Nucleic Acids Research, 2024, **52**, 4234–4256
https://doi.org/10.1093/nar/gkae088
Advance access publication date: 13 February 2024
Gene regulation, Chromatin and Epigenetics

Lineage-determining transcription factor-driven promoters regulate cell type-specific macrophage gene expression

Gergely Nagy[1,*], Dóra Bojcsuk[1], Petros Tzerpos[1], Tímea Cseh[1] and László Nagy[1,2,*]

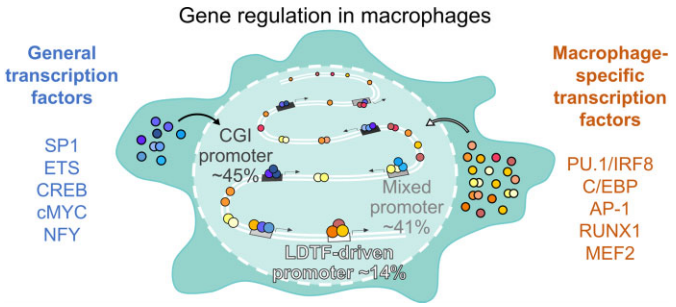
[1]Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Debrecen, Debrecen, Hungary
[2]Departments of Medicine and Biological Chemistry, Johns Hopkins University School of Medicine, Institute for Fundamental Biomedical Research, Johns Hopkins All Children's Hospital, St. Petersburg, FL, USA

[*]To whom correspondence should be addressed. Tel: +36 52 512 900 (Ext. 64616); Fax: +36 52 314 989; Email: nagygergely@med.unideb.hu
Correspondence may also be addressed to László Nagy. Email: lnagy@jhmi.edu

Abstract

Mammalian promoters consist of multifarious elements, which make them unique and support the selection of the proper transcript variants required under diverse conditions in distinct cell types. However, their direct DNA-transcription factor (TF) interactions are mostly unidentified. Murine bone marrow-derived macrophages (BMDMs) are a widely used model for studying gene expression regulation. Thus, this model serves as a rich source of various next-generation sequencing data sets, including a large number of TF cistromes. By processing and integrating the available cistromic, epigenomic and transcriptomic data from BMDMs, we characterized the macrophage-specific direct DNA-TF interactions, with a particular emphasis on those specific for promoters. Whilst active promoters are enriched for certain types of typically methylatable elements, more than half of them contain non-methylatable and prototypically promoter-distal elements. In addition, circa 14% of promoters—including that of *Csf1r*—are composed exclusively of 'distal' elements that provide cell type-specific gene regulation by specialized TFs. Similar to CG-rich promoters, these also contain methylatable CG sites that are demethylated in a significant portion and show high polymerase activity. We conclude that this unusual class of promoters regulates cell type-specific gene expression in macrophages, and such a mechanism might exist in other cell types too.

Graphical abstract



Introduction

Mammalian gene regulation is a complex process. This complexity starts with the fact that not only genes but any genomic regions flanking *cis*-regulatory elements can be transcribed (1–3). In most cases, transcription is aborted during pausing, but most genes and several non-coding regions show some degree of elongation (4). A key question is the identity and activity of promoters providing the transcriptional basis of cell type identity and specification. Based on transcriptomics studies, millions of transcription start sites (TSSs) have been located, although in most cases, the identity of *cis*-regulatory elements contributing to cell type-specific promoter activity remains uncertain and ill-defined (5–7). Some distinct sequences provide a clue about the location of TSSs, but these are typically infrequent such as TATA-box (up to 15%), indeterminate such as Initiator (up to 50%), or indefinite such as CpG islands (CGIs, up to 70%) (8–11). This means that based on DNA sequences alone, it is not possible to determine each promoter. However, by integrating the available data on DNA-protein interactions, one can characterize the active gene regulatory regions and among them, promoters.

Based on next-generation sequencing data targeting DNA-protein interactions, we already know several components of CGI promoters. Out of these, GC-box and ETS binding site (EBS) are the most common elements that are bound by SP1 and ETS family members, respectively (11–14). Some of these transcription factors (TFs) are likely to be present in all cell types. In addition, other ubiquitous and cell type-specific TFs bind to promoters or TSS-proximal enhancers beyond distal elements, such as NRF1, GFY, NFY (CCAAT-binding fac-

Received: March 6, 2023. **Revised:** January 18, 2024. **Editorial Decision:** January 20, 2024. **Accepted:** January 29, 2024

tor), and several bZIP and bHLH proteins (13,15,16). Altogether, there is a large pool of TFs, which supports cell- and condition-specific promoter and enhancer usage, but most of their interactions connecting *cis*-regulatory elements to polymerase activity are not known in detail. Considering the diversity and uniqueness of promoters, there is a significant gap in our knowledge regarding how the different combinations of *cis*-regulatory elements can serve as initiating points of transcription and how and what preinitiation complex can form in the lack of the well-known TFII binding sites such as TATA-box. In this study, we aimed to thoroughly and exhaustively map macrophage-specific direct TF binding sites and chromatin features to better understand how and which DNA-TF interactions initiate the cell type-specific gene transcription.

Murine bone marrow-derived macrophages (BMDMs) are an extensively used model for studying gene expression regulation in macrophages (1,11,17,18). This state represents differentiated macrophages, but these are not differentiated into a particular tissue-specific (e.g. liver, lung, peritoneum) or polarized (classical or alternative) state. Besides pathogen and cytokine signals, these cells are able to respond to a large number of other environmental stimuli such as proliferation signals, tissue damage, and lipid signal molecules. The required plasticity is provided by the collaboration of numerous lineage-determining and signal-dependent transcription factors (LDTFs and SDTFs, respectively) at the level of gene regulation. LDTFs are responsible for the generation and maintenance of a certain cell type, while SDTFs are the nuclear effectors of different signals. The major LDTFs of macrophages are PU.1 (*Spi1*) and FLI1 from the ETS family, C/EBPs from the bZIP family, and RUNX1 from the Runt family (13,19–21).

Besides these factors, there are other TFs that can be considered LDTFs in macrophages but also respond to the bacterial endotoxin lipopolysaccharide (LPS). These are the AP-1 (FOS/JUN; bZIP) and MEF2 family members, but IRF8—the heterodimerizing partner of PU.1 in several macrophage subtypes—has also been shown to have an LPS response independent of PU.1 (22–24). The classical SDTFs of macrophages are distinct IRFs and STATs that are all activated by cytokines and/or viral signals and have additional collaborating partners from other TF families (25). Interleukin-4 (IL-4) activates not only STAT6, but also cMYC (bHLH), EGR2, and the PPARγ/RXR heterodimer during alternative macrophage polarization (26–30). Similarly, the major SDTF of the LPS response is NFκB, but besides IRF8, AP-1, and MEF2, ATF2 (bZIP) is also induced during this kind of macrophage activation (25,31,32). There are additional bZIP proteins with signal-dependent activator functions in macrophages: ATF4 is induced upon amino acid deprivation, while NRF2 (*Nfe2l2*) is induced by tissue damage and oxidative stress (33,34). In addition to all of these activators, there are some signal-dependent repressors in macrophages too: BCL6 and ATF3 (bZIP) are both anti-inflammatory repressors, but while the former is inhibited, the latter is activated by LPS (31,32). BACH1 and MAFB are both signal-inhibited repressors (bZIP); the former is directly affected by heme, and the latter is inhibited by the proliferation signal of macrophage colony-stimulating factor (M-CSF/CSF-1) (35,36). These and additional TFs form an extended cell type-specific TF network, most of which interactions and their effects on gene expression are barely known.

In this study, we set out to systematically map all *cis*-regulatory elements with direct TF binding to characterize their epigenomic and transcriptional outcome in terminally differentiated unstimulated BMDMs. This was made feasible by the availability of a very large number (>40) of cistromic data, a quarter (>10) of which is generated by our laboratory. In order to achieve this, we collected all the available TF cistromes determined by ChIP-seq and representing sequence-specific DNA-protein interactions. In addition, we used DNA methylation (Bisulfite-seq), chromatin openness (ATAC-seq), histone modification (MNase-ChIP-seq), nascent transcriptomic (GRO-seq), and steady-state transcriptomic (RNA-seq and CAGE) data to probe the functional characteristics of the distinct TF binding patterns.

In contrast to previous studies using several, largely unexplored cistromes, here, we determined the exact recognition sequences for each TF and classified them based on the dominant TF(s). Importantly, certain elements of the same class having or lacking a CG dinucleotide show opposite characteristics. Methylatable but non-methylated elements are highly enriched in active promoters, while non-methylatable ones show typically promoter-distal distribution with less activity. By cataloguing all mapped *cis*-regulatory elements, we determined the characteristics of the most active promoters. Importantly, we identified a set of genes with promoters lacking all proximal elements, instead, these are composed exclusively of enhancer-specific ones. These promoters are CG-poor, bound by LDTFs, show moderate chromatin openness but a high number of acetylated H3K27 residues, and allow for the high level of expression of numerous macrophage-specific genes. This suggests that certain genomic regions with cell type-specific enhancer characteristics and bound by LDTFs are utilized as cell type-specific promoters to ensure proper gene expression patterns.

Materials and methods

Differentiation of BMDMs

Isolation and differentiation of bone marrow cells derived from male C57BL/6 mice were completed as described earlier (1,37). In detail, bone marrow was isolated by flushing femurs and tibiae with DMEM medium, and cells were purified through a Ficoll-Paque gradient. Cells were cultured and differentiated to macrophages for 6 days in DMEM medium containing 20% FBS and 30% conditioned medium of L929 cell line (as a source of M-CSF). Cells were seeded at 50 000 cells/cm^2 and were supplemented with the same medium at day 3 of differentiation.

Chromatin immunoprecipitation with sequencing (ChIP-seq)

ChIP was performed as described earlier (1,31,38). In detail, 10×10^6 adherent BMDMs were used as a starting input for each sample. Cells were first cross-linked with DSG (2 mM, Sigma-Aldrich) for 45 minutes and then with formaldehyde (1% v/v, Thermo Fisher Scientific) for 10 minutes at room temperature. Cells were washed and scraped in cold PBS. Cell pellets were resuspended and lysed on ice for 10 minutes (Lysis Buffer: 1% Triton X-100, 0.1% SDS, 150 mM NaCl, 1 mM EDTA and 20 mM Tris, pH 8.0) and sonicated with a Bioruptor sonicator in low strength for 10 minutes (30 s ON/30 s OFF). Immunoprecipitation was done overnight with 5 μg of antibodies for RXR (sc-774) and BACH1 (a gift from Dr Spilianakis, IMBB-FORTH, Greece), then BSA-blocked Protein A magnetic beads (Thermo Fisher Scientific)

were added for 2 h. Chromatin-bead complexes were washed five times; once with Wash Buffer 1 (1% Triton X-100, 0.1% SDS, 150 mM NaCl, 1 mM EDTA, 20 mM Tris, pH 8.0 and 0.1% NaDOC), twice with Wash Buffer 2 (1% Triton X-100, 0.1% SDS, 500 mM NaCl, 1 mM EDTA, 20 mM Tris, pH 8.0 and 0.1% NaDOC), once with Wash Buffer 3 (0.25 M LiCl, 0.5% NP-40, 1 mM EDTA, 20 mM Tris, pH 8.0 and 0.5% NaDOC), and once with TE buffer (1 mM EDTA and 20 mM Tris, pH 8.0). Immunoprecipitated chromatin was eluted, reverse-crosslinked overnight, and then treated with RNase A and Proteinase K. DNA was then column-purified with Qiagen DNA isolation kit. 1–10 ng of IP DNA was used for ChIP-seq library preparation with the TruSeq ChIP library kit (Illumina) according to the manufacturer's protocol. ChIP libraries were sequenced on an Illumina HiSeq 2500 platform.

Data collection

Besides our recent ChIP-seq data sets and our previously published ChIP-seq, ATAC-seq, GRO-seq and RNA-seq data sets, additional ChIP-seq, MNase-ChIP-seq, CAGE, and Bisulfite-seq data sets derived from terminally differentiated and unstimulated BMDMs were collected from NCBI's Sequence Read Archive (SRA) or Gene Expression Omnibus (GEO) as listed in Supplementary Table S1. In the case of availability of multiple ChIP-seq data sets for the same TF, data selection was done based on quality, considering the peak numbers, the specific signal-to-background ratio, and the specific motif enrichments. ATAC-seq data derived from splenic B cells and inguinal white adipose tissue (iWAT) cells were downloaded from SRA (Supplementary Table S1).

ChIP-seq analysis

Primary analysis

The primary analysis of raw sequence reads was carried out using the updated version of our ChIP-seq analysis command line pipeline (39). In detail, alignment to the mm10 mouse reference genome assembly was performed by the BWA v0.7.17 tool (40). BAM files were created by SAMtools v1.7 (41). Genome coverage (bedgraph) files were generated by makeUCSCfile.pl (HOMER v4.9.1) (13). The distribution of MNase-ChIP-seq reads at different genomic loci was determined by annotatePeaks.pl (HOMER). TF ChIP-seq peaks were predicted by MACS2 v2.1.1 (42), and artifacts were removed by intersectBed (bedtools v2.27.1) (43) according to the blacklist of ENCODE (44).

Filtering of TF ChIP-seq peaks

Peaks then were further filtered. Read density of the middle 100 bp of each peak was calculated by annotatePeaks.pl (HOMER). Density of the 1000th—or in the case of small bHLH cistromes, the 100th—peak was used for normalizing peak densities per sample, and peaks with lower than 1/3—or in the case of MEF2 cistromes, 1/2—normalized density were excluded. Cistromes with less than 1500 peaks after filtering were excluded from the further analyses.

Generation and clustering of aggregate cistromes

Cistromes of closely related TFs (basically TF families) were united to aggregate cistromes by mergeBed (bedtools). Peaks within aggregate cistromes were clustered based on their normalized densities (TF patterns) by Cluster 3.0 (45), and peaks within clusters were sorted based on one dominant TF. For clustering, the k-means clustering method was used assuming Euclidean distance with $k = 5$, 10, or 20 values, and clusters with similar patterns were united. For the co-localization frequency of TFs, the overlap of the cistromes was determined by intersectBed (bedtools). The distance of peaks relative to the closest TSS was determined by annotatePeaks.pl (HOMER). The average distance per 100 regions was determined in the order of the sorted aggregate cistromes.

ATAC-seq analysis

Primary analysis and integration with ChIP-seq data

The primary analysis of raw sequence reads was carried out as described above for ChIP-seq analysis. The distribution of ATAC-seq reads at different genomic loci was determined by annotatePeaks.pl (HOMER). Overlaps between ATAC-seq peaks and aggregate cistromes were determined using mergeBed and intersectBed (bedtools).

Comparison of cell type-specific open-chromatin regions

In order to compare the data derived from different cell types, a consensus peak set was generated by mergeBed (bedtools); fragment length was uniformly set to 150 bp; and densities determined by annotatePeaks.pl (HOMER) were decile normalized per sample. Peaks showing lower than 1/3 normalized density were filtered out. Overlaps of the cell type-specific ATAC-seq peaks were determined by intersectBed (bedtools).

Motif enrichment analysis

De novo motif enrichment analysis

The top 1000 peaks—or, in the cases of smaller peak sets, all peaks—were used for the *de novo* motif enrichment analysis. The top 1000 peaks of each cell type-specific or shared ATAC-seq peak subset were selected based on MACS2 peak scores. In the case of those BMDM-derived ATAC-seq peaks, which do not overlap with any of the aggregate cistromes, the top 2000 peaks were used. For promoter-specific motif optimization, the top 1000 ATAC-seq peaks with a given motif—enriched in the CREB1- and ELF1-specific clusters—were selected. For this, motif mapping was carried out by annotatePeaks.pl (HOMER). The top 1000 ChIP-seq peaks—of entire cistromes (RUNX1, RXR) or the clusters of aggregate cistromes—were selected based on the normalized read densities for each possible TF dominating a larger set of peaks. For example, separate analyses were carried out with the top peaks specific for both ELFs and those specific for ELF4 only. Similarly, C/EBPβ-, ATF3-, and JUNB-specific peaks also served as the basis for a motif search. The central 200 bp of ATAC-seq peaks and the central 100 bp of ChIP-seq peaks were used as target sequences, and the enrichment of 10-, 12- and 14-mers was determined by findMotifsGenome.pl (HOMER). For GFY motif optimization, 30-mers were set for the analysis. *P*-values were calculated by comparing the number of target and random (background) sequences carrying a certain motif.

Motif analyses

Motif mapping was carried out by annotatePeaks.pl (HOMER). The sequence of putative elements was obtained by the 'homerTools extract' command. The frequency of core sequences and half-sites was calculated for each TF family. The fold enrichment of bZIP-specific 8-mers was calculated relative to the frequencies expected based on the

frequencies of their constituent half-sites. Putative promoter-specific elements—determined within ATAC-seq peaks—were discriminated based on their CG content. Per position nucleotide frequencies were transformed to HOMER motif matrices. The distance distribution of putative elements relative to the closest TSS was determined by annotatePeaks.pl (HOMER).

