Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet

Coby Viner[1,2], James Johnson[3], Nicolas Walker[4], Hui Shi[4], Marcela Sjöberg[5], David J. Adams[5], Anne C. Ferguson-Smith[4], Timothy L. Bailey[3], and Michael M. Hoffman[1,2,6,*]

[1]*Department of Computer Science, University of Toronto, Toronto, ON, Canada*
[2]*Princess Margaret Cancer Centre, Toronto, ON, Canada*
[3]*Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD, Australia*
[4]*Department of Genetics, University of Cambridge, Cambridge, England*
[5]*Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge, England*
[6]*Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada*
[*]Correspondence: michael.hoffman@utoronto.ca

Abstract

Introduction. Many transcription factors initiate transcription only in specific sequence contexts, providing the means for sequence specificity of transcriptional control. A four-letter DNA alphabet only partially describes the possible diversity of nucleobases a transcription factor might encounter. For instance, cytosine is often present in a covalently modified form: 5-methylcytosine (5mC). 5mC can be successively oxidized to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC). Just as transcription factors distinguish one unmodified nucleobase from another, some have been shown to distinguish unmodified bases from these covalently modified bases. Modification-sensitive transcription factors provide a mechanism by which widespread changes in DNA methylation and hydroxymethylation can dramatically shift active gene expression programs.

Methods. To understand the effect of modified nucleobases on gene regulation, we developed methods to discover motifs and identify transcription factor binding sites in DNA with covalent modifications. Our models expand the standard `A`/`C`/`G`/`T` alphabet, adding `m` (5mC) `h` (5hmC), `f` (5fC), and `c` (5caC). We additionally add symbols to encode guanine complementary to these modified cytosine nucleobases, as well as symbols to represent states of ambiguous modification. We adapted the well-established position weight matrix model of transcription factor binding affinity to an expanded alphabet. We developed a program, Cytomod, to create a modified sequence. We also enhanced the MEME Suite to be able to handle custom alphabets. These versions permit users to specify new alphabets, anticipating future alphabet expansions.

Results. We created an expanded-alphabet sequence using whole-genome maps of 5mC and 5hmC in naive *ex vivo* mouse T cells. Using this sequence and ChIP-seq data from Mouse ENCODE and others, we identified modification-sensitive *cis*-regulatory modules. We elucidated various known methylation binding preferences, including the preference of ZFP57 and C/EBPβ for methylated motifs and the preference of c-Myc for unmethylated E-box motifs. We demonstrated that our method is robust to parameter perturbations, with transcription factors' sensitivities for methylated and hydroxymethylated DNA broadly conserved across a range of modified base calling thresholds. Hypothesis testing across different threshold values was used to determine cutoffs most suitable for further analyses. Using these known binding preferences to tune model parameters enables discovery of novel modified motifs.

Discussion. Hypothesis testing of motif central enrichment provides a natural means of differentially assessing modified versus unmodified binding affinity, without most of the limitations of a *de novo* analysis. This approach can be readily extended to other DNA modifications, provided genome-wide single-base resolution data is available. As more high-resolution epigenomic data becomes available, we expect this method to continue to yield insights into altered transcription factor binding affinities across a variety of modifications.

Introduction

Different cell types have widely varied gene expression, despite sharing the same genomic sequence. Epigenomic factors, including modifications to DNA; RNA; and proteins, modulate gene expression and contribute to the cellular regulatory program. Covalent cytosine modifications have an important regulatory role across a number of eukaryotic species, including both mice and humans.[1] The most well-studied cytosine modification involves the addition of a methyl group to the 5′ carbon of cytosine, creating 5-methylcytosine (5mC). Modified cytosine nucleobases do not substantively disrupt the overall structure of the DNA double helix, permitting transcription and replication to occur. However, these modifications alter various properties of the double-helix, including altering the conformation of both the major and minor grooves.[2] They can also lead to steric hindrance of transcription factor DNA binding domains, relative to the typical interactions of specific DNA motifs with their cognate binding sites.[3,4]

The demethylation cascade as functional genomic elements

5mC is the first of four modified cytosine nulceobases, that are involved in the demethylation of 5mC back to its unmodified form. This demethylation cascade occurs via successive oxidation of 5mC, to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC; Figure 1).[5,6] These oxidations are mediated by ten-eleven translocation (TET) enzymes.[7]

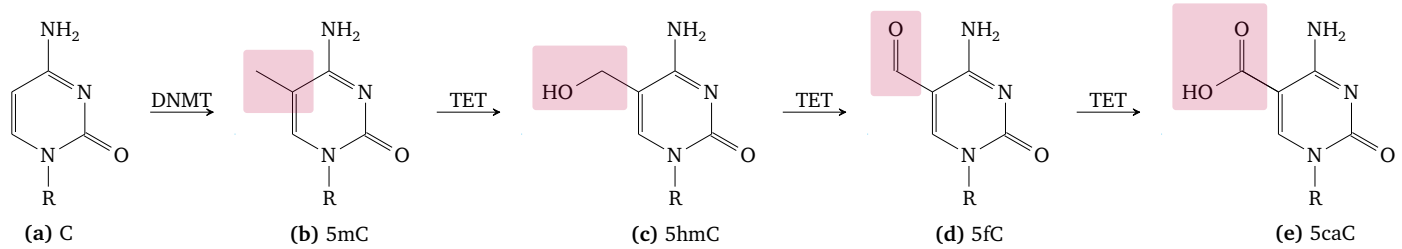


Figure 1. Stepwise epigenetic modification of cytosine. (a) Cytosine is methylated by DNA methyltransferase (DNMT) to create (b) 5-methylcytosine, which is oxidized by the ten-eleven translocation (TET) enzyme to create (c) 5-hydroxymethylcytosine, which is again oxidized by TET to create (d) 5-formylcytosine. Finally, 5fC can be oxidized to (e) 5-carboxylcytosine, which can then return to an unmodified cytosine via decarboxylation or thymine DNA glycosylase (TDG) mediated excision, followed by base excision repair. R indicates deoxyribose and the rest of a DNA molecule. Purple rectangles indicate functional groups changed in the reaction.

5mC has long been known to be involved in a diverse set of regulatory roles.[8,9] 5hmC is increasingly being implicated in regulatory processes,[10] and is now known to be a stable epigenetic modification,[11] with structural rationale for its reduced propensity of TET-mediated oxidation.[12] Far less is known about 5fC and 5caC, largely due to their considerably lesser abundance. They are far less abundant than 5hmC, itself around an order of magnitude less abundant than 5mC. The abundance of these modifications varies by cell type, with greater abundance observed in mouse embryonic stem cells (mESCs),[13–15] in which nearly 3% of cytosine bases were methylated,[5] while 0.055% were hydroxymethylated in a different mESC sample.[16] There have been only a few investigations into the genome-wide distributions and roles of 5fC (such as poised enhancers) and 5caC.[15,17,18] They are often regarded as mere intermediates of the demethylation cascade, largely due to their generally being two to three orders of magnitude less abundant than 5hmC, and capable of triggering a strong DNA damage response.[1,5,16] In mESCs, 5fC was found to account for 0.0014% of cytosine bases,[16] while 5caC accounted for a mere 0.000335%.[5] While it is by no means certain that they play a distinctive regulatory role across multiple tissue types, converging lines of evidence suggest that they too can be important modulators of gene expression.[10] 5fC alters the conformation of the DNA double helix[19] and is known to be stable in mESCs, not merely a demethylation intermediate.[20]

All of these modifications are (by far) most frequent at CpG dinucleotides, but non-CpG 5mC nucleobases are known to exist in non-negligible quantities, particularly within mESCs.[21,22] Mapping of these modifications is complicated by a few additional sources of biochemical complexities: strand biases,[13] concomitant modifications, and hemi-methylation.[23]

Modified nucleobases can substantially alter transcription factor recognition

Many transcription factors prefer specific motifs, enabling the sequence specificity of transcriptional control.[24] The position weight matrix (PWM) model allows for the computational identification of transcription factor binding sites, by characterizing a transcription factor's position-specific preference over the DNA alphabet.[25] Just as transcription factors distinguish one unmodified nucleobase from another, some transcription factors are known to distinguish between unmodified and modified bases. Despite these covalent modifications not altering base-pairing, they protrude into the major and minor grooves of DNA, and impact other aspects of DNA conformation. These changes can result in altered protein recognition.[2]

In particular, transcription factors often bind to novel motifs that differ from the unmodified core consensus sequences. MeCP2, one of many non–sequence-specific methyl-CpG binding proteins, has been shown to bind to 5hmC.[26] However, the role of non–sequence-specific modified nucleobase binding is limited to specific protein families.

It is more informative, but also more challenging, to elucidate and characterize sequence specific motifs. In 2013, Hu et al.[3] demonstrated that central CpG-methylated motifs have strong binding activity for certain transcription factors. Using protein binding microarrays, they showed that these motifs are often very different from the unmethylated sequences that those transcription factors usually bind. A few transcription factors have well-characterized modification preferences. These preferences can serve as a means of verifying a predictive framework; a working model is expected to be able to robustly yield the known preferences. Therefore an understanding of known modification-sensitivities informs the design of such a model. Since Hu et al.'s[3] work, other transcription factors have been shown to have methylation-sensitivity[27] and an instance of 5caC increasing binding affinity was found.[28] Both C/EBPα and C/EBPβ have increased binding activity when the central CpG of its canonical octamer (consensus: `TTGC|GCAA`) is methylated, formylated, or carboxylated, with both strands contributing to increase the effect and hemi-modification still demonstrating a reduced effect.[29] 5hmC was found to inhibit binding of C/EBPβ, but not C/EBPα.[29] c-Myc is a basic helix-loop-helix (bHLH) family transcription factor, which has been demonstrated to have a strong preference for unmethylated E-box motifs, often preferring the `CACGTG` hexamer.[30,31] It is one of many bHLH transcription factors that demonstrate such a preference.[32–36]

Recently, Quenneville et al.[37] demonstrated that ZFP57 has a preference for methylated motifs, specifically for the completely centrally-methylated `TGCCGC(R)` heptamer (red indicates methylation on the positive strand and blue on the negative strand). This was subsequently confirmed, and extended upon, by Strogantsev et al.[38], who additionally found that ZFP57 motifs with a final guanine residue as the core binding site are often concomitantly methylated at that second CpG site. This preference was also confirmed with crystallography and in solution with fluorescence polarization analyses, by Liu et al.[39], who additionally demonstrated that ZFP57 has successively decreasing affinity for the oxidized forms of 5mC. Xu et al.[40] recently applied a random forest to predict binding of transcription factors by combining genomic and methylation data. They did not attempt to predict the preference of factors for methylated DNA, but rather developed software to use profiles of 5mC or 5hmC bases to improve predictions of *in vivo* transcription factor binding events.

Stable modification-induced changes to DNA shape, and the existence of modification-sensitive transcription factors, motivate the development of a computational framework to elucidate and characterize altered motifs. We describe here methods to analyze covalent DNA modifications and their affects on transcription factor binding sites, by introducing an expanded epigenetic alphabet. We introduce Cytomod, software to integrate DNA modification information into a single genomic sequence and we detail the use of extensions to the MEME Suite[41] to analyze 5mC and 5hmC transcription factor binding site sensitivities.

Methods

An expanded epigenetic alphabet

To analyze DNA modifications' effects upon transcription factor binding, we developed a model of genome sequence that expands the standard `A/C/G/T` alphabet. Our model adds the symbols `m` (5mC), `h` (5hmC), `f` (5fC), and `c` (5caC). This allows us to more easily adapt existing computational methods, that work on a discrete alphabet, to work with epigenetic cytosine modification data.

Each symbol represents a base pair in addition to a single nucleotide, implicitly encoding a complementarity relation. Accordingly, we add four symbols to represent G when paired with modified C: `1` (G:5mC), `2` (G:5hmC), `3` (G:5fC), and `4` (G:5caC) (Table 1). This ensures that complementation remains a lossless operation. It also captures the fact that the properties of the base pairing of a guanine to a modified residue is altered by the presence of the modification.[2] We number these symbols in the same order in which the ten-eleven translocation (TET) enzyme acts on 5-methylcytosine and its oxidized derivatives (Figure 1).[6]

Many cytosine modification-detection assays only yield incomplete information of a cytosine's modification state. For example, conventional bisulfite sequencing alone determines if cytosine bases are modified to either 5mC or 5hmC, but cannot resolve between those two modifications.[6] Even with sufficient sequencing to disambiguate all modifications, statistical methods are needed to infer each modification from the data, resulting in additional uncertainty. To capture common instances of modification state uncertainty, we also introduce ambiguity codes: `z/9` for a cytosine of (completely) unknown modification state, `y/8` for a cytosine known to be neither hydroxymethylated nor methylated, `x/7` for a hydroxymethylated or methylated base, and `w/6` for formylated or carboxylated bases (Table 2). These codes are analogous to those defined by the Nomenclature Committee of the International Union of Biochemistry already in common usage, such as for unknown purines or pyrimidines (`R` or `Y`, respectively).[42,43]

Covalent Cytosine Modification			Complement	
Abbreviation	Name	Symbol	Name	Symbol
5mC	5-Methylcytosine	m	Guanine:5-Methylcytosine	1
5hmC	5-Hydroxymethylcytosine	h	Guanine:5-Hydroxymethylcytosine	2
5fC	5-Formylcytosine	f	Guanine:5-Formylcytosine	3
5caC	5-Carboxylcytosine	c	Guanine:5-Carboxylcytosine	4

Table 1. The expanded epigenetic alphabet. This includes the known modifications to cytosine and symbols for each guanine that is complementary to a modified nucleobase.

Ambiguous Nucleobase		Complement	
Symbol	Possible bases	Symbol	Possible bases
w	f, c	6	3, 4
x	m, h	7	1, 2
y	C, f, c	8	G, 3, 4
z	C, m, h, f, c	9	G, 1, 2, 3, 4

Table 2. Ambiguous bases for uncertain modification states. The MEME Suite recognizes these ambiguity codes in the same manner as the ambiguous bases already in common usage, such as `R` for `A` or `G` in the conventional DNA alphabet.

