell*, **152**, 327–339.
36. Rodríguez-Martínez,J.A., Reinke,A.W., Bhimsaria,D., Keating,A.E. and Ansari,A.Z. (2017) Combinatorial bZIP dimers display complex DNA-binding specificity landscapes. *Elife*, **6**, e19272.
37. Cai,D.H., Wang,D., Keefer,J., Yeamans,C., Hensley,K. and Friedman,A.D. (2008) C/EBPα:AP-1 leucine zipper heterodimers bind novel DNA elements, activate the PU.1 promoter and direct monocyte lineage commitment more potently than C/EBPα homodimers or AP-1. *Oncogene*, **27**, 2772–2779.
38. Jolma,A., Yin,Y., Nitta,K.R., Dave,K., Popov,A., Taipale,M., Enge,M., Kivioja,T., Morgunova,E. and Taipale,J. (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, **527**, 384–388.
39. Osada,S., Yamamoto,H., Nishihara,T. and Imagawa,M. (1996) DNA binding specificity of the CCAAT/enhancer-binding protein transcription factor family. *J. Biol. Chem.*, **271**, 3891–3896.
40. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
41. Miyamoto,K., Nguyen,K.T., Allen,G.E., Jullien,J., Kumar,D., Otani,T., Bradshaw,C.R., Livesey,F.J., Kellis,M. and Gurdon,J.B. (2018) Chromatin accessibility impacts transcriptional reprogramming in oocytes. *Cell Rep.*, **24**, 304–311.
42. Sardina,J.L., Collombet,S., Tian,T. V., Gómez,A., Di Stefano,B., Berenguer,C., Brumbaugh,J., Stadhouders,R., Segura-Morales,C., Gut,M. *et al.* (2018) Transcription factors drive Tet2-Mediated enhancer demethylation to reprogram cell fate. *Cell Stem Cell*, **23**, 727–741.
43. Thanos,D. and Maniatis,T. (1995) Virus induction of human IFNβ gene expression requires the assembly of an enhanceosome. *Cell*, **83**, 1091–1100.
44. Keilwagen,J., Posch,S. and Grau,J. (2019) Accurate prediction of cell type-specific transcription factor binding. *Genome Biol.*, **20**, 9.
45. Mohaghegh,N., Bray,D., Keenan,J., Penvose,A., Andrilenas,K.K., Ramlall,V. and Siggers,T. (2019) NextPBM: a platform to study cell-specific transcription factor binding and cooperativity. *Nucleic Acids Res.*, **47**, e31.
46. 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.
47. Lun,D.S., Sherrid,A., Weiner,B., Sherman,D.R. and Galagan,J.E. (2009) A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data. *Genome Biol.*, **10**, R142.
48. Lu,R., Mucaki,E.J. and Rogan,P.K. (2017) Discovery and validation of information theory-based transcription factor and cofactor binding site motifs. *Nucleic Acids Res.*, **45**, e27.
49. van Helden,J., Rios,A.F. and Collado-Vides,J. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.*, **28**, 1808–1818.