Sequence enrichment analysis

Sequence enrichment analyses were performed per cluster and aggregate cistrome (the filtered RUNX1 and RXR cistrome can be considered both) (see flowchart in Supplementary Figure S2E).

Determination of consensus sequences

Consensus sequences—determined manually based on the related *de novo* motif hits (Supplementary Table S2)—were transformed to HOMER motif matrices by seq2profile.pl (HOMER). For example, no ETS motif contained a T right downstream of the core 4-mer, so oligomers having this nucleotide at this position were not taken into account in further analyses. The length of flanking nucleotides was determined based on their nucleotide preferences according to the *de novo* motif hits. In the case of ETS and IRF elements with a tetrameric core, 8-mers including 2 × 2 flanking nucleotides were used as initial sequences. Similarly, 2 × 2 flanking nucleotides of the core 6-mer of RUNX elements were taken into account, and 2 nucleotides upstream and 1 nucleotide downstream of the core 6-mer of RXR were included in the initial set of sequences. In the case of MAF/antioxidant response element (MARE/ARE), only a 1-nucleotide extension of the short half-site was included, and in all other cases, a 1-nucleotide extension of both ends was applied (Supplementary Table S2).

Determination of oligomer enrichments

The enrichment of individual oligomers was calculated based on their frequency in each relevant set (cluster) of peaks and their flanking regions. More precisely, the central 100 bp of the peaks was used as target regions, and the flanking 2 × 500-bp regions were used as background (Supplementary Figure S2E, right). For this, the motif matrices representing the initial oligomers were mapped within the 1.1-kb (target + background) regions around the peaks by annotatePeaks.pl (HOMER). Putative elements of the target and background regions were separated by intersectBed (bedtools). The sequence of mapped oligomers was retrieved by the 'homerTools extract' command. The number of individual oligomers mapped in the target regions was compared to that in the background (Supplementary Figure S2E, right). Oligomers were considered enriched in a set of target regions if they reached a certain frequency—e.g. 1% in the case of 8-mers—and 3-fold enrichment over the background, but in the case of lower complexity or longer motifs, respectively stricter or more permissive frequency requirements were applied. In the case of RUNX- and bZIP-specific 10-mers, the target frequency threshold was reduced to 0.3%. In contrast, the low-complexity MEF2 and incomplete cAMP response element (CRE) oligomers required a higher, 5-fold enrichment threshold with at least 0.5% target frequency. In the bZIP-specific clusters of less than 1000 peaks, higher target frequency thresholds were applied (Supplementary Table S2, parentheses).

Filtering of the enriched oligomers

The enriched oligomers were subjected to further manual filtering. Those oligomers showing a high degree of overlap with sequences specific for other TF families were excluded. For example, the TTTCCTCA sequence is enriched as a bZIP-specific 8-mer similar to the C/EBP:ATF response element (CARE), but it is rather bound by AP-1 proteins and contains the PU.1-specific GAGGAAA sequence. Sequence similarities within the bZIP family were also considered. TPA response elements (TREs), being parts of MAREs, and incomplete CREs, being parts of complete ones, were excluded. In order to simplify the visualization of bZIP element-related results, the enrichments of oligomers with identical cores and different flanking nucleotides were re-calculated based on the summed oligomer frequencies.

Determination of TF preferences

All putative elements matching with the enriched oligomers were collected within the relevant aggregate cistrome. ChIP-seq read density for the related TFs was calculated at 50-bp regions around each putative element by annotatePeaks.pl (HOMER). Densities were normalized with the same values as for ChIP-seq peak filtering, and then the per oligomer median of the normalized TF densities was determined. In the case of ETS (+IRF8) and bZIP families, pairwise correlation analyses were performed on the per oligomer TF patterns, and the matrix of Pearson correlation coefficients was clustered by Cluster 3.0. Median TF densities were sorted in the order of the result of this hierarchical clustering. Sequence enrichments specific for TF patterns (clusters) were expressed as the product of the frequency and fold enrichment of each oligomer mapped in each cluster. The 'specific enrichment' of an oligomer within a cluster dominated by a TF then was coupled with the median density of the TF at the oligomer.

Classification of elements

The putative elements matching with the enriched oligomers were classified based on their DNA-TF interaction characteristics and CG content (Supplementary Table S3). The distance distribution of classes of putative elements relative to the closest TSS was determined by annotatePeaks.pl (HOMER).

GRO-seq analysis

Primary analysis

The primary analysis of raw sequence reads was carried out as described above for ChIP-seq analysis. Strand-specific genome coverages were determined by makeUCSCfile.pl (HOMER) and united to a single coverage (bedgraph) file with positive and negative signs representing the two strands. Read densities at different genomic loci were calculated by annotatePeaks.pl (HOMER).

Integration of GRO-seq and ATAC-seq data

GRO-seq read densities around ATAC-seq peaks were determined in 100-bp resolution (Supplementary Figure S1B). Around the central 100 bp, the higher out of the average density of the adjacent and subsequent 100-bp regions was considered the transcription initiation density. The average elongation density calculated for 100 bp was measured from 550 to 1550 bp relative to the peak centre in both directions (Supplementary Figure S1B). Densities for the related chromatin openness (measured at 100-bp peak centres) and tran-

scription initiation values and also for the related initiation and elongation values were calculated in 0.5-unit bins.

Putative elements matching with the enriched oligomers or the optimized promoter-specific motifs—mapped in the relevant regions—were also used as centres of chromatin openness and transcription initiation analyses. If two and only two putative elements were located within 50 bp distance relative to each other, they were united into a pair by mergeBed (bedtools). Elements farther than 50 bp from any other element were handled as single elements. ATAC-seq density within 100 bp around the single and double elements and the average GRO-seq density within the flanking 100-bp regions were determined using annotatePeaks.pl (HOMER). Median ATAC-seq and GRO-seq densities for each class of elements and each combination of classes represented by at least 19 loci were determined.

Calculation of gene coverages

In the initial comparison with RNA-seq data, reads per kilobase per million mapped reads (RPKM) values were calculated per gene according to the mm10 reference gene annotation. For this, genes were split into 500-bp fragments, an RPKM value was calculated for each fragment, and the median value was assigned to each gene. Reads mapping strand-specifically to the gene fragments were counted using intersectBed (bedtools), and identical reads were considered once.

De novo transcript prediction

De novo transcript prediction was carried out using the updated version of our GRO-seq analysis command line pipeline (Supplementary Figure S6A) (1). Shortly, strand-specific 'peaks' representing polymerase activity were determined based on the strand-specific genome coverage (bedgraph) file by PeakSplitter (EBI Bertone Group Software). Peak pairs representing divergent transcription (transcription initiation to both directions) were determined using intersectBed (bedtools). Consecutive peaks at the same strand were united to transcripts based on proximity and the mm10 reference gene annotation extended as described below for RNA-seq analysis. The 5′ end (TSS) of putative transcripts was determined as described below for promoter analysis. RPKM values were calculated for each transcript excluding—putative initiation—peaks showing >2.5-fold read enrichment relative to the whole gene body (Supplementary Figure S6A). Identical reads were considered once during the calculations.

RNA-seq analysis

Primary analysis

Paired-end reads were aligned to the mm10 mouse reference genome assembly by hisat2 v2.1.0 (46). BAM files were created by SAMtools. Genome coverage (wig) files were generated by bamCoverage2 (deepTools v3.0.2) (47). Gene expression levels determined in fragments per kilobase per million mapped fragments (FPKM) were calculated by StringTie v1.3.4d (48). The average FPKM values of two replicates were shown.

Integration of GRO-seq and RNA-seq data

Genes of the upper decile of gene expression values either based on GRO-seq (RPKM) or RNA-seq (FPKM) were selected for further classification. Non-protein-coding genes were determined using the Ensembl database (BioMart). Protein-coding genes were discriminated based on the fold difference between the expression levels measured by RNA-seq and GRO-seq. As the quotient of the median gene expression levels (FPKM/RPKM) is 22.23, two orders of magnitude to both directions were set as thresholds, meaning that the middle range is between 22.23 multiplied/divided by the square root of 10, and the further thresholds are orders of magnitude above/below relative to these values, respectively. According to this classification, 'very low turnover' genes show a fold difference higher than 703; 'low turnover' genes show a fold difference between 70.3 and 703; 'average turnover' genes show a fold difference between 7.03 and 70.3; 'high turnover' genes show a fold difference between 0.703 and 7.03; and 'very high turnover' genes show a fold difference <0.703. Gene ontology (GO) analyses were performed by ShinyGO (49).

Classification of transcription factors

Genes of the macrophage-specific TF families were collected based on the list of Zhou et al. (50) and in the case of the bHLH family, Skinner *et al.* (51). Besides the major TF families (ETS, IRF, bZIP, STAT, bHLH and NFκB), gene expression of the TFs in the upper decile and the related co-regulators was collected. TFs having or lacking a major cistrome, and those lacking any cistrome were discriminated.

De novo transcript prediction

De novo transcript prediction was carried out by StringTie using all reads from both replicates. The predicted transcripts were annotated using intersectBed (bedtools) and the mm10 reference gene annotation. Transcripts extended by at least 100 bases to the 5′ direction relative to the most upstream known TSS of an overlapping gene were added to the reference gene annotation.

Promoter analysis

TSS prediction based on GRO-seq data

GRO-seq peaks overlapping with any 5′ UTR within a first exon according to the mm10 reference gene annotation and/or the *de novo* RNA-seq predictions were considered putative initiation peaks—that follow promoters and may be followed by elongation. In addition, putative TSSs were selected out of the intragenic sites showing divergent transcription. More precisely, if a gene segment downstream of an intragenic divergent initiation site showed a coverage at least 50% higher than the previous one, it was considered the result of the activity of an alternative promoter (Supplementary Figure S6A). At first, the 5′ of the found initiation peaks ±150 bp was considered the 'TSS region' of the identified transcripts.

Promoter prediction and classification

Putative elements matching with the enriched oligomers or the optimized promoter-specific motifs—mapped in the relevant regions—were supplemented with those matching with the enriched oligomers and mapped in the common peaks of the relevant unfiltered cistromes and open-chromatin regions. Elements closer than 100 bp to each other were clustered by mergeBed (bedtools). Single or clustered elements with at least half of their length covered by any of the predicted TSS regions were considered putative promoters. Putative promoters composed exclusively of prototypically TSS-proximal elements, distal elements, or both kinds of elements including those showing bimodal distribution were discriminated. ATAC-seq

and GRO-seq read densities were determined as described for GRO-seq analysis (Supplementary Figure S1B).

CAGE (cap analysis of gene expression) analysis

Definition of TSSs

Per TSS transcription initiation frequency from BMDMs was downloaded as a bed file from the FANTOM5 database (CNhs14136). If there was overlap, the top TSS—according to transcription initiation frequency—was assigned to each TSS region (determined as described above for promoter analysis) using intersectBed (bedtools). Otherwise, the centre of the TSS regions was used. TSS clusters within ±100 bp of the top TSSs were selected based on CAGE data, and the width of the interquartile region (IQR) of these clusters—over 10 transcripts per million (TPM)—was collected per promoter type.

Data integration

Genes having exclusively those classes of promoters composed of prototypically TSS-proximal elements and those having exclusively those classes of promoters composed of prototypically TSS-distal elements were further filtered based on expression. The TSS of transcripts with an expression level >1 RPKM was collected. In addition, TSSs of (very) low turnover protein-coding genes <0.5 RPKM expression and those of protein-coding genes not expressed according to either GRO-seq or RNA-seq were collected. These four groups of promoters/TSSs then were characterized by CAGE, motif distribution, Bisulfite-seq, ATAC-seq, MNase-ChIP-seq, and GRO-seq data. TSS densities according to CAGE data were determined by annotatePeaks.pl (HOMER).

Bisulfite-seq analysis

The methyl-cytosine map of C57BL/6 BMDMs was downloaded from the GEO database (GSM2974655). Methylated CG dinucleotides were collected to a bed file, which was used to map the DNA methylation pattern of different genomic loci by annotatePeaks (HOMER). CG dinucleotides (independent of their methylation) and the identified and grouped TSS-proximal (EBS, enhancer [E]-box, GC-box, CCAAT-box and NRF1 element) and distal elements (PU-box and bZIP elements except for the CREB1-specific ones) were mapped by annotatePeaks (HOMER). The methylation pattern of promoters composed exclusively of proximal elements was modelled by Gaussian functions and a linear function.

Data visualization

Genome coverage (bedgraph and wig), gene annotation (gtf), and genome coordinate (bed) files were visualized by Integrative Genomics Viewer (IGV v2.16.2) (52). HOMER motif matrices were visualized by motif2Logo.pl (HOMER). Heat maps representing the distribution of reads, putative elements, CG dinucleotides, or TSSs at different genomic loci and those representing different enrichments, frequencies, and densities were visualized by Java TreeView v1.1.6r4 (53). Histograms representing the distribution of reads, nucleotides, CG dinucleotides and other sequences, or TSSs at different genomic loci, histograms representing the distribution of peaks relative to the closest TSS, violin plots representing the distance distribution of cistromes or putative elements relative to the closest TSS, violin plots representing ATAC-seq and GRO-seq densities, as well as scatter plots, box plots, bar charts, and a pie chart were visualized by GraphPad Prism v8.0.1. Proportional Venn diagrams were visualized by VennMaster (54), and the Chow-Ruskey Venn diagram was visualized by Intervene (55).

Results

The cell type-specific chromatin landscape is determined by specific LDTF combinations

Specific combinations of LDTFs result in different cell types. For example, the retroviral expression of PU.1 and C/EBPα can reprogram different kinds of fibroblasts into macrophage-like cells, while pre-B cells, which already express PU.1, can be transdifferentiated into macrophages by the overexpression of C/EBPα only (20,21). In order to compare the open-chromatin regions and the motif enrichments of the major LDTFs of murine BMDMs with those of other normal cell types, we used ATAC-seq data also from murine splenic B cells and inguinal white adipose tissue (iWAT) cells (56,57). The latter was chosen because macrophages and adipocytes share C/EBPα but not PU.1 (58,59). Despite the common LDTFs, the three cell types show significant differences in their open-chromatin regions (Figure 1A). Motif enrichment analysis within the regions common to all cell types ($n = 9162$) resulted in only promoter-specific motif hits (Supplementary Figure S1A). Pairwise intersections, in turn, show more specific hits, except for that of B cells and iWAT cells ($n = 1350$), which is enriched only for insulator (CTCF) elements and—the basically TSS-proximal—CCAAT-box. According to these results, BMDM- and/or B cell-specific regions both show PU.1 and RUNX motif enrichments (purple), BMDM- and/or iWAT-specific regions both have C/EBP and AP-1 (TRE) motifs enriched (red), and CRE and enhancer (E)-box motifs (black) enriched the most in macrophages (Supplementary Figure S1A). These results confirm that the cistrome of a certain LDTF can be diverse in different cell types, largely affected by the actual collaborating TFs (60,61). In addition, the macrophage-specific motif hits suggest the possible roles of multiple bZIP and—the E-box binding—bHLH proteins besides the well-known LDTFs.