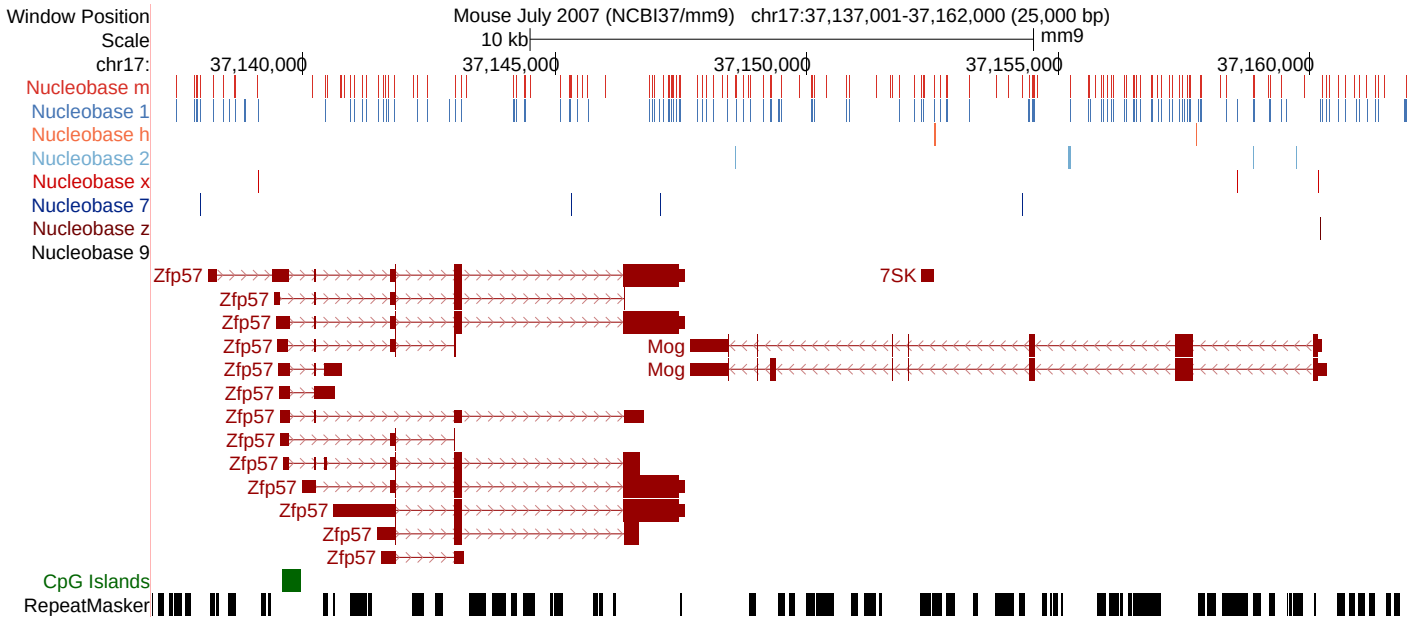


Figure 2. Differential cytosine modification status in naive mouse T-cells for a 25 kbp region (within cytoband 17qB1) surrounding *Zfp57* and *Mog*. UCSC Genome Browser[46] plot that includes RepeatMasker[48] regions, CpG islands,[49] GENCODE[50] genes, and calls for bases h (5hmC), m (5mC), x (5mC/5hmC), z (C with unknown modification state), 1 (G:5mC), 2 (G:5hmC), 7 (G:5mC/5hmC), and 9 (G:C with unknown modification state).

Creation of an expanded-alphabet genome sequence

Like most epigenomic data, the abundance and distribution of cytosine modifications is cell-type specific. Therefore, modified genomes need to be constructed for a particular cell-type and downstream analyses cannot necessarily be expected to generalize. Accordingly, we first need to construct a modified genome that pertains to the organism, assembly, and tissue type we wish to analyze. This modified genome uses the described expanded alphabet to encode cytosine modification state, using calls from single-base resolution modification data.

To do this, we created a Python program called Cytomod. It loads an unmodified assembly and then alters it using provided modification data. It relies upon Genomedata[44] and NumPy[45] to load and iterate over genome sequence data. Cytomod can take the intersection or union of replicates pertaining to a single modification type. It also allows one to provide a single replicate of each type, and potentially to run it multiple times to produce multiple independent replicates of modified genomes. It permits flagging of ambiguous input data, such as when only conventional bisulfite sequencing data is available and therefore the only modified bases are x/7. Cytomod additionally produces browser extensible data (BED) tracks for each cytosine modification, for viewing in the UCSC[46] (Figure 2) or Ensembl genome browsers.[47]

We used conventional and oxidative whole-genome bisulfite sequencing data generated for naive $CD4^+$ T cells, extracted from the spleens of C57BL/6J mice, aged 6–8 weeks. A fraction enriched in $CD4^+$ T cells was first obtained by depletion of non-$CD4^+$ T cells by magnetic labelling and then fluorescence-activated cell sorting was used to get the $CD4^+$, $CD62L^+$, $CD44^{low}$, and $CD25^-$ naive pool of T cells. This data was generated by the Ferguson-Smith and Adams labs at the University of Cambridge and the Sanger Institute for the BLUEPRINT project, as a part of Sjöberg et al.[51].

We refer to the combination of both conventional and oxidative whole-genome bisulfite sequencing as (ox)WGBS. We analyzed biological replicates separately, 2 of each sex. Unaligned, paired-end, BAM files output from the sequencer were subjected to a standardized internal quality check pipeline. We used MethPipe[52] (development version, commit 3655360) to process the data. All random chromosomes were

excluded, after alignment. We selected Bismark[53] for alignment, which has been demonstrated to work well.[54,55] The processing pipeline is as follows: sort the unaligned raw BAM files in name order (using SamBamba,[56] version 0.5.4); convert the files to FASTQ, splitting each paired-end (via version 2.23.0 of BEDTools[57] `bamtofastq`); align the FASTA files to NCBI m37/mm9 using Bismark[53] (version 0.14.3), which uses Bowtie[58] (version 2.2.4), in the default directional mode for a stranded library; sort the output aligned files by position (again via SamBamba `sort`); index sorted, aligned, BAMs (via version 1.2 of SAMtools[59] `index`); convert the processed BAM files into the format required by MethPipe, using `to-mr`; merge sequencing lanes (via direct concatenation of `to-mr` output files) for each specimen (biological replicate), for each sex, and each of WGBS and oxWGBS; sort the output as described in MethPipe's documentation (by position and then by strand); remove duplicates using MethPipe's `duplicate-remover`; run MethPipe's `methcounts` program; and finally run MLML,[60] which combines the conventional and oxidative bisulfite sequencing data to yield consistent estimations of cytosine modification state.

We then create modified genomes from the MLML outputs. MLML outputs maximum-likelihood estimates of the levels of 5mC, 5hmC, and C, which are between 0 and 1. These estimates are computed directly or via expectation maximization.[60] It outputs an indicator of the number of conflicts, which is an estimate of methylation or hydroxymethylation levels falling outside of the confidence interval computed from the input coverage and level. This value is 0, 1, or 2 in our case, since we have two inputs per run (WGBS and oxWGBS). An abundance of conflicts can indicate the presence of non-random error.[60] We assign `z/9` to all loci with any conflicts, regarding those loci as having unknown modification state. Our analysis pipeline accounts for cytosine modifications occurring in any genomic context, and additionally maintains the data's strandedness, allowing analyses of hemi-modification. We created modified genomes using a grid search, in increments of 0.01, for a threshold t, for the levels of 5mC (m) and 5hmC (h), as described in Figure 3.

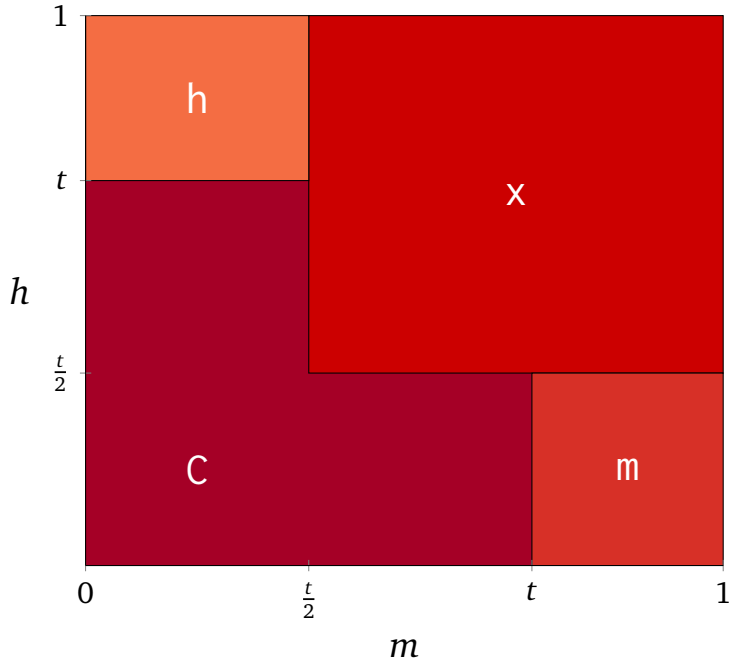


Figure 3. Conditions on the MLML[60] confidence levels of 5mC (m) and 5hmC (h) in relation to a threshold t, that lead to the calling of different modified nucleobases. These base assignments assume that MLML had no conflicts for the locus under consideration. If there were any conflicts, the base is assigned to z, irrespective of the values of m or h. Inequalities to call modifications are not strict. For example, $m = t/2 \Rightarrow$ x. The bases are depicted for the positive strand only, and are complemented when occurring on the negative strand, as outlined in Table 1 and Table 2.

We use half of the threshold value for assignment to x/7, since we consider that consistent with the use of the full threshold value to call a specific modification (since if t is sufficient to call 5mC or 5hmC alone, $m + h \geq t$ should be sufficient to call x/7).

We additionally analyzed base frequencies for each modification, both overall (Figure 4) and per genomic cytosine. These frequencies are computed genome-wide, for putative promoter regions and enhancer regions. To estimate promoter regions, we used GENCODE Release M1 (the last GENCODE version annotating NCBI m37/mm9), for the primary genome annotation, using a 2 kbp region upstream of the first transcription start site for each "known" GENCODE transcript. We create enhancer regions from a seven-state ChromHMM segmentation for ES-Bruce4.[61] We used segmentation state 3, "K4m1", which is highly enriched for H3K4me1.

Detection of altered transcription factor binding in modified genomic contexts

Next, we performed transcription factor binding site motif discovery, enrichment and modified-unmodified comparisons. Here, we use mouse assembly NCBI m37/mm9 for all analyses, since we wanted to be able to make use of all Mouse ENCODE[61] ChIP-seq data without re-alignment nor lift-over.[1] We updated the MEME Suite[41] to work with custom alphabets, such as our expanded epigenomic alphabet. We incorporated these modifications into MEME Suite version 4.11.0.

We characterize modified transcription factor binding sites using MEME-ChIP.[63] It allows us to rapidly assess the main software outputs we are interested in: Multiple EM (Expectation Maximization) for Motif Elicitation (MEME)[64] and Discriminative Regular Expression Motif Elicitation (DREME),[65] both for *de novo* motif elucidation; CentriMo,[66,67] for the assessment of motif centrality; SpaMo,[68] to assess Spaced Motifs (which is especially relevant for multi-partite motifs); and Find Individual Motif Occurrences (FIMO).[69]

CentriMo[66] is our main focus for the analysis of our results. It permits inference of the direct DNA binding affinity of motifs, by assessing a motif's local enrichment. In our case, we scan peak centres with PWMs, for the best match per region. The PWMs used are generated from MEME-ChIP, by loading the JASPAR 2014[70,71] core vertebrates database, in addition to any elucidated *de novo* motifs from MEME or DREME. The number of sequences at each position of the central peaks is counted and normalized to estimate probabilities of central enrichment. These are smoothed and plotted. A one-tailed binomial test is used to assess the significance of central enrichment.

If low complexity sequences are not masked out first, MEME-ChIP[63] can yield repetitive motifs. Existing masking algorithms are not designed to work with modified genomes, and we accordingly mask the assembly, prior to modification with Cytomod. This masking is only for downstream motif analyses. The unmasked modified genome output by Cytomod is always used for base frequency and distribution analyses. We use Tandem Repeat Finder (TRF)[72] (version 4.07b) to mask low complexity sequences and TRF masked genomes are always used with MEME-ChIP. We used the following parameters: `2 7 7 80 10 50 500 -h -m -ngs`, taken from the TRF parameter optimization results of Frith et al.[73].

We ran MEME-ChIP, using the published protocol for the command-line analysis of ChIP-seq data,[74] against Cytomod genome sequences for regions pertaining to chromatin immunoprecipitation-sequencing (ChIP-seq) peaks from transcription factors of interest. We employ positive controls, in two opposite directions, to assess the validity of our results. We use c-Myc as the positive control for an unmethylated binding preference.[30,31] ChIP-seq data for c-Myc was used from both a stringent streptavidin-based genome-wide approach with biotin-tagged Myc in mESCs from Krepelova et al.[75] (GEO: GSM1171648), as well as

[1] Specifically, we used the *Mus musculus* Illumina iGenome[62] packaging of the UCSC mm9 genome. This genome excludes all alternative haplotypes as well as all unreliably ordered, but chromosome-associated, sequences (the so-called "random" chromosomes). This was ideal for downstream analyses, but not sufficient for aligning data ourselves. This is because exclusion of these additional pseudo-chromosomes might deleteriously impact alignments, by resulting in the inclusion of spuriously unique reads. Therefore, the full UCSC mm9 build is used when aligning to a reference sequence.

murine erythroleukemia and CH12.LX Myc Mouse ENCODE samples (ENCFF001YJE and ENCFF001YHU). Conversely, both ZFP57 and C/EBPβ are used as positive controls for methylated binding preferences.[29,37–39] For C/EBPβ, we used Mouse ENCODE ChIP-seq data, conducted upon C2C12 cells (ENCFF001XUT) or myocytes differentiated from those cells (ENCFF001XUR and ENCFF001XUS). We used one replicate of ZFP57 peaks provided by Quenneville et al.[37]. We constructed a ZFP57 BED file using BEDTools[57] (version 2.17.0) to subtract the control influenza hemagglutinin (HA) ChIP-seq (GEO: GSM773065) from the target (HA-tagged ZFP57: GEO: GSM773066). Only target regions with no overlap with any features implicated by the control file were retained, yielding 11 231 of 22 031 features.

We also used ZFP57 ChIP-seq data from Strogantsev et al.[38] (GEO: GSE55382), consisting of 40 bp single-end reads from reciprocal F1 hybrid Cast/EiJ × C57BL/6J mESCs (BC8: sequenced C57BL/6J mother × Cast father and CB9: sequenced Cast mother × C57BL/6J father). We are not interested in allele-specificity and need it to correspond to the assembly we are using. We re-processed the data, aligning it to NCBI m37/mm9, in a similar manner to some of the Mouse ENCODE datasets, to maximize consistency for future Mouse ENCODE analyses. We obtained raw FASTQs using SRA Toolkit's `fastq-dump`; aligned them via Bowtie[58] (version 1.1.0; `bowtie -v 2 -k 11 -m 10 -t --best --strata`); sorted and indexed the BAM files (using Sambamba[56]); and called peaks, using the input as the negative enrichment set, via MACS 2,[76] with increased stringency ($q = 0.00001$), with parameters: `-q 0.00001 -f BAM -g mm`. This parameter list omits the previously explained target and control information, and parameters to set the output's base name and directory. This resulted in 90 478 BC8 and 56 142 CB9 peaks.

We used the ChIPQC[77] Bioconductor[78] package to assess the ChIP-seq data quality. We used the two control and two target runs for each of BC8 and CB9. We then used `ChIPQC(samples, consensus=TRUE, bCount=TRUE, summits=250, annotation="mm9", blacklist="mm9-blacklist.bed.gz", chromosomes =chromosomes)`. We set the `chromosomes` list to all the fully-aligned mouse chromosomes, excepting chrM. A blacklist of regions is used to filter out regions that appear uniquely mappable, but have been empirically found to show artificially elevated signal in short-read functional genomics data. We took the blacklist from the NCBI m37/mm9 ENCODE blacklist website (https://sites.google.com/site/anshulkundaje/projects/blacklists).[79] The fraction of reads in peaks (FRiP) was 13.7% and 9.12% for the BC8 and CB9 data respectively. We additionally performed peak calling at the default $q = 0.05$, which resulted in many more peaks (197 610 BC8 and 360 932 CB9 peaks) and respective FRiP values of 27.6% and 19.74%. The CB9 sample had a lesser fraction of reads in (overlapping) blacklisted regions (RiBL). At the default peak calling stringency, BC8 had 29.7% RiBL, while CB9 had only 8.38%.

