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

Coby Viner [1,2], Charles A. Ishak [2], James Johnson[3], Nicolas J. Walker [4], Hui Shi [4], Marcela K. Sjöberg-Herrera [5], Shu Yi Shen [2], Santana M. Lardo [6], David J. Adams [5], Anne C. Ferguson-Smith [4], Daniel D. De Carvalho [2,7], Sarah J. Hainer [6], Timothy L. Bailey [8], and Michael M. Hoffman [1,2,7,9,*]

[1] *Department of Computer Science, University of Toronto, Toronto, ON, Canada*
[2] *Princess Margaret Cancer Centre, University Health Network, 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 Sanger Institute, Wellcome Genome Campus, Cambridge, England*
[6] *Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA, USA*
[7] *Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada*
[8] *Department of Pharmacology, University of Nevada, Reno, Reno, NV, USA*
[9] *Vector Institute, Toronto, ON, Canada*
[*] To whom correspondence should be addressed (michael.hoffman@utoronto.ca).

July 29, 2022

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 successively oxidize to 5-hydroxymethylcytosine (5hmC), 5-formylmethylcytosine (5fC), and 5-carboxylmethylcytosine (5caC). Just as transcription factors distinguish one unmodified nucleobase from another, some distinguish between unmodified and covalently modified bases. Modification-sensitive transcription factors provide a mechanism for widespread changes in DNA methylation and hydroxymethylation to 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` DNA 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 (PWM) model of transcription factor binding affinity to an expanded DNA alphabet. Further, we developed a program, Cytomod, to create a modified sequence. We also enhanced the Multiple EM for Motif Elicitation (MEME) Suite, adding the capacity to handle custom alphabets. These versions permit users to specify new alphabets, anticipating future alphabet expansions.

Results. We created expanded-alphabet sequences, using whole-genome maps of 5mC and 5hmC in mouse T cells and 5mC in human K562 cells. Using this sequence and chromatin immunoprecipitation-sequencing (ChIP-seq) data from the Encyclopedia of DNA Elements (ENCODE) Project 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 determined cutoffs for modified

base calling using hypothesis testing across different thresholds. Using these known binding preferences to tune model parameters, we discovered novel modified motifs, across a wide array of transcription factors. We clustered modified-unmodified hypothesis pairs, revealing a complex interplay of modification preferences. After predicting methylated and hydroxymethylated binding preferences of OCT4, we validated them with the first conjoint cleavage under targets and release using nuclease (CUT&RUN) experiments across conventional, methylation-, and hydroxymethylation-enriched sequences.

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. Our approach readily extends to other DNA modifications, when provided with genome-wide single-base resolution data. 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. These modifications do, however, 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

Active demethylation of methylcytosine back to its unmodified form proceeds through successive oxidation of methylcytosine, to 5-hydroxymethylcytosine (5hmC), 5-formylmethylcytosine (5fC), and 5-carboxylmethylcytosine (5caC; Figure 1).[5,6] Ten-eleven translocation (TET) enzymes mediate these oxidations.[7]

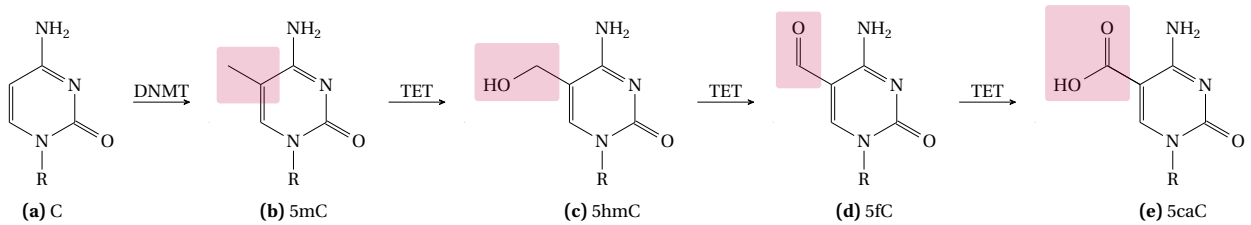


Figure 1. Stepwise epigenetic modification of cytosine. (a) DNA methyltransferase (DNMT) methylates cytosine to create **(b)** 5mC, which TET enzymes oxidize to create **(c)** 5hmC, which TET again oxidizes to create **(d)** 5fC. Finally, TET can further oxidize 5fC to **(e)** 5caC, which can then return to an unmodified cytosine through 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.

Widely known for its regulatory roles, 5mC occurs across diverse genomic contexts.[8,9] Regulatory processes increasingly implicate 5hmC.[10] Recognized as a stable modification,[11] 5hmC has structural rationale for its reduced propensity of TET-mediated oxidation.[12] We know less about 5fC and 5caC, largely due to their considerably lesser abundance. Far less abundant than 5hmC, these further oxidized bases occur with an order of magnitude less abundance than 5mC.[13] Methylation occurred at nearly 3% of mouse embryonic stem cells (mESCs) cytosine bases,[5] while hydroxymethylation occurred in 0.055%, in a different mESC sample.[14] Fewer works investigate the genome-wide distribution and role of 5fC (such as poised enhancers) and 5caC.[15–17] Often regarded as mere intermediates of the demethylation cascade, these nucleobases generally occur with two to three orders of magnitude less abundance than 5hmC, showing an ability to trigger a strong DNA damage responses.[1,5,14] In mESCs, 5fC accounts for 0.0014% of cytosine bases,[14] while 5caC accounts for a mere 0.000335%.[5] While their playing a distinctive and pan-tissue regulatory role remains uncertain, converging lines of evidence suggest that they too modulate gene expression.[10] Not merely a demethylation intermediate,[18] 5fC alters the conformation of the DNA double helix[19] and demonstrates stability in mESCs.[18]

All of these modifications occur most frequently, by far, at CpG dinucleotides. Non-CpG 5mC nucleobases, however, still exist in substantial quantities, particularly within mESCs.[20,21] Within a population of cells, at a given locus, unmodified nucleobases and different kinds of modified nucleobases can often overlap.[13] Our DNAmod database catalogues these and many other DNA modifications.[22]

In bulk data, usually one considers a base modified or not using some threshold, above which one "calls" a particular modified base or set of possible modifications. There exist several ways to perform modified base calling, generally first involving computing a proportion of modification, at a specific position. We use the maximum likelihood methylation levels (MLML)[23] method to do this. Then, we must decide the value sufficient to call a modification downstream. Additional biochemical complexities increase the difficulty of mapping these modifications. These complexities include strand biases,[13] concomitant modifications, and hemi-methylation.[24]

Modified nucleobases can substantially alter transcription factor recognition

Many transcription factors prefer specific motifs, enabling the sequence specificity of transcriptional control.[25] 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.[26] Just as transcription factors distinguish one unmodified nucleobase from another, some transcription factors 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. One of many such non–sequence-specific methyl-CpG binding proteins, MeCP2, binds 5hmC.[27] Non–sequence-specific modified nucleobase binding, however, has a role limited to specific protein families.[28] Elucidating and characterizing sequence-specific motifs provides greater information, but also poses a greater technical challenge.[28]

A few transcription factors have well-characterized modification preferences. These preferences can serve as a means of verifying a predictive framework. One would expect a working model to robustly yield the known preferences. Therefore, an understanding of known modification sensitivities informs the design of such a model.

Some specific transcription factors have an extensive history of study of their modification preferences. Both C/EBPα and C/EBPβ have increased binding activity, in the presence of central CpG methylation, formylation, or carboxylation of their canonical octamer (consensus: `TTGC|GCAA`). Both DNA strands contribute to increase that effect and hemi-modification still demonstrates a reduced effect.[29] 5hmC inhibits binding of C/EBPβ, but not C/EBPα.[29]

The basic helix-loop-helix (bHLH) family transcription factor c-Myc, has a strong preference for unmethylated E-box motifs, often preferring the `CACGTG` hexamer.[30,31] Many other bHLH transcription factors also demonstrate such a preference.[32–36]

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).[37,38] Additional methylation often occurs in ZFP57 motifs with a final guanine residue as the core binding site.[38] Crystallography and fluorescence polarization analyses further confirm this preference.[39] ZFP57 also has successively decreasing affinity for the oxidized forms of 5mC.[39]

Hu et al.[3] demonstrate that central CpG-methylated motifs have strong binding activity for multiple transcription factors. Their protein binding microarray data show that these motifs often differ

greatly from the unmethylated sequences that those transcription factors usually bind.

Other transcription factors also have methylation sensitivity.[40] Some may even increase binding in the presence of 5caC.[41] Among 327 *Arabidopsis* transcription factors, 248 (76%) exhibited sensitivity to covalent DNA modifications, with 14 preferring modified DNA.[42]

Models for the interaction between transcription factors and modified bases

The specific regulatory activity involved in the interplay between transcription factors binding to modified bases, versus the deposition and removal of these modifications, remains an open question. Transcription factors act as both readers and effectors of methylation.[28] Models for this regulatory interplay[28,43] highlight the need for additional genome-wide characterizations of transcription factor binding preferences.

Transcription factors may bind to a modified base to prevent its modification, or, in some instances, to increase the likelihood of its modification. Alternatively, transcription factors could bind to reverse an existing modification or the converse. This binding mode could reinforce transcriptional repression.[43] These scenarios could occur in different genic contexts, potentially mediated by different motif groups. Even factors within the same family may have differences in modified binding preferences, conferring additional specificity or assisting in stable protein-DNA complex formation. Additional characterization of modified transcription factors can help differentiate them. This would greatly increase our understanding of gene regulation.

Computational methods for base modification data in transcription factor binding prediction

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. Other work has profiled the methylation preferences of a large array of transcription factors, but has not done so at the scale of our current work. Mostly, when modeling the affinity of transcription factors for DNA sequences, previous work has not treated modified nucleobases as first-class objects akin to unmodified nucleobases, adding artificial distinctions unlikely to reflect the underlying biophysical interactions.

A comprehensive *in-vitro* analysis, coupled with selected follow-up crystal structures, revealed mechanistic bases for some 5mC interactions.[44] A random forest[45,46] combined genomic and methylation data[47] to predict transcription factor binding. Those predictions, however, did not attempt to predict the preference of factors for methylated DNA.[47] The MethMotif database enumerates methylated transcription factor motifs.[48]

Most recently, Grau et al.[49] also employed an expanded alphabet genome from whole genome bisulfite sequencing (WGBS) data, for 5mC only. Their focus differs from ours, however. They emphasize that their models go beyond PWMs, allowing for intra-motif dependencies. Their comparisons mainly focus on classification performance bench-marking in differentiating bound versus unbound sequences. Song et al.[50] demonstrate an *in vitro* method to assess modification-specific preferences of all cytosine states. They demonstrate distinct preferences of both symmetric and hemi-modifications. Most recently, Hernandez-Corchado et al.[51] developed a joint model of accessibility and methylation sequence. They used this model to explore a large number of chromatin immunoprecipitation-sequencing (ChIP-seq) datasets, assessing many transcription factor binding site preferences for 5mC.

Here, we describe methods to analyze covalent DNA modifications and their effects on transcription factor binding sites by introducing an expanded epigenetic alphabet. While others proposed

expanding the genomic alphabet in other ways,[52] we and Ngo et al.[53] first proposed expanding it in this context,[54,55] for facilitating bioinformatic analyses of cytosine modifications. Unlike our work, they focused on motif elucidation in this expanded alphabet, for which we leverage existing tools to focus on downstream biological consequences, as described later. We introduce Cytomod, software to integrate DNA modification information into a single genomic sequence and we detail the use of extensions to the Multiple EM for Motif Elicitation (MEME) Suite[56] to analyze 5mC and 5hmC transcription factor binding site sensitivities. We validate the main novel predicted transcription factor from our unbiased screen, OCT4. In so doing, we provide the first conjoint cleavage under targets and release using nuclease (CUT&RUN)[57,58] datasets across conventional, methylated-, and hydroxymethylated-enriched sequences, for OCT4. Overall, our results demonstrate that DNA methylation intermediates can directly affect transcription through the formation of *de novo* recognition motifs that regulate transcription factor binding.

Results

Expanded-alphabet genomes facilitate the analysis of modified base data

We created an expanded-alphabet sequence using oxidative (ox) and conventional WGBS maps of 5mC and 5hmC for naive *ex vivo* mouse $CD4^+$ T cells.[59] Then, we generated individual modified genomes across four replicates of combined ox and conventional 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. Using the thresholds discovered in the murine analyses, we created conventional WGBS maps of 5mC in the human K562 erythroid leukemia cell line,[60,61] using Encyclopedia of DNA Elements (ENCODE) data.

Hypothesis testing reveals altered transcription factor binding preferences

Expanded-alphabet analysis shows results consistent with known preferences

We used a hypothesis testing approach on the expanded-alphabet sequence to examine transcription factors' preferences for modified or unmodified DNA. First, we used this approach on three transcription factors with previously known methylation or hydroxymethylation sensitivities. Two of the tested transcription factors prefer methylated DNA: ZFP57[38] and C/EBPβ.[29] The remaining transcription factor, c-Myc, prefers unmethylated DNA.[30,31] Additionally, C/EBPβ has reduced affinity for hydroxymethylated DNA.[29]

We tested known unmodified transcription factor binding motifs against all possible 5mC and 5hmC modifications at all CpG dinucleotides. For each modified motif, we assessed its expected DNA binding affinity using the adjusted central enrichment p-value from CentriMo[62] (Methods). We conducted the same test for the unmodified version of the motif. To compare these binding affinities, we subtracted the $\log_{10}$-transformed p-value of the modified motif from the unmodified motif. Positive values for this difference represented a preference for the modified motif, while negative values represented a preference for the unmodified.

The expected transcription factor binding preferences held across all four biological replicates of WGBS and oxWGBS data. They held for all investigated modified nucleobase calling thresholds (from 0.01–0.99 inclusive, at 0.01 increments, representing modification confidence; Figure 2). No c-Myc log p-value differences exceeded zero. This implies that c-Myc favours unmodified E-box motifs over modified c-Myc motifs. ZFP57 favours binding to modified nucleobases, over their unmodified

counterparts. The often-found TGCm1m1 motif[38] had one of the greatest increases in predicted binding affinity for modified DNA. Two methylated motifs had the greatest increase in predicted binding affinity for C/EBPβ: TTGmGCAA and TTGC1TCA (see Table 1 and Table 3 for an overview of modified base notation).

We obtained the same results for multiple different ChIP-seq replicates for these three transcription factors (Figure S1). Perturbations of binding assessments, such as peak-calling stringency (Figure S2) and required degree of motif statistical significance (Figure S3), demonstrated the robustness of these results.

We found ZFP57 had a strong preference for methylated DNA, but also a noticeable preference for hydroxymethylated DNA. CentriMo quantifies these preferences,[62] both in terms of p-value significance, and in terms of the centrality of motif concentration (Methods). Hydroxymethylated CpGs had a substantially smaller increase in binding affinity than methylated motifs (Figure 2), but still greater than the completely unmethylated motif. Here, our expanded-alphabet methodology simply recapitulated a previous *in vitro* finding that ZFP57 has the greatest binding affinity for motifs containing 5mC, followed by 5hmC, and then by unmodified cytosine.[39]

Expanded-alphabet analysis enables comparisons across a wide array of transcription factors

We looked for shared preferences among families of transcription factors for modified or unmodified bases in both the mouse and the human data. Mostly, we defined families with TFClass[63,64] (Methods).

We and others[32–36] have found a consistent preference for unmethylated binding motifs across a broad selection of bHLH transcription factors. When we found bHLH motifs with a putatively modified preference against unmodified JASPAR motifs, often *de novo* motifs consistently had more significant unmodified p-values (Figure S6). This suggested that these transcription factors usually preferred the unmodified counterpart. The leucine-zipper subfamily of bHLH transcription factors, however, had a subset of motifs that preferred to bind in a modified context. For example, both USF1 and USF2 preferred to bind in unmodified and modified contexts, to differing extents, and had mixed binding preferences within motif clusters (Figure 5C; Discussion).

Many zinc finger family motifs displayed a propensity toward modified motifs, but not all. EGR1/ZIF268/NGFI-A, a Cis_2–His_2 zinc finger, had moderate evidence of a methylated binding preference, with multiple positively-scoring hypothesis pairs, including some > 100. Conversely, ZNF384/CIZ/TNRC1 of the same sub-family, present in both our mouse and human analyses, had only weak evidence of a modified preference, with only a single hypothesis pair scoring above 10 (Figure S1).