Transcription initiation is a diverse process in differentiated BMDMs

The genome-wide collection of open-chromatin regions provides a robust and unbiased basis to test genomic and epigenomic features corresponding to transcription initiation. Therefore, we started our analyses with the 42951 open-chromatin regions (peaks) determined based on our ATAC-seq data obtained from murine BMDMs. In order to examine the correlations between chromatin openness and transcription initiation, we compared ATAC- and GRO-seq densities at the predicted ATAC-seq peaks (Figure 1B, Supplementary Figure S1B). In this comparison, no statistical correlation and, as expected, no transcription initiation without open chromatin is detected. Interestingly, high initiation frequency can be reached at barely open regions, and naturally, also the opposite can be observed at highly open regions with low or no expression. Comparing how elongation follows transcription initiation based on GRO-seq data around the ATAC-seq peaks, we detected a broad distribution again ($r = 0.11$) (Supplementary Figure S1B, C). These results demonstrate that each open-chromatin region (putative gene regulatory region) is unique, and there is no statistical correlation between chromatin openness and transcription initiation, rather, these

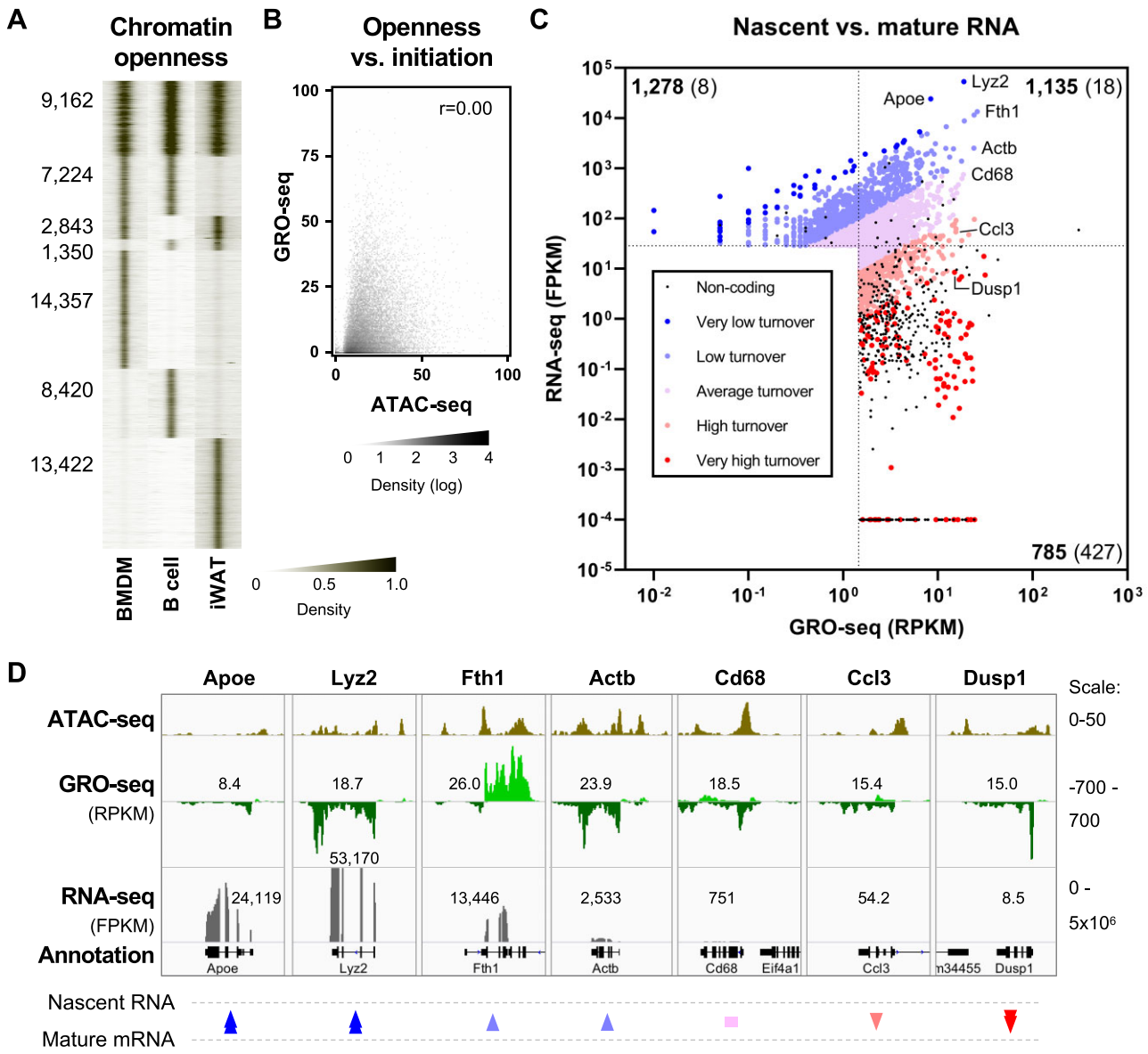


Figure 1. Correlations in BMDM gene regulation. (**A**) Read distribution plot depicts chromatin openness (ATAC-seq) in 1-kb windows around the different sets of ATAC-seq peaks derived from BMDM (SRR6246988), B cell (DRR277283), and iWAT (SRR6177788-89) samples. (**B**) Density plot represents correlations between the chromatin openness (ATAC-seq) and transcription initiation (GRO-seq) of BMDMs. (**C**) Scatter plot represents correlations between the nascent (GRO-seq) and mature RNA levels (RNA-seq) of BMDMs. The elevating turnover rate of protein-coding genes of the upper deciles of gene expression values is indicated as blue to red transition per order of magnitude. Non-protein-coding genes are represented by black dots. The number of protein-coding (bold) and non-coding genes (in parentheses) are shown for the top three quadrants determined by the upper deciles. (**D**) Genome browser view shows the chromatin openness (dark olive), nascent transcription (light and dark green, strand specifically), and mRNA level (dark grey) of representative genes selected based on their high transcription frequency and diverse mRNA levels. Colour code below fits the code introduced in Figure 1C.

are—at least in part—independent outcomes of DNA-protein and protein–protein interactions within the chromatin.

RNA turnover shows a broad distribution in BMDMs

The next level of gene expression is the amount of mature RNA molecules; therefore, our next question was whether there is a correlation between the nascent (GRO-seq) and mature RNA levels (RNA-seq) in our model system (unstimulated BMDMs) (Figure 1C). Again, the answer is no ($r = 0$); there can be several orders of magnitude difference in both directions between the gene expression values measured by the two techniques. Less than a third ($n = 1153$) of the genes in the upper decile either based on GRO-seq or RNA-seq are top genes based on both techniques (Figure 1C, top right). If we consider only protein-coding genes ($n = 1135$), the ratio is somewhat higher (35.5%) because of the large number of non-coding transcripts ($n = 427$, highlighted as black dots) that show frequent transcription but a low mature RNA level (Figure 1C, bottom right). Similar to most non-coding transcripts, several ($n = 59$) histone genes also show high turnover and low RNA levels (from the 131 'very high turnover' protein-coding transcripts; highlighted as red

dots), resulting in the enrichment of the GO term 'nucleosome assembly' in this gene set (Supplementary Figure S1D). On the other side, accumulated RNA molecules show the lowest ratio of non-coding transcripts (Figure 1C, top left), and the extremes ($n = 43$ genes, highlighted as blue dots) are enriched for immune-related and acetylcholine response genes (Supplementary Figure S1E). Overall, huge differences can be observed between the amounts of the produced and processed transcripts—the changes range between 5 and 8 orders of magnitude, respectively (Figure 1C).

In order to demonstrate the differences between the gene expression levels detected by GRO-seq and RNA-seq, we highlighted some genes with similarly high transcription frequency but different steady-state mRNA levels (Figure 1D). Beta-actin (*Actb*), which is frequently used as a representative housekeeping gene, has a relatively low mRNA turnover, while the macrophage-specific lysozyme M (*Lyz2*) is the highest expressed gene with very low mRNA turnover (62): *Lyz2* transcripts are ∼21-times more frequent than *Actb* transcripts. The mRNA level of *Fth1* is a quarter of this, while *Ccl3* and *Dusp1* show low expression levels relative to the polymerase activity on their gene bodies. Notably, the high initiation peak of *Dusp1* is not coupled with high chromatin openness, demonstrating the complexity of transcriptional regulation. In our further investigations, we aimed to identify what sequence characteristics are behind these, sometimes contradictious chromatin features.

The expression of LDTFs correlates with the size of their cistromes

Since the active TF and *cis*-regulatory element network is the key to gene regulation, first, we classified TFs based on their expression level and cistrome size using BMDM-derived RNA-seq and ChIP-seq data sets, respectively. Results from the former inform about the approximate expression level of protein products, while those from the latter confirm the expression and possible activity by showing the chromatin-bound fraction of the expressed TFs. We compiled 11 of our own and 30 additional publicly available cistromes with specific motif enrichment and compared the 'raw' (unfiltered) cistrome sizes with the gene expression values (Supplementary Figure S2A). Overall, there is no statistical correlation between these values; although three populations separate: (i) high-expressed TFs with large cistrome, (ii) high-expressed TFs with small cistrome and (iii) low-expressed TFs with small cistrome. These respectively represent major TFs, TFs to be activated, and minor TFs (referred to as groups I-III, respectively). In group II, among others, we found IRFs, STATs, and nuclear receptors, which are known SDTFs with immediate or quick response (25,63), while group I TFs, such as ETS and bZIP proteins (highlighted in blue and orange, respectively), are already active components of the chromatin.

Cistrome size (the number of ChIP-seq peaks) largely depends on technical issues such as the specificity and affinity of the antibody. In our case, a high-resolution sample could result in close to 100 000 peaks (Supplementary Figure S2B, Supplementary Table S1), most of which have a read density more than an order of magnitude lower than the top peaks—these can represent low-affinity, indirect, heterochromatic, or cell subpopulation-specific binding sites even if BMDMs form pretty homogenous populations. In contrast, low-resolution samples may show only the tip of the iceberg with mostly direct binding sites. In order to make samples comparable, the density of the 1000th peak was used for the normalization of all peak densities per sample, and peaks with less than 1/3 normalized density were not included in the further analyses (Supplementary Figure S2B, red line). This strict cut-off excluded 15 TFs (from groups II–III) and provided more comparable cistrome sizes and highly occupied binding sites, so a reasonable basis for the subsequent motif analyses. CTCF was also excluded from the TF list because of its unique role in insulator binding and looping instead of direct transcription regulation (64). After these filtering steps, cistrome size showed a higher correlation ($r = 0.26$) with gene expression, and 11 out of 26 TFs fell within the 95% confidence interval of the fitted line, including all C/EBPs, MEF2 proteins, and PU.1 (Spi1) (Supplementary Figure S2C). Using this approach, all AP-1 components (FOS/JUN), ATF2 and 2 ETS proteins show 'overestimated' cistrome size, while 3 cistromes seem slightly, and 5 strongly underestimated. This may mean that the former TFs have more high-affinity binding sites than the others, while the latter ones bind to a large number of low-affinity or indirect binding sites. Notably, unlike the other TFs, cMYC is probably unable to saturate its possible binding sites because of its lower expression (Supplementary Figure S2C, grey dashed line).

After selecting the available top cistromes, we grouped the high-expressed TFs based on their families using the extended TF list of Zhou *et al.* (Figure 2A, the red dashed line marks the upper decile of TF gene expression) (50). Families of the highest expressed TFs are all represented by relevant ChIP-seq data (highlighted in blue, orange, red or grey), although there are some additional TFs, whose genomic distribution would be interesting based on their expression level (highlighted in black). For example, several members of the bHLH family show high expression, but only 3 bHLH cistromes (of different subfamilies) are known from BMDMs (red and grey). In the case of ETS and IRF proteins, we have the most relevant cistromes (blue and red), as well as the bZIP family is highly represented (orange). It is also visible that IRFs (except for IRF8), STATs and NFκB proteins (grey) have small cistromes, which can be extended by TF activation, basically via phosphorylation (25). Notably, within these families, we can suppose that the unknown cistromes should show similarly small numbers of occupied *cis*-regulatory elements. For example, as IRF7 is a heterodimerizing partner of IRF3, it might have a similarly small cistrome as its partner; inhibitors of IRF2 (Irf2bp, green) show high mRNA levels; and the highly expressed NFκB inhibitors (Nfkbi, green) also take care of repression (25). Interestingly, both members of the TBP inhibitor DR1/DRAP1 heterodimer (green, column 'Other') show high expression, which might result in the suppression of classical TATA promoters (65). The last TFs with major cistromes are the RXR, RUNX1 and MEF2 proteins (red, column 'Other').

As highlighted above, in total, we collected 26 cistromes of significant TFs in terminally differentiated unstimulated BMDMs—13 of the bZIP family, 5 of the ETS family, 3 of the MEF2 family, 2 of the bHLH family, and those of IRF8, RXR and RUNX1. Notably, RXRα and β were detected with the same antibody, so hereinafter, they are simply referred to as RXR. Most of these TFs show average turnover, except for most bZIP proteins (especially AP-1 proteins) and cMYC, which show high turnover, and PU.1, which shows the lowest turnover (Supplementary Figure S2D). The selected cistromes of the 7 TF families then were processed as follows (Figure 2B,

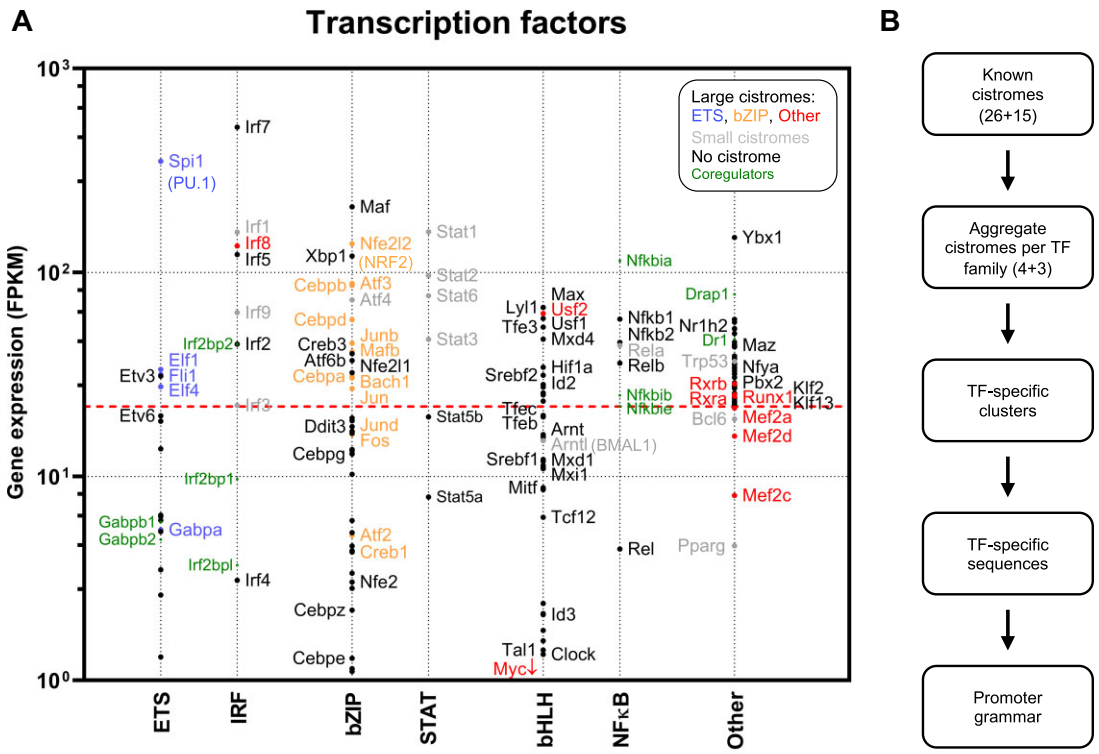


Figure 2. The major transcription factors of BMDMs. (**A**) Dot plot represents the gene expression level of the highest expressed TFs per TF family. The red dashed line represents the upper decile, below which only the expression levels of the top families are shown. TFs in blue, orange, or red have large cistrome (based on our filtering detailed in the Methods section), TFs in grey have small cistrome, TFs in black have no available cistrome, and genes in green are transcriptional co-regulators. (**B**) Flowchart represents the applied workflow of the analyses.

Supplementary Figure S2E): Because of the common or overlapping sequence preferences of TFs from the same family, we decided to combine their cistromes into a non-redundant, aggregate cistrome, and then, this was clustered to regions with characteristic TF and DNA sequence patterns. The goal of these analyses was to discriminate particular sequences specific for a certain TF or TF group, from which promoters and enhancers can be built up. Finally, sequence features of active regulatory regions were combined with the available epigenomic information to get to know how transcription is initiated (Figure 2B).

The large TF families have both promoter- and enhancer-specific members

In order to characterize the epigenomic features of the macrophage-specific cistromes, first, we compared the binding sites of the macrophage LDTF PU.1 (ETS) and its possible collaborating partners such as IRFs. Out of the 4 known IRF cistromes, besides IRF8 ($n = 39\ 097$), only IRF1 shows a relatively higher number (n = 6695) of binding sites, but by our strict filtering, these numbers were strongly reduced (to 5330 and 1103, respectively). As most IRF1 binding sites are occupied by both IRFs, we decided to continue with the larger cistrome (Supplementary Figure S3A). As expected, IRF8-dominated regions show motif enrichments specific for its heterodimer with PU.1—the ETS:IRF composite element (EICE) and IRF:ETS composite sequence (IECS) (66–68). The close collaboration between these TF families allowed us to analyse the IRF8 and ETS (PU.1, ELF1/4, FLI1 and GABPα) cistromes together in the frame of the aggregate ETS cistrome (Figure 3A). Besides the 5 specific clusters generated based on the normalized TF densities (left), we also plotted the local densities of Bisulfite-seq, ATAC-seq, and GRO-seq (middle), together with the distance distribution of regions relative to the closest TSS (right). The 5 clusters are well separated in all their characteristics, although ELFs show a common and an ELF4-specific subcluster, and the bottom of the FLI1-specific cluster is rather bound by GABPα and ELF1 than FLI1. GABPα typically co-localize with both ELFs and FLI1, and their common binding sites are the least methylated (Bisulfite-seq), highly open (ATAC-seq), the most transcribed (GRO-seq), and typically TSS-proximal (Figure 3A). In contrast, PU.1 and IRF8-specific regions are promoter-distal with a narrow unmethylated lane, less chromatin openness, and low transcription initiation frequency. ELF4 and FLI1-specific regions are between these extremes in all respects.