We additionally analyzed three ZFP57 ChIP-seq replicates (100 bp paired-end reads) pertaining to mESCs in pure C57BL/6J mice.[80] Each replicate is paired with an identically-conducted ChIP-seq in a corresponding sample, for which ZFP57 is not expressed (ZFP57-null controls). The same protocol as for the hybrid ZFP57 data was used, excepting that we used the ZFP57-null ChIP-seq data as the negative set for peak calling instead of the input and Bowtie was run in paired-end mode (using `-1` and `-2`). We additionally omitted the Bowtie arguments `--best --strata`, which do not work in paired-end mode and added `-y --maxbts 800`, the latter of which is what is set with `--best`'s value, instead of the default threshold of 125. We also set MACS to paired-end mode (via `-f BAMPE`). However, this resulted in very few peaks when processed with the same peak-calling stringency as the hybrid data (at most 1812 peaks) and FRiP values under 2%. Even when we used the default stringency threshold, there were at most 4496 peaks, with FRiP values of around 4.5%. Nonetheless, we still observed the expected preference for methylated motifs (Figure S1).

To directly compare various modifications of motifs to their cognate unmodified sequences, we adopted a hypothesis testing approach. Motifs of interest can be derived from a *de novo* result that merits further investigation, but are often formed from prior expectations of motif binding preferences from the literature, such as for c-Myc, ZFP57, and C/EBPβ. For every unmodified motif of interest, we can partially or fully change the base at a given motif position to some modified base (Table 3).

To directly compare modified hypotheses to their cognate unmodified sequences robustly, we try to

Modification description	Unmodified motif	Modified motif
Full CpG modification to 5mC	C NG	m1
Partial CpG hemi-modification to 5hmC	C NG	h NG
Full CpT modification to either 5mC or 5hmC	C NT	xT

Table 3. Illustrative examples of possible changes made to convert unmodified motifs to specific modified counterparts, for downstream hypothesis testing. Stacked letters are used like simple sequence logos. At these positions, N represents any base frequencies other than the base being modified. These are the other positions in the motif's position weight matrix. $\overset{\mathtt{C}}{\mathtt{N}}\mathtt{N} \rightarrow \mathtt{m}$ indicates that a position containing cytosine is modified by replacing all base frequencies at that position with m, with frequency 1. Conversely, $\overset{\mathtt{C}}{\mathtt{N}}\mathtt{N} \rightarrow \overset{\mathtt{h}}{\mathtt{N}}\mathtt{G}$ indicates that a position containing cytosine is modified by replacing the frequency apportioned to C with h, leaving the other base frequencies at that position unmodified. The second base of each dinulceotide is portrayed as having a frequency of one, however, it too could be composed of different bases of various frequencies, including the base shown.

minimize as many confounds as possible.

We fix the CentriMo central region width (via `--minreg 99 --maxreg 100`). We also compensate for the substantial difference in the background frequencies of modified versus unmodified bases. Otherwise, vastly lower modified base frequencies can yield higher probability and sharper CentriMo peaks, since when CentriMo scans with its "log-odds" matrix, it computes scores for nucleobase b with background frequency $\mathrm{f}(b)$ as

$$\log\left(\frac{\Pr(b)}{\mathrm{f}(b)}\right).$$

To compensate for this, we ensure that any motif pairs being compared have the same length and similar relative entropies. To do this, we use a larger motif pseudo-count (via `--motif-pseudo <count>`) for modified motifs. We compute the appropriate pseudo-count, as described below, and provide it to `iupac2meme`. We set CentriMo's pseudo-count to 0, since we have already applied the appropriate pseudo-count to the motif.

The relative entropy (or Kullback-Leibler divergence), D_{RE}, of a motif m of length $|m|$, with respect to a background model b over the alphabet A, of size $|A|$, is[81]

$$D_{\mathrm{RE}}(m,b) = \sum_{i=0}^{|m|-1}\sum_{j=0}^{|A|-1}\left(m_{ij}\log_2\left(\frac{m_{ij}}{b_j}\right)\right). \tag{1}$$

For each position, i, in the motif, the MEME Suite adds the pseudo-count parameter, α, times the background frequency for a given base, j, at the position: $m'_{ij} = m_{ij} + \alpha b_j$.

Accordingly, to equalize the relative entropies, we need only substitute m'_{ij} for each m_{ij} in Equation 1 and then isolate α. If we proceed in this fashion, however, our pseudo-count would depend upon the motif

frequency at each position and the background of each base in the motif. Instead, we can make a number of simplifying assumptions that apply in this particular case. First, the unmodified and modified motifs we are comparing differ only in the bases being modified, which in this case, are only C or G nucleobases, with a motif frequency of 1. Additionally, we set the pseudo-count of the unmodified motif to a constant 0.1 (CentriMo's default). Thus, the pseudo-count to use for a single modified base, is the value α, obtained by solving, for provided modified base background frequency b_m and unmodified base frequency b_u:

$$1 + \alpha b_m \log_2\left(\frac{1+\alpha b_m}{b_m}\right) = 1 + 0.1 b_u \log_2\left(\frac{1+0.1 b_u}{b_u}\right) \tag{2}$$

However, Equation 2 only accounts for a single modification, on a single strand. For complete modification, we also need to consider the potentially different background frequency of the modified bases' complement. Thus for a single complete modification, with modified positions m_1 and m_2 and corresponding unmodified positions u_1 and u_2, modified base background frequencies b_{m_1}, b_{m_2} and unmodified base frequencies b_{u_1}, b_{u_2}, we obtain

$$\begin{aligned} &1 + \alpha b_{m_1} \log_2\left(\frac{1+\alpha b_{m_1}}{b_{m_1}}\right) + 1 + \alpha b_{m_2} \log_2\left(\frac{1+\alpha b_{m_2}}{b_{m_2}}\right) \\ &\quad = 1 + 0.1 b_{u_1} \log_2\left(\frac{1+0.1 b_{u_1}}{b_{u_1}}\right) + 1 + 0.1 b_{u_2} \log_2\left(\frac{1+0.1 b_{u_2}}{b_{u_2}}\right). \end{aligned} \tag{3}$$

We numerically solve for α in Equation 3 for each modified hypothesis, using `fsolve` from SciPy.[82] Finally, we may have multiple modified positions. We always either hemi-modify or completely modify all modified positions, so the pseudocount to use is the product of modified positions and the α value from Equation 3.

The pseudo-count obtained in this fashion does not exactly equalize the two motif's relative entropies, since we do not account for the effect that the altered pseudo-count has upon all the other positions of the motif.

We then perform hypothesis testing for an unmodified motif and all possible 5mC/5hmC modifications of all CpGs for known modification-sensitive motifs for c-Myc, ZFP57, and C/EBPβ. These modifications consist of the six possible combinations for methylation and hydroxymethylation at a CpG, where a CpG is not permitted to be both hemi-methylated and hemi-hydroxymethylated. These six combinations are: `mG`, `C1`, `m1`, `hG`, `C2`, and `h2`. For c-Myc, the unmodified motif from which modified hypotheses were constructed is the standard E-box: `CACGTG`. For ZFP57, we tested the known binding motif, as both a hexamer (`TGCCGC`) and as extended heptamers (`TGCCGCR` and `TGCCGCG`).[37,38] We additionally tested motifs that we found to occur frequently in our *de novo* analyses, `C(C/A)TGm1(C/T)(A)`. We encoded this motif as the hexamer `MTGCGY` and heptamers, with one additional base for each side: `CMTGCGY` and `MTGCGYA`. This encoding permitted direct comparisons to the other known ZFP57-binding motifs of the same length. Finally, for C/EBPβ we tested the modifications of two octamers: its known binding motif (`TTGCGCAA`) and the chimeric C/EBP|CRE motif (`TTGCGTCA`).[29] These motifs were then assessed for their centrality within their respective ChIP-seq datasets, using CentriMo. We then compute the ratio of CentriMo central enrichment p-values, adjusted for multiple testing,[66] for each modified/unmodified motif pair. For numerical precision, we compute this ratio as the difference of their log values returned by CentriMo. This determines if the motif prefers a modified (positive) or unmodified (negative) binding site.

We conducted hypothesis testing across all four replicates of WGBS and oxWGBS data, for a grid search of modified base calling thresholds. These thresholds are based upon the levels output by MLML.[60] We interpret these values as our degree of confidence for a modification occurring at a given locus. We conducted our grid search from 0.01–0.99 inclusive, at 0.01 increments. Finally, the ratio of CentriMo p-values are plotted across the different thresholds, using Python libraries Seaborn[83] and Pandas.[84,85]

Results

We created an expanded-alphabet sequence using oxidative (ox) and conventional whole-genome bisulfite sequencing (WGBS) maps of 5mC and 5hmC for naive *ex vivo* mouse CD4$^+$ T cells.[51] We generated individual modified genomes across four replicates of (ox)WGBS data and for a variety of modified base calling thresholds. We used these modified genome sequences as the basis for the extraction of genomic regions implicated by ChIP-seq data for particular transcription factors.

The modification abundances obtained were as expected, with respect to the absolute abundance of nucleobases, including their modifications genome-wide, within promoter regions, and within enhancer regions (Figure 4). Genome-wide, at a 0.7 threshold, for the female 15-16 specimen, we find that 2.5% of cytosine residues are methylated, and that 5hmC abundance is 3.5–8.0% of 5mC abundance, depending upon the inclusion of ambiguous bases. These frequencies are consistent with previous results in other cell types.[5,14,16]

Additionally, 5hmC comprises 0.17% of cytosine or guanine bases genome-wide vs. 0.20% within enhancer regions. If ambiguous 5mC/5hmC (x/7) bases are included, this difference increases to 0.39% vs. 0.45%. These results are consistent with greater 5hmC abundance within enhancer regions.[86–89]

Hypothesis testing reveals altered modified transcription factor binding preferences

We conducted hypothesis testing across three transcription factors for which we can predict their expected methylation or hydroxymethylation sensitivities from the literature. Two of the tested transcription factors are expected to prefer methylated DNA: ZFP57[38] and C/EBPβ,[29] and one is known to prefer unmethylated DNA: c-Myc.[30,31] Additionally, C/EBPβ is known to have reduced affinity for hydroxymethylated DNA.[29]

We tested known unmodified transcription factor binding motifs against all possible 5mC and 5hmC modifications thereof, at all CpG dinucleotides. For each modified motif, we assessed its expected DNA binding affinity using its adjusted CentriMo central enrichment p-value.[66] We conducted the same test for the unmodified version of the motif, comparing their p-values as a ratio, using the difference of their log transformed values. Positive values for this difference represent a preference for the modified motif, while negative values represent the converse.

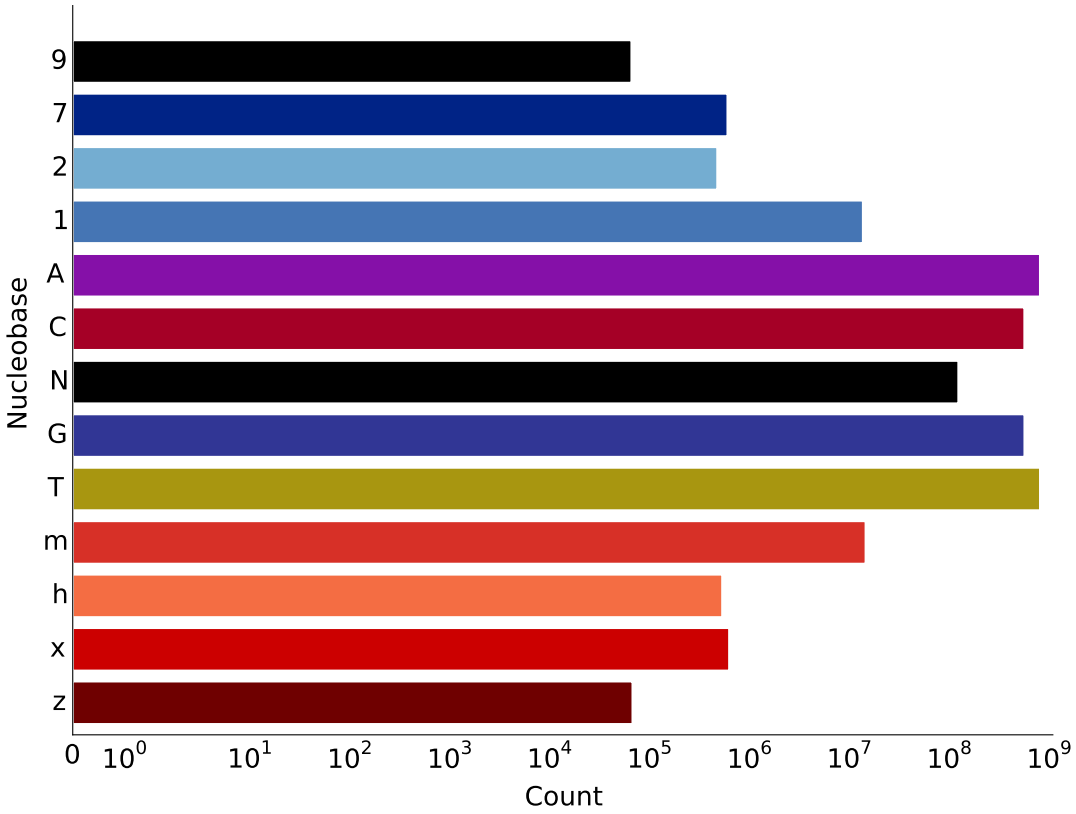
We find that the expected transcription factor binding preferences hold across all four (ox)WGBS replicates and for all investigated modified nucleobase calling thresholds (from 0.01–0.99 inclusive, at 0.01 increments, representing modification confidence; Figure 5). Our observation that all c-Myc log p-value differences are below zero, implies that modified c-Myc motifs are disfavoured compared with their unmodified E-box motifs. For the ZFP57 sample shown, all modifications are favoured, compared to their unmodified counterparts. One of the modified motifs which has the greatest increase in predicted binding affinity in the modified case is `TGCm1m1`, a motif that Strogantsev et al.[38] often found. Two methylated motifs had the greatest increase in predicted binding affinity for C/EBPβ: `TTGmGCAA` and `TTGC1TCA`. The same results are obtained for multiple different ChIP-seq replicates for these transcription factors (Figure S1). These results are robust in the face of perturbations, including peak calling stringency (Figure S2).

In addition to ZFP57 displaying a strong preference for methylated DNA, hydroxymethylated CpGs had a substantially lesser increase in binding affinity than methylated motifs (Figure 5), but still greater than the completely unmethylated motif. This recapitulates Liu et al.'s[39] *in vitro* finding that ZFP57 has the greatest binding affinity for motifs containing 5mC, followed by 5hmC, and then by unmodified cytosine.