C/EBPβ has dichotomous 5mC versus 5hmC binding preferences

C/EBPβ provides an excellent test case for the impact of modified bases on transcription factor binding because of its dichotomous preferences for 5mC versus 5hmC.[29] Our method recapitulated this preference, across all ChIP-seq datasets, for all replicates of ox and conventional WGBS. Methylated motif pairs generally had positive ratios, whereas hydroxymethylated motif pairs had negative (Figure S3).

A positive strand, hemi-methylated, motif (TTGmGTCA) presented an exceptional case. We observed a preference for the unmodified motif, over its hemi-methylated motif. Unlike the consensus C/EBPβ motif, this corresponds to the chimeric C/EBP|CRE octamer. This chimeric transcription factor has a more modest preference toward its methylated DNA motif.[29] Nonetheless, we would still have expected a weak preference for the hemi-methylated motif, over its unmodified counterpart, and

found the observed unmodified motif preference surprising. Additionally, we found greater enrichment for hemi-methylation than complete methylation, which contradicts findings of both strands contributing to increase the effect.[29] This may arise from technical issues with hemi-methylation in our modified sequence, or because our methods have greatest accuracy only within specific cell types or contexts.

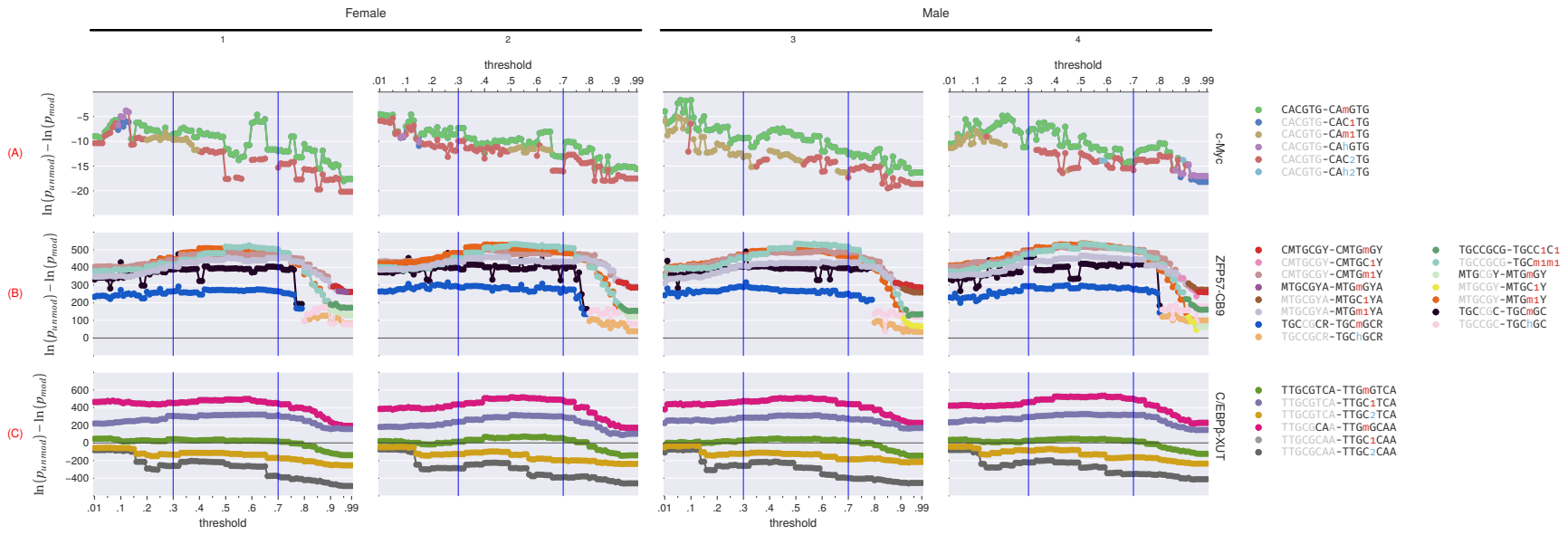


Figure 2. Relationship between unmodified versus modified motif statistical significance of central enrichment (from CentriMo[62]) and modified base calling thresholds across different WGBS and oxWGBS specimens, in mice.[59] We compare each unmodified motif, at each threshold, to its top three most significant modifications for c-Myc and C/EBPβ, but only the single most significant modification for ZFP57. The displayed motif pairs changes at individual thresholds, depending on which motif pairs stay in the top three. See Table 1 and Table 3 for an overview of modified base notation. Sign of value indicates preference for the unmodified (negative) motif or the modified (positive) motif. Rows: single ChIP-seq replicates for a particular transcription factor target, one each of: **(A)** c-Myc (mESCs; Krepelova et al.[65]), **(B)** ZFP57 (CB9 mESCs; Strogantsev et al.[38]), and **(C)** C/EBPβ (C2C12 cells; `ENCFF001XUT`). Columns: replicates of WGBS and oxWGBS (mouse CD4$^+$ T cells; Kazachenka et al.[59]).

Our methods yield suitable base-calling thresholds for 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 demonstrated 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 3) and 0.7 thresholds (Figure 4).

Despite robust findings with hypothesis testing, across almost the entire range of possible thresholds, we sometimes did not detect the expected binding preferences in a *de novo* context. The c-Myc and, to a lesser extent, ZFP57 CentriMo analyses, in a non-hypothesis testing context, did not demonstrate substantial enrichment nor depletion with respect to modified versus unmodified motifs. We consider potential explanations for this in Discussion.

Many transcription factors bind in modified and unmodified contexts, with variable motif preferences

Our analysis revealed that several factors bind in both modified and unmodified contexts (Figure 5). Unlike prior analyses which often aimed to binarize binding preferences, our results highlight that most factors can bind in both contexts, albeit to varying extents and with different motifs.

Protein binding microarray analyses have led to the conclusion that JUND is "uniformly inhibited by 5mC". Despite this, closer inspection of these prior analyses reveals that JUND has a small group of motifs with a slight preference for 5mC.[66] Specifically, at least 8 cytosine-containing motifs have 5mC z-scores above zero, with at least 3 such motifs having scores of close to 30. Similar findings applied to 5hmC.

The protein binding microarray data accord well with our results, for which almost all tested hypothesis pairs (37/44) showed an unmethylated preference. In our analysis, only 7 (inclusive of 1 hydroxymethylated motif) showed a slight preference for methylated binding (Figure 5A). JUN had similar results (Figure 5A; 3/37 had scores > 0, inclusive of 1 hydroxymethylated motif).

FEZF2 appeared to have two completely different motifs, with no overlap in their preferences for modified versus unmodified cytosine (Figure 5B). Indeed, removing modification information from modified-preferring FEZF2 motifs led to a single motif cluster, distinct from the main unmodified motif clusters. In summary, two distinct binding groups for FEZF2 appear to exist.

As mentioned above, both USF1 and USF2 preferred to bind in both unmodified and modified contexts (Figure 5C; Discussion). *In vivo* evidence suggests USF1/2 prefers 5hmC weakly, at least in some cellular contexts.[50]

C/EBPβ had the expected extreme preference for methylated motifs, and generally a preference against hydroxymethylated motifs, with some exceptions (Figure 5B). The very strong modified preferences of C/EBPβ appeared especially apparent, exhibiting a high rank (very low p-value) and substantial motif centrality.

As expected, ZFP57 showed a very strong and consistent preference for methylated motifs (Figure 5D). It had 328 of 393 total ZFP57 methylated motifs with a score > 0 (median: 171.2; max: 2292, exemplifying the strong preference), across all motifs in Figure 5D. This transcription factor also showed some apparent preference for hydroxymethylated motifs. It had 291 of 435 total ZFP57 hydroxymethylated motifs with a score > 0 (median: 152.4; max: 1379), across all motifs in Figure 5D. Most modified motifs that scored above zero, however, had at least one 5mC and one 5hmC nucleobase (Figure 5D, red motifs).

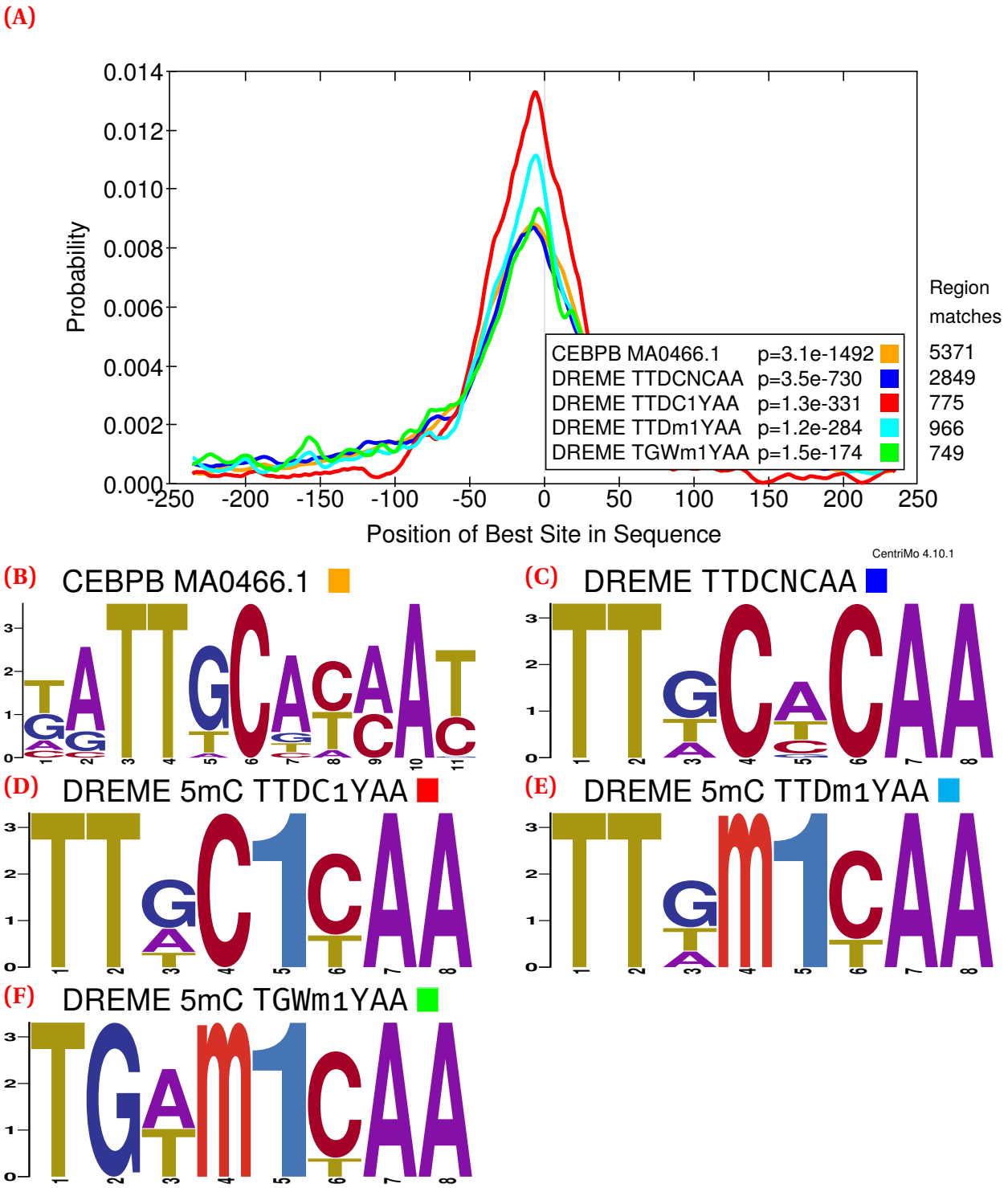


Figure 3. C/EBPβ (GSM915179; 11 434 ChIP-seq peaks) CentriMo analysis of *de novo* and JASPAR motifs (Methods). Depicts female replicate 2 of the combined WGBS and oxWGBS data[59] at a 0.3 modification threshold. **(A)** the CentriMo result with the JASPAR C/EBPβ motif (orange), top Discriminative Regular Expression Motif Elicitation (DREME) unmethylated C/EBPβ motif (blue), and DREME methylated motifs (red, cyan, and green). **(B)** Sequence logo of the JASPAR C/EBPβ motif. **(C)** Sequence logo of the top DREME unmodified motif. **(D)** Sequence logo of the top DREME methylated motif. **(E)** Sequence logo of the second DREME methylated motif. **(F)** Sequence logo of the third DREME methylated motif. Listed p-values computed by CentriMo.[62] For consistency, we depict the JASPAR sequence logo using MEME's relative entropy calculation and colouring.

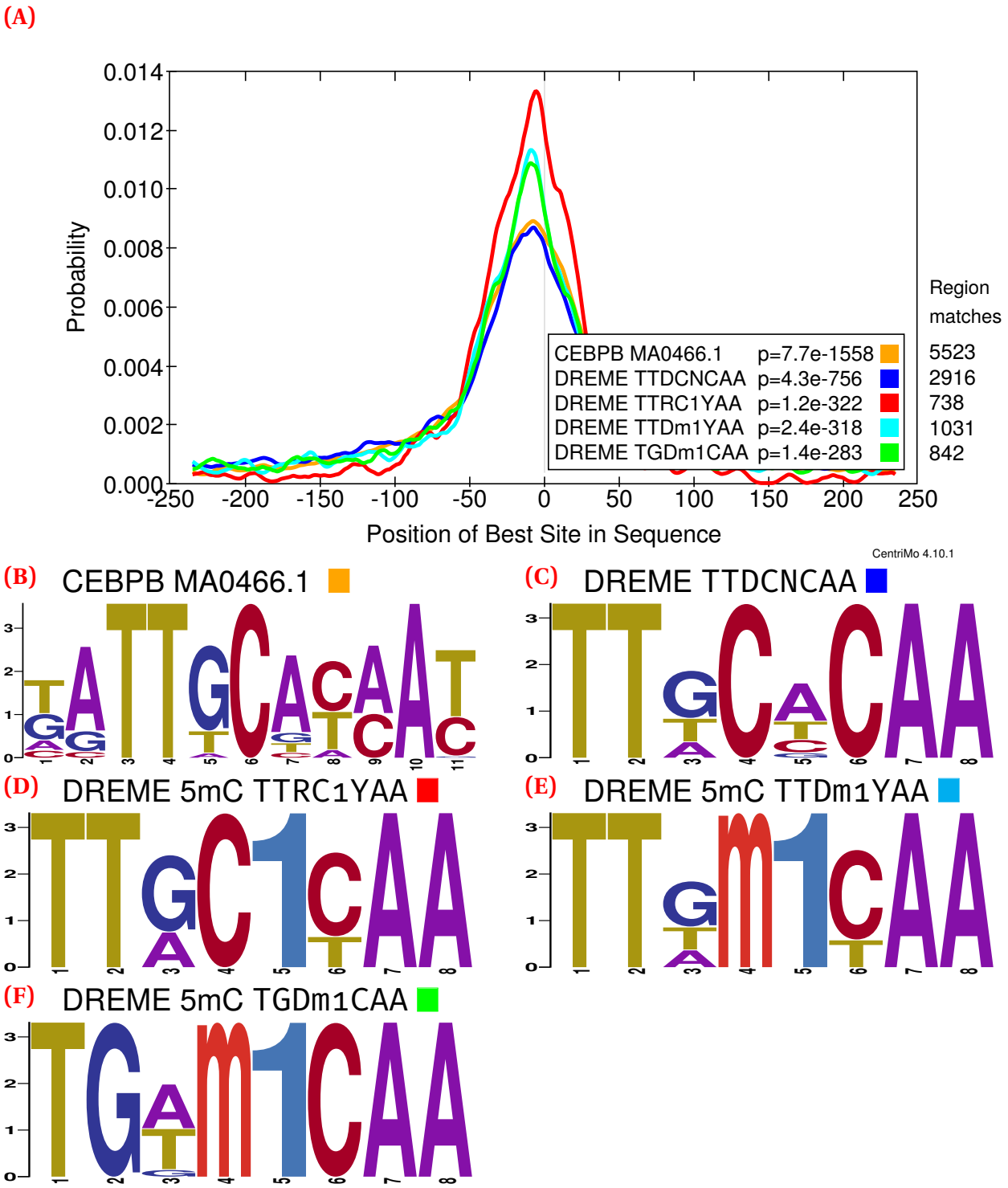


Figure 4. C/EBPβ (GSM915179; 11 434 ChIP-seq peaks) CentriMo analysis of *de novo* and JASPAR motifs (Methods). Depicts female replicate 2 of the combined WGBS and oxWGBS data[59] at a 0.7 modification threshold. **(A)** the CentriMo result with the JASPAR C/EBPβ motif (orange), top DREME unmethylated C/EBPβ motif (blue), and DREME methylated motifs (red, cyan, and green). **(B)** Sequence logo of the JASPAR C/EBPβ motif. **(C)** Sequence logo of the top DREME unmodified motif. **(D)** Sequence logo of the top DREME methylated motif. **(E)** Sequence logo of the second DREME methylated motif. **(F)** Sequence logo of the third DREME methylated motif. Listed p-values computed by CentriMo.[62] For consistency, we depict the JASPAR sequence logo using MEME's relative entropy calculation and colouring.