Next, we investigated the cistrome of the 13 bZIP proteins, which form a higher number of clusters and subclusters (Figure 3B). The C/EBP-dominated cluster can be separated into common and C/EBPβ/δ-specific regions, but both are of promoter-distal characteristics and can be bound by other bZIP proteins except for CREB1. According to these results, CREB1 is a promoter-specific TF—not only because of its TSS-proximal enrichment, but also the high transcriptional activity, the high chromatin openness, and the lack of DNA methylation in these regions. CREB1-specific regions show the presence of most of the examined bZIP proteins, but a mutual exclusion can be observed with both C/EBP and ATF3 proteins. In addition, BACH1 and MAFB also show low affinity

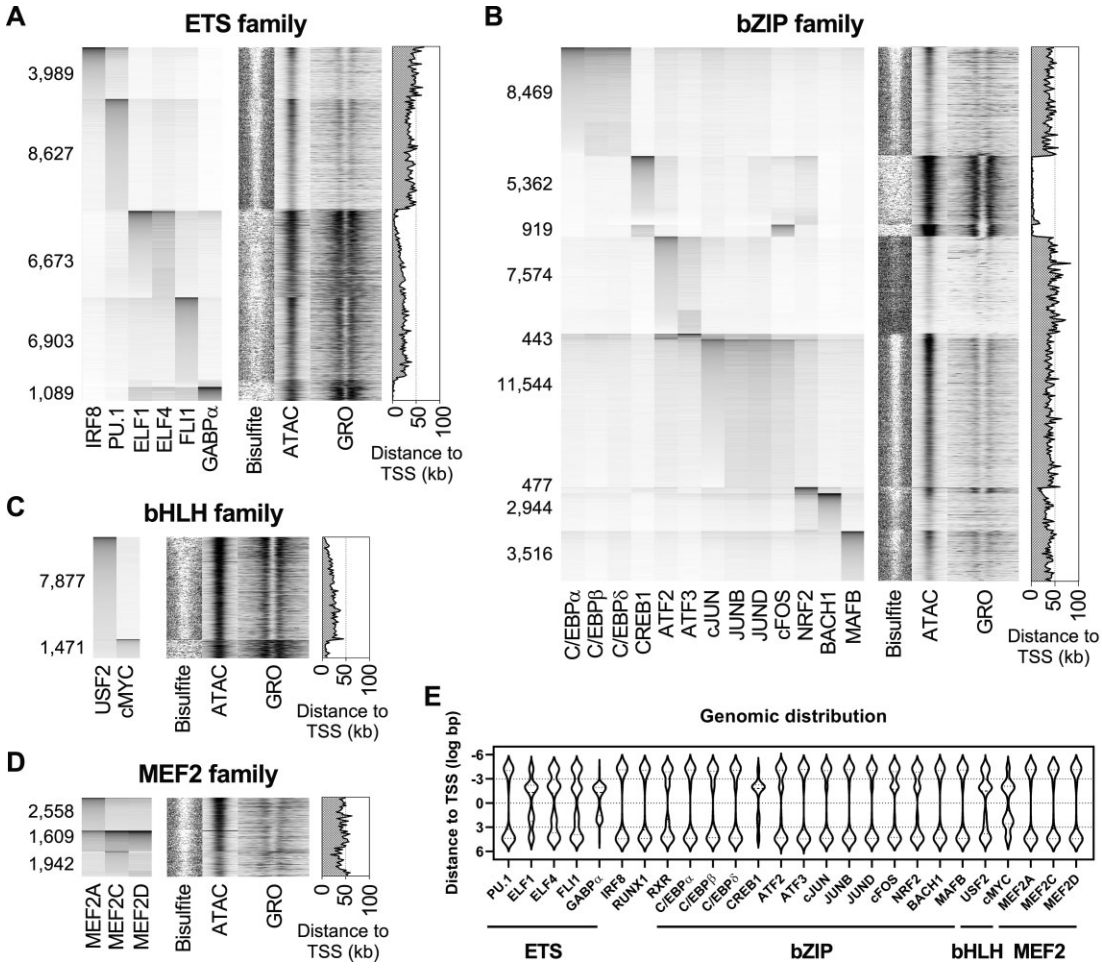


Figure 3. Characterizing the major cistromes of BMDMs. (A-D) Cistromes of related TFs were filtered and united to aggregate cistromes; those of the ETS family with IRF8 (**A**) and the bZIP (**B**), bHLH (**C**) and MEF2 families (**D**). Heat maps represent the aggregate cistromes clustered and sorted based on TF densities (left). The scale of densities ranges from 0 to 3. The extent of DNA methylation (Bisulfite-seq, scale: 0–0.1) and chromatin openness (ATAC-seq, scale: 0–20) in 1-kb windows and nascent transcription (GRO-seq, scale: 0–20) in 2-kb windows for each region (middle) and the genomic distribution of the regions relative to the closest TSS (right) are depicted. For the latter, the average distance of 100 consecutive regions on the heat maps was calculated. (**E**) Violin plot represents the genomic distribution of the indicated cistromes relative to the closest TSS.

to these regions. Interestingly, there is a cFOS-dominated set of regions without significant JUN binding but with frequent CREB1 binding and TSS-proximal characteristics. The cluster of ATF2/3 can be further divided too, but both subclusters represent highly DNA-methylated, promoter-distal, and closed chromatin regions. Classical AP-1 (FOS/JUN) binding sites are also promoter-distal ones, but—in contrast to the ATF binding sites—they are bound by other bZIP proteins and show enhancer characteristics—moderate openness and transcription initiation with a narrow unmethylated lane—like PU.1 and IRF8. The 477 regions (and a similar number of regions within the 'CREB1-specific' cluster) enriched for NRF2 only are closer to promoters and more active like those of FLI1, and the BACH1- and MAFB-specific clusters behave like that of C/EBP; however, relative to MAFB, BACH1 occupies more methylated, less open and less active regions.

The next aggregate cistrome, which can be analysed, is that of bHLH proteins, although there are huge differences between the raw cistrome sizes: while USF2 can be detected at 96 817 genomic regions, the number of cMYC and BMAL1 binding sites is around 3000. In order to make the cistromes comparable and an aggregate cistrome assembled, the small cistromes were normalized with the 100th peaks' density. This way, the size of the filtered cMYC cistrome is 1998, while BMAL1 with 881 peaks was left out from the further analyses. Otherwise, BMAL1-specific regions are represented by the other two bHLH proteins, and its E-box motif enrichment does not show special characteristics (Supplementary Figure S3B) (69). In contrast, E-boxes of cMYC are not only rather palindromic (CACGTG) but are also enriched for C/G flanks, the opposite of which is true for the E-boxes of USF2, preferring the asymmetrical CACATG sequence with 5′ T and/or 3′ A flanking nucleotides (18). When we compared only the filtered aggregate cistrome of USF2 and cMYC, it became visible that cMYC and strong USF2 binding sites behave similarly (Figure 3C). These are closer to TSSs and show a smaller number of methylated cytosines, higher openness, and higher polymerase activity. If we get farther from TSSs, we see weaker USF2 enrichment and more methylated, less open, and transcriptionally less active regions.

MEF2 family members were also problematic in the sense that our otherwise strict filtering still resulted in many, presumably indirect binding sites according to the motif enrichment results. Because of this, we used 1/2 threshold in-

stead of 1/3 to generate the aggregate cistrome (Figure 3D). Although most MEF2 binding sites are promoter-distal and not so active transcriptionally, there are differences in their characteristics—MEF2A-specific regions are less methylated and more open than the others, and regions with all 3 TFs are more open than the MEF2C- or MEF2D-specific ones.

During these cistromic analyses, we characterized how proximity to active promoters and the range of low methylated genomic regions correlate, and how this is coupled with specific TF binding. In order to make the major characteristics of the 26 TFs more comparable, we plotted their genomic distribution relative to the closest TSS (Figure 3E). This indicates that most (n = 16) of the examined TFs bind almost exclusively to promoter-distal regions. In contrast, CREB1 and GABPα bind almost exclusively to TSS-proximal regions (either up- or downstream), cMYC, and ELF1 mostly to TSS-proximal regions, and there are three TFs with bimodal distribution: ELF4, FLI1 and USF2. The remaining TFs bind to DNA typically far from promoters, but they also have a small population of TSS-proximal regions—these are cFOS, NRF2 and ATF2, as it is clear also based on their distribution relative to CREB1 (Figure 3B). In addition, RXR also falls into this category, at least hundreds of nucleotides away from the TSSs. Interestingly, the limit between TSS-proximal and distal regions is strictly around 1 kb (Figure 3E, dotted lines). We also plotted the distribution of ATAC-seq densities per cistrome, and besides those TFs enriched in TSS-proximal sites—with CREB1 in the lead –, NRF2 and AP-1 proteins, especially cFOS showed high chromatin openness (Supplementary Figure S3C). In contrast, the lowest openness can be observed at the ATF2/3, BACH1, PU.1, IRF8, RXR and C/EBP binding sites.

Finally, in order to examine also the possible interfamilial interactions, we compared all cistromes with each other (Supplementary Figure S3D). Naturally, we see high co-localization frequencies within certain (sub)families such as MEF2, C/EBP, and JUN, and most TFs co-localize with each other in some proportion. Importantly, the TSS-proximal TFs (GABPα, CREB1, cMYC, and ELF1) are more or less out of line and bind together rather than with the other TFs. For example, GABPα, besides the ETS proteins FLI1 and ELFs, co-localizes with CREB1 and also ATF2, NRF2, USF2 and cMYC to a lesser extent. These results are consistent with the previous findings (Figure 3A–D), and no significant, previously unseen interfamilial interactions could be detected (Supplementary Figure S3D).

There are two major classes of ETS binding sites (EBSs)

After characterizing the top cistromes, we were interested to know how these are determined by the code of the low methylated non-coding genome. Using the clusters of the aggregate cistromes (Figure 3A–D) and their motif enrichments, we determined the most enriched sequences (oligomers) that are specific for any of the particular TF patterns. The initial oligomers were selected by considering the core sequences, their possible variants and extensions (Supplementary Table S2), and their enrichment was calculated within ±50 bp relative to the peaks' centre (Supplementary Figure S2E).

First, we determined the EBSs, which have a common GGAA, AGAA or GGAT core, and their specificity is mostly determined by the flanking 2 × 2 nucleotides (70). Out of the tested octamers, 78 enriched in any of the clusters of the aggregate ETS cistrome (Figure 3A). Based on the median ChIP-seq densities calculated for each sequence, 3 groups of elements separated: i) methylatable EBSs with frequent ELF/FLI1 binding and the presence of GABPα, ii) intermediate EBSs with any ETS proteins but GABPα, and iii) PU-boxes with frequent PU.1 binding and some ELF/FLI1 and IRF8 binding (Figure 4A). (i) Methylatable EBSs all have a C right upstream of their core, thus forming the methylatable CG site itself (blue), (ii) the 2 kinds of intermediate EBSs are those with the FLI1-specific 5′ CA (purple) or PU.1-specific 5′ dinucleotides (black) and (iii) PU-boxes typically have an RR (R = A/G) 5′ extension, or in the case of a 3′ GT extension, T can substitute one of the Rs. Based on this comparison, the GGAT core is PU.1-specific—otherwise, it is specific for SPDEF (70)—, and the AGAA core is rather PU.1-, and sometimes also ELF4-specific. Besides the 5′ flanking nucleotides, the other side is similarly important. In general, the 3′ GT makes sequences strong EBSs: the top GABPα-, PU.1- and ELF-specific sequences carry this 3′ extension. In contrast, 3′ AY (Y = C/T) sequences are preferred by FLI1, but the 3′ CT can turn the 5′ CA PU.1-specific instead of FLI1-specific (Figure 4A).

In order to examine the correlations between the sequence preferences of ETS proteins and the enrichment of sequences, we also plotted the cluster-specific sequence enrichments (fold enrichment × frequency) in the function of the per sequence median TF binding frequency (Figure 4B). This clearly demonstrates that methylatable elements (blue) are typically not enriched in the PU.1 cistrome, while GABPα binds only to CG-containing EBSs, and FLI1 and ELF proteins bind to both methylatable and non-methylatable sites, but they have higher affinity to the former ones. Overall, promoter-specific ETS proteins, especially GABPα show a DNA methylation-dependent DNA binding.

In addition to the ETS proteins, we also plotted the EBS sequence enrichments and the per sequence median binding frequencies for the PU.1 partner IRF8 (Supplementary Figure S4A). Notably, the most enriched sequences do not show the most frequent binding: IRF8 prefers several lower-affinity and unusual PU-boxes such as AGAGAAGT and AGGGAAAT. Next, we also examined the direct binding sites of IRF8, of which the characteristic core is GAAA (67,68), but according to the motif enrichments (Supplementary Figure S3A), some mismatches (resulting in e.g. AAAA or GAGA) might be allowed. Out of the tested octamers—containing 2 × 2 flanking nucleotides, which can also be characteristic—, 24 enriched in the IRF8-specific cluster (Figure 3A, Supplementary Figure S4B). Except for GTGAGAGT, all enriched sequences have a GAAA core, and the most enriched IRF binding site is GTGAAACT, while the highest affinity one is GTGAAAGT. Importantly, the flanking nucleotides essentially only allow the formation of EICEs, and the number of IECS is minimal, so the detected A-rich sequences upstream of PU-boxes may be a specific extension for a more frequent PU.1 binding (Supplementary Figure S3A).

Unlike other bZIP proteins, CREB1 binds exclusively to methylatable elements

Our next aim was to investigate whether there is a bZIP protein with a CG preference. These proteins form a diverse family with distinct sequence requirements (66,71). Their dimer binding sites can be octamers, heptamers, or extended

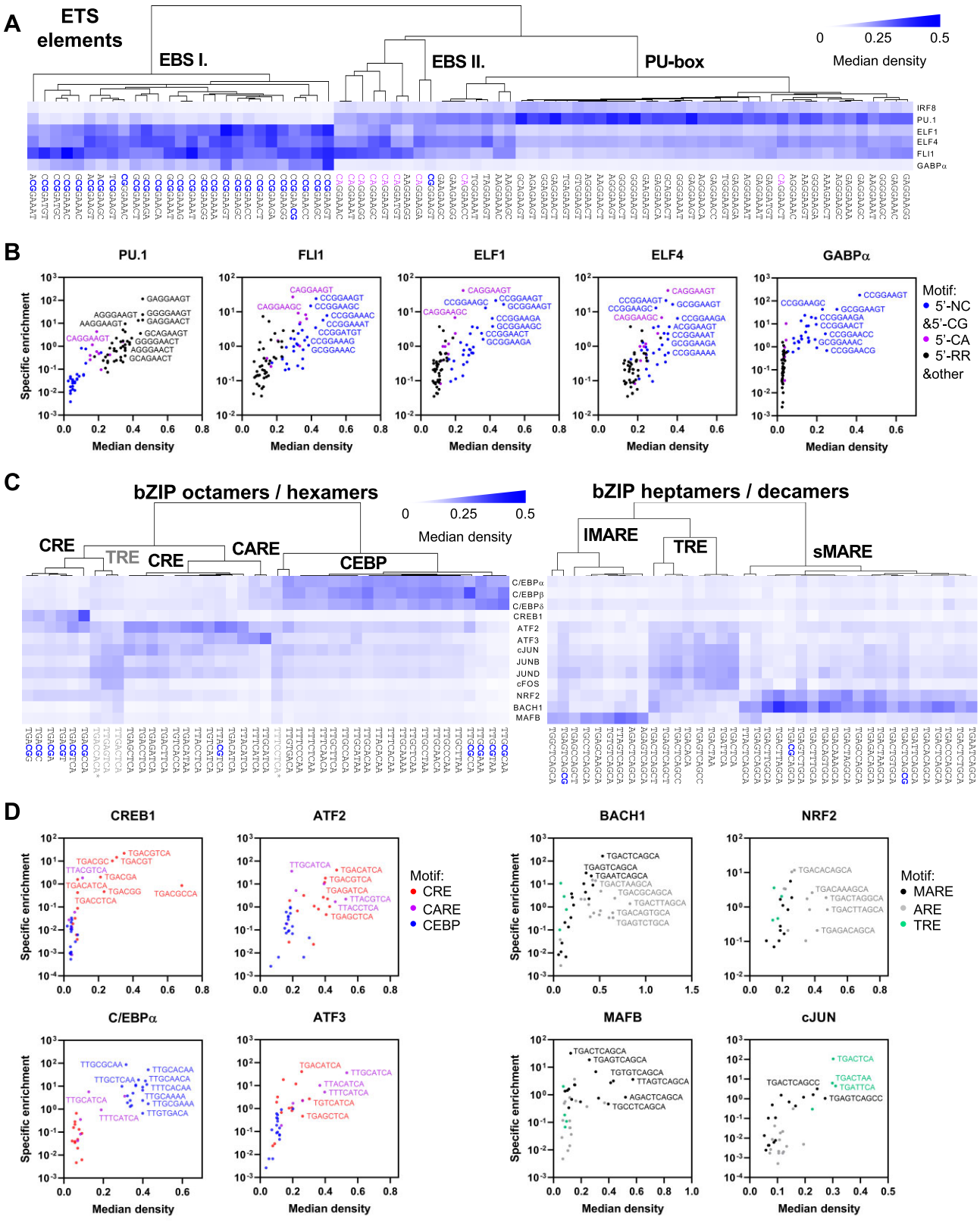


Figure 4. Distinction of DNA elements. (**A**) Heat map represents the per sequence median density of the indicated IRF8 and ETS proteins at ETS-specific elements. Hierarchical clustering was performed based on the TF-specific patterns. Methylatable CG dinucleotides are highlighted in blue, and FLI1-specific 5′ CA dinucleotides are highlighted in purple. (**B**) Scatter plots represent correlations between the per sequence median density of the indicated ETS proteins and the enrichment (fold enrichment x frequency) of the individual sequences within the cluster specific for the given TF. (**C**) Heat maps represent the per sequence median density of the indicated bZIP proteins at 'octamer/hexamer' (left) and 'heptamer/decamer' elements (right). Hierarchical clustering was performed based on the TF-specific patterns. CG dinucleotides are highlighted in blue, sequences out of place are highlighted in grey, and sequences specific to a TF family other than bZIP are marked with asterisks. (**D**) Scatter plots represent correlations between the per sequence median density of the indicated bZIP proteins and the enrichment (fold enrichment x frequency) of the individual sequences within the cluster specific for the given TF.

