Computational models for probing the *in vivo* effect of DNA methylation on transcription factor binding

Aldo Hernández Corchado

Department of Human Genetics
McGill University
Montreal, Quebec
August 2021

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

DEDICATION

This thesis is dedicated to my parents and siblings for their unwavering support.

Thank you for everything.

ABSTRACT

Background: Cytosine methylation, particularly 5-methylcytosine in cytosine-phosphate-guanine (CpG) sites, has long been considered to primarily repress the binding of transcription factors (TFs) *in vivo*. This DNA modification is known to change the local structural features of DNA and, when occurring on binding sites, change the binding affinity of TFs. In contrast to the conventional repressive role of methylation, recent high-throughput *in vitro* studies of TF-DNA interactions have revealed that cytosine methylation has a heterogeneous effect on TF binding, with the direction of this effect depending on the specific TF and the position where methylation appears. Expanding these *in vitro* observations to *in vivo* TF binding preferences, however, is a challenging task, since confounding factors like DNA accessibility and regional DNA methylation make it difficult to isolate the effect of individual CpG sites. As a result, the *in vivo* methylation preferences of most TFs remain uncharacterized.

Methodology: In order to infer the effect of CpG methylation on TF binding *in vivo*, we developed Joint Accessibility-Methylation-Sequence (JAMS) models. JAMS creates quantitative models that connect the DNA accessibility, regional methylation level, sequence, and base-resolution methylation to the strength of the binding signal observed in ChIP-seq of a TF. Furthermore, by jointly modeling both the control and pull-down signal in a ChIP-seq experiment, JAMS is able to isolate the TF-specific effects from background effects, revealing how methylation of specific CpGs within a binding site alters the TF binding affinity *in vivo*.

Key results: Using the transcription factor CTCF as a model, we show that JAMS can quantitatively model the TF binding strength and learn the accessibility-methylation-sequence determinants of TF binding. In addition, JAMS can faithfully recapitulate cell type-specific CTCF binding based on differential accessibility and methylation across cell lines. We show that even in the absence of any change in DNA accessibility, changes in the methylation level of specific CpGs within the CTCF binding site drive its differential binding across cell lines. Systematic application of JAMS to 2368 ChIP-seq experiments covering 260 TFs revealed that 45% of TFs are inhibited by

methylation of their potential binding sites. In contrast, 6% prefer to bind to methylated sites and 1% show mixed effects. The other 48% either do not bind to CpG-containing sequences or are indifferent to CpG methylation. Comparison of these *in vivo* models to *in vitro* data confirmed high precision of the methyl-preferences inferred by JAMS. Finally, among the CpG-binding proteins from the ZF-KRAB family of TFs, we observed a disproportionately high preference for methylated sequences (24%), highlighting the role of CpG methylation in determining the genome-wide binding profiles of the TFs from this family.

RÉSUMÉ

Contexte: La méthylation des cytosines, et en particulier des 5-methylcytosine sur les sites cytosine phosphate-guanine (CpG), a longtemps été considéré comme réprimant principalement la liaison des facteurs de transcription (TF) *in vivo*. Cette modification de l'ADN est connue pour changer les caractéristiques structurelles locales de l'ADN, ainsi que l'affinité des liaisons des TF, lorsqu'elle apparait sur leurs sites de liaison. Contrairement au rôle répressif de la méthylation, beaucoup d'études *in vitro* récentes, sur l'interaction FT-ADN ont révélé que la méthylation des cytosines a un effet hétérogène sur la liaison des FT. Ces études ont notamment montré que la liaison d'un TF dépends du TF lui-même et de la position où apparait la méthylation. L'extrapolation de ces observations *in vitro*, à la préférence de liaison des TF *in vivo*, est cependant difficile à réaliser. En effet, des facteurs comme l'accessibilité de l'ADN, et la méthylation régionale de l'ADN rend difficile l'isolation de l'effet des CpG seuls. En conséquence, l'étude des préférences de méthylation sur la plupart des FT reste inexploré *in vivo*.

Méthode: Dans le but de conclure sur l'effet de la méthylation des CpG sur les liaisons des FT *in vivo*, nous avons développé le modèle JAMS (pour Joint Accessibility-Methylation-Sequence). JAMS crée des modèles quantitatifs mettant en relation des données sur l'accessibilité de l'ADN, le degré de méthylation locale, les séquences, et, la méthylation à l'échelle des bases, avec l'intensité du signal de liaison observée dans le ChIP-sep d'un TF. De plus, en modélisant conjointement le signal pull-down et le signal contrôle dans les expériences ChIP-seq, JAMS est capable d'isoler les effets spécifiques des TF par rapport aux effets du background. Cela montre comment la méthylation de certaines CpG d'un site de liaison modifie l'affinité de liaison du TF *in vivo*.