Our final control in the unmodified-preferring direction, c-Myc, also had the anticipated results, with a strong preference for unmodified motifs (Figure 5E). While the c-Myc results indicated a small number of preferences for modified DNA, these preferences were exceptionally weak and clustered with much stronger preferences for unmodified DNA. Specifically, all analyzed data included 67 c-Myc motifs with the expected negative score (median: −219.6; min: −3167) and only 4 with a positive score (median: 12.96; max: 24.91).

While a number of OCT4 motifs bound in an unmodified context, some OCT4 motifs showed clear preferences for both methylated and hydroxymethylated motifs. While others reported evidence of this methylated preference,[44] we did not know of either of these modified preferences during the course of our analysis. We remain unaware of any reports of its hydroxymethylated preference. Interestingly, those hydroxymethylated motifs appeared to predominantly cluster either on their own, or with the canonical OCT4 homo- and hetero-dimer.

More than half of our assessed transcription factors bound almost or entirely exclusively in unmodified contexts (Figure S6). Specifically, if limiting to transcription factors without even a single slightly positive motif, 49.3 % of factors had all tested hypothesis pairs score below zero. This varied across our overall dataset, with other factors having some occasional modified preferences.

Modified-base CUT&RUN validates our predictions for OCT4

We validated our OCT4 predictions, performing CUT&RUN[57,58] in mESCs, with conventional, bisulfite-converted, and hmC-Seal-seq library preparations. These three sets of library preparations allowed characterizing the modification preferences of OCT4-bound fragments, across both methylated and hydroxymethylated contexts.

We found considerable evidence for our prediction that OCT4 has a strong preference to bind in a hydroxymethylated context. When comparing unconverted, conventional CUT&RUN to hmC-Seal-seq CUT&RUN, OCT4 and similar *de novo* motifs had much more significant results in the hydroxymethylated context (Figure 6; top DREME motif: $p = 4.5 \times 10^{-149}$; top OCT4 motif: $p = 2.5 \times 10^{-13}$) than in the bisulfite context (top DREME motif: $p = 4.1 \times 10^{-104}$; top OCT4 motif: $p = 1.1 \times 10^{-8}$), all with comparable or greater motif centrality. These motifs also had at least similar preferences for binding in a methylated context (Figure S5).

Our predicted cluster containing the canonical POU5F1B JASPAR motif (`MA0792.1`), had similarly enriched 5hmC motifs, as in our hmC CUT&RUN. Overall, our findings suggest that OCT4 specifically binds hydroxymethylated nucleobases, in concert with methylated and unmodified binding sites.

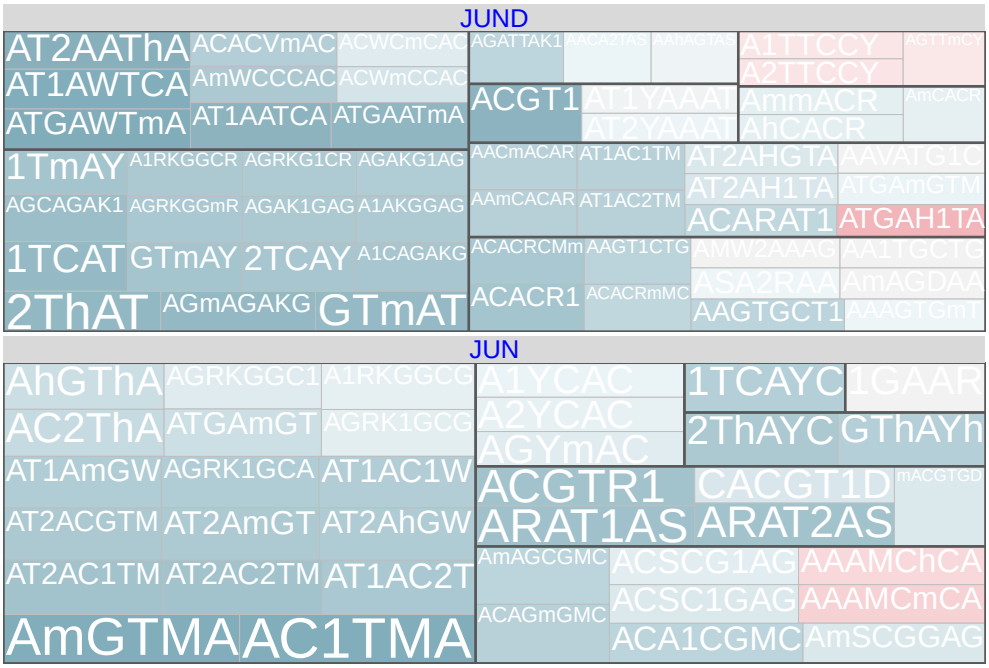
Methods

An expanded epigenetic alphabet

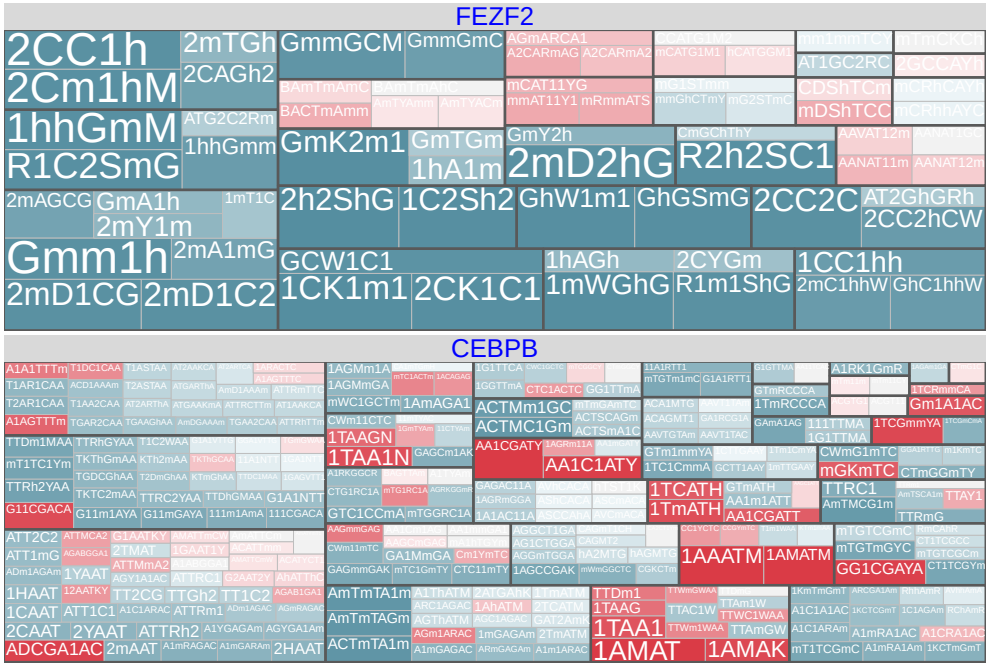
To analyze DNA modifications' effects upon transcription factor binding, we developed a model of genome sequence that expands the standard `A/C/G/T` alphabet. Our model adds the symbols `m` (5mC), `h` (5hmC), `f` (5fC), and `c` (5caC). This allows us to more easily adapt existing computational methods,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

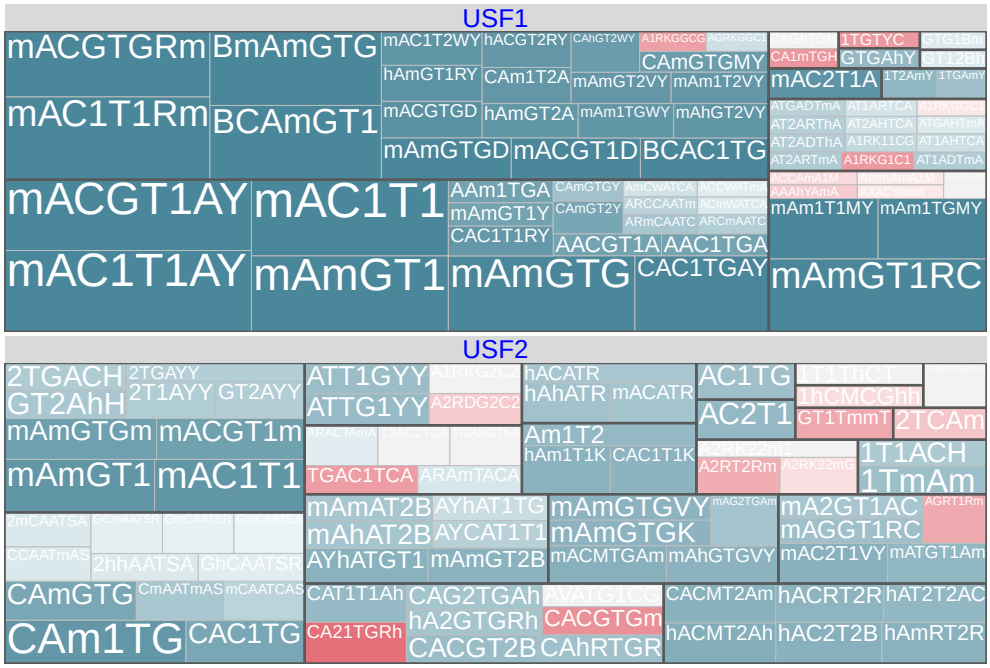
(A)


JUND
AT2AAThA
ACACVmAC
ACWCmCAC
AGATTAK1
A1TTCCY
AT1AWTCA
AmWCCCAC
ACWmCCAC
A2TTCCY
ATGAWTmA
AT1AATCA
ATGAATmA
ACGT1
AT1YAAAT
AmmACR
AT2YAAAT
AhCACR
1TmAY
A1RKGGCR
AGRKG1CR
AGAKG1AG
AACmACAR
AT1AC1TM
AT2AHGTA
AGCAGAK1
AGRKGGmR
AGAK1GAG
A1AKGGAG
AAmCACAR
AT1AC2TM
AT2AH1TA
ACARAT1
ATGAH1TA
1TCAT
GTmAY
2TCAY
A1CAGAKG
ACACRCMm
AAGT1CTG
AMW2AAAG
AA1TGCTG
2ThAT
AGmAGAKG
GTmAT
ACACR1
ACACRmMC
ASA2RAA
AmAGDAA
AAGTGCT1
JUN
AhGThA
AGRKGGC1
A1RKGGCG
A1YCAC
1TCAYC
1GAAR
AC2ThA
ATGAmGT
AGRK1GCG
A2YCAC
2ThAYC
GThAYh
AGYmAC
AT1AmGW
AGRK1GCA
AT1AC1W
ACGTR1
CACGT1D
AT2ACGTM
AT2AmGT
AT2AhGW
ARAT1AS
ARAT2AS
AmAGCGMC
ACSCG1AG
AAAMChCA
AT2AC1TM
AT2AC2TM
AT1AC2T
ACSC1GAG
AAAMCmCA
ACAGmGMC
AmGTMA
AC1TMA
ACA1CGMC
AmSCGGAG

(B)

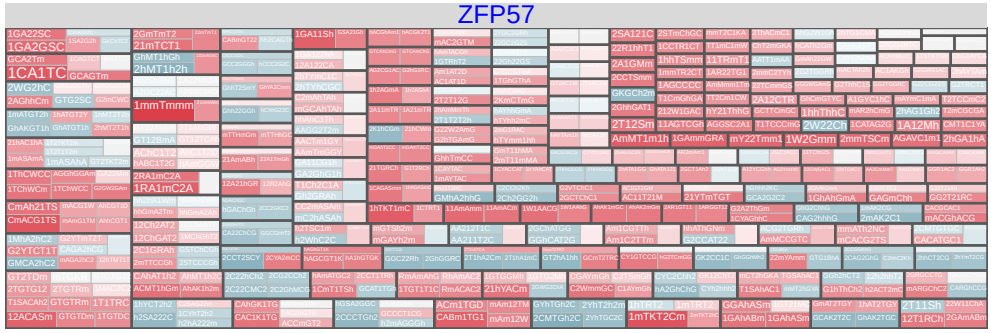

FEZF2
2CC1h
2mTGh
GmmGCM
GmmGmC
AGmARCA1
2CC1hM
2CAGh2
AT1GC2RC
1hhGmM
ATG2C2Rm
BACTmAmm
mCAT11YG
mmAT11Y1
mRmmATS
mDShTCC
1hhGmm
GmK2m1
GmTGm
GmY2h
R1C2SmG
1hA1m
2mD2hG
R2h2SC1
AANAT11m
AANAT12m
2mAGCG
GmA1h
2mY1m
2h2ShG
1C2Sh2
GhW1m1
GhGSmG
2CC2C
AT2GhGRh
2CC2hCW
Gmm1h
2mA1mG
GCW1C1
1hAGh
2CYGm
1CC1hh
2mD1CG
2mD1C2
1CK1m1
2CK1C1
1mWGhG
R1m1ShG
2mC1hhW
GhC1hhW
CEBPB
A1A1TTTm
T1AR1CAA
T2AR1CAA
A1AGTTTm
TTDm1MAA
mT1TC1Ym
TTRh2YAA
G11CGACA
ATT2C2
ATT1mG
ADm1AGAm
1HAAT
1CAAT
2CAAT
ADCGA1AC
2YAAT
2mAAT
1AGMm1A
1AGMmGA
mWC1GCTm
1AmAGA1
CWm11CTC
1TAAGN
1TAA1N
GAGCm1AK
GTC1CCmA
mTGGRC1A
AAGmmGAG
CWm11mTC
GA1MmGA
GAGmmGAK
mTC1GmTY
CTC11mTY
AmTmTA1m
AmTmTAGm
ACTmTA1m
1GGTTmA
CTC1ACTC
GG1TTmA
ACTMm1GC
ACTMC1Gm
ACTSCAGm
ACTSmA1C
AA1CGATY
AA1C1ATY
GAGAC11A
1AGRmGGA
1A1AC11A
AGGCT1GA
AG1CTGGA
AGGmTGGA
1AGCCGAK
1TCATH
1TmATH
GTmATH
AA1m1ATT
AA1CGATT
1AAATM
1AMATM
TTDm1
1TAAG
1TAA1
1AMAT
1AMAK
A1RK1GmR
1TCRmmCA
Gm1A1AC
1TmRCCCA
1TCGmmYA
1G1TTMA
CWmG1mTC
mGKmTC
CTmGGmTY
TTRC1
AmTMCG1m
TTAY1
TTRmG
mTGTCGmC
mTGTmGYC
GG1CGAYA
CT1TCGYm
1TAC1W
TTAm1W
TTWC1WAA
TTAmGW
A1C1A1AC
A1C1ARAm
A1mRA1AC
A1CRA1AC
mT1TCGmC
A1mRA1Am
1KCTmGmT

(C)


USF1
mACGTGRm
BmAmGTG
mAC1T2WY
hACGT2RY
CAhGT2WY
A1RKGGCG
AGRKGGC1
CAmGTGMY
hAmGT1RY
CAm1T2A
mAmGT2VY
mAmT1T2VY
mAC1T1Rm
BCAmGT1
mACGTGD
hAmGT2A
mAm1TGWY
mAhGT2VY
mAmGTGD
mACGT1D
BCAC1TG
mACGT1AY
mAC1T1
AAm1TGA
CAmGTGY
AmCWATCA
ACCWATmA
mAmGT1Y
CAmGT2Y
ARCCAATm
ACmWATCA
ARmCAATC
ARCmAATC
CAC1T1RY
AACGT1A
AAC1TGA
mAC1T1AY
mAmGT1
mAmGTG
CAC1TGAY
AGhTGH
1TGTYC
GTG1Bm
CA1mTGH
GTGAhY
CT1T2B
mAC2T1A
1T2AmY
1TGAmY
ATGADTmA
AT1ARTCA
A1RKGGC
AT2ARThA
AT2AHTCA
ATGAHTmA
AT2ADThA
A1RK11CG
AT1AHTCA
AT2ARTmA
A1RKG1C1
AT1ADTmA
ACCAmA1M
AmmAmA1M
AAAhYAmA
AAACYAmA
mAm1T1MY
mAm1TGMY
mAmGT1RC