heptamers (decamers or tridecamers) (Supplementary Table S2). According to the motif enrichments, octamers can be formed by C/EBP- and CREB/ATF-type half-sites, resulting in a C/EBP dimer binding site, a C/EBP:ATF response element (CARE), or a CRE (72,73). Based on the enrichments of the mapped octamers composed of the most frequent half-sites, C/EBP binding sites are typically non-palindromic (Supplementary Figure S4C): one of their half-sites is characteristic (TTDC, D = G/A/T), while the other is non-canonical (TTDD/HHAA, H = A/C/T) (72). In the case of CRE sequences (TGAN/NTCA, N = any nucleotides), both palindromic and non-palindromic octamers can be frequent, but methylatable CREs—with CG in the middle—are underrepresented, and sequences with TA in the middle are even rarer. Otherwise, the TGAT (ATCA) sequence forms one of the half-sites of CAREs (73). The other group of bZIP elements is the heptamers, which are the shorter forms of CREs, composed of 2 'partially overlapping' CRE-type half-sites and called TRE or AP-1 binding sites (TGANTCA) (71). MAF proteins bind the extended forms of TREs: small MAFs with a CNC (NFE2 or BACH) partner require one extended half-site (resulting in a decamer, TGANTCAGCA), while dimers of large MAFs have a long MARE extended to both directions (resulting in a tridecamer) (66,74). There is an additional 'variant' of 'short' MAREs, of which the extended half-site is degenerate (TGANNNNGCA—NNN means any trimer but TCA now)—this is called antioxidant response element (ARE) (75).

We started our sequence enrichment analyses with the octameric elements, and as bZIP proteins prefer certain flanking nucleotides as shape motifs (72), we calculated the enrichments for any sequence matching with the NTKDNNHMAN consensus (K = T/G, M = C/A), representing all the characteristic C/EBP, CARE, and CRE sequences (Supplementary Table S2). We found 39 octamers enriched, 4 of which are non-specific hits showing the characteristics of PU.1, AP-1, and RUNX elements with—direct or indirect—AP-1 binding (Figure 4C, left, grey). More than half of the octamers are dominated by C/EBP proteins, even if they could be a CARE based on their composition (TTGCTTCA), and 13 out of 18 have the canonical TTGC half-site. There are two CAREs (TTKCATCA), which are bound both by ATF3 and C/EBPs, and there are 2 unusual CAREs, which can be bound by C/EBPs and ATF2 (TTGCGCCA and TTGTGACA). The remaining sequences are methylatable (CG is highlighted in blue, n = 3) and non-methylatable CREs (n = 10), although the TTACGTCA sequence behaves like the latter group, bound by ATF, JUN, and NRF2 proteins, especially ATF2, cJUN and JUND. There is an additional CARE-like sequence, TTACATCA, which is equally bound by the examined 2 ATF proteins. Notably, the TTAC half-site, depending on the other half-site, can behave both as a C/EBP and ATF2 binding site. We observed that CREB1-specific sites are underrepresented by the octamers, but additional methylatable 'half-sites' (TGACGN) could be detected in higher amounts (Supplementary Table S2). These behave similarly to the complete CREB1-specific CREs, including the relatively frequent ATF2, NRF2, cJUN and JUND binding (Figure 4C, left). Cluster-specific enrichments also show that even if different CREs are frequent at CREB1-specific sites, only methylatable ones are bound by CREB1 with higher frequency (Figure 4D, left). While ATF2 binds to any CREs and shows the enrichment of certain CAREs, ATF3 is the real CARE binder, and it has also a high affinity to CREs with the same CATCA sequence. All C/EBP proteins prefer their specific dimer binding sites over CARE (and CRE) both in enrichment and binding frequency (Figure 4D, left; Supplementary Figure S4D, left), and all JUNs and NRF2 proteins bind various CREs to a certain extent (Supplementary Figure S4D, right). Overall, the CG content of CREs is of more importance than that of C/EBP elements because the former can be discriminated by specific TFs.

The other group of bZIP elements is the simple and extended heptamers, from which 4 TREs and 36 (M)AREs enriched (Figure 4C, right; Supplementary Table S2). Notably, as long MAREs are found variable and mostly incomplete, they are represented by decamers in further analyses. Comparing the protein binding at the enriched sequences, it is clear that in general, these cannot be bound by C/EBP and CREB1 proteins, and MARE-like sequences ending with GCY behave as TREs with AP-1 binding—in the further steps, we excluded these decamers and used the enriched TRE sequences instead. More than half of the decamers are CNC protein binding sites (short MAREs), typically with more frequent BACH1 binding even if these are called AREs, and 9 MAREs are dominated by MAFB (Figure 4C, right). Cluster-specific enrichments show that even if the consensus MARE shows the highest enrichment at BACH1-dominated sites, AREs show more frequent BACH1 binding (Figure 4D, right). In the case of NRF2, AREs are not only enriched but also highly occupied. In contrast, MAFB requires one complete and one imperfect MAF half-site, which latter can be both non-extended and non-TGA-containing. Although MAREs with TGA half-site are enriched at MAFB-specific binding sites, those with trinucleotides other than TGA result in higher-affinity MAFB binding. Unsurprisingly, AP-1 and ATF proteins all prefer TREs, and (M)AREs are less enriched and less bound by these TFs (Figure 4D, right; Supplementary Figure S4E).

Flanking nucleotides determine E-box specificity

In order to discriminate specific E-boxes, we investigated the most enriched oligomers of the bHLH cistromes. According to the sequences represented by our motif enrichments (Supplementary Figure S3B), E-boxes are basically hexamers with a variable middle part (CNNNNG instead of CACGTG). This diversity and the nature of the flanking nucleotides may provide the specific binding by the numerous bHLH dimers (51,76). Based on our analyses, 63 octamers (representing 18 hexamers) enriched in any of the two bHLH clusters (Figure 3C, Supplementary Table S2), and based on USF2 and cMYC binding, they almost perfectly separate depending on the presence or absence of a 5′ TCA sequence—a flanking T followed by CA core nucleotides (Supplementary Figure S4F). In the presence of 5′ TCA, USF2 shows a much higher recruitment than cMYC, while cMYC prefers C/G flanking nucleotides as the motif enrichments also indicated (Supplementary Figure S3B). Importantly, 2/3 of the identified E-box oligomers can be methylated, and these are higher-affinity binding sites than the non-methylatable ones, in general (Supplementary Figure S4F).

MEF2, RUNX and RXR elements are typically unmethylated

We also applied sequence enrichment analysis for the remaining cistromes to make our oligonucleotide list more complete. MEF2 proteins bind to a variable AT-rich octamer that

is flanked with typically non-A/T nucleotides (23), thus the used raw consensus sequence of MEF2 binding sites was NYAAAAATAN. Using this sequence pool allowing 2 additional mismatches resulted in 42 decamers enriched in any of the 3 clusters of the aggregate MEF2 cistrome (Figure 3D, Supplementary Figure S4G, Supplementary Table S2). The resulting sequences show similar enrichments and affinities to all examined family members, so protein-protein interactions should affect their specific genomic distribution, too.

The consensus sequence of RUNX1 binding sites is the ACCACA hexamer (77), but motif enrichments show some alternative variants and the possible significance of at least 2 flanking nucleotides in both directions. Within the filtered RUNX1 cistrome, 135 decamers (39 octamers) showed an enrichment (Supplementary Table S2). The most enriched octamers can be described as RACCACAR, but there are some less frequent and less typical sequences with higher affinity binding (grey) (Supplementary Figure S4H). We found one single methylatable hit (AACCGCAG) (blue), but this does not show major differences compared to the non-methylatable sequences.

RXR is the heterodimerizing partner of class II nuclear receptors that bind to various direct repeats of the AGGTCA consensus sequence (63,66). As this hexamer also shows some variety and RXR can have a 5′ nucleotide preference (60), we started our analysis with the set of NNRGKKSAN nonamers (S = C/G) (Supplementary Table S2). 67 out of the possible sequences enriched (representing 7 hexamers), but this enrichment is below those of other TFs, probably because of the repetitive nature of the nuclear receptor elements (Supplementary Figure S4I). The most enriched sequences match with the extended ARAGGTCAN consensus, while those showing the highest RXR density are the AGAGTTCA sequences (grey), suggesting the binding of the RXR partner vitamin D receptor (VDR) as it prefers this kind of half-site (66,78).

Altogether, the elements of MEF2, RUNX1 and RXR typically do not contain CG dinucleotides, which further strengthens the observed correlations between the CG content of elements and their promoter proximity (like in the case of GABPα and CREB1).

TSS-proximal and distal elements contribute differentially to determine chromatin openness and transcription initiation

Our next goal was to characterize how the hundreds of oligomers specific for macrophage-determining TFs affect chromatin functions. For this, first, we classified the oligomers based on their CG content and the TF patterns they show and determined their genomic distribution relative to the closest TSS (Figure 5A, Supplementary Table S3). Notably, in certain cases, the presence or absence of CG did not make a major difference: Certain methylatable CREs, (M)AREs and RUNX1 and C/EBP elements, similarly to their CG-free forms, generally show a promoter-distal distribution (Supplementary Figure S5A). In contrast, methylatable EBSs, E-boxes, and CREB1 sites—CREB1-specific CREs (cCREs) and CRE half-sites (hCREs)—all show a TSS-proximal (light green) or bimodal distribution (medium green), and in the lack of CG, the majority of the elements are farther than 1 kb from any TSS (dark green) (Figure 5A).

In order to approach the additional general and macrophage-specific elements and possible DNA-protein interactions, we compared the aggregate cistromes with the set of open-chromatin regions based on ATAC-seq (Supplementary Figure S5B). More than a third (n = 15462) of the (n = 42951) ATAC-seq peaks show low or no binding by the major TFs, and these are enriched for the TSS-proximal GC-box and a smaller number of CTCF binding sites and PU-boxes (Supplementary Figure S5C). Chromatin openness here is approximately half of the openness measured at sites that are highly occupied by the examined TFs (Supplementary Figure S5D, left). Relative to these sites, ELF1- and CREB1-specific sites show even higher openness (Supplementary Figure S5D, right) and more promoter-specific motif hits, including those of SP1, NFY, GFY and NRF1 (Supplementary Figure S5E, asterisks). In order to improve these motifs, we used the highest ATAC-seq peaks that carry them for an additional *de novo* motif enrichment analysis. Besides the targeted ones, an additional PBX motif was also enriched, and as PBX2 is a high-expressed TF in BMDMs (Figure 2A, column 'Other'), we included it in the further analyses. Most of the identified promoter-specific motifs can be methylated, or at least a part of the mapped elements contains a CG sequence (Figure 5B, asterisks). This way, SP1 motifs can be separated into GC- and 'GT'-boxes; CCAAT-box can have a CG both right up- and downstream; ideally both half-sites of the NRF1 element can be methylated; and both the GFY and PBX elements can be methylated at several sites. The ability to be methylated correlates with the genomic distribution to some degree also in the case of these elements (Figure 5C).

After having the most important elements of BMDMs, we turned to the initial question of how TF-bound non-coding sequences determine chromatin openness and transcription initiation. First, we selected those elements, around which there is no other element within 50 bp, and plotted the per element class median ATAC-seq and GRO-seq densities measured at the specific single sites (Figure 5D, left). Except for the CREB1-specific sites (cCRE and hCRE), all methylatable elements (blue) separated well from the non-methylatable ones (red). The former elements are associated with both higher openness and higher transcription initiation frequency, while there are two populations of the non-methylatable ones: (i) bZIP, MEF2 and RUNX elements do not contribute much to transcription initiation, although TREs and RUNX elements contribute to openness to a higher degree, while (ii) the others occupy an intermediate position on the plot, in a similar pattern to those of their methylatable forms (if they have). EBSs and E-boxes are better 'initiators' than 'openers' (top left), while the promoter-specific elements identified by ATAC-seq rather contribute to chromatin openness (right). We also tested pairs of elements within 50 bp and got a distribution similar to that in Figure 1B (Figure 5D, right; purple indicates pairs with one methylatable element). According to this, two ETS elements, if at least one of them is methylatable, can lead to significant transcription initiation with low openness, while TREs—and also the non-methylatable CCAAT-boxes—in several combinations are good openers with minimal transcription. SP1 elements, in contrast, are not only good openers; depending on their pair, they can lead to high initiation frequency. Overall, a more or less additive effect can be observed, and this way, the top values of the single elements can be exceeded by the pairs (dotted lines).

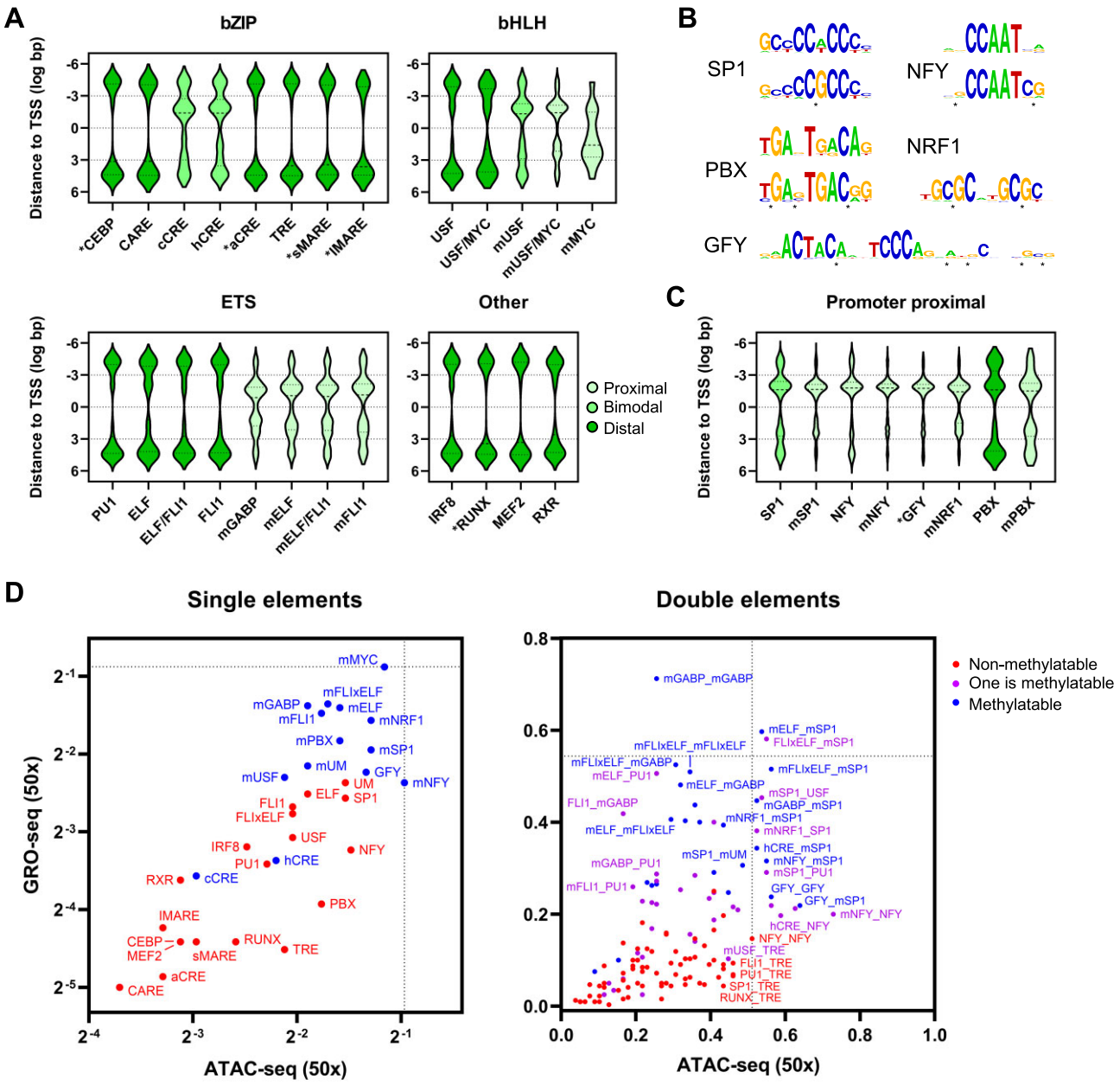


Figure 5. Characterizing TSS-proximal and distal elements. (**A**) Violin plots represent the genomic distribution of the indicated classes of elements relative to the closest TSS. 'm' initials denote classes of only methylatable elements. Asterisk indicates that some of the elements within the given class contain a CG dinucleotide. cCRE: CREB1-specific CRE; aCRE: ATF/JUN-specific CRE; hCRE: CRE half-site with CG; s/lMARE: short/long MARE; dark, medium, and light green colours highlight TSS-distal, bimodal, and proximal distribution, respectively. (**B**) Motif logos represent promoter-specific sequences having or lacking at least one CG dinucleotide. Asterisks represent possible CG sites. (**C**) Violin plot represents the genomic distribution of promoter-specific elements relative to the closest TSS. Asterisk indicates that not all GFY elements contain a CG dinucleotide. The colour code is the same as in (A). (**D**) Dot plots represent correlations between the median chromatin openness (ATAC-seq) and median polymerase activity (GRO-seq) for the indicated classes of elements (left) or their combinations (right). Elements farther than 50 bp from other elements were considered single elements (left), and pairs of elements within 50 bp but more than 50 bp away from other elements were considered double elements.