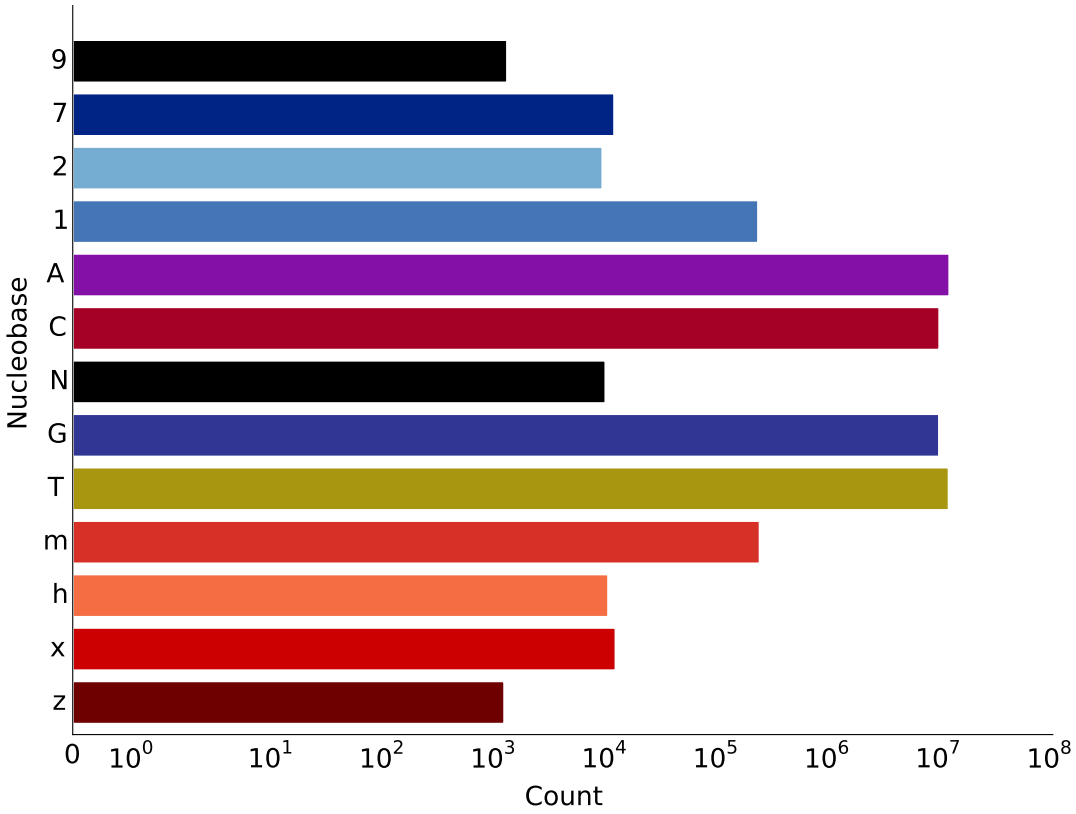
Elucidation of dichotomous binding preferences—C/EBPβ

C/EBPβ is of particular interest because of its dichotomous binding preferences for 5mC versus 5hmC.[29] Our method is able to recapitulate this preference, across all replicates of (ox)WGBS and ChIP-seq data, with methylated motif pairs generally having positive ratios, whereas hydroxymethylated motif ratios are

(A)



(B)



(C)

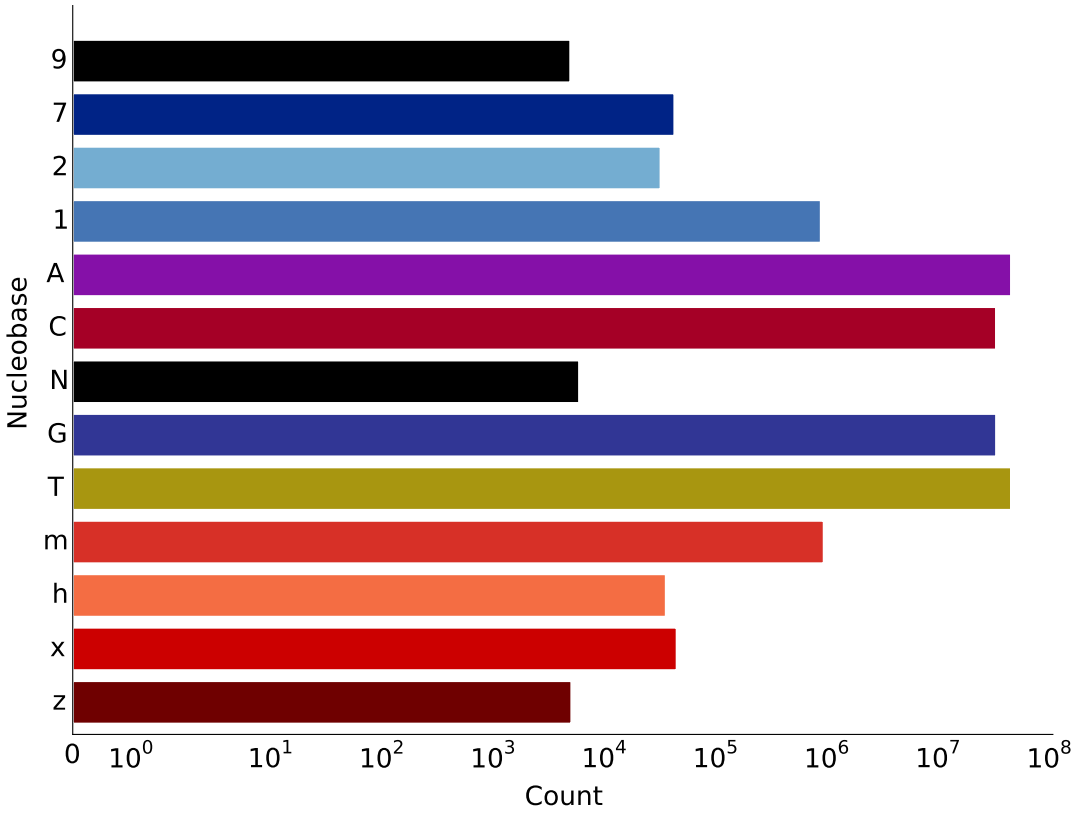


Figure 4. Modified nucleobase counts for a single female replicate (# 2) of naive T-cell mouse (ox)WGBS data at a modified base calling threshold of 0.7. Counts are shown (A) genome-wide, (B) within promoter regions, and (C) within enhancer regions.

negative (Figure S3). One exceptional case is for a positive strand, hemi-methylated, motif (TTGmGTCA), which is often disfavoured compared to the unmodified motif. This motif is not the consensus C/EBPβ motif, but rather the chimeric C/EBP|CRE octamer. While Sayeed et al.[29] demonstrated that this chimeric transcription factor had a more modest preference toward its methylated DNA motif, we would still have expected a weak preference for this motif, over its unmodified counterpart, as opposed to the unmodified motif preference observed. Additionally, we find hemi-methylation to have greater enrichment than complete methylation, which contradicts their finding of both strands contributing to increase the effect.[29] This may be due to technical issues with hemi-methylation in our modified sequence and requires further investigation.

Suitable thresholds for *de novo* and downstream analyses

The grid search for transcription factor binding thresholds at 0.01 increments allowed us to determine suitable thresholds (0.3 and 0.7) for further investigation (Figure S1). Overall, this grid search demonstrates the suitability of a wide-range of thresholds, likely useful for assessing future datasets. *De novo* analyses of C/EBPβ confirmed the preference for methylated DNA, with methylated motifs having much greater central enrichment than their unmethylated counterparts, at both the 0.3 (Figure 6) and 0.7 thresholds (Figure 7).

Despite robust findings with hypothesis testing, across almost the entire range of possible thresholds, we were, however, often unable to detect the expected binding preferences in a *de novo* context. The c-Myc and, to a lesser extent, ZFP57 CentriMo runs, in a non-hypothesis testing context, did not demonstrate substantial enrichment nor depletion with respect to modified vs. unmodified motifs. An example of this is shown in Figure S4 for c-Myc. We consider potential explanations for this in the Discussion.

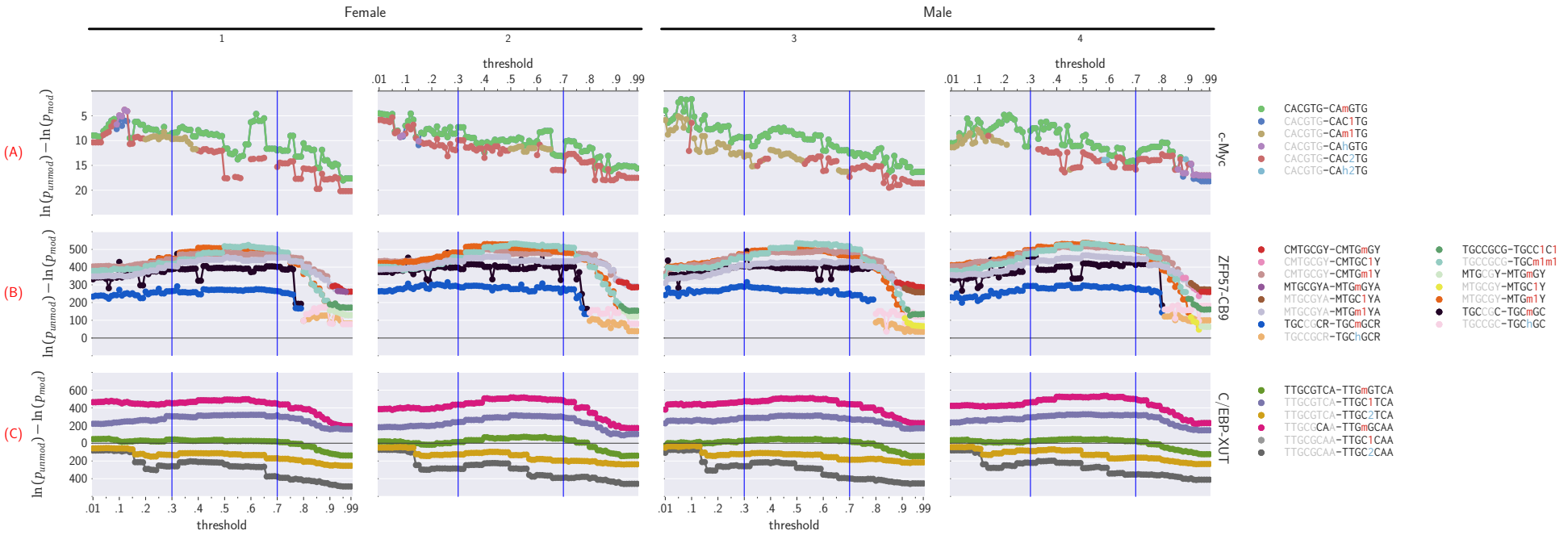
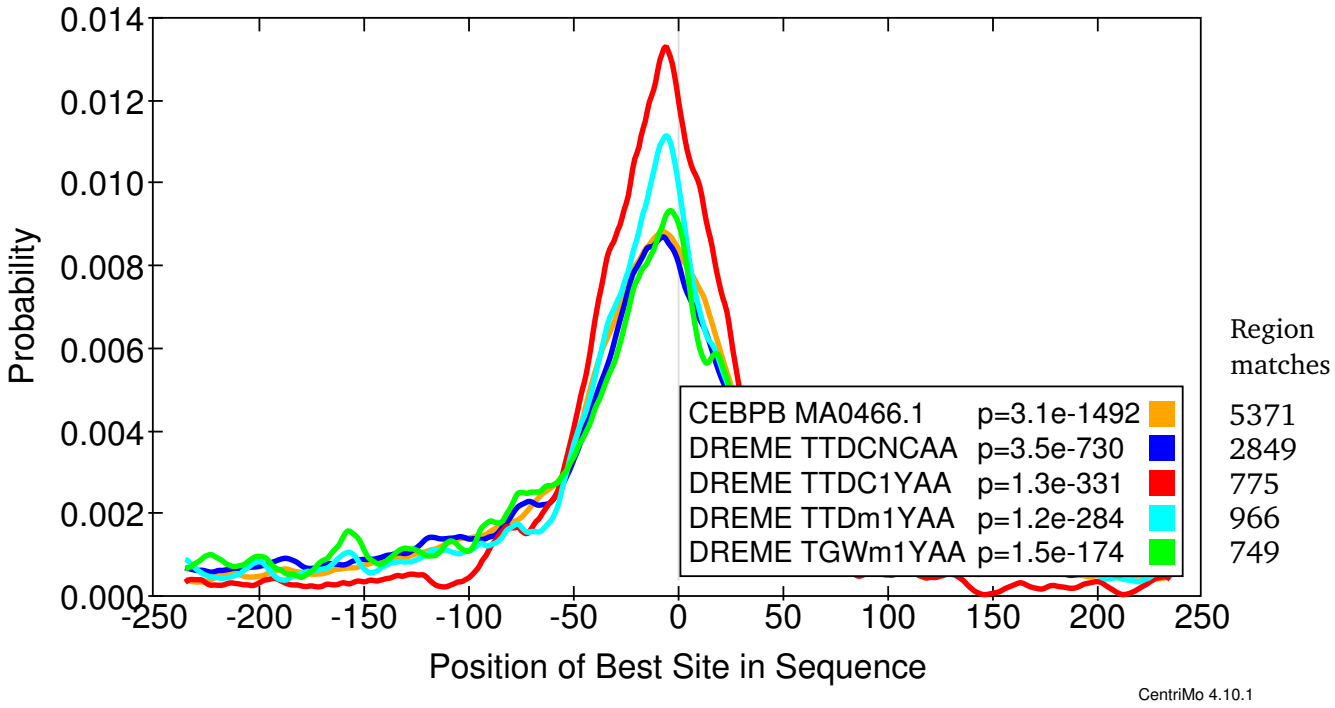


Figure 5. Relationship between unmodified versus modified motif statistical significance of central enrichment (from CentriMo[66]) and modified base calling thresholds across different (ox)WGBS specimens.[51] Each unmodified motif, at each threshold, is compared to its top three most significant modifications for c-Myc and C/EBPβ, but top one most significant modification for ZFP57. Since at most the top three motif pairs are selected per threshold value, specific thresholds can result in new motif pairs, while previously enriched pairs may no longer be present at a given threshold. Each column pertains to a particular (ox)WGBS replicate. Each row of plots pertains to a single ChIP-seq replicate for a particular transcription factor target, one each of: c-Myc (Krepelova et al.[75]), ZFP57 (CB9; Strogantsev et al.[38]), and C/EBPβ (ENCFF001XUT). Negative values correspond to a preference for the unmodified motif, while positive values correspond to a preference for the modified motif.