USF2
2TGACH
2TGAYY
2T1AYY
GT2AYY
GT2AhH
mAmGTGm
mACGT1m
mAmGT1
mAC1T1
2mCAATSA
CCAATmAS
2hhAATSA
GhCAATSR
CAmGTG
CmAATmAS
mCAATCAS
CAm1TG
CAC1TG
ATT1GYY
A1RKG2C2
hACATR
hAhATR
mACATR
ATTG1YY
A2RDG2C2
mACATAmA
CACAGTC1
TGAmGTm
Am1T2
hAm1T1K
CAC1T1K
TGAC1TCA
ARAmTACA
AC1TG
1hCMCGhh
AC2T1
GT1TmmT
2TCAm
A2RT2Rm
A2RK22mG
1T1ACH
1TmAm
mAmAT2B
AYhAT1TG
mAhAT2B
AYCAT1T1
AYhATGT1
mAmGT2B
mAmGTGVY
mAG2TGAm
mAmGTGK
mACMTGAm
mAhGTGVY
mA2GT1AC
AGRT1Rm
mAGGT1RC
mAC2T1VY
mATGT1Am
CAT1T1Ah
CAG2TGAh
AVATG1CG
CACMT2Am
hACRT2R
hAT2T2AC
CA21TGRh
hA2GTGRh
CACGTGm
CACGT2B
CAhRTGR
hACMT2Ah
hAC2T2B
hAmRT2R

(D)


ZFP57

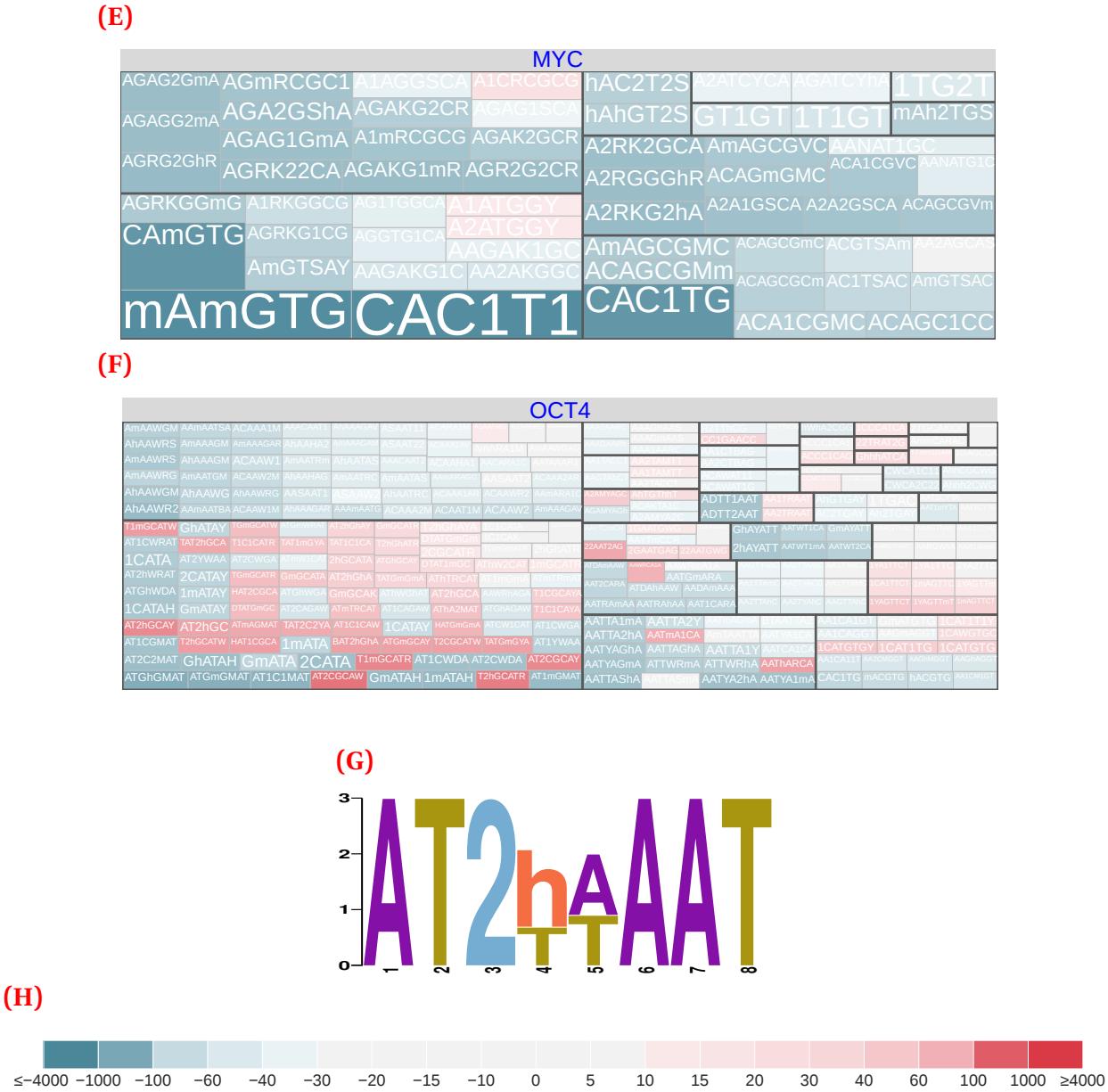


Figure 5. Modified versus unmodified motifs, combining score and cluster information, for selected transcription factors. These plots come from non–spike-in calibrated data, for the 500 bp regions surrounding peak summits. We clip scores above or below ±4000 and plot them at threshold to maximize dynamic range where most scores occur. Some combinations of the displayed hypothesis pairs had multiple data points (for example, multiple identical hypothesis pairs, but for different data sub-types or stringencies). We aggregated these data points by plotting the maximum score. An asymmetric, diverging, colour scale further highlights modification-preferring motifs. We depict a larger selection of transcription factors in Figure S6. **(A)** JUND and JUN. **(B)** FEZF2 and C/EBPβ. **(C)** USF1 and USF2. **(D)** ZFP57. **(E)** c-Myc. **(F)** OCT4. **(G)** The most highly significant and centrally enriched DREME motif, for OCT4 hmC-Seal CUT&RUN in mESCs (replicate 1). See Table 1 and Table 3 for an overview of modified base notation. **(H)** Colour scale.

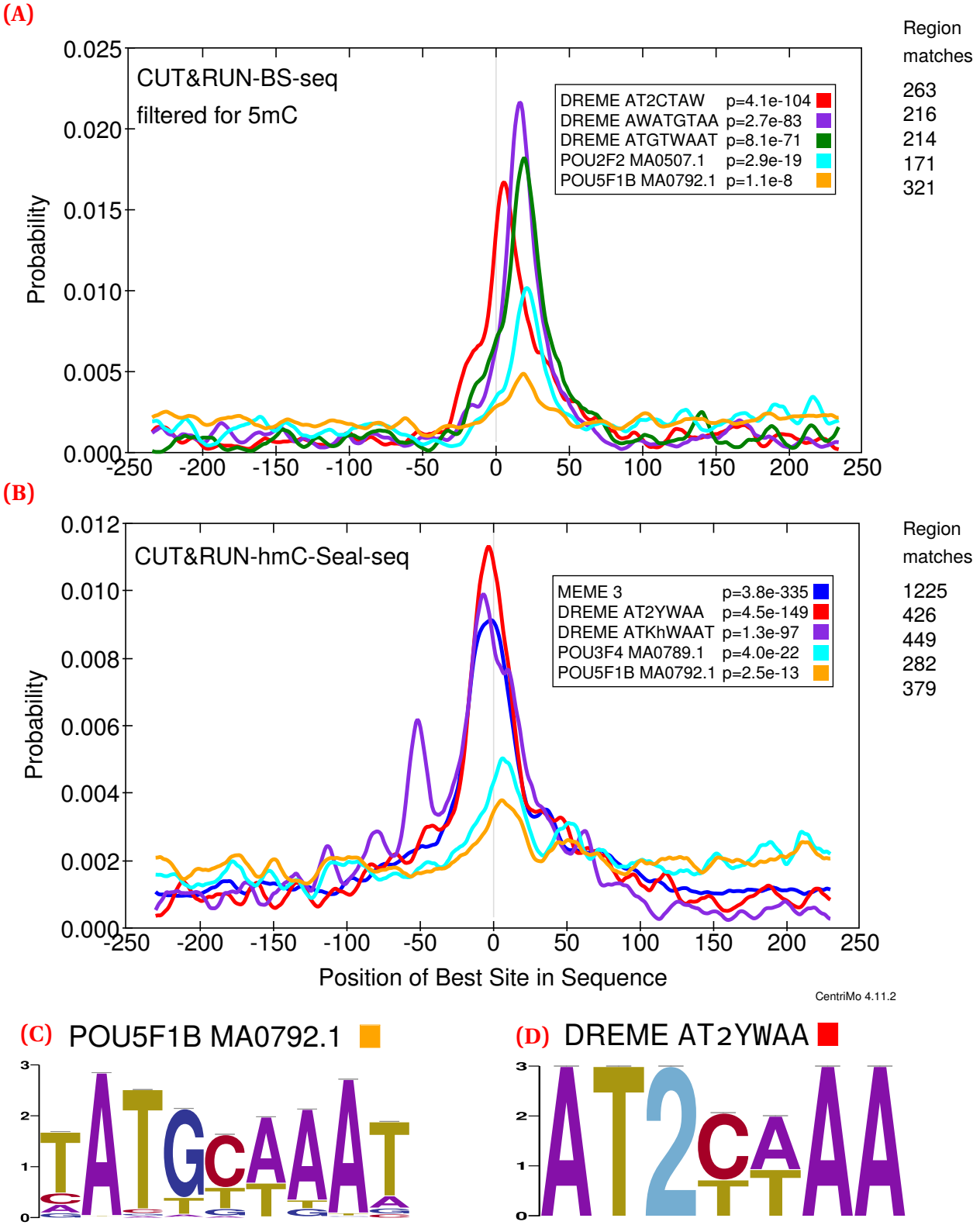

(A)
CUT&RUN-BS-seq
filtered for 5mC
Probability
DREME AT2CTAW p=4.1e-104
DREME AWATGTAA p=2.7e-83
DREME ATGTWAAT p=8.1e-71
POU2F2 MA0507.1 p=2.9e-19
POU5F1B MA0792.1 p=1.1e-8
Region matches
263
216
214
171
321
(B)
CUT&RUN-hmC-Seal-seq
Probability
MEME 3 p=3.8e-335
DREME AT2YWAA p=4.5e-149
DREME ATKhWAAT p=1.3e-97
POU3F4 MA0789.1 p=4.0e-22
POU5F1B MA0792.1 p=2.5e-13
Region matches
1225
426
449
282
379
Position of Best Site in Sequence
CentriMo 4.11.2
(C) POU5F1B MA0792.1
(D) DREME AT2YWAA

Figure 6. CentriMo[62] results for replicate 1 of OCT4 CUT&RUN in mESCs. Motifs include the top three DREME motifs with colour indicating rank: first (red); second (purple), third, where applicable (dark green). Motifs also include the top non-POU5F1 JASPAR motif (cyan), and the JASPAR POU5F1B motif (orange). Both of these motifs come from the JASPAR 2020[67] core vertebrate set. We generated these results using 500 bp regions centred upon the summits of MACS 2[68] peaks generated from those CUT&RUN fragments ≤ 120 bp. We called peaks using IgG controls and without any spike-in calibration (Methods). Listed p-values computed by CentriMo.[62] For consistency, we depict the JASPAR sequence logo using MEME's relative entropy calculation and colouring. **(A)** bisulfite-converted (methylated; 2797 CUT&RUN peaks) sequences. **(B)** hmC-Seal (3974 CUT&RUN peaks) sequences. Also depicts the top MEME[69] *de novo* motif (blue). **(C)** Sequence logo of the POU5F1B motif (`MA0792.1`). **(D)** Sequence logo of the top DREME *de novo* motif.

complementary guanine,[2] which this also captures. We number these symbols in the same order in which the TET enzyme acts on 5mC 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 modification of cytosine bases to either 5mC or 5hmC, but cannot resolve between those two modifications.[6] Even with sufficient sequencing to disambiguate all modifications, we require statistical methods 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 neither hydroxymethylated nor methylated cytosine, `x/7` for a hydroxymethylated or methylated cytosine, and `w/6` for a formylated or carboxylated cytosine (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 (`R`) or pyrimidines (`Y`).[70,71]

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 complementary to a modified nucleobase.

Cytomod: method for creation of an expanded-alphabet genome sequence

Like most epigenomic data, abundance and distribution of cytosine modifications varies by cell-type. Therefore, we require modified genomes for a particular cell-type and would not necessarily expect downstream analyses 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.

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.

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[72] and NumPy[73] 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 possessing conventional bisulfite sequencing data, therefore yielding only x/7 as modified bases. Cytomod additionally produces browser extensible data (BED)[74,75] tracks for each cytosine modification, for viewing in the University of California, Santa Cruz (UCSC)[74] (Figure 7) or Ensembl genome browsers.[76]

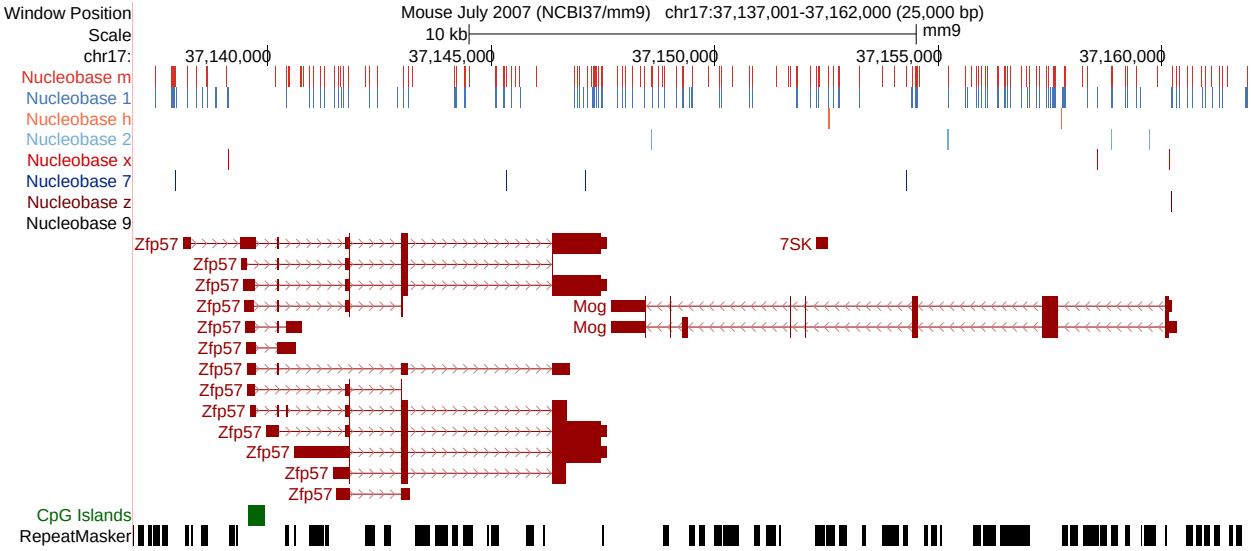


Figure 7. Differential cytosine modification status in naive mouse T cells for a 25 kbp region (within cytoband 17qB1) surrounding *Zfp57* and *Mog*. This UCSC Genome Browser[74] display includes RepeatMasker[77] regions, CpG islands,[78] GENCODE[79] genes, and calls for modified nucleobases h (5hmC), 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).

We subjected the unaligned, paired-end, binary alignment/map (BAM) files output from the sequencer to a standardized internal quality check pipeline. Widely known to work well, we selected Bismark[80] for alignment.[81,82] We use the following processing pipeline: sort the unaligned raw BAM files in name order using Sambamba[83] (version 0.5.4); convert the files to FASTQ,[84] splitting each paired-

end (using BEDTools[85] version 2.23.0 `bamtofastq`); align the FASTA files to NCBI m37/mm9 using Bismark[80] (version 0.14.3), which used Bowtie 2,[86–88] in the default directional mode for a stranded library; sort the output aligned files by position (using Sambamba `sort`); index sorted, aligned, BAMs (SAMtools[89] version 1.2 `index`); convert the processed BAM files into the format required by MethPipe, using `to-mr`; merge sequencing lanes (using direct concatenation of `to-mr` output files) for each specimen (biological replicate), for each sex, and each of WGBS and oxidativeWGBS (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.[23] After alignment, we excluded all random chromosomes. We use modifications called beyond a specified threshold (as described below) as input for Cytomod (with Genomedata[72] version 1.36.dev-r0).

MLML[23] outputs maximum-likelihood estimates of the levels of 5mC, 5hmC, and C, between 0 and 1. It outputs an indicator of the number of conflicts—an estimate of methylation or hydroxymethylation levels falling outside of the confidence interval computed from the input coverage and level. An abundance of conflicts can indicate the presence of non-random error.[23] 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. It additionally maintains the data's strandedness, allowing analyses of hemi-modification.