Certain macrophage-specific promoters are composed exclusively of prototypically distal elements

After characterizing single and double elements, we turned our attention to promoters, from which both initiation and elongation take place. In order to determine BMDM-specific promoters, we used the *de novo* transcript predictions determined by GRO-seq and RNA-seq coverages (Supplementary Figure S6A). Then, we assigned the groups of putative elements to the predicted TSSs and classified them based on their composition (Figure 6A). Most of these promoter(-proximal) regions contain the characteristic elements only (black) or in combination with prototypically distal elements (grey). Elements with bimodal distribution are also included in this latter category. Importantly, 14.4% of TSS-proximal regions lack any of the characteristic promoter elements, instead, macrophage-specific, otherwise distal elements are present (red). From now on, we call this unusual group of regulatory regions

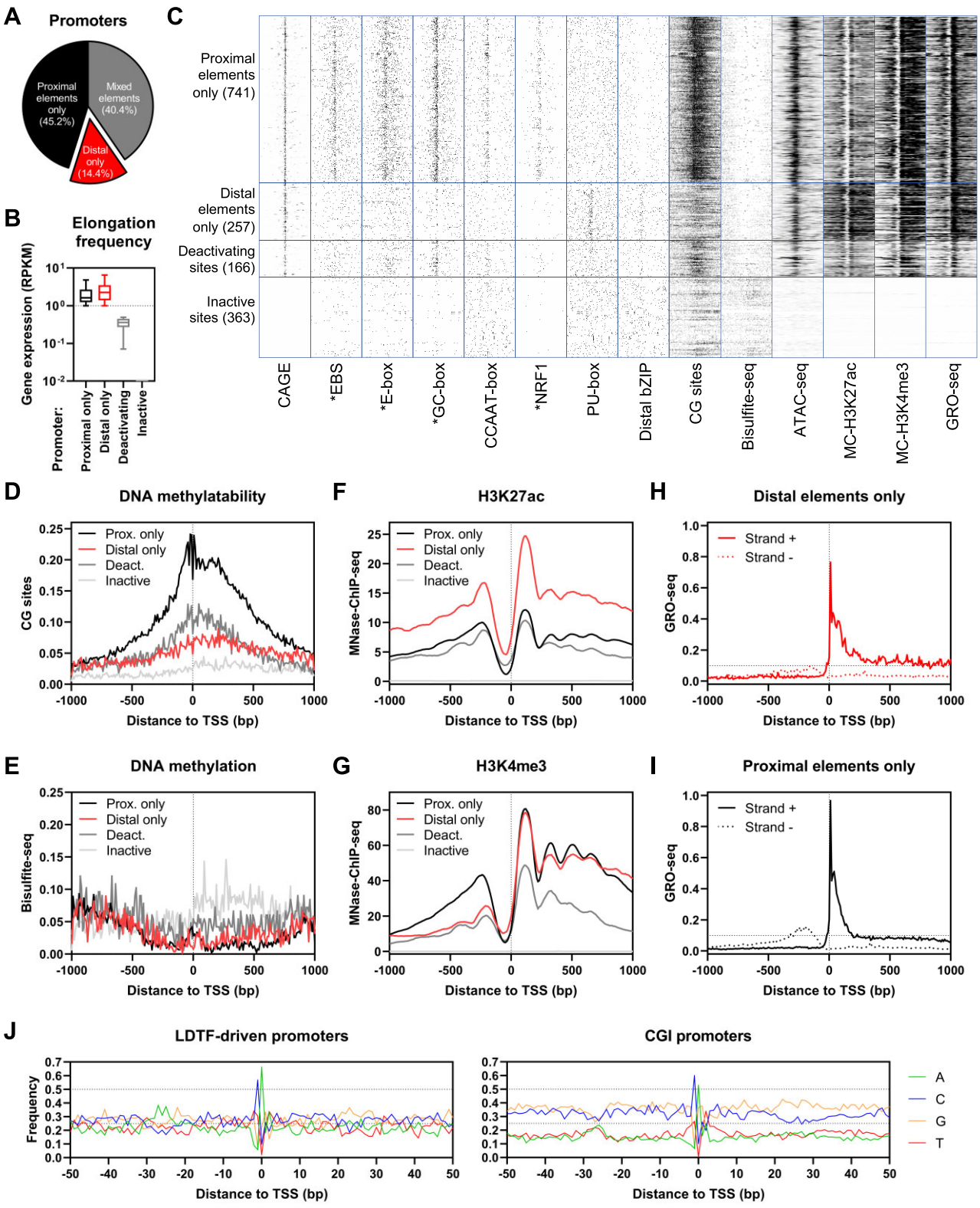


Figure 6. Identification of macrophage-specific promoter classes. (**A**) Pie chart represents the ratio of putative promoters with the indicated composition. (**B**) Box plot represents the levels of nascent transcripts (GRO-seq) originating from the different classes of promoters indicated. Whiskers are plotted according to the Tukey method. (**C**) The distribution of TSSs (CAGE), the identified motifs, and CG sites and the extent of DNA methylation (Bisulfite-seq), chromatin openness (ATAC-seq), H3K27ac and H3K4me3 modifications (MC, MNase-ChIP-seq), and nascent transcription (GRO-seq) are depicted in 1-kb windows around the top TSSs at the indicated promoter classes (TSS selection is detailed in the Methods section). (D–I) Histograms show the distribution of CG sites (**D**), the extent of DNA methylation (**E**), the H3K27ac (**F**) and H3K4me3 modifications (**G**) and nascent transcription (H, I) around the top TSSs at the indicated promoter classes. (**J**) Line plots represent per-position nucleotide frequencies around the top TSSs at the promoters with distal elements only (left) or proximal elements only (right).

'LDTF-driven' promoters. Unsurprisingly, mixed promoters are the most active regarding openness and transcription initiation frequency, while 'distal' elements result in the lowest initiation frequency, and clearly promoter elements are in between (Supplementary Figure S6B).

For a more thorough characterization of LDTF-driven promoters, we collected all highly expressed genes (over 1 RPKM gene body enrichment according to GRO-seq) having only those promoters with 'distal' elements only. As control sets, promoters of highly expressed genes with proximal elements only and those of deactivating (low turnover) and completely silent (inactive) genes were selected (Figure 6B). TSSs—where it was possible—were determined based on BMDM-derived CAGE data, which showed a sharp TSS distribution in most active promoters (Supplementary Figure S6C, Figure 6C, left). In promoters with exclusively proximal elements ($n = 741$) not only the elements but also their environment is CG-rich, while LDTF-driven promoters ('distal elements only', $n = 257$) contain much fewer CG sites (Figure 6C, middle). Importantly, the extent of demethylation (Bisulfite-seq) and chromatin openness (ATAC-seq) is especially lower at LDTF-driven promoters, but in the transcriptional output (GRO-seq), the opposite can be observed (Figure 6B, C, right). Notably, trimethylation of H3K4 residues follows the GRO-seq patterns, while the high H3K27ac signal is specific for LDTF-driven promoters (Figure 6C, right). Deactivating genes (n = 166) can have both CG-rich and CG-poor promoters, and their chromatin is clearly less active, while inactive promoters (n = 363) show the lowest demethylation level and zero activity (Figure 6C, bottom).

During these analyses, a pattern in CG distribution grabbed our attention because a shift could be observed towards the genes (Figure 6D). A more thorough analysis showed that at least two curves can model the distribution of CG sequences around TSSs (Supplementary Figure S6D). We identified a peak right upstream of the TSS (at -30 bp), and a much higher and broader one mostly within the 5′ UTR (around +120 bp). In addition, an ascending line was required for more accurate modelling. As no motifs are enriched in the 5′ UTRs, these results suggest a function other than the classical DNA-TF interactions. In any case, demethylation of 5′ UTRs seems to be a process required for gene activation because 5′ UTRs following deactivating and inactive promoters show higher and higher levels of methylated CGs (Figure 6E). Importantly, at these sites, promoters are less methylated than 5′ UTRs, and the difference is more pronounced at completely inactive promoters. Furthermore, this phenomenon is in line with chromatin openness because differences can be observed not only at promoters but also at the 5′ of the genes (Supplementary Figure S6E). These results suggest that gene inactivation and DNA methylation start at CG sites of 5′ UTRs that cannot be bound by TFs.

Then, we focused on nucleosome positioning and histone modifications using MNase-ChIP-seq data and found the expected nucleosome loss at active promoters and a series of peaks (nucleosomes) both up- and downstream of the TSSs (Figure 6F, G). H3K27ac behaves as an LDTF-driven promoter marker showing a signal approximately twice as high as of the other active promoters (Figure 6F). In the sense direction, promoters with extreme composition ('Prox. only' and 'Distal only') show a similar extent of H3K4 trimethylation, while there is a difference in the antisense direction (Figure 6G): CG-rich ('Prox. only') promoters have significantly more K4 trimethylated H3 histones than the LDTF-driven ('Distal only') and the less active ones. This correlates well with the transcriptional activities observed on the two strands because CG-rich promoters show higher activity in the antisense direction (79). Furthermore—as expected—CG-rich promoters show higher initiation frequency on the sense strand but reach lower gene body enrichment relative to the LDTF-driven promoters, in general (Figure 6H, I).

Next, we aimed to test whether LDTF-driven promoters show any specific positioning of the elements that make them up. For this, per position frequencies of nucleotides were plotted around TSSs in LDTF-driven promoters and also CGI promoters as a comparison set (Figure 6J). The most characteristic enrichment in both promoter classes is that of C/T (-1) and A/G (+1) right at the TSSs, which form the core of Initiator (79). In general, CGI promoters contain approximately twice as many C/G nucleotides as LDTF-driven promoters, and A/T nucleotides show a moderately elevated frequency only at the position of the TATA-box (around -30). Although A/T enrichment is more pronounced in LDTF-driven promoters, this still represents only ∼5% of TATA-boxes (Supplementary Figure S6F). We also checked the motif composition of these promoters and found that non-methylatable ETS (mostly PU)-boxes dominate in these regions, but any other LDTF alone or in combination can bind directly to LDTF-driven promoters (Supplementary Figure S6G). Although the mostly unidirectional transcription suggests a fixed position of a motif at these promoters, even the most frequent ETS core does not show any specific enrichment (Supplementary Figure S6H).

LDTF-driven promoters regulate macrophage-specific gene expression

In order to further characterize LDTF-driven promoters, we also investigated the function of their targeted genes. The 257 promoters belong to a total of 217 genes, out of which 198 are protein-coding (Figure 7A). Among these genes, we identified several of those essential to 'macrophageness'. One of the major myeloid-specific genes is colony-stimulating factor 1 receptor (*Csf1r*), whose protein product allows the response to macrophage colony-stimulating factor (MCSF) (80). In addition, our list includes several genes involved in bacterial pattern recognition (*Cd14*, *Trem2*) (81,82) or antiviral response (*Irf5*, *Isg15*, *Ifi203*, *Ly6e*) (83–86), as well as those encoding chemokines (*Ccl3*, *Ccl4*, *Ccl9*) (87) or macrophage-specific proteases (*Lyz2*, *Mmp12*, *Ctss*, *Psmb8*) (62,88–90). Importantly, the top 20 biological processes obtained by GO analysis were all related to immune functions (Figure 7B), while genes with CGI promoters are related mostly to other, more general cellular functions such as nitrogen metabolism (Supplementary Figure S7A).

Discussion

The widespread use of murine BMDMs in functional genomics studies allowed us to thoroughly examine multiple levels of gene regulation, including the extent of DNA methylation (Bisulfite-seq), chromatin openness (ATAC-seq), histone modifications (MNase-ChIP-seq), DNA-TF interactions (ChIP-seq), polymerase activity (GRO-seq), and mature RNA production (RNA-seq). The remarkably high number of available TF cistromes ($n = 41$ as listed in Supplementary Table S1) served as the basis of our analyses, allowing the mapping of

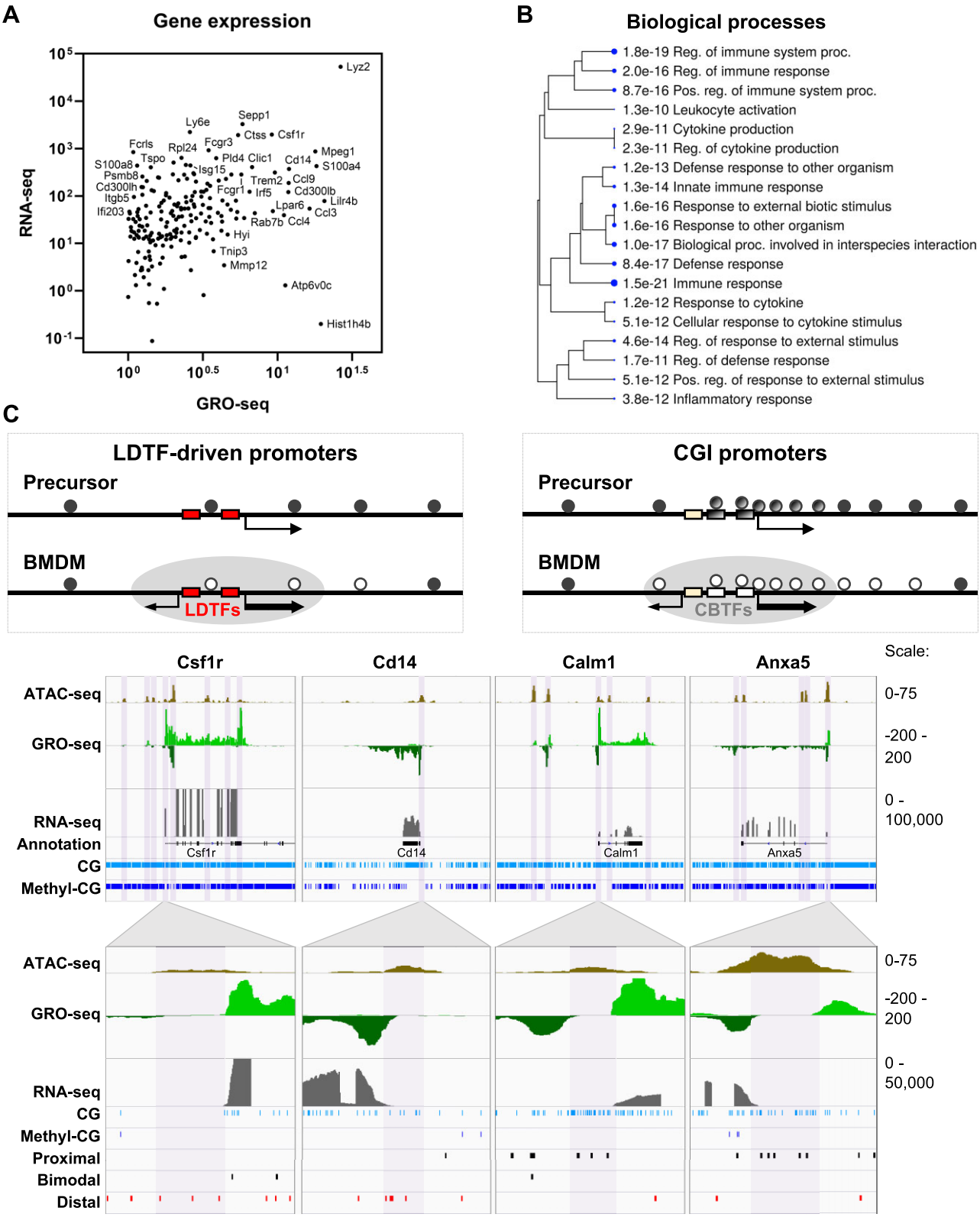


Figure 7. Characteristics of LDTF-driven promoters. (**A**) Scatter plot represents correlations between the nascent (GRO-seq) and mature mRNA levels (RNA-seq) associated with LDTF-driven promoters. (**B**) GO terms (Biological processes) specific to LDTF-driven promoters are shown. (**C**) Putative models (top) and examples (bottom) for LDTF-driven (left) and CGI promoters (right) are shown. Model: Arrows represent TSSs and transcription initiation frequencies. Greyscale, white, red, and light gold boxes represent respectively methylatable, non-methylated, and prototypically TSS-distal or proximal non-methylatable elements. Black, greyscale and white circles represent respectively methylated, potentially methylated, and non-methylated cytosines. LDTF: lineage-determining TF, CBTF: CGI-binding TF. Genome browser views show the chromatin openness (dark olive), nascent transcription (light and dark green, strand specifically), mRNA level (dark grey), the density of CG (light blue) and methyl-CG dinucleotides (blue), and the distribution of elements (black and red boxes) of representative genes (middle) and promoters (bottom).

the most significant elements that correspond to the localization of transcription initiation. Taking advantage of this rich data source and using millions of sequences selected based on specific *de novo* motifs, we identified the most enriched oligomers and compared their binding frequencies by the active TFs (Supplementary Table S2). Although we excluded rare, unusual, and low-affinity elements that may also play important roles (91), this approach provides exact sequences with exact specificity instead of motifs representing mixes of sequences with different characteristics. Importantly, some of these characteristics have significant functional consequences for transcription regulation, which have not been shown in detail before because mixing up different elements obscures causality.