Résultats: En utilisant le facteur de transcription CTCF comme exemple, nous avons montré que JAMS est capable de modéliser quantitativement les forces de liaison d'un TF et d'intégrer le critère déterminant d'accessibilité des séquences méthylées pour une liaison du TF. De plus, JAMS peut reproduire fidèlement les

liaisons spécifiques du CTCF aux cellules, en se basant sur différents degrés d'accessibilité et la méthylation entre différentes lignées cellulaire. Nous avons montré que même en l'absence de quelques changements dans l'accessibilité de l'ADN, des changements dans le degré de méthylation des CpGs spécifiques dans les sites de liaison du CTFT, amène à une liaison différente entre différentes lignées cellulaires. L'application systémique de JAMS à 2368 expériences ChIP-seq comprenant 260 TF a révélé que 45% des TF sont inhibés par la méthylation sur leurs sites de liaison potentiels. Cependant, 6% préfèrent se lier aux sites méthylés et 1% ont montré des effets mitigés. Le reste des 48% de TF ne se sont pas liés à des séquences contenant des CpG ou sont indifférents à la méthylation des CpG. Les données résultantes de la comparaison de ces modèles *in vivo* aux modèles *in vitro* confirment la haute précision des préférences de méthylations proposées par JAMS. Finalement, parmi les liaisons de la protéine CpG à la famille de TF ZF-KRAB, nous avons observé une préférence disproportionnellement grande des TF pour les séquences méthylé (24%). Cette différence met en évidence le rôle de la méthylation des CpG dans la détermination des motifs de liaisons des TF de cette famille à l'échelle du génome.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

5mC	5-methylcytosine
bHLH	Basic helix-loop-helix
C2H2-ZF	The Cys2His2 zinc-finger protein family
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CpGs	Cytosines followed by guanine residues
DHS	DNase I hypersensitive site
DNA	Deoxyribonucleic acid
DNase-seq	DNase I hypersensitive site sequencing
DNMTs	DNA methyltransferases
EMSA	Electrophoretic mobility shift assay
ENCODE	ENCyclopedia Of DNA Elements
EREs	Endogenous repeat elements
FDR	False discovery rate
GEO	Gene Expression Omnibus
H3K4	Histone 3 at lysine 4
HDACs	Histone deacetylases
HDMTs	Histone demethylases
HMTs	Histone methyltransferases
JAMS	Joint Accessibility-Methylation-Sequence
KRAB-ZFPs	Krüppel-associated box domain zinc finger proteins
LFC	Log fold change
MBD	Methyl-CpG-binding domain
PBM	Protein-Binding Microarrays
PCR	Polymerase chain reaction
PFM	Position frequency matrix
PPM	Position probability matrix
PWM	Position weight matrices
RCADE	Recognition Code-Assisted Discovery of regulatory Elements

RF	Random forests
SELEX	Systematic evolution of ligands by exponential enrichment
SEM	Standard error of mean
SMiLE-seq	Selective microfluidics-based ligand enrichment followed by sequencing
TCGA	The Cancer Genome Atlas
TET	Ten-Eleven-Translocation
TFBS	Transcription factor binding sites
TFs	Transcription factors
WGBS	Whole genome bisulfite sequencing
ZF	Zinc finger

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to acknowledge **Dr. Hamed Shateri Najafabadi** for his constant support and mentoring during the duration of this project. His in-depth feedback on analysis and written works helped me become a better scientist.

I would like to thank **Dr. Senthilkumar Kailasam** for his guidance, especially during the start of the project.

I would also like to thank my supervisory committee, **Profs. Guillaume Bourque** and **Yasser Riazalhosseini** for their comments and feedback on the project.

Notes of sincere gratitude to:

The lab members of the Computational and Statistical Genomics Lab for the discussion, and comments, especially to **Ariel**, **Gabrielle**, **Rached**, and **Pubudu**.

Marie for translating the abstract to French.

My **father**, **mother**, **sister** and **brother** for their constant support and giving me every opportunity I could have asked for.

Mitacs Globalink Graduate Fellowship for financial support.

FORMAT OF THE THESIS

This thesis was prepared in adherence to the traditional thesis format outlined by McGill University's Faculty of Graduate and Postdoctoral Studies. This thesis is composed of six chapters. Chapter 1 is a comprehensive review of the literature relevant to this thesis. Chapter 2 is a presentation of the materials and methods. Chapter 3 is a presentation of results, which is in preparation for submission to a peer-reviewed journal. Chapter 4 is a discussion of the results presented in this thesis, while Chapter 5 is a conclusion discussing future research directions. Chapter 6 includes all the references of the thesis. Finally, appendices include supplementary figures, and tables, and copyright clearances.

CONTRIBUTION OF AUTHORS

Detailed contributions are as follows:

Aldo Hernández Corchado (A.H.C.) analyzed the data, developed figures and wrote the thesis.

Hamed S. Najafabadi (H.S.N.) and A.H.C. developed the computational and statistical methods.

H.S.N. conceived and directed the study and edited the thesis and figures.