Mouse expanded-alphabet genome sequences

We used conventional and oxidative WGBS data generated for naive $CD4^+$ T cells, extracted from the spleens of C57BL/6J mice, aged 6 weeks–8 weeks. The dataset authors obtained a fraction enriched in $CD4^+$ T cells, by depletion of non-$CD4^+$ T cells by magnetic labelling, followed by fluorescence-activated cell sorting to get the $CD4^+$, $CD62L^+$, $CD44^{low}$, and $CD25^-$ naive pool of T cells. We previously published these data[59] as part of the BLUEPRINT project[90] (`GSE94674`; `GSE94675`). We analyzed biological replicates separately, 2 of each sex.

For our mouse datasets, we aligned sequencing reads with Bowtie 2[86–88] version 2.2.4. We used MethPipe[91] (development version, commit 3655360) to process the data.

We used our mouse datasets to calibrate our modified base calling thresholds. MLML[23] combines the conventional and oxidative bisulfite sequencing data to yield consistent estimations of cytosine modification state. In our case, with two inputs per mouse run (WGBS and oxWGBS), we obtain values of 0, 1, or 2. 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 8.

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. Namely, if t suffices to call 5mC or 5hmC alone, $m + h \geq t$ ought suffice to call `x/7`.

Human expanded-alphabet genome sequence

Using publicly available ENCODE WGBS data (`ENCFF557TER` and `ENCFF963XLT`), we created K562 (RRID: `CVCL_0004`) modified genome for the GRCh38/hg38 assembly at 0.3 and 0.7 stringencies. WGBS data alone does not differentiate 5mC and 5hmC. As the data cannot differentiate between these states, one might represent them as `x`. Nonetheless, we represent modified bases from the WGBS data as `m`, both for convenience, and because, in most cases, these positions are just methylated. We processed these datasets, as previously described. We aligned human datasets with Bowtie 2[86–88] version 2.2.3 and processed with MethPipe[91] release version 3.4.2.

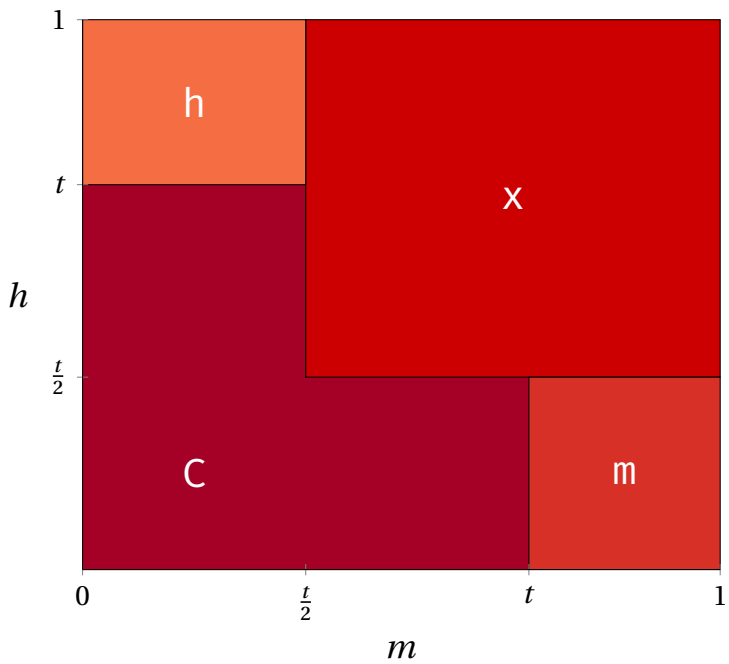


Figure 8. Conditions on the MLML[23] confidence levels of 5mC (m) and 5hmC (h) in relation to a threshold t, that lead to the calling of different modified nucleobases. We call a modification if m or h equal or exceed the threshold. These base assignments assume that MLML had no conflicts for the locus under consideration. If any conflicts occur, we use z as the base, irrespective of the values of m or h. We depict bases for the positive strand only, and complement those occurring on the negative strand, as outlined in Table 1 and Table 2.

Detection of altered transcription factor binding in modified genomic contexts

Following creation of expanded-alphabet genome sequences, we performed transcription factor binding site motif discovery, enrichment and modified-unmodified comparisons. Here, we use mouse assembly NCBI m37/mm9 for all murine analyses, since we wanted to make use of all Mouse ENCODE[92] ChIP-seq data (RRID: CVCL_0188) without re-alignment nor lift-over. Specifically, we used the *Mus musculus* Illumina (San Diego, CA, USA) iGenome[93] packaging of the UCSC NCBI m37/mm9 genome. This assembly excludes all alternative haplotypes as well as all unreliably ordered, but chromosome-associated, sequences (the so-called "random" chromosomes). While ideal for downstream analyses, that assembly does not suffice for aligning data ourselves. Exclusion of these additional pseudo-chromosomes might deleteriously impact alignments, by resulting in the inclusion of spuriously unique reads. Therefore, we used the full UCSC NCBI m37/mm9 build when aligning to a reference sequence. For our human datasets, we used GRCh38/hg38, with all K562 ENCODE datasets.

We used all K562 peak calls, processed as outlined below, from Human ENCODE ChIP-seq data, from preliminary data processed for Karimzadeh et al.[94]. We briefly recapitulate the processing steps here. First, they align the raw reads with Bowtie 2[86–88] (version 2.2.3). Then, they then de-duplicate reads using SAMtools[89] (version 0.1.19) and filter for those with a mapping quality of greater than 10 using option `-bq 10`. Finally, they call peaks without any control, using MACS 2[68] (version 2.1.0) `callpeak` and options `--qvalue 0.001 --format 'BAM' --gsize 'hs' --nomodel --call-summits`.

We updated the MEME Suite[56] to work with custom alphabets, such as our expanded epigenomic alphabet. Additionally, we created the `MEME:Alphabet` Perl module to assist with its internal functionality. Others can use this module to rapidly obtain suitable expanded-alphabet definitions, making it easier to extend older code bases. We incorporated these modifications into MEME Suite version 4.11.0.

We characterize modified transcription factor binding sites using MEME-ChIP.[95] It allows us to rapidly assess the main software outputs of interest: MEME[69] and DREME,[96] both for *de novo* motif elucidation; CentriMo,[62,97] for the assessment of motif centrality; SpaMo,[98] to assess spaced motifs (especially relevant for multipartite motifs); and Find Individual Motif Occurrences (FIMO).[99]

We mainly focus upon CentriMo[62] 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. We generate the PWMs used from MEME-ChIP, by loading the JASPAR 2014[100,101] core vertebrates database, in addition to any elucidated *de novo* motifs from MEME or DREME. CentriMo counts and normalizes the number of sequences at each position of the central peaks to estimate probabilities of central enrichment. CentriMo smooths and plots these estimates, using a one-tailed binomial test to assess the significance of central enrichment.[62]

MEME-ChIP[95] can yield repetitive motifs, without masking of low complexity sequences. Existing masking programs do not support modified genomes, and we accordingly mask the assembly, prior to modification with Cytomod. We use this masking only for downstream motif analyses. We use tandem repeat finder (TRF)[102] (version 4.07b in mice and 4.09 in humans) to mask low complexity sequences. We used the following parameters: `2 5 5 80 10 30 200 -h -m -ngs`, from published TRF parameter optimizations.[103] For version 4.09, to ensure compatibility with GRCh38/hg38 or larger future genomes, we increased the maximum "expected" tandem repeat length to 12 000 000, adding `-l 12`.

We ran MEME-ChIP, against Cytomod genome sequences for regions pertaining to ChIP-seq peaks from transcription factors of interest. For this analysis, we used the published protocol for the command-line analysis of ChIP-seq data.[104] We employed positive controls, in two opposite direc-

tions, to assess the validity of our results. We use c-Myc as the positive control for an unmethylated binding preference.[30,31] For this control, we used ChIP-seq data from a stringent streptavidin-based genome-wide approach with biotin-tagged Myc in mESCs from Krepelova et al.[65] (GSM1171648). Also, we used murine erythroleukemia and CH12.LX Myc Mouse ENCODE samples (ENCFF001YJE and ENCFF001YHU). Conversely, we used both ZFP57 and C/EBPβ 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). Also, we used one replicate of ZFP57 peaks provided by Quenneville et al.[37]. When processing this replicate, we used the same parameters as for our other ZFP57 samples, except for employing default MACS stringency ($q = 0.05$). The reduced peak calling stringency allowed us to ensure sufficient peaks for this older, lower-coverage, dataset. We constructed a ZFP57 BED file using BEDTools[85] (version 2.17.0) to subtract the control influenza hemagglutinin (HA) ChIP-seq (GSM773065) from the target (HA-tagged ZFP57: GSM773066). We retain only target regions with no overlap with any features implicated by the control file, yielding 11 231 of 22 031 features.

Modified binding preferences of ZFP57

We used ZFP57 ChIP-seq data, provided by Strogantsev et al.[38] (GSE55382), to examine the modified binding preferences of that transcription factor. Strogantsev et al.[38] derived these 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 additionally used OCT4 and n-Myc ChIP-seq data, from Yin et al.[44] (GSE94634). For these transcription factors, we used mouse data for three conditions per factor. These conditions consist of a wild-type and two triple-knockout samples, for each of TET1–3 and DNMT1/3A/3B. Both transcription factors have two replicates, per condition. OCT4 has two antibodies, both replicated, across all 3 conditions. Because of pooling of both replicates for one antibody, in the TET triple-knockout condition, this leads to 11 OCT4 samples. The n-Myc came from only a single antibody, resulting in 6 samples. We used the provided IgG samples, as negative controls for peak calling for this dataset. As discussed previously,[44] we also called peaks for matching samples against each of their mouse, rabbit, and goat IgG samples.

We re-processed the ZFP57 data to obtain results for NCBI m37/mm9. We performed this reprocessing similarly to some of the Mouse ENCODE datasets, to maximize consistency for future Mouse ENCODE analyses. We obtained raw FASTQs using Sequence Read Archive (SRA) Toolkit's `fastq-dump`. Then, we aligned the FASTQs using Bowtie[86] (version 1.1.0; `bowtie -v 2 -k 11 -m 10 -t --best --strata`). We sorted and indexed the BAM files using Sambamba.[83] Finally, we called peaks, using the input as the negative enrichment set, using MACS 2[68] (version 2.0.10) `callpeak`, with increased stringency ($q = 0.00001$), with parameters: `--qvalue 0.00001 --format 'BAM' --gsize 'mm'`. This resulted in 90 478 BC8 and 56 142 CB9 peaks.

We used the ChIPQC[105] Bioconductor[106] package to assess the ChIP-seq data quality. We used the two control and two target runs for each of BC8 and CB9. Then, we used `ChIPQC(samples, consensus=TRUE, bCount=TRUE, summits=250, annotation="mm9", blacklist="mm9-blacklist.bed.gz", chromosomes=chromosomes)`. We set the utilized list of mouse chromosomes to only the canonical 19 autosomal and 2 sex chromosomes. Using a blacklist, we filtered out regions that appeared uniquely mappable, but empirically show artificially elevated signal in short-read functional genomics data. We obtained the blacklist file from the NCBI m37/mm9 ENCODE blacklist website (https://sites.google.com/site/anshulkundaje/projects/blacklists).[107] The BC8 data had 13.7% fraction of reads in peaks (FRiP) and the CB9 data had 9.12% FRiP. Additionally, we performed

peak calling at the default $q = 0.05$. This resulted in many more peaks for both BC8 (197 610 peaks; 27.6% FRiP) and CB9 (360 932 peaks; 19.7% FRiP). The CB9 sample had a smaller fraction of overlapping reads in blacklisted regions (RiBL). At the default peak calling stringency, BC8 had an RiBL of 29.7%, while CB9 had only 8.38%. This likely accounts for our improved results with the CB9 replicate (Figure S1).

Additionally, we analyzed three ZFP57 ChIP-seq replicates (100 bp paired-end reads) pertaining to mESCs in pure C57BL/6J mice.[108] We paired each replicate with an identically-conducted ChIP-seq in a corresponding sample, which lacked ZFP57 expression (ZFP57-null controls).

For the pure C57BL/6J data, we used the same protocol as for the hybrid data, except for the following three differences. First, instead of input as the negative control, we used the ZFP57-null ChIP-seq data. We ran Bowtie in paired-end mode (using `-1` and `-2`). Second, we omitted the Bowtie arguments `--best --strata`, which do not work in paired-end mode. Instead, we added `-y --maxbts 800`, the latter of which we set with `--best`'s value, in lieu of the default threshold of 125. Third, we set MACS to paired-end mode (option `-f BAMPE`). This resulted in very few peaks, however, 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, we obtained at most 4496 peaks, with FRiP values of around 4.5%. Nonetheless, we still observed the expected preference for methylated motifs (Figure S1).

Comparing motif modifications, using hypothesis testing

To directly compare various modifications of motifs to their cognate unmodified sequences, we adopted a hypothesis testing approach. One can derive motifs of interest from a *de novo* result that merits further investigation. Often, however, researchers identify motifs of interest using prior expectations of motif binding preferences in 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 tried to minimize as many confounds as possible. We fixed the CentriMo central region width (options `--minreg 99 --maxreg 100`). We also compensated 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 ensured that any motif pairs compared have the same length and similar relative entropies. To do this, we used a larger motif pseudocount for modified motifs (using CentriMo option `--motif-pseudo`). We computed the appropriate pseudocount, as described below, and provided it to `iupac2meme`. We set CentriMo's pseudocount to 0, since we had already applied the appropriate pseudocount to the motif. We seek to normalize the average relative entropies of the PWM columns between two motifs.

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[109]

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

Modification description	Unmodified motif	Modified motif
Full CpG modification to 5mC	$\overset{\mathtt{C}}{\mathtt{N}}\mathtt{G}$	m1
Full CpG modification to 5hmC	$\overset{\mathtt{C}}{\mathtt{N}}\mathtt{G}$	h2
Partial CpG modification to 5hmC	$\overset{\mathtt{C}}{\mathtt{N}}\mathtt{G}$	$\overset{\mathtt{h}}{\mathtt{N}}\mathtt{2}$
Partial CpG hemi-modification to 5hmC	$\overset{\mathtt{C}}{\mathtt{N}}\mathtt{G}$	$\overset{\mathtt{h}}{\mathtt{N}}\mathtt{G}$
Full CpT modification to either 5mC or 5hmC	$\overset{\mathtt{C}}{\mathtt{N}}\mathtt{T}$	xT

Table 3. Illustrative examples of possible changes made to convert unmodified motifs to specific modified counterparts, for downstream hypothesis testing. We use stacked letters like simple sequence logos. At these positions, N represents any base frequencies other than the base being modified. These make up the other positions in the motif's PWM. $\overset{\mathtt{C}}{\mathtt{N}} \rightarrow \mathtt{m}$ indicates that a position containing cytosine is modified by setting all base frequencies other than m to 0 and setting the frequency of m to 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. We portray the second base of each dinucleotide as having a frequency of 1. This second base, however, could also comprise different bases of various frequencies, including the base shown.

For each position, i, in the motif, the MEME Suite adds the pseudocount parameter, α, times the background frequency for a given base, j, at the position: $m'_{ij} = m_{ij} + \alpha b_j$. This omits the effective number of observed sites, which the MEME Suite also accounts for, essentially setting it to 1.

Accordingly, to equalize the relative entropies, we needed only substitute m'_{ij} for each m_{ij} in Equation 1 and then isolate α. In this process, we solve for α, by equating D_{RE} for the unmodified motif with that of the modified motif, substituting as above, while holding α for the unmodified motif constant. If we proceed in this fashion, however, our pseudocount 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 compare differ only in the modified bases, which in this case, comprise only C or G nucleobases, with a motif frequency of 1. Additionally, we set the pseudocount of the unmodified motif to a constant 0.1 (CentriMo's default). Thus, the pseudocount 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:

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

Equation 2 only accounts for a single modification, however, 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 obtained

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

We numerically solved for α in Equation 3 for each modified hypothesis, using `fsolve` from SciPy.[110] Finally, we may have multiple modified positions. We always either hemi-modify or completely modify all modified positions, so the pseudocount is the product of modified positions and the α value from Equation 3.

The pseudocount 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 pseudocount has upon all the other positions of the motif. It does, however, exactly equalize the relative entropies per column (RE/col, as defined by Bailey et al.[109]) of the modified versus unmodified motifs, which suffices to ensure correctly normalized comparisons.