Now, we demonstrate that macrophage-specific TF binding sites with a methylatable CG dinucleotide are typically (i) located proximal to TSSs and (ii) bound by promoter-specific TFs and (iii) their specific interactions result in high chromatin openness and high polymerase activity relative to the TSS-distal DNA-TF interactions independent of DNA methylation. Importantly, we also demonstrate significant exceptions to these regularities: promoters composed exclusively of non-methylatable, macrophage-specific, and prototypically promoter-distal elements can also be responsible for the regulation of macrophage-specific gene expression. This class of promoters is not unprecedented in the literature on macrophage gene expression: for example, both *Csf1r* and *Tlr9* have been described with TATA-less and non-CGI promoters that contain purine-rich elements (Figure 7C, Supplementary Figure S7B) (92,93). However, we thoroughly characterized, contrasted with more traditional ones, and greatly extended the list of these promoters (Supplementary Table S4).

Although >80% of the open-chromatin regions could be covered by the identified elements, the major limitation of our approach is the missing DNA-TF interactions. Most of the highest expressed TFs are coupled with a cistrome, but there are still many, whose binding sites would be informative regarding the gene regulation of macrophages. Based on gene expression levels, there are several candidates to fill this gap. For example, >20 bHLH genes are at the top of the TF list, suggesting the presence of different bHLH dimers in BMDMs, and the large number of identified E-box variants further strengthens the notion about an extensive bHLH network. Regardless, members of the MiTF/TFE family have similar binding sites to those of USF1/2, as well as the partner and competitors of cMYC (respectively MAX and MAD proteins) may play important roles in unstimulated macrophages (76,94–97). There are additional, highly expressed TFs with possible binding sites such as PBX2, XBP1, or YBX1. PBX motifs were identified in open-chromatin regions. A part of CREB1 binding sites may function as the XBP1-specific X-boxes, suggesting a possible collaboration between CREB1 and XBP1 (bZIP) (98). Furthermore, there can be some kind of interaction between NFY and YBX1, as well, because they have identical elements: CCAAT- and Y-boxes, respectively (99). In the case of the TFs listed above, we cannot be sure that these are permanently active components of the chromatin because several TFs have small cistromes in spite of their high gene expression levels. These characteristics are true of several well-known macrophage-specific SDTFs from the IRF, STAT, NFκB and nuclear receptor families (25,63). Some bZIP proteins can also be classified here as they can be sensitive to either endoplasmic reticulum (ER) stress (CREB3, ATF6) or amino acid deprivation (ATF4) (33,100). Interestingly, while motifs of SP1, GFY and NRF1 showed high enrichment, no TATA-box could be detected in any of the motif enrichment analyses. This does not mean that all identified promoters lack this element, rather these elements are inactive, e.g. suppressed by the highly expressed TBP inhibitor DR1/DRAP1 heterodimer (65).

Most of the extensively examined TFs have largely promoter-distal cistromes, and only 4 TFs (GABPα, ELF1, CREB1 and cMYC) have a preference for promoters vs. distal regions. Importantly, there is a sharp demarcation between proximal and distal binding sites around 1 kb relative to the closest TSS (this is the separating line between all bimodal distributions observed). Active promoters detected by TF binding, in general, show a low extent of DNA methylation in a broad, up to 1 kb lane, as well as high chromatin openness and high transcription initiation and elongation rates. In contrast, approximately only half of the promoter-distal regions are open and active. The length of the non-methylated lane, in this case, is a few hundred bp, except for the ATF2/3-specific binding sites, where only minimal demethylation can be observed. Notably, there is a well-visible difference between the binding sites of signal-activated (ATF3) and signal-inhibited repressors (BACH1 and MAFB). In the latter case, there is a more active or at least more accessible chromatin environment showing some demethylation, openness, and transcription. Overall, patterns of demethylation, co-localizing TFs, chromatin openness, and polymerase activity highly correlate with each other, which is in agreement with the general DNA methylation patterns of vertebrates according to which all cytosines—out of the active *cis*-regulatory regions and most CGIs—are methylated (101,102).

Based on our and others' results, the feature of sequences to be methylated serves as a switch, so independent of the clustering and composition of elements, their DNA environment should be demethylated to become active (101). Except for the transcriptional repressor Kaiso (103,104), TFs show low or no affinity to methylated sequences, so demethylation probably originates from certain non-methylatable elements, which recruit DNA demethylases releasing the neighbouring DNA segments—including methylatable elements—from inhibition (Supplementary Figure S7C). This suggests that each promoter should carry at least one strong, non-methylatable element, for example, a TATA-, CCAAT- or GT-box besides the CG-containing elements. In contrast, promoter-distal elements are less strictly regulated by DNA methylation, they contain CG dinucleotides less frequently, and their function depends primarily on the concentration of their binding TFs (Supplementary Figure S7C). Likely this explains the high mRNA levels of LDTFs compared to the basically CGI-binding TFs such as GABP, CREB1, cMYC, as well as the SP1 family members. In line with this, elements of the well-known LDTFs of macrophages (PU.1, IRF8, C/EBP, AP-1, MEF2 and RUNX1) are typically non-methylatable, either they do not contain a CG dinucleotide, or it does not have a significant effect. Despite this, promoters that consist exclusively of the elements of LDTFs appoint the TSS(s) of hundreds of macrophage-specific genes. This means that not only promoter-distal regions, but also promoters can be activated by LDTFs prior to demethylation and independent of CGI-binding TFs. Although these contain less CG sites, their methylation level is similar to those measured at CGI promot-

ers (Figure 7C, Supplementary Figure S7B). Notably, CG sites also extend to the 5′ UTR(s) of these genes like in the case of CGI promoters (Figure 7C). As the extent of demethylation around gene regulatory sites is a feature discriminating promoters from enhancers (101), CGs in 5′ UTRs may participate in directing polymerases towards the genes, thus facilitating elongation, while this does not happen at most enhancers.

Considering the composition of the identified macrophage-specific promoters, we termed them as LDTF-driven promoters. There are two evolutionary ways leading to this composition: (i) certain groups of promoter-distal elements (enhancers) gained promoter properties (got 'promoterized') to initiate cell type-specific gene transcripts or (ii) certain promoters have lost their characteristic elements to be controlled exclusively by LDTFs. These processes are in line with the results of CAGE experiments from different mammals, according to which immune-related promoters are not highly conserved, suggesting a rapid evolution of the immune system (6). We propose that this class of promoters can be a general feature in cell type-specific gene expression and affects not only myeloid or immune cells but any other cell type to a certain extent.

Data availability

The data underlying this article are available in the Sequence Read Archive at https://www.ncbi.nlm.nih.gov/sra, and can be accessed with SRR23683735 and SRR25923453.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

The authors would like to thank Lajos Széles, Zsolt Czimmerer, and Krisztián Bene for the helpful discussions. This study makes use of its own and publicly available data generated by various labs, which were cited in Supplementary Table S1.

Funding

Nuclear Receptor Research Laboratory was supported by grants from the Hungarian Scientific Research Fund [OTKA PD135102 to G.N., PD137902 to D.B., KKP129909]. Funding for open access charge: OTKA PD135102, KKP129909.

Conflict of interest statement

None declared.

References

1. Daniel,B., Nagy,G., Hah,N., Horvath,A., Czimmerer,Z., Poliska,S., Gyuris,T., Keirsse,J., Gysemans,C., Van Ginderachter,J.A., *et al.* (2014) The active enhancer network operated by liganded RXR supports angiogenic activity in macrophages. *Genes Dev.*, **28**, 1562–1577.
2. Hah,N., Murakami,S., Nagari,A., Danko,C.G. and Lee Kraus,W. (2013) Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res.*, **23**, 1210–1223.
3. Li,W., Notani,D., Ma,Q., Tanasa,B., Nunez,E., Chen,A.Y., Merkurjev,D., Zhang,J., Ohgi,K., Song,X., *et al.* (2013) Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*, **498**, 516–520.
4. Wagner,E.J., Tong,L. and Adelman,K. (2023) Integrator is a global promoter-proximal termination complex. *Mol. Cell*, **83**, 416–427.
5. Balwierz,P.J., Carninci,P., Daub,C.O., Kawai,J., Hayashizaki,Y., Van Belle,W., Beisel,C. and van Nimwegen,E. (2009) Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol.*, **10**, R79.
6. Forrest,A.R.R., Kawaji,H., Rehli,M., Baillie,J.K., De Hoon,M.J.L., Haberle,V., Lassmann,T., Kulakovskiy,I.V., Lizio,M., Itoh,M., *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.
7. Adiconis,X., Haber,A.L., Simmons,S.K., Levy Moonshine,A., Ji,Z., Busby,M.A., Shi,X., Jacques,J., Lancaster,M.A., Pan,J.Q., *et al.* (2018) Comprehensive comparative analysis of 5′-end RNA-sequencing methods. *Nat. Methods*, **15**, 505–511.
8. Gershenzon,N.I. and Ioshikhes,I.P. (2005) Synergy of human pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics*, **21**, 1295–1300.
9. Baumann,M., Pontiller,J. and Ernst,W. (2010) Structure and basal transcription complex of RNA polymerase II core promoters in the mammalian genome: an overview. *Mol. Biotechnol.*, **45**, 241–247.
10. Sandelin,A., Carninci,P., Lenhard,B., Ponjavic,J., Hayashizaki,Y. and Hume,D.A. (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat. Rev. Genet.*, **8**, 424–436.
11. Curina,A., Termanini,A., Barozzi,I., Prosperini,E., Simonatto,M., Polletti,S., Silvola,A., Soldi,M., Austenaa,L., Bonaldi,T., *et al.* (2017) High constitutive activity of a broad panel of housekeeping and tissue-specific cis-regulatory elements depends on a subset of ETS proteins. *Genes Dev.*, **31**, 399–412.
12. FitsGerald,P.C., Shlyakhtenko,A., Mir,A.A. and Vinson,C. (2004) Clustering of DNA sequences in human promoters. *Genome Res.*, **14**, 1562–1574.
13. 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.
14. Nagy,G., Dániel,B., Jónás,D., Nagy,L. and Barta,E. (2013) A novel method to predict regulatory regions based on histone mark landscapes in macrophages. *Immunobiology*, **218**, 1416–1427.
15. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
16. Hong,C.P., Choe,M.K. and Roh,T.-Y. (2012) Characterization of chromatin structure-associated histone modifications in breast cancer cells. *Genomics Inform.*, **10**, 145–152.
17. Purbey,P.K., Scumpia,P.O., Kim,P.J., Tong,A.J., Iwamoto,K.S., McBride,W.H. and Smale,S.T. (2017) Defined sensing mechanisms and signaling pathways contribute to the global inflammatory gene expression output elicited by ionizing radiation. *Immunity*, **47**, 421–434.
18. Link,V.M., Duttke,S.H., Chun,H.B., Holtman,I.R., Westin,E., Hoeksema,M.A., Abe,Y., Skola,D., Romanoski,C.E., Tao,J., *et al.* (2018) Analysis of genetically diverse macrophages reveals local and domain-wide mechanisms that control transcription factor binding and function. *Cell*, **173**, 1796–1809.
19. Lichtinger,M., Ingram,R., Hannah,R., Müller,D., Clarke,D., Assi,S.A., Lie-A-Ling,M., Noailles,L., Vijayabaskar,M.S., Wu,M., *et al.* (2012) RUNX1 reshapes the epigenetic landscape at the onset of haematopoiesis. *EMBO J.*, **31**, 4318–4333.
20. Feng,R., Desbordes,S.C., Xie,H., Tillo,E.S., Pixley,F., Stanley,E.R. and Graf,T. (2008) PU.1 and C/EBPalpha/beta convert

fibroblasts into macrophage-like cells. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 6057–6062.
21. Di Tullio,A., Vu Manh,T.P., Schubert,A., Månsson,R. and Graf,T. (2011) CCAAT/enhancer binding protein alpha (C/EBP(alpha))-induced transdifferentiation of pre-B cells into macrophages involves no overt retrodifferentiation. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 17016–17021.
22. Guha,M. and Mackman,N. (2001) LPS induction of gene expression in human monocytes. *Cell. Signal.*, **13**, 85–94.
23. Cilenti,F., Barbiera,G., Caronni,N., Iodice,D., Montaldo,E., Barresi,S., Lusito,E., Cuzzola,V., Vittoria,F.M., Mezzanzanica,L., *et al.* (2021) A PGE2-MEF2A axis enables context-dependent control of inflammatory gene expression. *Immunity*, **54**, 1665–1682.
24. Mancino,A., Termanini,A., Barozzi,I., Ghisletti,S., Ostuni,R., Prosperini,E., Ozato,K. and Natoli,G. (2015) A dual cis-regulatory code links IRF8 to constitutive and inducible gene expression in macrophages. *Genes Dev.*, **29**, 394–408.
25. Platanitis,E. and Decker,T. (2018) Regulatory networks involving STATs, IRFs, and NFκB in inflammation. *Front. Immunol.*, **9**, 2542.
26. Czimmerer,Z., Daniel,B., Horvath,A., Rückerl,D., Nagy,G., Kiss,M., Peloquin,M., Budai,M.M., Cuaranta-Monroy,I., Simandi,Z., *et al.* (2018) The transcription factor STAT6 mediates direct repression of inflammatory enhancers and limits activation of alternatively polarized macrophages. *Immunity*, **48**, 75–90.
27. Daniel,B., Nagy,G., Horvath,A., Czimmerer,Z., Cuaranta-Monroy,I., Poliska,S., Hays,T.T., Sauer,S., Francois-Deleuze,J. and Nagy,L. (2018) The IL-4/STAT6/PPARγ signaling axis is driving the expansion of the RXR heterodimer cistrome, providing complex ligand responsiveness in macrophages. *Nucleic Acids Res.*, **46**, 4425–4439.
28. Daniel,B., Nagy,G., Czimmerer,Z., Horvath,A., Hammers,D.W., Cuaranta-Monroy,I., Poliska,S., Tzerpos,P., Kolostyak,Z., Hays,T.T., *et al.* (2018) The nuclear receptor pparγ controls progressive macrophage polarization as a ligand-insensitive epigenomic ratchet of transcriptional memory. *Immunity*, **49**, 615–626.
29. Daniel,B., Czimmerer,Z., Halasz,L., Boto,P., Kolostyak,Z., Poliska,S., Berger,W.K., Tzerpos,P., Nagy,G., Horvath,A., *et al.* (2020) The transcription factor EGR2 is the molecular linchpin connecting STAT6 activation to the late, stable epigenomic program of alternative macrophage polarization. *Genes Dev.*, **34**, 1474–1492.
30. Piccolo,V., Curina,A., Genua,M., Ghisletti,S., Simonatto,M., Sabò,A., Amati,B., Ostuni,R. and Natoli,G. (2017) Opposing macrophage polarization programs show extensive epigenomic and transcriptional cross-talk. *Nat. Immunol.*, **18**, 530–540.
31. Barish,G.D., Yu,R.T., Karunasiri,M., Ocampo,C.B., Dixon,J., Benner,C., Dent,A.L., Tangirala,R.K. and Evans,R.M. (2010) Bcl-6 and NF- B cistromes mediate opposing regulation of the innate immune response. *Genes Dev.*, **24**, 2760–2765.
32. Nguyen,H.C.B., Adlanmerini,M., Hauck,A.K. and Lazar,M.A. (2020) Dichotomous engagement of HDAC3 activity governs inflammatory responses. *Nature*, **584**, 286–290.
33. Halaby,M.J., Hezaveh,K., Lamorte,S., Ciudad,M.T., Kloetgen,A., MacLeod,B.L., Guo,M., Chakravarthy,A., Medina,T.D.S., Ugel,S., *et al.* (2019) GCN2 drives macrophage and MDSC function and immunosuppression in the tumor microenvironment. *Sci. Immunol.*, **4**, eaax8189.
34. Eichenfield,D.Z., Troutman,T.D., Link,V.M., Lam,M.T., Cho,H., Gosselin,D., Spann,N.J., Lesch,H.P., Tao,J., Muto,J., *et al.* (2016) Tissue damage drives co-localization of NF-κb, Smad3, and Nrf2 to direct rev-erb sensitive wound repair in mouse macrophages. *eLife*, **5**, e13024.
35. Ogawa,K., Sun,J., Taketani,S., Nakajima,O., Nishitani,C., Sassa,S., Hayashi,N., Yamamoto,M., Shibahara,S., Fujita,H., *et al.* (2001) Heme mediates derepression of Maf recognition element through direct binding to transcription repressor Bach1. *EMBO J.*, **20**, 2835–2843.
36. Soucie,E.L., Weng,Z., Geirsdóttir,L., Molawi,K., Maurizio,J., Fenouil,R., Mossadegh-Keller,N., Gimenez,G., Vanhille,L., Beniazza,M., *et al.* (2016) Lineage-specific enhancers activate self-renewal genes in macrophages and embryonic stem cells. *Science*, **351**, aad5510.
37. Barish,G.D., Downes,M., Alaynick,W.A., Yu,R.T., Ocampo,C.B., Bookout,A.L., Mangelsdorf,D.J. and Evans,R.M. (2005) A nuclear receptor atlas: macrophage activation. *Mol. Endocrinol.*, **19**, 2466–2477.
38. Patsalos,A., Tzerpos,P., Halasz,L., Nagy,G., Pap,A., Giannakis,N., Lyroni,K., Koliaraki,V., Pintye,E., Dezso,B., *et al.* (2019) The BACH1–HMOX1 regulatory axis is indispensable for proper macrophage subtype specification and skeletal muscle regeneration. *J. Immunol.*, **203**, 1532–1547.
39. Barta,E. (2011) Command line analysis of ChIP-seq results. *EMBnet.journal*, **17**, 13.
40. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
41. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
42. 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.
43. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
44. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
45. de Hoon,M.J.L., Imoto,S., Nolan,J. and Miyano,S. (2004) Open source clustering software. *Bioinformatics*, **20**, 1453–1454.
46. Kim,D., Paggi,J.M., Park,C., Bennett,C. and Salzberg,S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**, 907–915.
47. Ramírez,F., Dündar,F., Diehl,S., Grüning,B.A. and Manke,T. (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.*, **42**, W187–W191.
48. Pertea,M., Pertea,G.M., Antonescu,C.M., Chang,T.C., Mendell,J.T. and Salzberg,S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.
49. Ge,S.X., Jung,D. and Jao,R. (2020) ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics*, **36**, 2628–2629.
50. Zhou,Q., Liu,M., Xia,X., Gong,T., Feng,J., Liu,W., Liu,Y., Zhen,B., Wang,Y., Ding,C., *et al.* (2017) A mouse tissue transcription factor atlas. *Nat. Commun.*, **8**, 15089.
51. Skinner,M.K., Rawls,A., Wilson-Rawls,J. and Roalson,E.H. (2010) Basic helix-loop-helix transcription factor gene family phylogenetics and nomenclature. *Differentiation*, **80**, 1–8.
52. Thorvaldsdóttir,H., Robinson,J.T. and Mesirov,J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.
53. Saldanha,A.J. (2004) Java treeview–extensible visualization of microarray data. *Bioinformatics*, **20**, 3246–3248.
54. Kestler,H.A., Muller,A., Gress,T.M. and Buchholz,M. (2005) Generalized Venn diagrams: a new method of visualizing complex genetic set relations. *Bioinformatics*, **21**, 1592–1595.
55. Khan,A. and Mathelier,A. (2017) Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. *BMC Bioinf.*, **18**, 287.