(A)

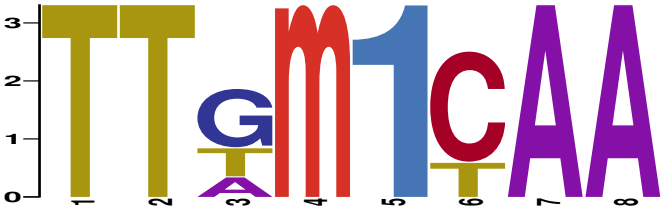


(B)

(C)

(D)

(E)



(F)

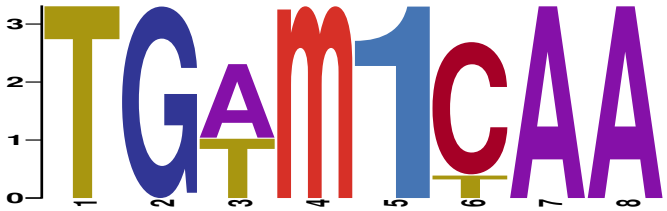
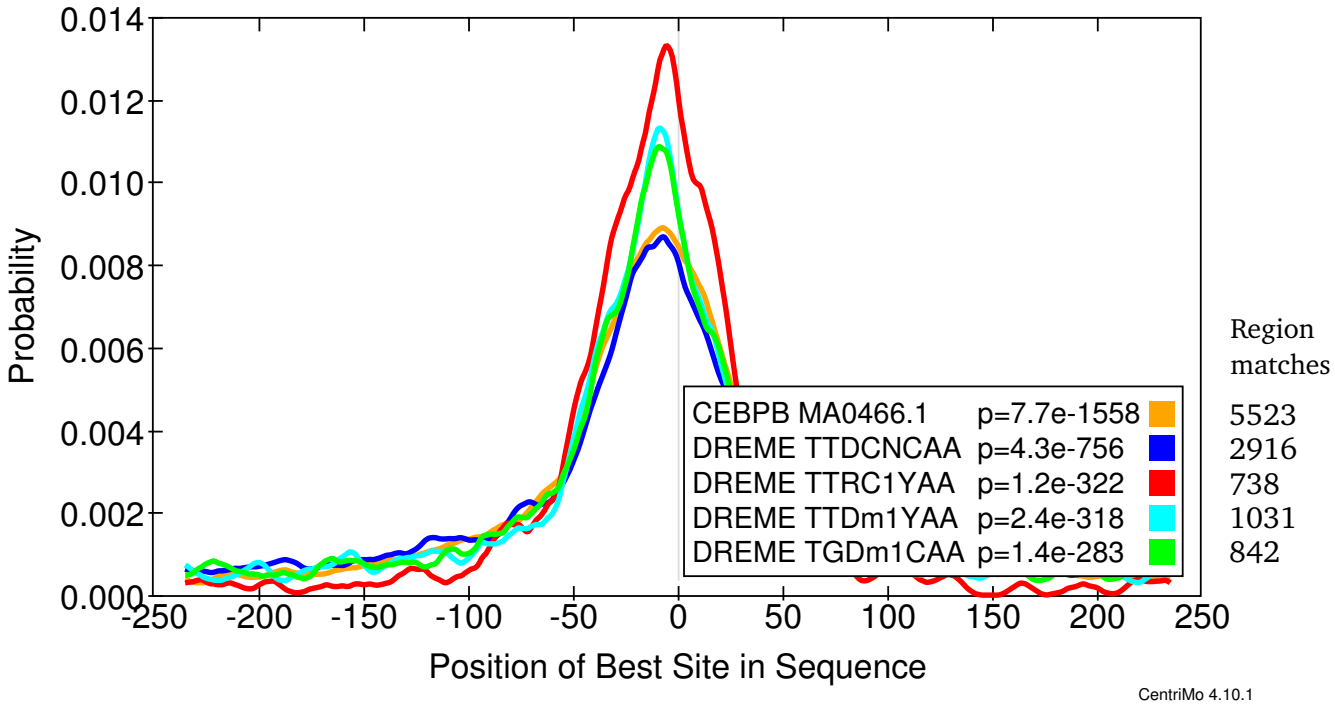


Figure 6. C/EBPβ (GEO: GSM915179; 11 434 ChIP-seq peaks) CentriMo analysis upon *de novo* and JASPAR motifs (see Methods) using female replicate 2 of the (ox)WGBS data[51] at a 0.3 modification threshold. (A) the CentriMo result with the JASPAR C/EBPβ motif in orange (logo in (B)), the main unmethylated C/EBPβ motif in blue (logo in (C)) and DREME methylated motifs in red, cyan, and green (respective logos in (D), (E), and (F)). We depict the JASPAR sequence logo using MEME's relative entropy calculation and colouring for consistency.

(A)

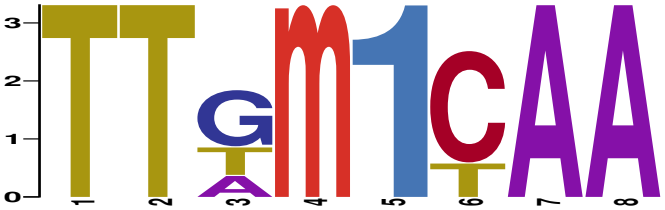


(B)

(C)

(D)

(E)



(F)

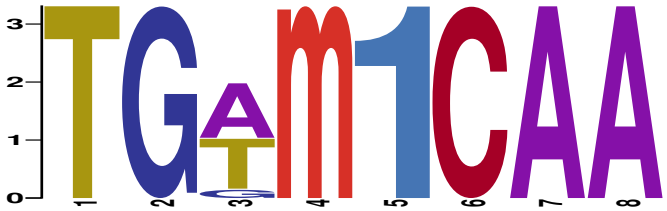


Figure 7. C/EBPβ (GEO: GSM915179; 11 434 ChIP-seq peaks) CentriMo analysis upon *de novo* and JASPAR motifs (see Methods) using female replicate 2 of the (ox)WGBS data[51] at a 0.7 modification threshold. (A) the CentriMo result with the JASPAR C/EBPβ motif in orange (logo in (B)), the main unmethylated C/EBPβ motif in blue (logo in (C)) and DREME methylated motifs in red, cyan, and green (respective logos in (D), (E), and (F)). We depict the JASPAR sequence logo using MEME's relative entropy calculation and colouring for consistency.

Discussion

We have added expanded alphabet capabilities to the widely-used MEME Suite,[41] a set of software tools for the sequence-based analysis of motifs. This included extending several of its core tools, including: MEME,[64] DREME,[65] and CentriMo,[66] used in a unified pipeline via MEME-ChIP.[63] We undertook further extension of all downstream analysis tools and pipelines, and most of the MEME Suite[41] can now be used with arbitrary alphabets. We have processed maps of cytosine modifications in *ex vivo* mouse T-cells to yield genome sequences that use the expanded alphabet. We then used the extended software on the modified genome sequences in regions defined by ChIP-seq data to confirm previously known transcription factor binding preferences using our expanded-alphabet models.

Hypothesis testing, with equal central region widths and relative entropies, leads to more interpretable results than the standard CentriMo analyses, in that it permits a direct comparison of centrality p-values. We often observed the expected pattern in many replicates of conventional CentriMo runs with *de novo* motifs, such as with C/EBPβ (Figure 6 and Figure 7) and ZFP57 (Figure S5). However, there were instances in which the expected motif binding preference was not obvious from *de novo* CentriMo analyses, such as for c-Myc (Figure S4) and other ZFP57 CentriMo results, despite the hypothesis testing robustly corroborating its expected preference for unmethylated DNA (Figure S1).

We suspect that the inability of *de novo* analyses to elucidate modified binding preferences is primarily due to such analyses not having any means of integrating modified and unmodified motifs. Our *de novo* analyses are also unable to compensate for the large differences in modified versus unmodified background frequencies. *De novo* elucidation involves some form of optimization or heuristic selection of sites, and is an inherently variable process. Modified motifs have particular characteristics that differ from most unmodified motifs. Most notably, they are necessarily different from the overall and likely local sequence backgrounds, as a result of the low frequency of modifications. Conversely, an unmodified genome sequence has a comparably uniform nucleobase background, and unmodified motifs are usually found within local sequence of highly similar properties to the motifs themselves.[90] Accordingly, modified motifs can get lost within a background of irrelevant unmodified motifs or no comparable sets of motifs may be found, without specifically accounting for these confounds. Also, modified motifs that a *de novo* analysis finds might not be comparable to any unmodified counterpart. This could occur due to their being of substantially different lengths, often being shorter. It is also difficult to compare motifs having sequence properties that often indicate a poor-quality motif, such as repetitious motifs, or off-target motifs, such as zingers.[91] Hypothesis testing, with relative entropy normalization, can be used to mitigate these concerns.

This method is robust in the face of parameter perturbations. In particular, changes in the modified base calling threshold, across a broad range, consistently led to the same expected results, across three transcription factors and a number of ChIP-seq and bisulfite sequencing replicates (Figure S1). Furthermore, modification of peak calling stringency for a set of ZFP57 datasets, did not negatively impact our detection of its affinity for methylated DNA (Figure S2). The consistency of our controls provides confidence in the ability of this method to detect and accurately characterize the effect of modified DNA on transcription factor binding. This is instrumental in applying this method to a diverse array of ChIP-seq data, towards the elucidation of novel binding preferences.

There is an inherent trade-off between a lower threshold, yielding more modified loci but potentially introducing false positives, and a higher one, which may be too stringent to detect modified base binding preferences. We selected a lower threshold of 0.3, based primarily on the observation of increased variance and decreased apparent preference for unmethylated DNA for c-Myc below this threshold, across multiple replicates (Figure S1). We also selected an upper threshold of 0.7, based primarily on the rapid decrease in relative affinity for methylated over unmethylated motifs in ZFP57 (Figure 5) and, to a lesser extent, C/EBPβ (Figure S3).

We found that there is often an enrichment for hemi-modified, as opposed to completely-modified binding sites. Motifs with hemi-(hydroxy)methylation were often more centrally enriched than those with

complete modification of a central CpG dinucleotide (Figure 6 and Figure 7). This is surprising, because numerous *in vitro* experiments have demonstrated that for transcription factors preferring modified DNA, each modification is often additive, resulting in completely modified motifs having greatest affinity.[29,39] It is possible that the hemi-(hydroxy)methylation events we detect are the result of asymmetric binding affinities for 5mC (5hmC). ZFP57, for example, is known to have asymmetric recognition of 5mC, with the negative strand methylation being more important than the positive strand methylation with respect to the TGCC**G**C motif.[39] Further work is needed to determine if this is due to technical artifacts (either at the level of the bisulfite sequencing data or the methods used) or if this reflects an actual biological preference.

There are few high-quality single-base resolution datasets of 5hmC, 5fC, and 5caC. We had previously attempted analyses using modification data, from assays like MeDIP,[92] that did not employ single-base resolution methods.[14] We found that without single-base resolution, it was difficult to create a discrete genome sequence with a reasonable abundance of the modification under consideration without biasing the sequence, thereby making downstream analyses of transcription factor binding uninformative. It is essential to have single-base resolution data, for any modifications that one wishes to analyze. Additionally, many datasets which do meet this criteria use some form of reduced representation approach, in which CpGs are enriched, allowing for much cheaper sequencing, while still capturing many DNA modifications. The use of reduced representation bisulfite sequencing data can lead to confounding factors, due to the non-uniform distribution of methylated sites surveyed. We accordingly recommend that enrichment approaches be avoided for use with these methods, at least until these confounds are better addressed.

The ChIP-seq data we used was not from the same cell type as the (ox)WGBS data. While transcription factor binding models created from one cell type are often assumed to be consistently useful across different cell types, in some cases, they are not. Nonetheless, we consistently observed the expected preferences in transcription factor binding for the expected modification affinities, across multiple ChIP-seq replicates, often in different cell types.

The MEME Suite's new custom alphabet capability permits further downstream analyses of modified motifs. For example, one can find individual motif occurrences with FIMO[69] or conduct pathway analyses with Gene Ontology for MOtifs (GOMO).[93] Alternatively, FIMO results can be used for pathway analyses via GREAT,[94] and downstream pathway analysis tools, such as Enrichment Map,[95,96] can be used for further interpretation of the results. This permits inference of implicated genomic regions and biological pathways, which can then be subjected to further analysis.

This approach can be readily extended to other DNA modifications, since we designed all of our software with this in mind. A number of DNA modifications can now be detected at high resolution, with many known to occur endogenously across diverse organisms,[1] such as 5-hydroxymethyluracil (5hmU), 5-formyluracil (5fU), 8-oxoguanine (8-oxoG), and 6-methyladenine (6mA).[97–99] We provide recommendations in Appendix A for the nomenclature of these modified nucleobases, among others.

We provide a framework to readily apply motif analyses on sequences containing DNA modifications. Consistent reproduction of known transcription factor binding affinities suggests that these methods produce biologically meaningful results and can predict the modification sensitivity of other transcription factors. We intend to apply these methods to analyze all Mouse ENCODE factors toward the identification of novel epigenetic binding preferences.

Acknowledgements

We thank William Stafford Noble and Charles E. Grant for useful discussions and contributions to the MEME Suite. We thank Andrew D. Smith, Meng Zhou, Ben Decato, and Egor Dolzhenko for their work on MethPipe[52,60] and for actively providing support. We thank Michael Waskom for his visualization work on the Seaborn[83] Python package and for actively providing support. We thank Carl Virtanen and Zhibin Lu for technical assistance.

This research was enabled by support provided by: Globus,[100,101] Compute Canada (specifically, WestGrid, SHARCNET, and SciNet[102]), and the Princess Margaret Computational Biology Resource Centre.

This work was supported by the Canadian Cancer Society (703827 to M.M.H.), the Natural Sciences and Engineering Research Council of Canada (RGPIN-2015-03948 to M.M.H. and an Alexander Graham Bell Canada Graduate Scholarship to C.V.), the Ontario Ministry of Training, Colleges and Universities (Ontario Graduate Scholarship to C.V.), the Ontario Institute for Cancer Research through funding provided by the Government of Ontario (CSC-FR-UHN to John E. Dick), the University of Toronto McLaughlin Centre (MC-2015-16 to M.M.H.), and the Princess Margaret Cancer Foundation.

References

1. Breiling, A. & Lyko, F. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics & Chromatin* **8,** 24 (2015).
2. Dantas Machado, A. C., Zhou, T., Rao, S., Goel, P., Rastogi, C., Lazarovici, A., Bussemaker, H. J. & Rohs, R. Evolving insights on how cytosine methylation affects protein-DNA binding. *Briefings in Functional Genomics* **14,** 61–73 (2014).
3. Hu, S., Wan, J., Su, Y., Song, Q., Zeng, Y., Nguyen, H. N., Shin, J., Cox, E., Rho, H. S., Woodard, C., Xia, S., Liu, S., Lyu, H., Ming, G.-L., Wade, H., Song, H., Qian, J. & Zhu, H. DNA methylation presents distinct binding sites for human transcription factors. *ELife* **2,** e00726 (2013).
4. Lercher, L., McDonough, M. a., El-Sagheer, A. H., Thalhammer, A., Kriaucionis, S., Brown, T. & Schofield, C. J. Structural insights into how 5-hydroxymethylation influences transcription factor binding. *Chemical Communications* **50,** 1794–6 (2014).
5. Ito, S., Shen, L., Dai, Q., Wu, S. C., Collins, L. B., Swenberg, J. a., He, C. & Zhang, Y. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333,** 1300–3 (2011).
6. Booth, M. J., Raiber, E.-A. & Balasubramanian, S. Chemical methods for decoding cytosine modifications in DNA. *Chemical Reviews* **115,** 2240–54 (2014).
7. Kohli, R. M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502,** 472–9 (2013).
8. Watt, F. & Molloy, P. L. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes & Development* **2,** 1136–43 (1988).
9. Varley, K. E., Gertz, J., Bowling, K. M., Parker, S. L., Reddy, T. E., Pauli-Behn, F., Cross, M. K., Williams, B. A., Stamatoyannopoulos, J. A., Crawford, G. E., Absher, D. M., Wold, B. J. & Myers, R. M. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Research* **23,** 555–67 (2013).
10. Song, C.-X. & He, C. Potential functional roles of DNA demethylation intermediates. *Trends in Biochemical Sciences* **38,** 480–4 (2013).
11. Bachman, M., Uribe-Lewis, S., Yang, X., Williams, M., Murrell, A. & Balasubramanian, S. 5-hydroxymethylcytosine is a predominantly stable DNA modification. *Nature Chemistry* **6,** 1049–55 (2014).
12. Hu, L., Lu, J., Cheng, J., Rao, Q., Li, Z., Hou, H., Lou, Z., Zhang, L., Li, W., Gong, W., Liu, M., Sun, C., Yin, X., Li, J., Tan, X., Wang, P., Wang, Y., Fang, D., Cui, Q., Yang, P., He, C., Jiang, H., Luo, C. & Xu, Y. Structural insight into substrate preference for TET-mediated oxidation. *Nature* **527,** 118–22 (2015).
13. Yu, M., Hon, G. C., Szulwach, K. E., Song, C. X., Zhang, L., Kim, A., Li, X., Dai, Q., Shen, Y., Park, B., Min, J. H., Jin, P., Ren, B. & He, C. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149,** 1368–80 (2012).
14. Song, C.-X., Yi, C. & He, C. Mapping recently identified nucleotide variants in the genome and transcriptome. *Nature Biotechnology* **30,** 1107–16 (2012).