Then, we performed 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 both hemi-methylated and hemi-hydroxymethylated. These six combinations are: `mG`, `C1`, `m1`, `hG`, `C2`, and `h2`. For c-Myc, we constructed modified hypotheses from the standard unmodified 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 occurred 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]

Using CentriMo, we assessed motifs for their centrality within their respective ChIP-seq datasets. Then, we computed the ratio of CentriMo central enrichment p-values, adjusted for multiple testing,[62] for each modified/unmodified motif pair. For numerical precision, we computed 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 mouse WGBS and oxWGBS data, for a grid search of modified base calling thresholds. The levels output by MLML,[23] allowed us to obtain these thresholds. 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, we plotted the ratio of CentriMo p-values across the different thresholds, using Python libraries Seaborn[111] and Pandas.[112] Then we extended this combinatorial hypothesis testing approach, across all JASPAR and *de novo* motifs from our Mouse and Human ENCODE datasets, at our 0.3 and 0.7 selected thresholds.

Assessment of transcription factor familial preferences

We used TFClass[63,64] (downloaded on November 8, 2017) to categorize and group analyzed transcription factors into families and super-families. We used Pronto[113] (version 0.2.1) to parse the TFClass

ontology files.

For transcription factors either not categorized at time of analysis or that yielded inexact matches, we manually assigned them to families and super-families. We performed curation by searching one or more of GeneCards,[114] Genenames.org,[115] UniProt,[116] Gene3D,[117] InterPro,[118] Pfam,[119] SMART,[120] and SUPERFAMILY.[121]

We manually re-named a number of transcription factors in the family assignment to match the names used elsewhere in our data, largely removing hyphens for consistency, creating a group for POLR2A, and adding a number of missing transcription factors. The following factors (with asterisks denoting any suffix and slashes denoting synonymous factors) underwent this manual annotation: RAD21, REC8, SCC1, ZC3H11*, CHD*, NELFE, PAH2, SIN3*, PIAS*, ZMIZ*, KLHL, HCFC*, EP300, TCF12, TIF1*/TRIM24/TRIM28/TRIM33, SMC*, KAT2A/GCN5, and SMARCA4/BRG1.

We aggregated all hypothesis testing results across either mouse or K562 datasets distinctly. Grouped by modification type (`m` or `h`), we aggregated across stringency (0.3 or 0.7), replicate of origin, and unmodified hypothesis. When comparing a modified hypothesis pair to its unmodified counterpart, different replicates of data may produce different scores. In this instance, to aggregate multiple similar hypothesis tests, we took the maximum absolute value score. For each transcription factor, we retained only the most statistically significant ("top-1") or top three most statistically significant ("top-3") hypothesis pairs. We omitted hypothesis pairs that lack statistical significance (p-value ≥ 0.05).

Motif clustering of modified binding preferences

We used Regulatory Sequence Analysis Tools (RSAT) `matrix-clustering`[122] to hierarchically cluster similar motifs. For each transcription factor, we clustered each of its unmodified motifs, alongside their modified counterparts. These motifs matched the set of hypothesis pairs for that factor.

We partitioned each transcription factor's motifs into unmodified-preferring ($score < -\epsilon$), modified-preferring ($score > \epsilon$), and those without any substantive preference ($-\epsilon \leq 0 \leq \epsilon$). Here, we set $\epsilon = 5$, to ignore any near-neutral preferences.

After this, we removed duplicate hypothesis pairs, selecting only those with the scores furthest away from zero. Then, we plotted these clusters, annotated by their score, in a treemap[123] plot. We created this plot using R[124] (version 3.5.1) ggplot2's[125] treemapify[126] and Python (version 2.7.15) Pandas (version 0.22.0)[112] data structures through rpy2[127] (version 2.8.6).

We designed a colour scheme to highlight motifs with strong preferences. To do this, we used white for scores of 80, above or below 0 (namely, from −80–80). This represented an expansion of disregarded motif from the $\epsilon = 5$ threshold used above. We also kept the colour ramp linear within a score of 20 on either side of 0. Outside of this range around the centre, the ramp becomes logarithmic.

To further highlight the rarer and lower scores occurring for motifs preferring to bind in modified contexts, we altered the mid-point of the colour scheme. To do this, we re-centred the colour scheme, shifting it −10, thereby biasing it towards modified contexts, in shades of red. The re-centring offset skews the entire colour scheme toward red hues, including moving the white regions accordingly.

Validation of our OCT4 findings, using CUT&RUN

After finding that OCT4 had a number of both methyl- and hydroxymethyl-preferring motifs, we performed CUT&RUN[57,58] on mESCs, targeting OCT4. We also performed CUT&RUN for a matched IgG, for use as a background during peak calling. We performed CUT&RUN on mESCs, targeting OCT4. We subjected the resultant DNA to 3 workflows: conventional library preparation for sequencing

on an Illumina platform, bisulfite sequencing, and Nano-hmC-Seal-seq.[128] Using a NovaSeq 6000 (Illumina), we sequenced the resulting libraries from the 3 workflows, using a paired-end 2 × 150 bp read configuration (Princess Margaret Genomics Centre, Toronto, ON, Canada).

Cell lines

Using feeder-free conditions, we grew male E14 murine embryonic stem cells[129] on 10 cm plates gelatinized with 0.2 % porcine skin gelatin type A (Sigma, St. Louis, MO, USA) at 37 °C and 5 % CO_2. We cultured cells in N2B27+2i media.[130] Briefly, this media contains DMEM/F12[131] (Sigma) and Neurobasal media (ThermoFisher, Waltham, MA, USA), supplemented with 0.5× B27 (Invitrogen, Waltham, MA, USA), 1× N-2 Supplement, 50 µmol/l 2-mercaptoethanol (ThermoFisher), 2 mmol/l glutamine (ThermoFisher), Leukemia Inhibitory Factor (LIF), 3 µmol/l CHIR99021 glycogen synthase kinase (GSK) inhibitor (p212121, Boston, MA, USA), and 1 µmol/l PD0325091 mitogen-activated protein kinase/extracellular signal-regulated kinase kinase (MEK) inhibitor (p212121). We passaged cells every 48 h using trypsin (Gibco, Waltham, MA, USA) and split them at a ratio of ≈ 1 : 8 with fresh medium. We conducted routine anti-mycoplasma cleaning (LookOut DNA Erase spray, Sigma) and screened cell lines by PCR to confirm no mycoplasma presence.

CUT&RUN assay

We performed CUT&RUN as described elsewhere[132–134] using recombinant Protein A-micrococcal nuclease (pA-MN). Briefly, we extracted nuclei from ≈ 4 500 000 embryonic stem cells using a nuclear extraction buffer comprised of 20 mmol/l 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES)-KOH,[135] pH 7.9, 10 mmol/l KCl, 0.5 mmol/l spermidine, 0.1 % Triton X-100, 20 % glycerol, and freshly added protease inhibitors. We bound the nuclei to 500 µl pre-washed lectin-coated concanavalin A magnetic beads (Polysciences, Warrington, PA, USA). Then, we washed beads in binding buffer (20 mmol/l HEPES-KOH, pH 7.9, 10 mmol/l KCl, 1 mmol/l $CaCl_2$ $MnCl_2$). We pre-blocked immobilized nuclei with blocking buffer (20 mmol/l HEPES, pH 7.5, 150 mmol/l NaCl, 0.5 mmol/l spermidine, 0.1 % bovine serum albumin (BSA), 2 mmol/l EDTA, fresh protease inhibitors). We washed the nuclei once in wash buffer (20 mmol/l HEPES, pH 7.5, 150 mmol/l NaCl, 0.5 mmol/l spermidine, 0.1 % BSA, fresh protease inhibitors). Following this, we incubated nuclei in wash buffer containing primary antibody (anti-Oct4, Diagenode (Denville, NJ, USA) cat. no. C15410305 (lot A2089P) or anti-IgG, Abcam (UK) cat. no. ab37415; RRID: AB_2631996) for 1 h at 4 °C with rotation. Then, we incubated in wash buffer containing recombinant pA-MN for 30 min at 4 °C with rotation.

Using an ice-water bath, we equilibrated samples to 0 °C and added 3 mmol/l $CaCl_2$ to activate pA-MN cleavage. Then, we performed sub-optimal digestion, at 0 °C for 30 min. As described in Step 31 of Skene et al.[58], we intentionally conducted digestion at a temperature lower than optimal, to prevent otherwise unacceptable background cleavage levels.[57,58] We chelated digestion with 2XSTOP+ buffer (200 mmol/l NaCl, 20 mmol/l EDTA, 4 mmol/l ethylene glycol-bis(β-aminoethyl ether)-*N,N,N′,′*-tetraacetic acid (EGTA), 50 µg/ml RNase A, 40 µg/ml glycogen, and 1.5 pg MNase-digested *Saccharomyces cerevisiae* mononucleosome-sized DNA spike-in control).

After RNase A treatment and centrifugation, we released and separated genomic fragments. We digested protein using proteinase K. Finally we purified DNA using phenol:chloroform:isoamyl alcohol extraction, followed by ethanol precipitation.

Library preparation for bisulfite sequencing

We prepared our bisulfite sequencing library using 30 ng of CUT&RUN DNA. We used the Ultra II Library Preparation Kit (New England Biolabs (Canada) (NEB), cat. no. E7645L) following manufacturer's protocol, with some modifications. In brief, after end-repair and A-tailing, we ligated NEBNext methylated adapters for Illumina (NEB, cat. no. E7535) at a final concentration of 0.04 µmol/l onto the DNA, followed by incubation at 20 °C for 20 min. Post adapter incubation, we subjected adapters to USER enzyme digestion at 37 °C for 15 min prior to clean up using AMPure XP Beads (Beckman Coulter, cat. no. A63881).

We bisulfite-converted adapter-ligated CUT&RUN DNA using Zymo Research EZ DNA Methylation Kit (Zymo, Irving, CA, USA, cat. no. D5001) following the alternative protocol for the Infinium Methylation Assay (Illumina). Briefly, we added 5 µl of M-Dilution buffer to purified adapter-ligated DNA, and adjusted total sample volume to 50 µl with sterile molecular grade water. We incubated samples at 37 °C for 15 min, prior to the addition of 100 µl of CT Conversion Reagent. We further incubated samples, prior to purification, at (95 °C for 30 s, 50 °C for 60 min) for 16 cycles, then at 4 °C for at least 10 min, following the manufacturer's protocol. Finally, we eluted samples in 23 µl molecular grade water.

We amplified the bisulfite-converted DNA using 2× HiFi HotStart Uracil+ ReadyMix (KAPA, Wilmington, MA, USA, cat. no. KK2801), and unique dual index primers (NEB, cat. no. E6440S) in a final volume of 50 µl. We performed this using the following PCR program: 98 °C for 45 s, followed by 17 cycles of: 98 °C for 15 s, 65 °C for 30 s, 72 °C for 30 s, and final extension at 72 °C for 60 s. We purified and dual size selected amplified libraries using AMPure (Beckman Coulter, ON, Canada) XP Beads at 0.6× to 1.0× bead ratio, eluted in a volume of 20 µl.

Library preparation for hmC-Seal sequencing

We performed library preparation using the NEB Ultra II Library Preparation Kit (NEB, cat. no. E7645L) on 30 ng CUT&RUN DNA, as per the manufacturer's protocol, with the below modifications. In brief, after end-repair and A-tailing, we purified adapter ligated DNA using AMPure XP Beads at a 0.9× ratio and eluted in 11.5 µl sterile water.

We added three spike-in DNA controls to the adapter ligated DNA to assess specific enrichment of modified DNA fragments. Controls consisted of 0.2 ng/µl working stocks of unmethylated and methylated *Arabidopsis* DNA spike-in controls from the Diagenode DNA methylation control package (cat. no. C02040012). They also included the 5hmC spike-in control DNA (amplified from the *APC* promoter) from the Active Motif Methylated DNA standard kit (Active Motif, cat. no. 55008). We combined 0.3 ng of each spike-in DNA in a final volume of 4.5 µl per experimental sample. We mixed the adapter ligated DNA with the spike-in DNA mix. Then we aliquoted 1.6 µl of this mix into a separate PCR tube and stored it at −20 °C, as an input control.

We 5hmC-glucosylated the remaining 14.4 µl CUT&RUN DNA mixed with spike-in controls, as previously described,[128] with the below modifications. Briefly, we 1 : 1 diluted a 3 µmol/l stock of uracil diphosphate (UDP)-azide-glucose (Active Motif, cat. no. 55020) in 1× phosphate-buffered saline (PBS) to establish a working stock of 1.5 µmol/l for Mastermix preparation. We prepared a 20 µl glucosylation Mastermix per experimental sample consisting of the following: 14.4 µl CUT&RUN DNA mixed with spike-in controls, 50 µmol/l HEPES (pH 8.0), 25 mmol/l $MgCl_2$, 0.1 mmol/l UDP-azide-glucose and 1 U of T4 Phage β-glucosyltransferase (NEB, cat. no. M0357L). We incubated the mix for 1 h at 37 °C, to promote glucosylation.

Then, we performed biotinylation of azide-labelled 5hmC residues of the glucosylated DNA fragments. In sterile water, we prepared 20 mmol/l dibenzocyclooctyne-PEG4-biotin conjugate (Bio-

science, cat. no. CLK-A105P4-10) and stored it in one-time use aliquots at −20 °C, to avoid freeze-thaw. We mixed 20 µl of glucosylated DNA with 1.8 µmol/l dibenzocyclooctyne-PEG4-biotin in a final reaction volume of 22 µl, then incubated 2 h at 37 °C to promote biotinylation. Then, we prepared MicroSpin P-30 Gel Columns (Bio-Rad, Hercules, CA, USA, cat. no. 7326223), following the manufacturer's protocol, and used them to purify total DNA fragments from reaction components. Briefly, we loaded the sample onto the column, then centrifuged 4 min at 1000 × g to elute purified DNA sample in 22 µl of Tris buffer.

To specifically capture biotinylated 5hmC DNA fragments, we prepared 2× binding and washing (B&W) buffer (10 mmol/l Tris-HCl, 10 mmol/l EDTA, 2 mol/l NaCl). Using 20 µg of MyOne Streptavidin C1 Dynabeads (ThermoFisher, cat. no. 65001), we re-suspended in 0.2 ml of 1× B&W buffer per experimental sample to wash beads. We subjected beads to 3 total washes, then re-suspended to a final volume of 22 µl per sample in 2× B&W buffer. We added 22 µl of purified total DNA fragments to 22 µl of washed beads, then incubated 15 min under gentle rotation to promote streptavidin-biotin binding.

To isolate beads containing streptavidin-bound biotinylated DNA fragments, we incubated them on magnet for 3 min. Then, we washed the beads 3 times with 1× B&W buffer to remove non-biotinylated DNA fragments lacking 5hmC. Finally, we re-suspended the beads in 50 µl of low TE buffer.

We conducted quantitative PCR (qPCR) to compare enrichment of 5hmC spike-in after biotin enrichment relative to input sample stored earlier. qPCR in the form of a 10 µl reaction, consisted of 1× SYBR Fast qPCR Mastermix (KAPA, cat. no. KK4601), 1 µl of template DNA, and primers at a final concentration of 0.3 µmol/l. We set up different reactions for each primer set to detect each spike-in control DNA separately. We used template-specific forward and reverse primers. For 5hmC, we quantified spike-in DNA fragment from the Active Motif Methylated DNA standard kit. We also quantified the methylated or unmethylated *Arabidopsis* DNA spike-in controls from the Diagenode DNA Methylation Control package kit. We amplified with the following PCR program: 98 °C for 30 s, followed by 40 cycles of 98 °C for 15 s and 60 °C for 15 s (with image capture), ending with melt curve analysis.

To generate bead-free template for library DNA amplification, we established a PCR reaction mix containing 0.3 µmol/l of unique dual index primers (NEB, cat. no. E6440S), 1× NEBUltra II Q5 MM, and DNA/bead template for a final volume of 100 µl per sample. We split samples into 2× 50 µl reactions and amplified using the following PCR program: 98 °C for 30 s, followed by 5 cycles of 98 °C for 10 s, 60 °C for 75 s, ending with a hold at 4 °C. Then, we transferred reaction tubes to a magnetic rack and transferred bead-free supernatant to new PCR tubes. To amplify DNA libraries for a maximum of 16 cycles total (including initial 5 cycles), we used the same PCR conditions for the bead-free template. Then, we dual size selected DNA libraries using AMPure XP beads at 0.7× to 1.0× ratio, as described in the library preparation for bisulfite sequencing.