56. Rajbhandari,P., Thomas,B.J., Feng,A.C., Hong,C., Wang,J., Vergnes,L., Sallam,T., Wang,B., Sandhu,J., Seldin,M.M., *et al.* (2018) IL-10 signaling remodels adipose chromatin architecture to limit thermogenesis and energy expenditure. *Cell*, **172**, 218–233.
57. Kawakami,R., Kitagawa,Y., Chen,K.Y., Arai,M., Ohara,D., Nakamura,Y., Yasuda,K., Osaki,M., Mikami,N., Lareau,C.A., *et al.* (2021) Distinct Foxp3 enhancer elements coordinate development, maintenance, and function of regulatory T cells. *Immunity*, **54**, 947–961.
58. Rajakumari,S., Wu,J., Ishibashi,J., Lim,H.W., Giang,A.H., Won,K.J., Reed,R.R. and Seale,P. (2013) EBF2 determines and maintains brown adipocyte identity. *Cell Metab.*, **17**, 562–574.
59. Hepler,C., Weidemann,B.J., Waldeck,N.J., Marcheva,B., Cedernaes,J., Thorne,A.K., Kobayashi,Y., Nozawa,R., Newman,M.V., Gao,P., *et al.* (2022) Time-restricted feeding mitigates obesity through adipocyte thermogenesis. *Science*, **378**, 276–284.
60. Nagy,G., Daniel,B., Cuaranta-Monroy,I. and Nagy,L. (2020) Unraveling the hierarchy of *cis* and *trans* factors that determine the DNA binding by PPARγ. *Mol. Cell. Biol.*, **40**, e00547-19.
61. Bojcsuk,D., Nagy,G. and Bálint,B.L. (2020) Alternatively constructed estrogen receptor alpha-driven super-enhancers result in similar gene expression in breast and endometrial cell lines. *Int. J. Mol. Sci.*, **21**, 1630.
62. Clausen,B.E., Burkhardt,C., Reith,W., Renkawitz,R. and Förster,I. (1999) Conditional gene targeting in macrophages and granulocytes using LysMcre mice. *Transgenic Res.*, **8**, 265–277.
63. Mangelsdorf,D.J., Thummel,C., Beato,M., Herrlich,P., Schütz,G., Umesono,K., Blumberg,B., Kastner,P., Mark,M., Chambon,P., *et al.* (1995) The nuclear receptor superfamily: the second decade. *Cell*, **83**, 835–839.
64. Nagy,G., Czipa,E., Steiner,L., Nagy,T., Pongor,S., Nagy,L. and Barta,E. (2016) Motif oriented high-resolution analysis of ChIP-seq data reveals the topological order of CTCF and cohesin proteins on DNA. *BMC Genomics*, **17**, 637.
65. Mermelstein,F., Yeung,K., Cao,J., Inostroza,J.A., Erdjument-Bromage,H., Eagelson,K., Landsman,D., Levitt,P., Tempst,P. and Reinberg,D. (1996) Requirement of a corepressor for Dr1-mediated repression of transcription. *Genes Dev.*, **10**, 1033–1048.
66. Nagy,G. and Nagy,L. (2020) Motif grammar: the basis of the language of gene expression. *Comput. Struct. Biotechnol. J.*, **18**, 2026–2032.
67. Meraro,D., Gleit-Kielmanowicz,M., Hauser,H. and Levi,B.-Z. (2002) IFN-stimulated gene 15 is synergistically activated through interactions between the myelocyte/lymphocyte-specific transcription factors, PU.1, IFN regulatory Factor-8/IFN consensus sequence binding protein, and IFN regulatory Factor-4: characterization of a new subtype of IFN-stimulated Response element. *J. Immunol.*, **168**, 6224–6231.
68. Tamura,T., Thotakura,P., Tanaka,T.S., Ko,M.S.H. and Ozato,K. (2005) Identification of target genes and a unique cis element regulated by IRF-8 in developing macrophages. *Blood*, **106**, 1938–1947.
69. Ayer,D.E., Kretzner,L. and Eisenman,R.N. (1993) Mad: a heterodimeric partner for Max that antagonizes Myc transcriptional activity. *Cell*, **72**, 211–222.
70. Wei,G.H., Badis,G., Berger,M.F., Kivioja,T., Palin,K., Enge,M., Bonke,M., Jolma,A., Varjosalo,M., Gehrke,A.R., *et al.* (2010) Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J.*, **29**, 2147–2160.
71. Amoutzias,G.D., Veron,A.S., Weiner,J., Robinson-Rechavi,M., Bornberg-Bauer,E., Oliver,S.G. and Robertson,D.L. (2007) One billion years of bZIP transcription factor evolution: conservation and change in dimerization and DNA-binding site specificity. *Mol. Biol. Evol.*, **24**, 827–835.
72. 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.
73. Shan,J., Zhang,F., Sharkey,J., Tang,T.A., Örd,T. and Kilberg,M.S. (2016) The C/ebp-Atf response element (CARE) location reveals two distinct Atf4-dependent, elongation-mediated mechanisms for transcriptional induction of aminoacyl-tRNA synthetase genes in response to amino acid limitation. *Nucleic Acids Res.*, **44**, 9719–9732.
74. Motohashi,H., Shavit,J.A., Igarashi,K., Yamamoto,M. and Engel,J.D. (1997) The world according to Maf. *Nucleic Acids Res.*, **25**, 2953–2959.
75. Itoh,K., Chiba,T., Takahashi,S., Ishii,T., Igarashi,K., Katoh,Y., Oyake,T., Hayashi,N., Satoh,K., Hatayama,I., *et al.* (1997) An Nrf2/small maf heterodimer mediates the induction of phase II detoxifying enzyme genes through antioxidant response elements. *Biochem. Biophys. Res. Commun.*, **236**, 313–322.
76. Aksan,I. and Goding,C.R. (1998) Targeting the microphthalmia basic helix-loop-helix-leucine zipper transcription factor to a subset of E-box elements in vitro and in vivo. *Mol. Cell. Biol.*, **18**, 6930–6938.
77. Meyers,S., Downing,J.R. and Hiebert,S.W. (1993) Identification of AML-1 and the (8;21) translocation protein (AML-1/ETO) as sequence-specific DNA-binding proteins: the runt homology domain is required for DNA binding and protein-protein interactions. *Mol. Cell. Biol.*, **13**, 6336–6345.
78. Tuoresmäki,P., Väisänen,S., Neme,A., Heikkinen,S. and Carlberg,C. (2014) Patterns of genome-wide VDR locations. *PLoS One*, **9**, e96105.
79. Carninci,P., Sandelin,A., Lenhard,B., Katayama,S., Shimokawa,K., Ponjavic,J., Semple,C.A.M., Taylor,M.S., Engström,P.G., Frith,M.C., *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.
80. Dai,X.M., Ryan,G.R., Hapel,A.J., Dominguez,M.G., Russell,R.G., Kapp,S., Sylvestre,V. and Stanley,E.R. (2002) Targeted disruption of the mouse colony-stimulating factor 1 receptor gene results in osteopetrosis, mononuclear phagocyte deficiency, increased primitive progenitor cell frequencies, and reproductive defects. *Blood*, **99**, 111–120.
81. Park,B.S. and Lee,J.O. (2013) Recognition of lipopolysaccharide pattern by TLR4 complexes. *Exp. Mol. Med.*, **45**, e66.
82. Kober,D.L. and Brett,T.J. (2017) TREM2-Ligand interactions in health and disease. *J. Mol. Biol.*, **429**, 1607–1629.
83. Takaoka,A., Yanai,H., Kondo,S., Duncan,G., Negishi,H., Mizutani,T., Kano,S.I., Honda,K., Ohba,Y., Mak,T.W., *et al.* (2005) Integral role of IRF-5 in the gene induction programme activated by toll-like receptors. *Nature*, **434**, 243–249.
84. Sadler,A.J. and Williams,B.R.G. (2008) Interferon-inducible antiviral effectors. *Nat. Rev. Immunol.*, **8**, 559–568.
85. Choubey,D., Duan,X., Dickerson,E., Ponomareva,L., Panchanathan,R., Shen,H. and Srivastava,R. (2010) Interferon-inducible p200-Family proteins as novel sensors of cytoplasmic DNA: role in inflammation and autoimmunity. *J. Interf. Cytokine Res.*, **30**, 371.
86. Mar,K.B., Rinkenberger,N.R., Boys,I.N., Eitson,J.L., McDougal,M.B., Richardson,R.B. and Schoggins,J.W. (2018) LY6E mediates an evolutionarily conserved enhancement of virus infection by targeting a late entry step. *Nat. Commun.*, **9**, 3603.
87. Maurer,M. and Von Stebut,E. (2004) Macrophage inflammatory protein-1. *Int. J. Biochem. Cell Biol.*, **36**, 1882–1886.
88. Shapiro,S.D., Kobayashi,D.K. and Ley,T.J. (1993) Cloning and characterization of a unique elastolytic metalloproteinase produced by human alveolar macrophages. *J. Biol. Chem.*, **268**, 23824–23829.
89. Shi,G.P., Munger,J.S., Meara,J.P., Rich,D.H. and Chapman,H.A. (1992) Molecular cloning and expression of human alveolar macrophage cathepsin S, an elastinolytic cysteine protease. *J. Biol. Chem.*, **267**, 7258–7262.

90. Morel,S., Lévy,F., Burlet-Schiltz,O., Brasseur,F., Probst-Kepper,M., Peitrequin,A.L., Monsarrat,B., Van Velthoven,R., Cerottini,J.C., Boon,T., *et al.* (2000) Processing of some antigens by the standard proteasome but not by the immunoproteasome results in poor presentation by dendritic cells. *Immunity*, **12**, 107–117.
91. Farley,E.K., Olson,K.M., Zhang,W., Rokhsar,D.S. and Levine,M.S. (2016) Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 6508–6513.
92. Hume,D.A., Summers,K.M. and Rehli,M. (2016) Transcriptional regulation and macrophage differentiation. *Microbiol. Spectr.*, **4**, https://doi.org/10.1128/microbiolspec.MCHD-0024-2015.
93. Schroder,K., Lichtinger,M., Irvine,K.M., Brion,K., Trieu,A., Ross,I.L., Ravasi,T., Stacey,K.J., Rehli,M., Hume,D.A., *et al.* (2007) PU.1 and ICSBP control constitutive and IFN-gamma-regulated Tlr9 gene expression in mouse macrophages. *J. Leukoc. Biol.*, **81**, 1577–1590.
94. Corre,S. and Galibert,M.D. (2005) Upstream stimulating factors: highly versatile stress-responsive transcription factors. *Pigment Cell Res*, **18**, 337–348.
95. Madden,S.K., de Araujo,A.D., Gerhardt,M., Fairlie,D.P. and Mason,J.M. (2021) Taking the Myc out of cancer: toward therapeutic strategies to directly inhibit c-myc. *Mol. Cancer*, **20**, 3.
96. Pogenberg,V., Ögmundsdóttir,M.H., Bergsteinsdóttir,K., Schepsky,A., Phung,B., Deineko,V., Milewski,M., Steingrímsson,E. and Wilmanns,M. (2012) Restricted leucine zipper dimerization and specificity of DNA recognition of the melanocyte master regulator MITF. *Genes Dev.*, **26**, 2647–2658.
97. Carroll,P.A., Freie,B.W., Mathsyaraja,H. and Eisenman,R.N. (2018) The MYC transcription factor network: balancing metabolism, proliferation and oncogenesis. *Front. Med.*, **12**, 412–425.
98. Yoshida,H., Matsui,T., Yamamoto,A., Okada,T. and Mori,K. (2001) XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. *Cell*, **107**, 881–891.
99. Dolfini,D. and Mantovani,R. (2013) Targeting the Y/CCAAT box in cancer: YB-1 (YBX1) or NF-Y&quest. *Cell Death Differ.*, **20**, 676–685.
100. Asada,R., Kanemoto,S., Kondo,S., Saito,A. and Imaizumi,K. (2011) The signalling from endoplasmic reticulum-resident bZIP transcription factors involved in diverse cellular physiology. *J. Biochem.*, **149**, 507–518.
101. Schübeler,D. (2015) Function and information content of DNA methylation. *Nature*, **517**, 321–326.
102. Kreibich,E., Kleinendorst,R., Barzaghi,G., Kaspar,S. and Krebs,A.R. (2023) Single-molecule footprinting identifies context-dependent regulation of enhancers by DNA methylation. *Mol. Cell*, **83**, 787–802.
103. Prokhortchouk,A., Hendrich,B., Jørgensen,H., Ruzov,A., Wilm,M., Georgiev,G., Bird,A. and Prokhortchouk,E. (2001) The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. *Genes Dev.*, **15**, 1613–1618.
104. Buck-Koehntop,B.A., Stanfield,R.L., Ekiert,D.C., Martinez-Yamout,M.A., Dyson,H.J., Wilson,I.A. and Wright,P.E. (2012) Molecular basis for recognition of methylated and specific DNA sequences by the zinc finger protein Kaiso. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 15229–15234.

Received: March 6, 2023. **Revised:** January 18, 2024. **Editorial Decision:** January 20, 2024. **Accepted:** January 29, 2024