15. Song, C. X., Szulwach, K. E., Dai, Q., Fu, Y., Mao, S. Q., Lin, L., Street, C., Li, Y., Poidevin, M., Wu, H., Gao, J., Liu, P., Li, L., Xu, G. L., Jin, P. & He, C. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153,** 678–91 (2013).

16. Booth, M. J., Marsico, G., Bachman, M., Beraldi, D. & Balasubramanian, S. Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nature Chemistry* **6,** 435–40 (2014).

17. Shen, L., Wu, H., Diep, D., Yamaguchi, S., D'Alessio, A. C., Fung, H. L., Zhang, K. & Zhang, Y. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153,** 692–706 (2013).

18. Lu, X., Han, D., Zhao, B. S., Song, C.-X., Zhang, L.-S., Doré, L. C. & He, C. Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics. *Cell Research* **25,** 386–89 (2015).

19. Raiber, E.-A., Murat, P., Chirgadze, D. Y., Beraldi, D., Luisi, B. F. & Balasubramanian, S. 5-formylcytosine alters the structure of the DNA double helix. *Nature Structural & Molecular Biology* (2014).

20. Bachman, M., Uribe-Lewis, S., Yang, X., Burgess, H. E., Iurlaro, M., Reik, W., Murrell, A. & Balasubramanian, S. 5-formylcytosine can be a stable DNA modification in mammals. *Nature Chemical Biology* **11,** 555–7 (2015).

21. Ramsahoye, B. H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A. P. & Jaenisch, R. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proceedings of The National Academy of Sciences of The United States of America* **97,** 5237–42 (2000).

22. Ziller, M. J., Müller, F., Liao, J., Zhang, Y., Gu, H., Bock, C., Boyle, P., Epstein, C. B., Bernstein, B. E., Lengauer, T., Gnirke, A. & Meissner, A. Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLOS Genetics* **7,** e1002389 (2011).

23. Arita, K., Ariyoshi, M., Tochio, H., Nakamura, Y. & Shirakawa, M. Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature* **455,** 818–21 (2008).

24. Li, J. J., Bickel, P. J. & Biggin, M. D. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* **2,** e270 (2014).

25. Berg, O. G. & von Hippel, P. H. Selection of DNA binding sites by regulatory proteins. *Journal of Molecular Biology* **193,** 723–43 (1987).

26. Mellén, M., Ayata, P., Dewell, S., Kriaucionis, S. & Heintz, N. MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell* **151,** 1417–30 (2012).

27. Gustems, M., Woellmer, A., Rothbauer, U., Eck, S. H., Wieland, T., Lutter, D. & Hammerschmidt, W. c-Jun/c-Fos heterodimers regulate cellular genes via a newly identified class of methylated DNA sequence motifs. *Nucleic Acids Research* **42,** 3059–72 (2014).

28. Golla, J. P., Zhao, J., Mann, I. K., Sayeed, S. K., Mandal, A., Rose, R. B. & Vinson, C. Carboxylation of cytosine (5caC) in the CG dinucleotide in the E-box motif (CGCAG|GTG) increases binding of the Tcf3|Ascl1 helix-loop-helix heterodimer 10-fold. *Biochemical and Biophysical Research Communications* **449,** 248–55 (2014).

29. Sayeed, S. K., Zhao, J., Sathyanarayana, B. K., Golla, J. P. & Vinson, C. C/EBPβ (CEBPB) protein binding to the C/EBP|CRE DNA 8-mer TTGC|GTCA is inhibited by 5hmC and enhanced by 5mC, 5fC, and 5caC in the CG dinucleotide. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1849,** 583–89 (2015).

30. Prendergast, G. C. & Ziff, E. B. Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region. *Science* **251,** 186–9 (1991).

31. Guccione, E., Martinato, F., Finocchiaro, G., Luzi, L., Tizzoni, L., Dall' Olio, V., Zardo, G., Nervi, C., Bernard, L. & Amati, B. Myc-binding-site recognition in the human genome is determined by chromatin context. *Nature Cell Biology* **8,** 764–70 (2006).

32. Murre, C., McCaw, P. S. & Baltimore, D. A new DNA binding and dimerization motif in immunoglobulin enhancer binding, *daughterless*, *MyoD*, and *myc* proteins. *Cell* **56,** 777–83 (1989).

33. Fisher, F. & Goding, C. R. Single amino acid substitutions alter helix-loop-helix protein specificity for bases flanking the core CANNTG motif. *The EMBO Journal* **11,** 4103–9 (1992).

34. Bendall, A. J. & Molloy, P. L. Base preferences for DNA binding by the bHLH-Zip protein USF: effects of $MgCl_2$ on specificity and comparison with binding of Myc family members. *Nucleic Acids Research* **22,** 2801–10 (1994).

35. Atchley, W. R. & Fitch, W. M. A natural classification of the basic helix-loop-helix class of transcription factors. *Proceedings of The National Academy of Sciences of The United States of America* **94,** 5172–76 (1997).

36. Boyd, K. E. & Farnham, P. J. Coexamination of site-specific transcription factor binding and promoter activity in living cells. *Molecular and Cellular Biology* **19,** 8393–99 (1999).

37. Quenneville, S., Verde, G., Corsinotti, A., Kapopoulou, A., Jakobsson, J., Offner, S., Baglivo, I., Pedone, P. V., Grimaldi, G., Riccio, A. & Trono, D. In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Molecular Cell* **44,** 361–72 (2011).

38. Strogantsev, R., Krueger, F., Yamazawa, K., Shi, H., Gould, P., Goldman-Roberts, M., McEwen, K., Sun, B., Pedersen, R. & Ferguson-Smith, A. C. Allele-specific binding of ZFP57 in the epigenetic regulation of imprinted and non-imprinted monoallelic expression. *Genome Biology* **16,** 112 (2015).

39. Liu, Y., Toh, H., Sasaki, H., Zhang, X. & Cheng, X. An atomic model of Zfp57 recognition of CpG methylation within a specific DNA sequence. *Genes & Development* **26,** 2374–79 (2012).

40. Xu, T., Li, B., Zhao, M., Szulwach, K. E., Street, R. C., Lin, L., Yao, B., Zhang, F., Jin, P., Wu, H. & Qin, Z. S. Base-resolution methylation patterns accurately predict transcription factor bindings *in vivo*. *Nucleic Acids Research* **43,** 2757–66 (2015).

41. Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W. & Noble, W. S. MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research* **37,** W202–8 (2009).

42. Nomenclature Committee of the International Union of Biochemistry (NC-IUB). Nomenclature for incompletely specified bases in nucleic acid sequences. *European Journal of Biochemistry* **150,** 1–5 (1985).

43. IUPAC-IUB Commission on Biochemical Nomenclature (CBN). Abbreviations and symbols for nucleic acids, polynucleotides and their constituents. *European Journal of Biochemistry* **15,** 203–8 (1970).

44. Hoffman, M. M., Buske, O. J. & Noble, W. S. The Genomedata format for storing large-scale functional genomics data. *Bioinformatics* **26,** 1458–59 (2010).

45. Van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering* **13,** 22–30 (2011).

46. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. & Haussler, D. The human genome browser at UCSC. *Genome Research* **12,** 996–1006 (2002).

47. Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Birney, E., Harrow, J., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S. J., Cunningham, F., Aken, B. L., Zerbino, D. R. & Flicek, P. Ensembl 2016. *Nucleic Acids Research* **44,** D710–6 (2016).

48. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends in Genetics* **16,** 418–20 (2000).

49. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *Journal of Molecular Biology* **196,** 261–82 (1987).

50. Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.-K., Chrast, J., Lagarde, J., Gilbert, J. G. R., Storey, R., Swarbreck, D., Rossier, C., Ucla, C., Hubbard, T., Antonarakis, S. E. & Guigo, R. GENCODE: producing a reference annotation for ENCODE. *Genome Biology* **7 Suppl 1,** S4.1–9 (2006).

51. Sjöberg, M., Walker, N., Vallejos, C. A., Booth, M. J., Wong, K., Velasco-Herrera, M. D. C., Vanhille, L., Kazachenka, A., Gunning, R., Shi, H., Rashid, M., Mamanova, L., Ashcroft, A. S., Bachman, M., Kołodziejczyk, A. A., Corish, J. A., Gray, D., Spicuglia, S., Balasubramanian, S., Teichmann, S. A., Richardson, S., Marioni, J. C., Adams, D. J. & Ferguson-Smith, A. C. A DNA methylation signature characterises functional genes in the adaptive immune system. In preparation.

52. Song, Q., Decato, B., Hong, E. E., Zhou, M., Fang, F., Qu, J., Garvin, T., Kessler, M., Zhou, J. & Smith, A. D. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLOS One* **8,** e81148 (2013).

53. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27,** 1571–72 (2011).

54. Kunde-Ramamoorthy, G., Coarfa, C., Laritsky, E., Kessler, N. J., Harris, R. A., Xu, M., Chen, R., Shen, L., Milosavljevic, A. & Waterland, R. A. Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. *Nucleic Acids Research* **42,** e43 (2014).

55. Tran, H., Porter, J., Sun, M. A., Xie, H. & Zhang, L. Objective and comprehensive evaluation of bisulfite short read mapping tools. *Advances in Bioinformatics* **2014,** 472045 (2014).

56. Tarasov, A. & Prins, P. Sambamba. http://lomereiter.github.io/sambamba/ (2014).

57. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–42 (2010).

58. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10,** R25 (2009).

59. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–79 (2009).

60. Quy, J., Zhouy, M., Song, Q., Hong, E. E. & Smith, A. D. MLML: consistent simultaneous estimates of DNA methylation and hydroxymethylation. *Bioinformatics* **29,** 2645–46 (2013).

61. Yue, F. et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515,** 355–64 (2014).

62. Illumina. iGenomes. https://support.illumina.com/sequencing/sequencing_software/igenome.html (2016).

63. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27,** 1696–7 (2011).

64. Bailey, T. L. & Elkan, C. *Fitting a mixture model by expectation maximization to discover motifs in biopolymers* in *Proceedings of the International Conference on Intelligent Systems for Molecular Biology* (eds Altman, R., Brutlag, D., Karp, P., Lathrop, R. & Searls, D.) **2** (1994), 28–36.

65. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27,** 1653–9 (2011).

66. Bailey, T. L. & Machanick, P. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Research* **40,** e128 (2012).

67. Lesluyes, T., Johnson, J., Machanick, P. & Bailey, T. L. Differential motif enrichment analysis of paired ChIP-seq experiments. *BMC Genomics* **15,** 752 (2014).

68. Whitington, T., Frith, M. C., Johnson, J. & Bailey, T. L. Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Research* **39,** e98 (2011).

69. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27,** 1017–8 (2011).

70. Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research* **32,** D91–4 (2004).

71. Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C.-y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A. & Wasserman, W. W. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research* **42,** D142–7 (2014).

72. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27,** 573–80 (1999).

73. Frith, M. C., Hamada, M. & Horton, P. Parameters for accurate genome alignment. *BMC Bioinformatics* **11,** 80 (2010).

74. Ma, W., Noble, W. S. & Bailey, T. L. Motif-based analysis of large nucleotide data sets using MEME-ChIP. *Nature Protocols* **9,** 1428–50 (2014).

75. Krepelova, A., Neri, F., Maldotti, M., Rapelli, S. & Oliviero, S. Myc and Max genome-wide binding sites analysis links the Myc regulatory network with the polycomb and the core pluripotency networks in mouse embryonic stem cells. *PLOS One* **9,** e88933 (2014).

76. Zhang, Y., Liu, T., Meyer, C. a., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. & Liu, X. S. Model-based analysis of ChIP-Seq (MACS). *Genome Biology* **9,** R137 (2008).

77. Carroll, T. S., Liang, Z., Salama, R., Stark, R. & de Santiago, I. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Frontiers in Genetics* **5,** 1–11 (2014).

78. Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H. & Zhang, J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5,** R80 (2004).

79. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

80. Shi, H., Strogantsev, R., Takahashi, N., Kazachenka, A., Lorincz, M. C., Hemberger, M. & Ferguson-Smith, A. C. Epigenetic regulation of unique genes and repetitive elements by KRAB zinc finger protein ZFP57 in embryonic stem cells. In preparation.

81. Bailey, T. L. & Elkan, C. *Fitting a mixture model by expectation maximization to discover motifs in bipolymers* tech. rep. CS94-351 (University of California, San Diego, San Diego, 1994). 33 pp.

82. Jones, E., Oliphant, T., Peterson, P., et al. SciPy: open source scientific tools for Python. 2001–2016.

83. Waskom, M., Botvinnik, O., Hobson, P., Warmenhoven, J., Cole, J. B., Halchenko, Y., Vanderplas, J., Hoyer, S., Villalba, S., Quintero, E., Miles, A., Augspurger, T., Yarkoni, T., Evans, C., Wehner, D., Rocher, L., Megies, T., Coelho, L. P., Ziegler, E., Hoppe, T., Seabold, S., Pascual, S., Cloud, P., Koskinen, M., Hausler, C., kjemmett, Milajevs, D., Qalieh, A., Allan, D. & Meyer, K. seaborn: v0.6.0 (June 2015). 2015.

84. McKinney, W. *Python for data analysis: data wrangling with Pandas, NumPy, and IPython* (2012).

85. McKinney, W. *Data structures for statistical computing in Python* in *Proceedings of the 9th Python in Science Conference* (eds van der Walt, S. & Millman, J.) (2010), 51–56.

86. Choi, I., Kim, R., Lim, H.-W., Kaestner, K. H. & Won, K.-J. 5-hydroxymethylcytosine represses the activity of enhancers in embryonic stem cells: a new epigenetic signature for gene regulation. *BMC Genomics* **15,** 670 (2014).

87. Putiri, E. L., Tiedemann, R. L., Thompson, J. J., Liu, C., Ho, T., Choi, J.-H. & Robertson, K. D. Distinct and overlapping control of 5-methylcytosine and 5-hydroxymethylcytosine by the TET proteins in human cancer cells. *Genome Biology* **15,** R81 (2014).