Sample sequencing

We performed library preparation of 5 ng of CUT&RUN DNA, following the NEB Ultra II Library Preparation Kit (cat. no. E7645L) manufacturer's protocol. We used different NEB dual indices for each sample (Table 4). We sequenced all libraries on a NovaSeq 6000 sequencing system using a SP flow cell run in standard mode, with paired-end 2 × 150 bp read length configuration. This allowed us to obtain the desired number of reads per sample (Table 4).

Sample name	Technique	i7 Index	i5 Index	Target reads (millions of single reads)
OCT-4_Pool_1	CUT&RUN	TCTAGGAG	AGGTCACT	40
OCT-4_Pool_2	CUT&RUN	TGCGTAAC	GATAGGCT	40
IgG_Pool_1	CUT&RUN	CTTGCTAG	GGAGATGA	20
OCT-4_Pool_1_BS	CUT&RUN-BS	AGCGAGAT	GATACTGG	180
OCT-4_Pool_2_BS	CUT&RUN-BS	TATGGCAC	TCTCGCAA	180
IgG_Pool_1_BS	CUT&RUN-BS	GAATCACC	CTTCGTTC	40
Oct-4_hmeseal_rep_1	CUT&RUN-5hmC	GTAAGGTG	GCAATTCG	150
Oct-4_hmeseal_rep_2	CUT&RUN-5hmC	CGAGAGAA	TCTCTTCC	150

Table 4. CUT&RUN samples used in our experiments, with their sequencing technique, indices, and target read details. Target reads represent the number of single-end equivalent Illumina passing-filter read estimates we sought to obtain.

Data processing

We performed base calls using Real-Time Analysis (RTA) (version 3.4.4). Using bcl2fastq (version 2.20), we converted Binary Base Call (BCL) files to FASTQ files.

We processed the CUT&RUN sequences as follows. Before alignment, we trimmed adapter sequences with fastp (version 0.19.4).[136] We assessed sequencing data quality using FastQC (version 0.11.8),[137] Picard[138] (version 2.6.0) `CollectInsertSizeMetrics`, QualiMap[139] (version 2.2) `bamqc`, Preseq[140] (version 2.0.0) `bound_pop` and `lc_extrap`, DeepTools[141] (version 3.1.3), and MultiQC[142] (version 1.7). For tools requiring Java, we used Java SE 8 Update 45. For tools requiring, Python we used version 2.7.12, except as otherwise noted.

We aligned reads to GRCm38/mm10 with Bismark[80] (version 0.22.3; for 5mC or 5hmC sequences). Bismark used Bowtie 2[86–88] (version 2.4.1; and for conventional sequences), SAMtools[89] (version 1.10), and BEDTools[85] (version 2.29.2). We used Bismark's default parameters, save those controlling output destinations and use of multiple cores, and parameters passed to Bowtie 2, as described below.

We used Bowtie 2 parameters as recommended,[143] excepting increasing alignment sensitivity, and specifying implied or default parameters. Therefore, we used the parameters `-D 20 -R 6 -N 1 -L 18 -i S,1,0.25` for increased sensitivity, slightly more so than the `--very-sensitive-local` preset. We used `-I 10` for a minimum fragment length of 10 bp and `-X 700` for a maximum fragment length of 700 bp|, as recommended.[58,143] This range of fragment lengths included those we selected for during library preparation (30 bp–280 bp). We also used the parameters `--local --phred33 --no-unal --no-discordant --no-mixed`. For alignments used for calculating the spike-in coefficient, we did not permit dovetailing (`--no-dovetail`) nor overlaps (`--no-overlap`), as recommended.[58]

For post-processing, we used Sambamba[83] (version 0.7.1), including marking duplicates. Where applicable, we performed spike-in calibration as described by Meers et al.[143].

For our final OCT4 results, we did not use our *Saccharomyces cerevisiae* spike-in calibrated data. In the unmodified context, the spike-in calibrated data made little difference. In the modified context, insufficient modified bases in the spike-in probably prevented us from properly calibrating.

We called peak summits using MACS 2[68] (version 2.1.2). We ensured that the input only included reads with insert sizes ≤120 bp, as recommended,[58,143] by using DeepTools[141] (version 3.1.3) `alignmentSieve`.

For data not calibrated with spike-in, we used MACS 2 `callpeak`, specifying treatment and control

inputs and outputs as usual. We used the additional MACS 2 parameters, `--buffer-size 1000000 --format 'BAMPE' --gsize 'mm' --qvalue 0.05 --call-summits --bdg --SPMR`.

For spike-in calibrated data, we used advanced MACS sub-commands, constructed to yield a peak calling scheme that worked well for CUT&RUN datasets. Specifically, we used `pileup` on BAMPE input, then `bdgopt` to multiply by the scaling factor defined by the spike-in calibration. Then, we added a pseudocount of 1.0 to mimic the default workflow, using `bdgcmp`, specifying `--pseudocount '0.0' --method 'qpois'`, followed by `bdgpeakcall`, with `--cutoff` $-\ln(0.05)/\ln(10)$. This cutoff parameter represented the usual q-value cutoff of 0.05 converted to $-\log_{10}$ space.

For bisulfite-converted data, we extracted and called peaks only upon methylated reads (filtering through Sambamba using Bismark's added "XM:Z:" tag). We regarded all hmC-Seal-seq reads as completely hydroxymethylated.

Finally, we used MEME-ChIP (version 4.11.2.1, with Perl version 5.18.1),[95] with DREME,[96] as previously described, on a TRF-masked genome (same parameters as before, using version 4.9). For this particular processing, we used SAMtools[89] version 1.3.1 and BEDTools[85] version 2.27.1.

Discussion

We have added expanded alphabet capabilities to the widely-used MEME Suite,[56] a set of software tools for the sequence-based analysis of motifs. This included extending several of its core tools, including: MEME,[69] DREME,[96] and CentriMo,[62] used in a unified pipeline through MEME-ChIP.[95] We undertook further extension of all downstream analysis tools and pipelines, and most of the MEME Suite[56] now supports arbitrary alphabets. We have processed maps of cytosine modifications in naive *ex vivo* mouse CD4$^+$ T cells to yield genome sequences that use the expanded alphabet. Then, we 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 3 and Figure 4)and ZFP57 (Figure S4). We encountered instances, however, in which *de novo* CentriMo analyses did not make obvious the expected motif binding preference. This occurred for c-Myc 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 primarily arises from such analyses not having any means of integrating modified and unmodified motifs. Our *de novo* analyses cannot compensate for the large differences in modified versus unmodified background frequencies. *De novo* elucidation involves some form of optimization or heuristic selection of sites—an inherently variable process. Modified motifs have particular characteristics that differ from most unmodified motifs. Most notably, they necessarily differ 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 usually appear within local sequence of highly similar properties to the motifs themselves.[144] Accordingly, without specifically accounting for these confounds, modified motifs can get lost within a background of irrelevant unmodified motifs or one might not find comparable sets of motifs. Also, modified motifs that a *de novo* analysis finds might not be comparable to any unmodified counterpart. This may arise from the potential pairs of motifs having substantially different lengths, often with the modified motifs having significantly shorter length. Comparing motifs having sequence

properties that often indicate a poor-quality motif also remains difficult. These properties include repetitious motifs, or off-target motifs, such as zingers.[145] Hypothesis testing, with relative entropy normalization, can mitigate these concerns. Possibly, however, we often simply observe bona fide, but non-canonical, motifs. Non-canonical binding sites have far more abundance and importance than generally appreciated.[146]

Our 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 confidence has proven instrumental in applying this method to a diverse array of ChIP-seq data, towards the elucidation of novel binding preferences.

There exists an inherent trade-off between a lower threshold, yielding more modified loci but potentially introducing false positives, and a higher one, which may prove 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 2) and, to a lesser extent, C/EBPβ (Figure S3).

We found that motifs enriched often for hemi-modified, as opposed to completely-modified binding sites. These hemi-modified motifs often had more central enrichment than those with complete modification of a central CpG dinucleotide (Figure 3; Figure 4). This appears surprising, because numerous *in vitro* experiments have demonstrated that for transcription factors preferring modified DNA, often each modification has an additive effect, resulting in completely modified motifs having greatest affinity.[29,39] The hemi-methylation events we detect may arise from asymmetric binding affinities for 5mC (and 5hmC). ZFP57, for example, has known asymmetric recognition of 5mC, with the negative strand methylation having more importance than the positive strand methylation with respect to the `TGCC`**`G`**`C` motif.[39] Also, there exists evidence of a preference for hemi-methylation of the C/EBP half-site `|GmAA`.[66] This could indicate that the hemi-methylated motifs observed in some of our analyses, especially for C/EBPβ may not represent merely technical artifacts, but represent bona fide preferences, even if the underlying CpG is completely methylated (though it also may not be). This would accord with similar findings in an independent analysis.[50] Further work could determine whether the hemi-methylated motifs arise from technical artifacts (either at the level of the bisulfite sequencing data or the methods used) or whether they reflect an actual biological preference.

There exist few high-quality single-base resolution datasets of 5hmC, 5fC, and 5caC. We had previously attempted analyses using modification data, from assays like methylation DNA immunoprecipitation (MeDIP),[147] that did not employ single-base resolution methods.[148] A lack of single-base resolution makes it 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 single-base resolution datasets use some form of reduced representation approach that enriches CpGs, 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. Accordingly, we recommend avoiding enrichment approaches for use with these methods.

The ChIP-seq data we used did not come from the same cell type as the WGBS and oxWGBS data. While transcription factor binding models created from one cell type are often assumed 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.

We found that USF1 and USF2 appear to have a subset of 5mC- and 5hmC-preferring motifs (Figure 5C). This contradicted previous *in vitro* work,[149] where USF1 uniformly prefers to bind neither 5mC nor 5hmC. Additionally, the *in vitro* work implies that while TCF3 can undergo a conformational shift to bind to 5hmC, USF1 cannot. The *in vitro* work largely derived from structural preferences, however: it "only focused on the most obvious, steric and hydrogen bonding effects …[more] complex methods are required to explain [their] subtler [protein binding microarray] results".[149] Those results included, for USF1, a general inhibition of binding with 5hmC. Nonetheless, some of the C versus 5hmC and 5mC versus 5hmC z-score plots depict a few significant motif pairs with near equal preference for versus against hydroxymethylation.

Our results for USF1 only had a single weakly positive-scoring motif containing a hydroxymethylated base: `AAAhYAmA`. This motif had only a slight preference to bind over its modified counterpart. This preference may instead have arisen primarily from the methylated base near the end of the motif. The low score may simply represent no strong preference or a technical artifact. All our other modified-preferring motifs for USF1 preferred to bind in a methylated context, with most showing only weak preferences. This stood in contrast to the many methylated motifs which instead showed strong preferences for their unmodified counterparts.

USF2, however, has a non-negligible number of motifs which appear able to bind in a hydroxymethylated context. This suggests that, unlike USF1, USF2 may sometimes undergo a similar conformational change to TCF3, to permit binding in this context.

We performed some assessment of transcription factor preferences across different families, but did not endeavour to draw clear conclusions from these analyses. Despite previous findings for specific families, like for basic helix-loop-helix factors,[32–36] most findings in this area are equivocal.[48–50,150]

Two main reasons likely account for widespread equivocal results and militate against attempts to discriminate based upon family. First, the existence of myriad transcription factors with different motif groups that prefer unmethylated versus methylated binding, tends to elude individual, let alone familial, binary categorization. Even closely related factors, like USF1/2, often have different degrees of preferences and variable most-significant motifs. Second, there exists substantial observation bias in the transcription factors for which one can find binding data. Accordingly, we can only observe a limited and biased subset of transcription factors, often particularly depleted in zinc finger factors. This bias may account, at least partially, for the often observed enrichment of unmodified-preferring factors.[28,44] For an unbiased assessment of transcription factors across families, we need data on a less biased set of transcription factors.[44,151,152]

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[99] or conduct pathway analyses with Gene Ontology for MOtifs (GOMO).[153] Alternatively, one can use FIMO results for pathway analyses through tools like Genomic Regions Enrichment of Annotations Tool (GREAT)[154] or biological enrichment of hidden sequence targets (BEHST).[155] For further interpretation of the results, one can use downstream pathway analysis tools, such as Enrichment Map.[156,157] This permits inference of implicated genomic regions and biological pathways, which one can then subject to further analysis.

One can readily extend the approach here to other DNA modifications, since we designed all of our software with this in mind. Technology now allows detection of a number of DNA modifications[22] 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).[158–160] We provide recommendations in Appendix A for the nomenclature of these modified nucleobases, among others. We use these recommendations, as initially outlined in our original preprint of this work,[54] in our database of DNA modifications, DNAmod.[22] The Global Alliance for Genomics and Health (GA4GH)[161] has also adopted these recommendations for use in sequence alignment/map (SAM) and BAM formats.[89]

We provide a framework to readily apply motif analyses on sequences containing DNA modifications. Our consistent reproduction of known transcription factor binding affinities, as well as our validation of our prediction for OCT4, suggests that these methods meaningfully predict the modification sensitivity of transcription factors. Establishment of *de novo* recognition motifs by the oxidative forms of 5mC may reveal why transcription factors bind specific repetitive element loci, as opposed to every genome-wide iteration of the element. Our work provides an initial foundation towards a better understanding of this facet of motif specificity.

It is important to characterize modified binding affinities *in vivo*, as opposed to more abundant *in vitro* approaches, such as high-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX) and DNA affinity purification sequencing (DAP-seq).[42] While the *in vitro* approaches contribute to our improved understanding of the underlying biophysics, only *in vivo* analyses can directly further our understanding of actual binding events that lead to differences in gene expression. Our work contributes to this effort, and bolsters it by providing the first transcription factor CUT&RUN dataset directly assessing unmodified and modified binding states. These data represent a unique type of experiment needed to fully understand the role of methylation and hydroxymethylation in transcription factor binding.

Availability

Cytomod is available at: https://github.com/hoffmangroup/cytomod. Persistent availability is ensured by Zenodo, in which we have deposited the version of our code we used (https://doi.org/10.5281/zenodo.6345378). We also provide additional source code, containing other analysis scripts (https://github.com/hoffmangroup/2022modTFBSs; also archived at https://doi.org/10.5281/zenodo.6347792). All source code is licensed under a GNU General Public License, version 3 (GPLv3), except for CentriMo, which retains its original license. We have additionally archived our full set of scores for every assessed hypothesis pair (https://doi.org/10.5281/zenodo.6345400). We have deposited all CUT&RUN sequencing data and peak calls generated for this work in GEO (`GSE198458`).

Authors' contributions

Conceptualization, M.M.H.; Data Curation, C.V., N.J.W., and H.S.; Formal Analysis, C.V. and T.L.B.; Investigation, C.V. and C.A.I.; Methodology, C.V., T.L.B., and M.M.H.; Software, C.V., J.J., T.L.B.; Visualization, C.V.; Validation, C.V., C.A.I., S.Y.S., S.M.L., and S.J.H.; Writing — Original Draft, C.V., C.A.I., S.M.L, and S.J.H.; Writing — Review & Editing, C.V., C.A.I., M.K.S.-H., S.Y.S., D.J.A., A.C.F.-S., D.D.C., S.J.H., T.L.B., and M.M.H.; Resources, M.K.S.-H., D.J.A., A.C.F.-S., D.D.C., M.M.H.; Funding Acquisition, M.M.H.; Project Administration, C.V. and M.M.H.; Supervision, M.M.H.

Acknowledgements

We thank William Stafford Noble and Charles E. Grant for useful discussions and contributions to the MEME Suite. We thank Mehran Karimzadeh, for providing us with the K562 subset of his processed Human ENCODE ChIP-seq peak calls. We thank Andrew D. Smith, Meng Zhou, Benjamin E. Decato, and Egor Dolzhenko for their work on MethPipe[23,91] and for rapidly working to clarify and address any issues. We thank Jaime Castro-Mondragón and Jacques van Helden for their work on RSAT matrix-clustering[122] and for their assistance with modifications to permit clustering in an expanded-alphabet context. We thank Michael L. Waskom for his visualization work on the Seaborn[111] Python package and for actively providing support. We thank Nicholas Khuu for assistance with library preparation and DNA sequencing. We thank Neil Weingarden for coordination and consultation regarding DNA sequencing. We thank Carl Virtanen and Zhibin Lu for technical assistance. This research was enabled by support provided by: Globus,[162,163] Compute Canada (specifically, WestGrid, SHARCNET, and SciNet[164]), and the Bioinformatics and HPC Core, University Health Network.