88. Sérandour, A. a., Avner, S., Oger, F., Bizot, M., Percevault, F., Lucchetti-Miganeh, C., Palierne, G., Gheeraert, C., Barloy-Hubler, F., Péron, C. L., Madigou, T., Durand, E., Froguel, P., Staels, B., Lefebvre, P., Métivier, R., Eeckhoute, J. & Salbert, G. Dynamic hydroxymethylation of deoxyribonucleic acid marks differentiation-associated enhancers. *Nucleic Acids Research* **40,** 8255–65 (2012).

89. Stroud, H., Feng, S., Morey Kinney, S., Pradhan, S. & Jacobsen, S. E. 5-hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biology* **12,** R54 (2011).

90. Dror, I., Golan, T., Levy, C., Rohs, R. & Mandel-Gutfreund, Y. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Research* **25,** 1268–80 (2015).

91. Worsley Hunt, R. & Wasserman, W. W. Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. *Genome Biology* **15,** 412 (2014).

92. Weber, M., Davies, J. J., Wittig, D., Oakeley, E. J., Haase, M., Lam, W. L. & Schübeler, D. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genetics* **37,** 853–62 (2005).

93. Buske, F. A., Bodén, M., Bauer, D. C. & Bailey, T. L. Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinformatics* **26,** 860–6 (2010).

94. McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M. & Bejerano, G. GREAT improves functional interpretation of *cis*-regulatory regions. *Nature Biotechnology* **28,** 495–501 (2010).

95. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment Map: a network-based method for gene-set enrichment visualization and interpretation. *PLOS One* **5,** e13984 (2010).

96. Isserlin, R., Merico, D., Voisin, V. & Bader, G. D. Enrichment Map – a Cytoscape app to visualize and explore OMICs pathway enrichment results. *F1000Research* **3,** 141 (2014).

97. Heyn, H. & Esteller, M. An adenine code for DNA: a second life for N6-Methyladenine. *Cell* **161,** 710–13 (2015).

98. Hardisty, R. E., Kawasaki, F., Sahakyan, A. B. & Balasubramanian, S. Selective chemical labeling of natural T modifications in DNA. *Journal of The American Chemical Society* **137,** 9270–2 (2015).

99. Zarakowska, E., Gackowski, D., Foksinski, M. & Olinski, R. Are 8-oxoguanine (8-oxoGua) and 5-hydroxy-methyluracil (5-hmUra) oxidatively damaged dna bases or transcription (epigenetic) marks? *Mutation Research - Genetic Toxicology and Environmental Mutagenesis* **764-765,** 58–63 (2014).

100. Foster, I. Globus Online: accelerating and democratizing science through cloud-based services. *IEEE Internet Computing* **15,** 70–73 (2011).

101. Allen, B., Pickett, K., Tuecke, S., Bresnahan, J., Childers, L., Foster, I., Kandaswamy, G., Kettimuthu, R., Kordas, J., Link, M. & Martin, S. Software as a service for data scientists. *Communications of the ACM* **55,** 81 (2012).

102. Loken, C., Gruner, D., Groer, L., Peltier, R., Bunn, N., Craig, M., Henriques, T., Dempsey, J., Yu, C.-H., Chen, J., Dursi, L. J., Chong, J., Northrup, S., Pinto, J., Knecht, N. & Zon, R. V. SciNet: lessons learned from building a power-efficient top-20 system and data centre. *Journal of Physics: Conference Series* **256,** 012026 (2010).

103. Perez, F. & Granger, B. E. IPython: a system for interactive scientific computing. *Computing in Science & Engineering* **9,** 21–29 (2007).

104. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14,** 178–92 (2013).

105. Chen, K., Zhao, B. S. & He, C. Nucleic acid modifications in regulation of gene expression. *Cell Chemical Biology* **23,** 74–85 (2016).

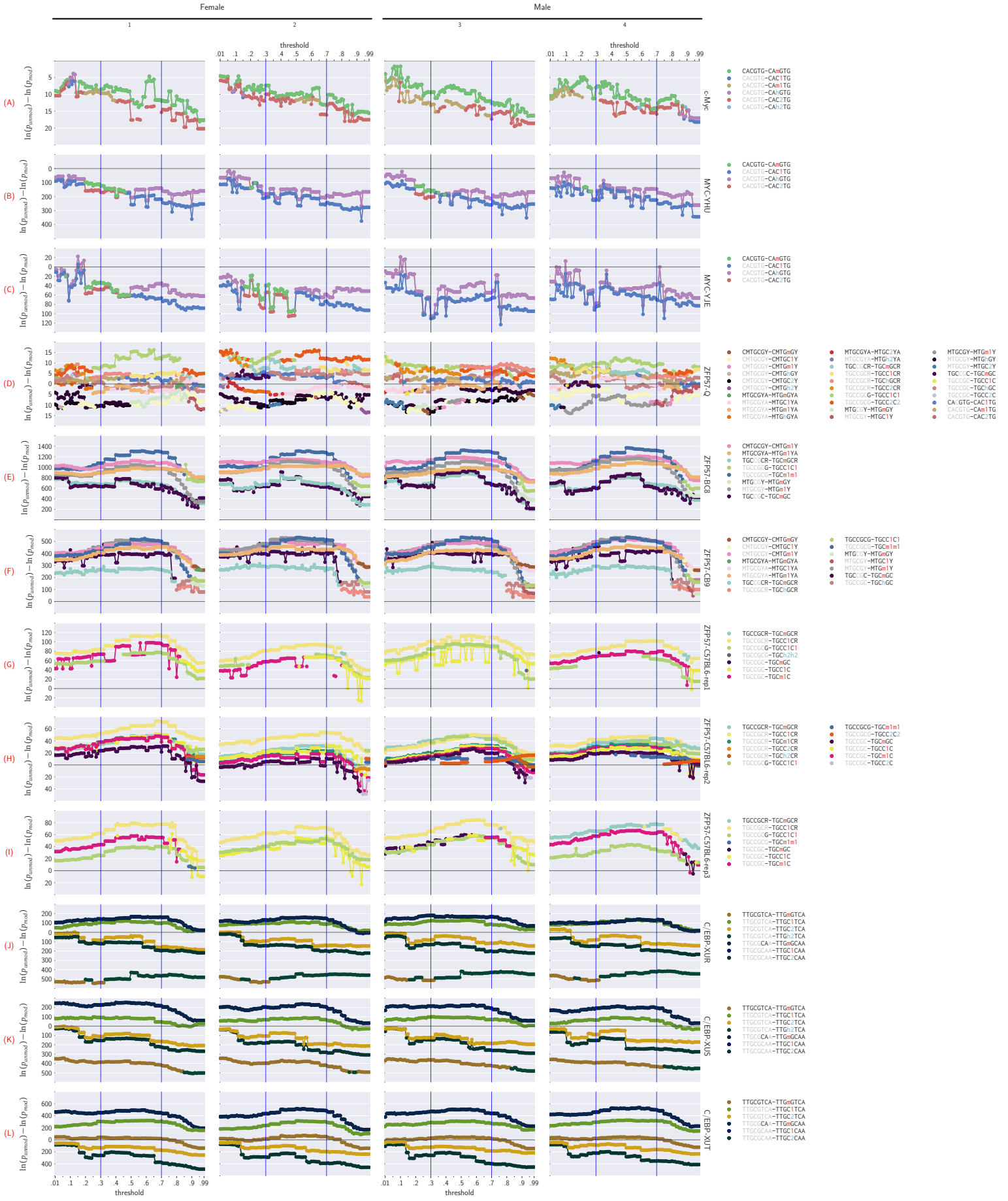


Figure S1 ***(preceding page).*** **Relationship between unmodified versus modified motif statistical significance of central enrichment (from CentriMo[66]) and modified base calling thresholds across different (ox)WGBS specimens.[51] Each unmodified motif, at each threshold, is compared to its top three most significant modifications for c-Myc, C/EBPβ, and replicate two of the C57BL/6 ZFP57 samples, but top one most significant modification for all other ZFP57 samples. Since at most the top three motif pairs are selected per threshold value, specific thresholds can result in new motif pairs, while previously enriched pairs may no longer be present at a given threshold. Each column pertains to a particular (ox)WGBS replicate. Each row of plots pertains to a particular ChIP-seq dataset: c-Myc (3), ZFP57 (6), or C/EBPβ (3). In order (from top–bottom): Krepelova et al.[75], ENCFF001YHU, ENCFF001YJE, Quenneville et al.[37], Strogantsev et al.[38] BC8 cells, Strogantsev et al.[38] CB9 cells, Shi et al.[80] (3 replicates), ENCFF001XUR, ENCFF001XUS, ENCFF001XUT. All ZFP57 peaks, excepting for Quenneville et al.[37], were called using a stringent MACS[76] peak calling threshold, $q = 0.00001$. Negative values correspond to a preference for the unmodified motif, while positive values correspond to a preference for the modified motif. The C/EBPβ results are additionally depicted alone in Figure S3.**

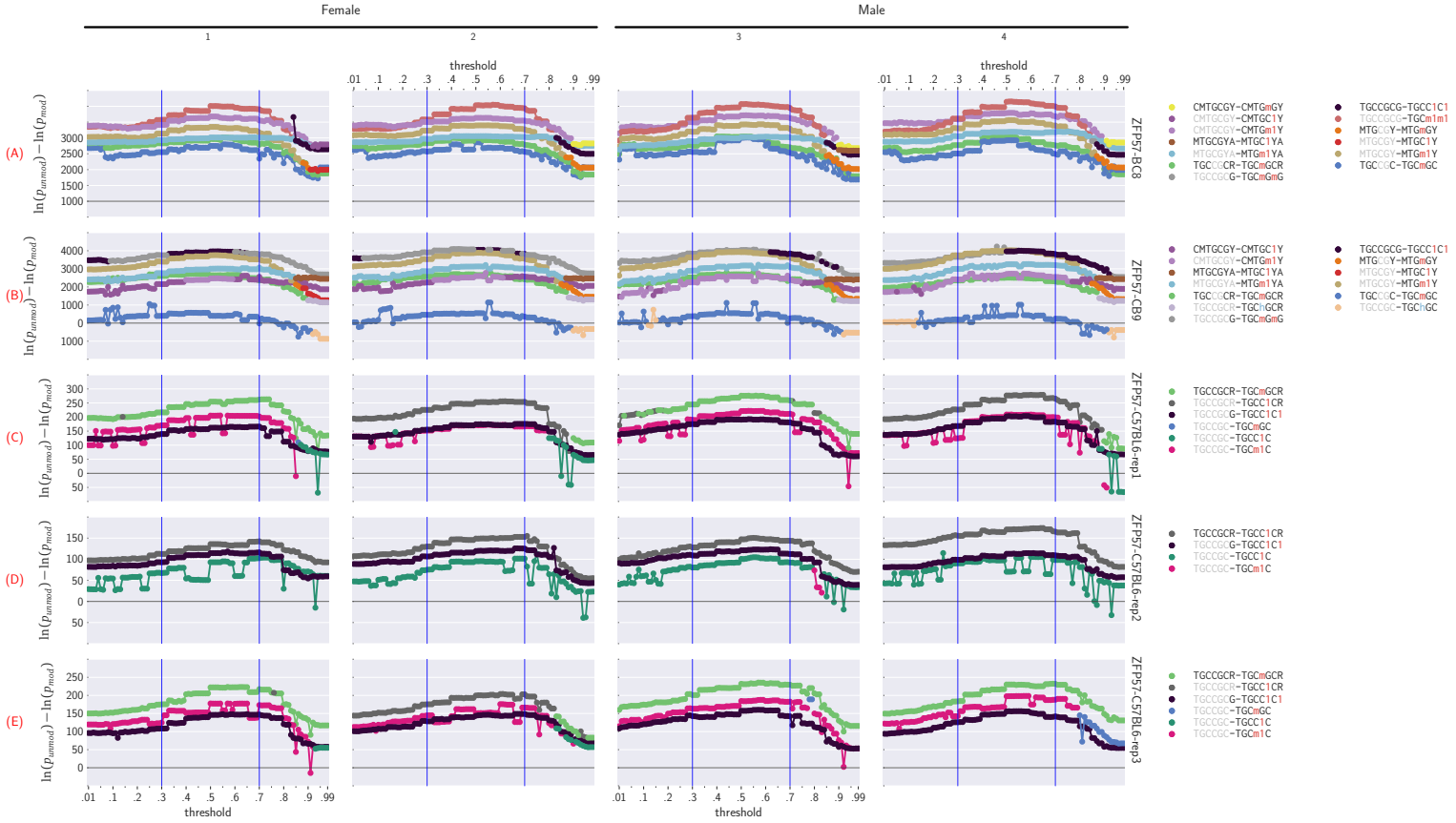


Figure S2. **Relationship between unmodified versus modified motif statistical significance of central enrichment (from CentriMo[66]) and modified base calling thresholds across different (ox)WGBS specimens.[51] Each unmodified motif, at each threshold, is compared to its top most significant modification. Since only the top motif pair is selected per threshold value, specific thresholds can result in new motif pairs, while previously enriched pairs may no longer be present at a given threshold. Each column pertains to a particular (ox)WGBS replicate. Each row of plots pertains to a particular ZFP57 ChIP-seq dataset: Strogantsev et al.[38] BC8 cells, Strogantsev et al.[38] CB9 cells, Shi et al.[80] (3 replicates). All ZFP57 peaks were called using the default MACS[76] peak calling threshold, $q = 0.05$. Negative values correspond to a preference for the unmodified motif, while positive values correspond to a preference for the modified motif.**