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2015-03948 to M.M.H. and Alexander Graham Bell Canada Graduate Scholarships to C.V.), the Canadian Institutes of Health Research (201512MSH-360970 to M.M.H. and Postdoctoral Fellowship MFE-164724 to C.A.I.), the Ontario Ministry of Training, Colleges and Universities (Ontario Graduate Scholarships to C.V.), the Canadian Cancer Society (703827 to M.M.H.), the Ontario Ministry of Research, Innovation and Science (ER-15-11-223 to M.M.H.), 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.), the Princess Margaret Cancer Foundation, the Chilean National Agency for Research and Development, ANID (CONICYT/FONDECYT/REGULAR No. 1171004 to M.K.S.-H.), the BLUEPRINT project[90] (HEALTH-F5-2011-282510 to A.C.F.-S. and D.J.A.), the Wellcome Trust (WT095606RR to A.C.F.-S.), the Medical Research Council, United Kingdom (MR/J001597/1 to A.C.F.-S.), and the National Institutes of Health (R35GM133732 to S.J.H. and R01GM103544 to T.L.B.).

Conflict of interest statement. N.J.W. is an inventor on patent applications for technologies that measure and analyze DNA modifications, filed by Cambridge Epigenetix Ltd., for which he also holds stock options. S.Y.S., D.D.De C., and M.M.H. are inventors on patent applications related to cell-free DNA methylation analysis technologies, licensed to Adela. D.D.De C. is co-founder of, serves in a leadership role at, and owns equity in Adela.

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,** 1794–96 (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,** 472–79 (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,** 480–84 (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. 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).

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. 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).

17. 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).

18. 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,** 555–57 (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* **22,** 44–49 (2015).

20. 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).

21. 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).

22. Sood, A. J., Viner, C. & Hoffman, M. M. DNAmod: the DNA modification database. *Journal of Cheminformatics* **11,** 30 (2019).

23. 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).

24. 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).

25. 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).

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

27. 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).

28. Zhu, H., Wang, G. & Qian, J. Transcription factors as readers and effectors of DNA methylation. *Nature Reviews Genetics* **17,** 551–65 (2016).

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,** 186–89 (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., Wells, J., Gutman, J., Bartley, S. M. & Farnham, P. J. c-Myc target gene specificity is determined by a post-DNA binding mechanism. *Proceedings of The National Academy of Sciences of The United States of America* **95,** 13887–92 (1998).

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. 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).

41. 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).

42. O'Malley, R. C., Huang, S.-s. C., Song, L., Lewsey, M. G., Bartlett, A., Nery, J. R., Galli, M., Gallavotti, A. & Ecker, J. R. Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell* **165,** 1280–92 (2016).

43. Blattler, A. & Farnham, P. J. Cross-talk between site-specific transcription factors and DNA methylation states. *The Journal of Biological Chemistry* **288,** 34287–94 (2013).

44. Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P. K., Kivioja, T., Dave, K., Zhong, F., Nitta, K. R., Taipale, M., Popov, A., Ginno, P. A., Domcke, S., Yan, J., Schübeler, D., Vinson, C. & Taipale, J. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356,** eaaj2239 (2017).

45. Breiman, L. Random forests. *Machine Learning* **45,** 5–32 (2001).

46. Denisko, D. & Hoffman, M. M. Classification and interaction in random forests. *Proceedings of The National Academy of Sciences of The United States of America* **115,** 1690–92 (2018).

47. 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).

48. Xuan Lin, Q. X., Sian, S., An, O., Thieffry, D., Jha, S. & Benoukraf, T. MethMotif: an integrative cell specific database of transcription factor binding motifs coupled with DNA methylation profiles. *Nucleic Acids Research* **47,** D145–D154 (2019).

49. Grau, J., Schmidt, F. & Schulz, M. H. Widespread effects of DNA methylation and intra-motif dependencies revealed by novel transcription factor binding models. *bioRxiv*:348193.

50. Song, G., Wang, G., Luo, X., Cheng, Y., Song, Q., Wan, J., Moore, C., Song, H., Jin, P., Qian, J. & Zhu, H. An all-to-all approach to the identification of sequence-specific readers for epigenetic DNA modifications on cytosine. *Nature Communications* **12,** 795 (2021).

51. Hernandez-Corchado, A. & Najafabadi, H. S. Toward a base-resolution panorama of the in vivo impact of cytosine methylation on transcription factor binding. *Genome Biology* **23,** 151 (2022).

52. Henry, A. A. & Romesberg, F. E. Beyond A, C, G and T: augmenting nature's alphabet. *Current Opinion in Chemical Biology* **7,** 727–33 (2003).

53. Ngo, V., Wang, M. & Wang, W. Finding *de novo* methylated DNA motifs. *bioRxiv*:043810. https://www.biorxiv.org/content/10.1101/043810v1.

54. Viner, C., Johnson, J., Walker, N., Shi, H., Sjöberg, M., Adams, D. J., Ferguson-Smith, A. C., Bailey, T. L. & Hoffman, M. M. Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet. *bioRxiv*:043794. https://www.biorxiv.org/content/10.1101/043794v1.

55. Ngo, V., Wang, M. & Wang, W. Finding *de novo* methylated DNA motifs. *Bioinformatics* **35** (ed Birol, I.) 3287–93 (2019).

56. 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–W208 (2009).

57. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *ELife* **6** (ed Reinberg, D.) e21856 (2017).

58. Skene, P. J., Henikoff, J. G. & Henikoff, S. Targeted *in situ* genome-wide profiling with high efficiency for low cell numbers. *Nature Protocols* **13,** 1006–19 (2018).

59. Kazachenka, A., Bertozzi, T. M., Sjoberg-Herrera, M. K., Walker, N., Gardner, J., Gunning, R., Pahita, E., Adams, S., Adams, D. & Ferguson-Smith, A. C. Identification, characterization, and heritability of murine metastable epialleles: implications for non-genetic inheritance. *Cell* **175,** 1259–71 (2018).

60. Lozzio, C. & Lozzio, B. Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. *Blood* **45,** 321–34 (1975).

61. Andersson, L. C., Nilsson, K. & Gahmberg, C. G. K562—a human erythroleukemic cell line. *International Journal of Cancer* **23,** 143–47 (1979).

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

63. Wingender, E., Schoeps, T., Haubrock, M., Krull, M. & Dönitz, J. TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Research* **46,** D343–D347 (2018).

64. Wingender, E., Schoeps, T. & Dönitz, J. TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Research* **41,** D165–D170 (2013).

65. 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).

66. Syed, K. S., He, X., Tillo, D., Wang, J., Durell, S. R. & Vinson, C. 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) enhance the DNA binding of CREB1 to the C/EBP half-site tetranucleotide GCAA. *Biochemistry* **55,** 6940–48 (2016).

67. Fornes, O., Castro-Mondragon, J. A., Khan, A., van der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D., Santana-Garcia, W., Tan, G., Chèneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W. W. & Mathelier, A. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **48,** D87–D92 (2020).

68. 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).

69. 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.

70. 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).

71. 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).

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

73. 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).

74. 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).

75. Niu, J., Denisko, D. & Hoffman, M. M. *The browser extensible data (BED) format* File Format Standard 1 (Global Alliance for Genomics & Health (GA4GH), 2022), 8.

76. 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–D716 (2016).

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

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

79. 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,** S4 (2006).

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

81. 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).

82. 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).

83. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31,** 2032–34 (2015).

84. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* **38,** 1767–71 (2010).

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

86. 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).

87. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–59 (2012).

88. Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* **35** (ed Hancock, J.) 421–32 (2018).

89. 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).

90. Adams, D., Altucci, L., Antonarakis, S. E., Ballesteros, J., Beck, S., Bird, A., Bock, C., Boehm, B., Campo, E., Caricasole, A., Dahl, F., Dermitzakis, E. T., Enver, T., Esteller, M., Estivill, X., Ferguson-Smith, A., Fitzgibbon, J., Flicek, P., Giehl, C., Graf, T., Grosveld, F., Guigo, R., Gut, I., Helin, K., Jarvius, J., Küppers, R., Lehrach, H., Lengauer, T., Lernmark, Å., Leslie, D., Loeffler, M., Macintyre, E., Mai, A., Martens, J. H., Minucci, S., Ouwehand, W. H., Pelicci, P. G., Pendeville, H., Porse, B., Rakyan, V., Reik, W., Schrappe, M., Schübeler, D., Seifert, M., Siebert, R., Simmons, D., Soranzo, N., Spicuglia, S., Stratton, M., Stunnenberg, H. G., Tanay, A., Torrents, D., Valencia, A., Vellenga, E., Vingron, M., Walter, J. & Willcocks, S. BLUEPRINT to decode the epigenetic signature written in blood. *Nature Biotechnology* **30,** 224–26 (2012).

91. 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).

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

93. Illumina. iGenomes. https://support.illumina.com/sequencing/sequencing_software/igenome.html.

94. Karimzadeh, M. & Hoffman, M. M. Virtual ChIP-seq: predicting transcription factor binding by learning from the transcriptome. *Genome Biology* **23,** 126 (2022).

95. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27,** 1696–97 (2011).

96. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27,** 1653–59 (2011).

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

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

99. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27,** 1017–18 (2011).

100. 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–D94 (2004).

101. 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–D147 (2014).

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

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

104. 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).

105. 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** (2014).

106. Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., Pagès, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L. & Morgan, M. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods* **12,** 115–21 (2015).

107. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

108. 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 the KRAB zinc finger protein ZFP57. *bioRxiv*:611400.

109. 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, 1994).

110. Jones, E., Oliphant, T., Peterson, P., et al. SciPy: open source scientific tools for Python.

111. 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., Emmett, K., Milajevs, D., Qalieh, A., Allan, D. & Meyer, K. Seaborn: v0.6.0 (june 2015). 2015.

112. 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.

113. Larralde, M. Pronto: release v0.2.1. 2016.

114. Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., Sirota-Madi, A., Olender, T., Golan, Y., Stelzer, G., Harel, A. & Lancet, D. GeneCards version 3: the human gene integrator. *Database* **2010** (2010).

115. Gray, K. A., Yates, B., Seal, R. L., Wright, M. W. & Bruford, E. A. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Research* **43,** D1079–D1085 (2015).

116. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research* **43,** D204–D212 (2015).

117. Lam, S. D., Dawson, N. L., Das, S., Sillitoe, I., Ashford, P., Lee, D., Lehtinen, S., Orengo, C. A. & Lees, J. G. Gene3D: expanding the utility of domain assignments. *Nucleic Acids Research* **44,** D404–D409 (2016).

118. Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., Sangrador-Vegas, A., Scheremetjew, M., Rato, C., Yong, S.-Y., Bateman, A., Punta, M., Attwood, T. K., Sigrist, C. J. A., Redaschi, N., Rivoire, C., Xenarios, I., Kahn, D., Guyot, D., Bork, P., Letunic, I., Gough, J., Oates, M., Haft, D., Huang, H., Natale, D. A., Wu, C. H., Orengo, C., Sillitoe, I., Mi, H., Thomas, P. D. & Finn, R. D. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research* **43,** D213–D221 (2015).

119. Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J. & Bateman, A. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* **44,** D279–D285 (2016).

120. Letunic, I., Doerks, T. & Bork, P. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Research* **43,** D257–D260 (2015).

121. Gough, J., Karplus, K., Hughey, R. & Chothia, C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of Molecular Biology* **313,** 903–19 (2001).

122. Castro-Mondragon, J. A., Jaeger, S., Thieffry, D., Thomas-Chollier, M. & van Helden, J. RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Research* **45,** e119 (2017).

123. Shneiderman, B. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics* **11,** 92–99 (1992).

124. R Core Team. *R: a language and environment for statistical computing* R Foundation for Statistical Computing (Vienna, Austria, 2016).

125. Wickham, H. *ggplot2: elegant graphics for data analysis* (Springer, 2016).

126. Wilkins, D. *treemapify: draw treemaps in ggplot2* R package version 2.4.0 (2017).

127. Gautier, L. rpy2: a simple and efficient access to R from Python. 2018.

128. Han, D., Lu, X., Shih, A. H., Nie, J., You, Q., Xu, M. M., Melnick, A. M., Levine, R. L. & He, C. A highly sensitive and robust method for genome-wide 5hmC profiling of rare cell populations. *Molecular Cell* **63,** 711–19 (2016).

129. Hooper, M., Hardy, K., Handyside, A., Hunter, S. & Monk, M. HPRT-deficient (Lesch–Nyhan) mouse embryos derived from germline colonization by cultured cells. *Nature* **326,** 292–95 (1987).

130. Mulas, C., Kalkan, T., von Meyenn, F., Leitch, H. G., Nichols, J. & Smith, A. Defined conditions for propagation and manipulation of mouse embryonic stem cells. *Development* **146,** dev173146 (2019).

131. Dulbecco, R. & Freeman, G. Plaque production by the polyoma virus. *Virology* **8,** 396–97 (1959).

132. Hainer, S. J., Bošković, A., McCannell, K. N., Rando, O. J. & Fazzio, T. G. Profiling of pluripotency factors in single cells and early embryos. *Cell* **177,** 1319–1329.e11 (2019).

133. Hainer, S. J. & Fazzio, T. G. High-resolution chromatin profiling using CUT&RUN. *Current Protocols in Molecular Biology* **126,** e85 (2019).

134. Patty, B. J. & Hainer, S. J. Transcription factor chromatin profiling genome-wide using uliCUT&RUN in single cells and individual blastocysts. *Nature Protocols* **16,** 2633–66 (2021).

135. Good, N. E., Winget, G. D., Winter, W., Connolly, T. N., Izawa, S. & Singh, R. M. M. Hydrogen ion buffers for biological research. *Biochemistry* **5,** 467–77 (1966).

136. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34,** i884–90 (2018).

137. Andrews, S. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc. 2018.

138. Wysoker, A., Tibbetts, K., Fennell, T., et al. Picard tools. http://broadinstitute.github.io/picard/.

139. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32,** 292–94 (2016).

140. Daley, T. & Smith, A. D. Predicting the molecular complexity of sequencing libraries. *Nature Methods* **10,** 325–27 (2013).

141. Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F. & Manke, T. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* **44,** W160–W165 (2016).

142. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32,** 3047–48 (2016).

143. Meers, M. P., Tenenbaum, D. & Henikoff, S. Peak calling by sparse enrichment analysis for CUT&RUN chromatin profiling. *Epigenetics & Chromatin* **12,** 42 (2019).

144. 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).

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

146. Chumpitaz-Diaz, L., Samee, M. A. H. & Pollard, K. S. Systematic identification of non-canonical transcription factor motifs. *BMC Molecular and Cell Biology* **22,** 44 (2021).

147. 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).

148. Song, C.-X., Yi, C. & He, C. Mapping recently identified nucleotide variants in the genome and transcriptome. *Nature Biotechnology* **30,** 1107–16 (2012).

149. Khund-Sayeed, S., He, X., Holzberg, T., Wang, J., Rajagopal, D., Upadhyay, S., Durell, S. R., Mukherjee, S., Weirauch, M. T., Rose, R. & Vinson, C. 5-hydroxymethylcytosine in E-box motifs ACAT|GTG and ACAC|GTG increases DNA-binding of the B-HLH transcription factor TCF4. *Integrative Biology* **8,** 936–45 (2016).

150. Lin, Q. X. X., Thieffry, D., Jha, S. & Benoukraf, T. TFregulomeR reveals transcription factors' context-specific features and functions. *Nucleic Acids Research* **48,** e10–e10 (2019).

151. Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R. & Weirauch, M. T. The human transcription factors. *Cell* **172,** 650–65 (2018).

152. Najafabadi, H. S., Mnaimneh, S., Schmitges, F. W., Garton, M., Lam, K. N., Yang, A., Albu, M., Weirauch, M. T., Radovani, E., Kim, P. M., Greenblatt, J., Frey, B. J. & Hughes, T. R. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nature Biotechnology* **33,** 555–62 (2015).

153. Buske, F. A., Bodén, M., Bauer, D. C. & Bailey, T. L. Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinformatics* **26,** 860–66 (2010).

154. 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).

155. Chicco, D., Bi, H. S., Reimand, J. & Hoffman, M. M. BEHST: genomic set enrichment analysis enhanced through integration of chromatin long-range interactions. *bioRxiv*:168427.

156. 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).

157. 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).

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

159. 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,** 9270–72 (2015).

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

161. Rehm, H. L. et al. GA4GH: international policies and standards for data sharing across genomic research and healthcare. *Cell Genomics* **1,** 100029 (2021).

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

163. 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).

164. 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,** 12026 (2010).

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

166. Kulikowska, E., Kierdaszuk, B. & Shugar, D. Xanthine, xanthosine and its nucleotides: solution structures of neutral and ionic forms, and relevance to substrate properties in various enzyme systems and metabolic pathways. *Acta Biochimica Polonica* **51,** 493–531 (2004).

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

168. 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).

169. Tange, O. GNU Parallel: the command-line power tool. *;login: The USENIX Magazine* **36,** 42–47 (2011).