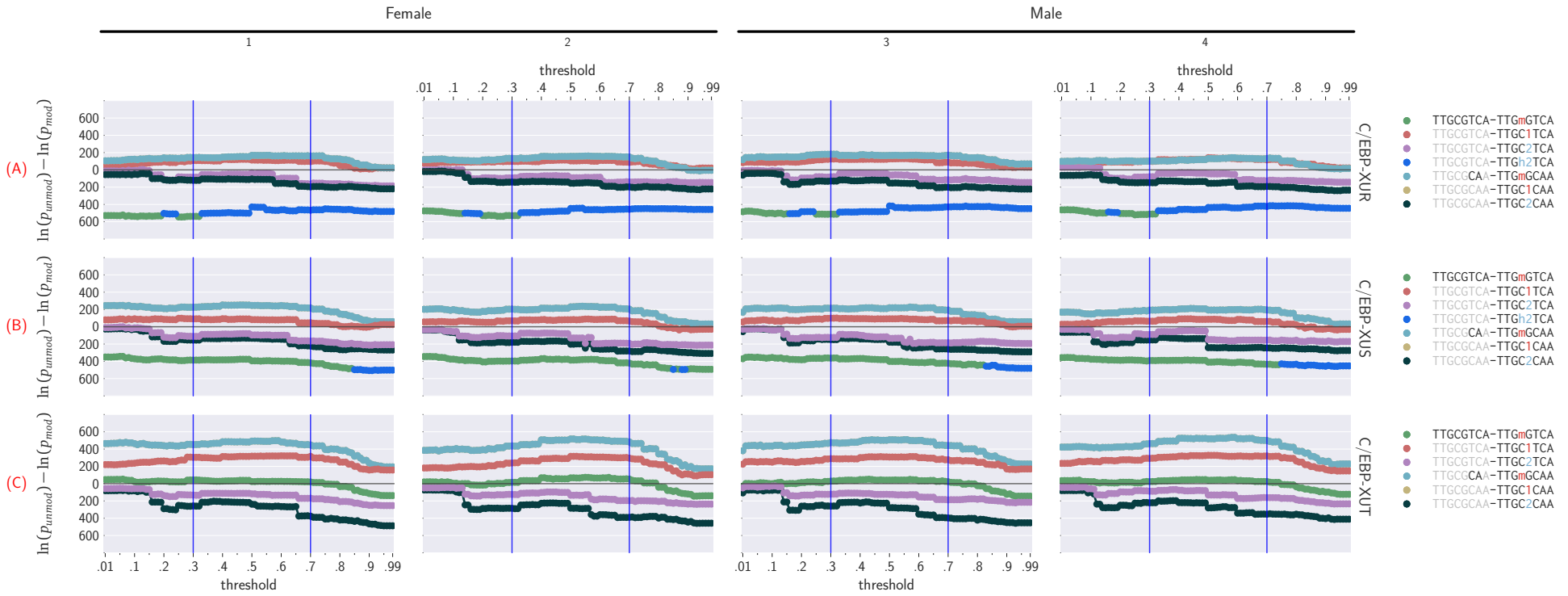
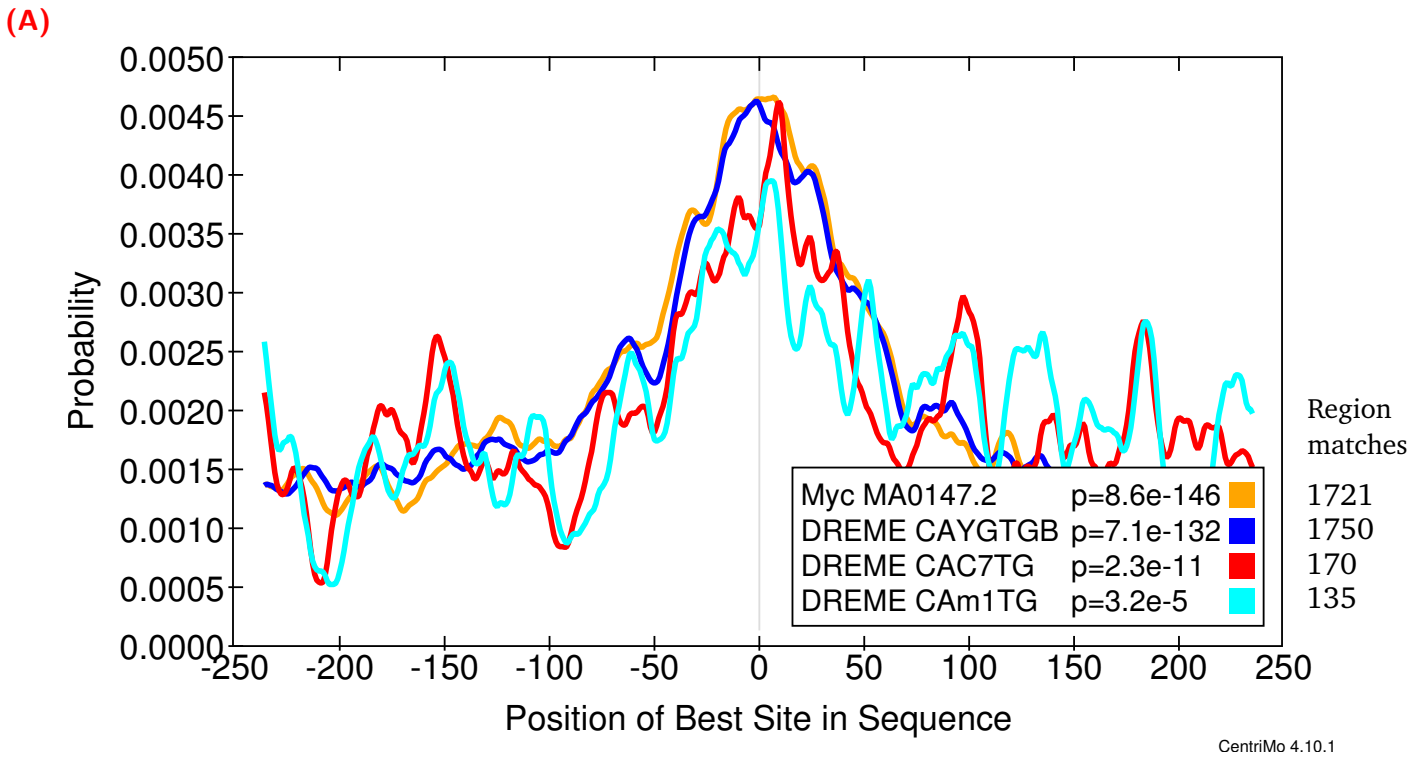


Figure S3. **Relationship between unmodified versus modified C/EBPβ statistical significance of central enrichment (from CentriMo[66]) and modified base calling thresholds across different (ox)WGBS specimens.[51] Each unmodified motif, at each threshold, is compared to its top three most significant modifications. Since only the top three motif pairs are selected per threshold value, specific thresholds can result in new motif pairs, while previously enriched pairs may no longer be present at a given threshold. Each column pertains to a particular (ox)WGBS replicate. Each row of plots pertains to a particular C/EBPβ ChIP-seq dataset: ENCFF001XUR, ENCFF001XUS, ENCFF001XUT. Negative values correspond to a preference for the unmodified motif, while positive values correspond to a preference for the modified motif. These results are additionally depicted with all of the other tested transcription factors in Figure S1.**



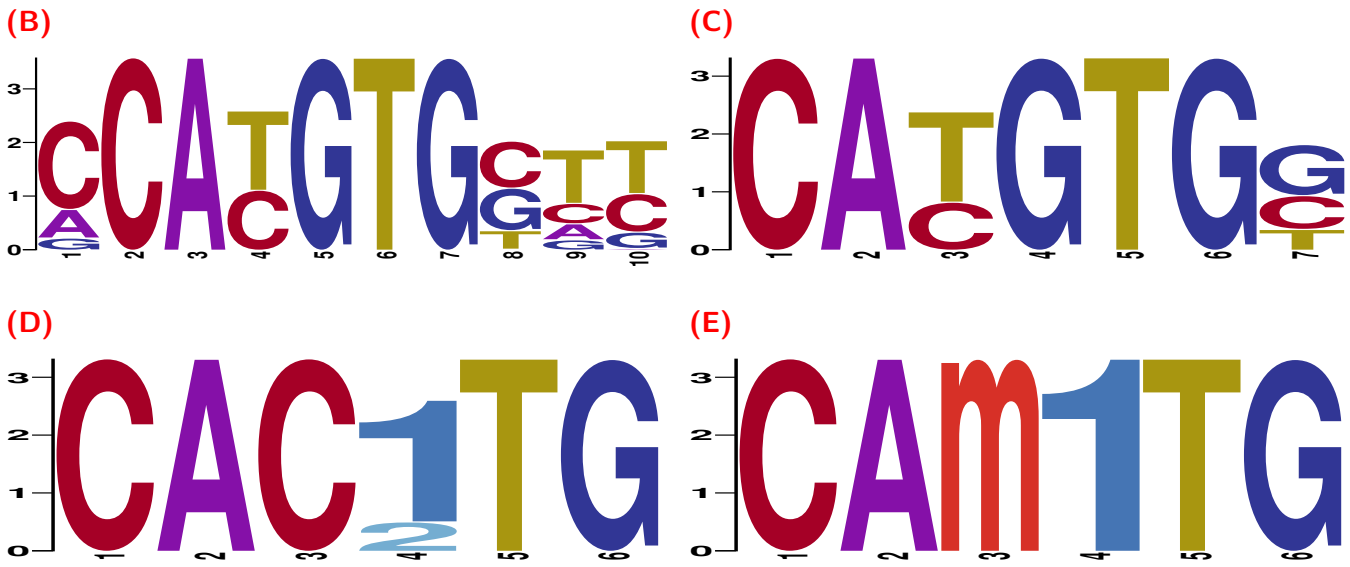
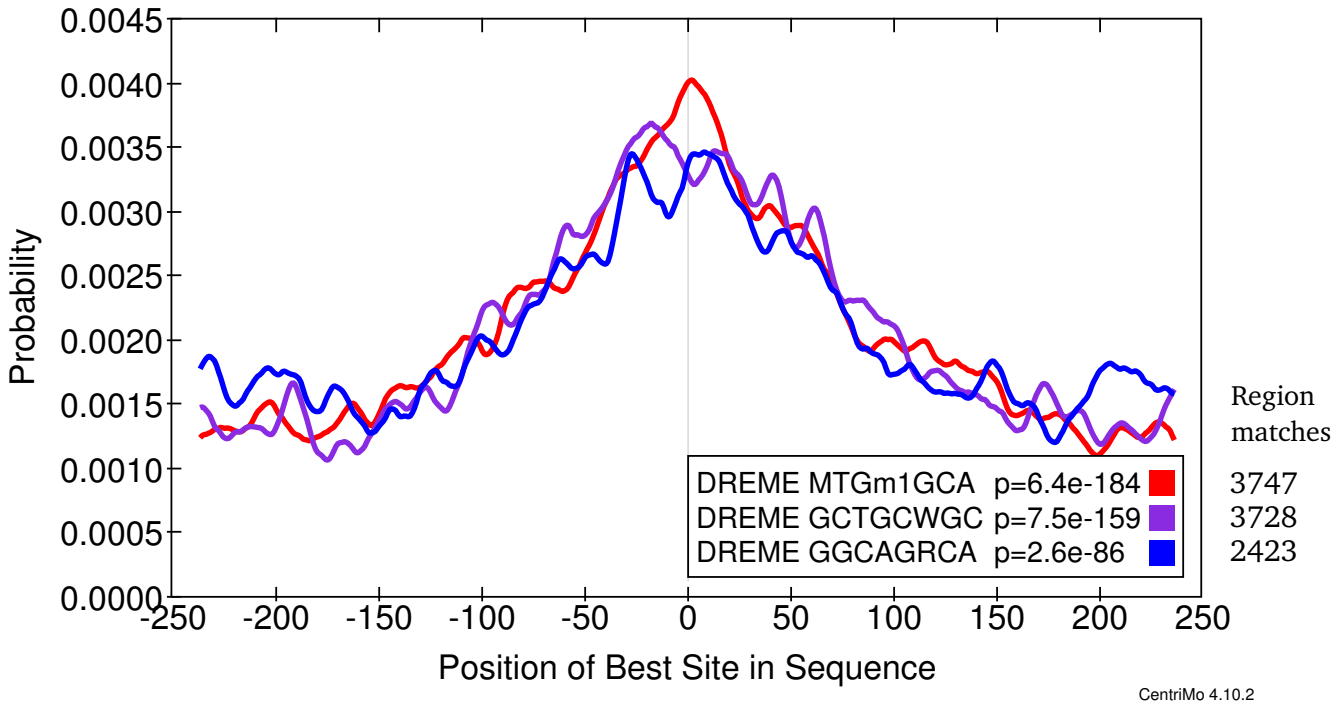


Figure S4. c-Myc (ENCFF001YJE; 6297 ChIP-seq peaks) CentriMo analysis upon *de novo* and JASPAR motifs (see Methods) using female replicate 2 of the (ox)WGBS data[51] at a 0.7 modification threshold. (A) the CentriMo result with the c-Myc JASPAR motif in orange (logo in (B)), a top E-box unmodified motif in blue (logo in (C)), and DREME (hydroxy)methylated motifs in red and cyan (logos in (D) and (E), respectively). We depict the JASPAR sequence logo using MEME's relative entropy calculation and colouring for consistency.

(A)



(B)

(C)

(D)

Figure S5. **Strogantsev et al.[38] CB9 ZFP57 (mm9 aligned; 72 592 ChIP-seq peaks) CentriMo analysis upon *de novo* and JASPAR motifs (see Methods) using female replicate 2 of the (ox)WGBS data[51] at a 0.7 modification threshold. (A) the CentriMo result with an expected ZFP57 modified motif in red (logo in (B), reverse complement shown) and top unmodified motifs in purple (logo in (C)) and blue (logo in (D)).**

Appendix A Recommendations for modified nucleobase nomenclature

Interest in different covalent DNA modifications and improvements in sequencing technologies are expected to create a greater need for computational analyses of modified sequence data. In order to encourage standardization, we recommend symbols for various modified nucleobases (Table S1). We use lower case letters (a–z) for specific nucleobase forms and numerals (0–9) to specify complements without any information loss. The list is not comprehensive, but may provide guidance for those who need to select symbols for these bases. This list also reserves specific symbols in an attempt to reduce contradictory definitions. All upper-case letters of the Latin alphabet are considered to be reserved for allocation by IUPAC, in addition to those already specified.[43] We use lower-case letters, which may have different meanings in upper-case, since there are insufficient unassigned letters to restrict ourselves to those. Many applications, including the current implementation of the MEME Suite, only support the Latin letters and numerals.

For any abbreviations of covalently modified nucleobases, we recommend that they be referred to as <position><modification><base>, where <position> is the position of the modified atom, <modification> is the modification, and <base> is the nucleobase being modified, such as "5mC". In particular, we recommend that no punctuation be used to demarcate the number of the atom from its modification and that the numeral always appear before the base being modified. For example, others have occasionally abbreviated 6-methyladenine as m6A, but we recommend the use of 6mA instead when the modification occurs in DNA rather than in RNA. This distinction in the nomenclature of DNA vs. RNA modifications can be seen in a recent review by Chen et al.[105].

The core symbols for cytosine modifications (Table 1) have been incorporated into Table S1. While we specified a set of ambiguity codes for our usage in this work (Table 2), we do not recommend general definitions. Instead, we suggest that the latter portions of the lower-case Latin alphabet and numerals be reserved for this purpose. This increases the likelihood that sufficient symbols will be available within the alphanumeric alphabet for a variety of use-cases. As implemented in this work, we recommend that ambiguity codes be assigned starting from the end of their character set, beginning with the most equivocal ambiguity code.

Nucleobase			Complement	
Abbreviation	Name	Symbol	Name	Symbol
Covalent modifications of cytosine				
5mC	5-Methylcytosine	m	Guanine:5-Methylcytosine	1
5hmC	5-Hydroxymethylcytosine	h	Guanine:5-Hydroxymethylcytosine	2
5fC	5-Formylcytosine	f	Guanine:5-Formylcytosine	3
5caC	5-Carboxylcytosine	c	Guanine:5-Carboxylcytosine	4
Covalent modifications of thymine				
5hmU	5-Hydroxymethyluracil	g	Adenine:5-Hydroxymethyluraci	
5fU	5-Formyluracil	e	Adenine:5-Formyluracil	
5caU	5-Carboxyluracil	b	Adenine:5-Carboxyluracil	
Covalent modifications of adenine				
6mA	6-Methyladenine	a	Thymine:6-Methyladenine	
Covalent modifications of guanine				
8oxoG	8-Oxoguanine	o	Adenine:8-Oxoguanine (mismatch)	
Reserved synthetic bases				
Xao	Xanthosine	n	Cytosine:Xanthosine	

Table S1. Recommendations for the nomenclature of modified nucleobases, grouped by the unmodified nucleobase. Numeral symbols for complements are provided for some bases, to ensure they can be differentiated from bases complementary to their unmodified forms. These numerals could be re-assigned for work with other sets of covalent modifications, if necessary.