CHAPTER 1. INTRODUCTION

Transcription factors (TFs) are key regulators of gene expression. Each TF usually recognizes a specific sequence motif; however, TF binding is affected by several other variables as well. One of these variables is DNA methylation, which has traditionally been viewed as having a repressive effect on TF binding [1]. However, this traditional view is gradually changing, as more examples are reported of TFs that bind to methylated sequences. These include studies that have reported increased binding of specific TFs to methylated DNA *in vitro* [2], in addition to reports indicating that, for some TFs, a large fraction of their *in vivo* binding sites is highly methylated [3,4]. While it is tempting to view these anecdotal cases as exceptions rather than a general trend, a recent systematic analysis of TF CpG methylation preferences revealed that, in fact, a large fraction of TFs may bind to methylated CpGs *in vitro*. Based on this study, the effect of methylation is dependent on its position in the binding site, and is heterogeneous within and across TF families [5]. While this study provides *in vitro* evidence for widespread recognition of methylated CpGs by TFs, a comparable systematic analysis of *in vivo* methylation preferences of TFs is still lacking. This is primarily because observing the specific *in vivo* effect of intra-motif CpG methylation is confounded by binding site-specific factors such as DNA accessibility, regional methylation level, and binding site sequence [6–8]. Experimental control of these confounding factors is complicated and resource-exhaustive [9–11], highlighting the need for computational methods to untangle, from these confounding variables, the base-resolution relationship between TF binding affinity and intra-motif CpG methylation.

In this study, we introduce Joint Accessibility-Methylation-Sequence (JAMS) models, a statistical framework for deconvolving the individual contribution of various factors, including intra-motif CpG methylation, on the *in vivo* strength of TF binding as observed by ChIP-seq. We show that JAMS models are reproducible and generalizable, can capture known CpG methyl-preferences of TFs, and can even predict differential TF

binding across cell lines based on changes in intra-motif CpG methylation. Finally, we apply JAMS to a large compendium of ChIP-seq experiments to systematically explore the CpG methylation preferences of TFs across different families.

In this work, I use the term *in vitro* to refer to experimental conditions where naked DNA is tested for its ability to bind a specific TF outside the cell (e.g., SELEX, PBM, and SmiLE-seq experiments). DNA in these experiments is usually synthetic, with a random sequence [5,12,13]. On the other hand, I use the term *in vivo* for experiments performed in cell lines and tissues (i.e. ChIP-seq), where the resulting signal can be confounded by different DNA-binding proteins (e.g., other TFs or MBD proteins) and chromatin state [14].

1.1 Transcription factors

Transcription factors (TFs) are proteins that bind to DNA in a sequence-specific manner and regulate the transcription [15]. TFs play a major role in regulating gene expression through various mechanisms, which range from recruiting other activating/repressive proteins to simply obstructing the binding of other factors to DNA [16]. Most TFs in eukaryotes are believed to exert their function through recruiting cofactors, i.e. proteins or protein complexes that act as activators or repressors of transcription [17]. However, other TFs may have other mechanisms of action, some of which are shown in **Fig. 1** [18].

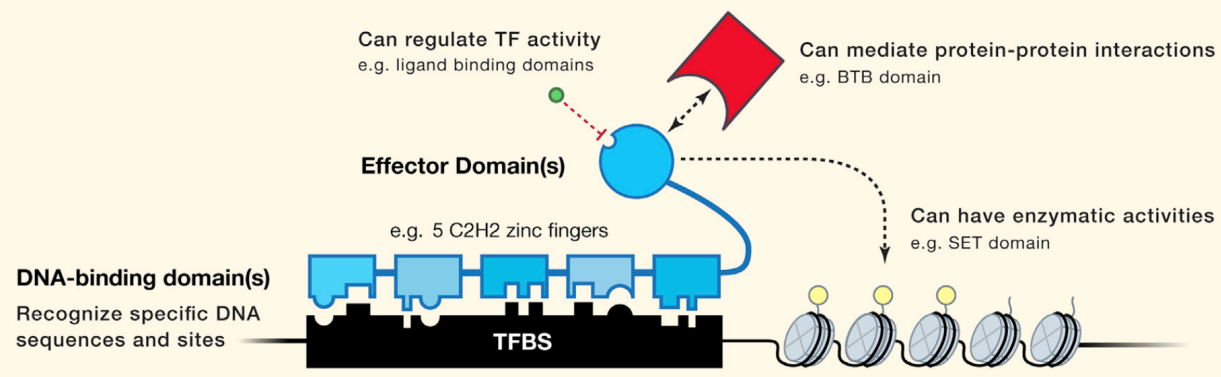


Figure 1: Schematic of a TF and its domains. Adapted from Lambert, S. A., et al. (2018). Cell 172(4): 650-665 with permission (DOI:https://doi.org/10.1016/j.cell.2018.01.029) [18].

Most TFs have at least one DNA binding domain (DBD) that recognizes its targets with sequence specificity: i.e. binds with higher affinity to a specific set of sequences than to other sequences [19]. The sequences preferred by a given TF are regularly summarized as a motif model. Motifs are normally represented as position weight matrices (PWM) and visualized using sequence logos (**Fig. 2**) [20,21]. PWMs and sequence logos provide the means to understand how TFs recognize their specific targets, and even to predict the binding of TFs to a given DNA sequence [20,22]. Overall, there are >1,600 known or likely TFs encoded by the human genome, 1107 of which have a known motif, 104 have a homologous TF in other organisms with a known motif, and 428 have no motifs associated with them **[18].**

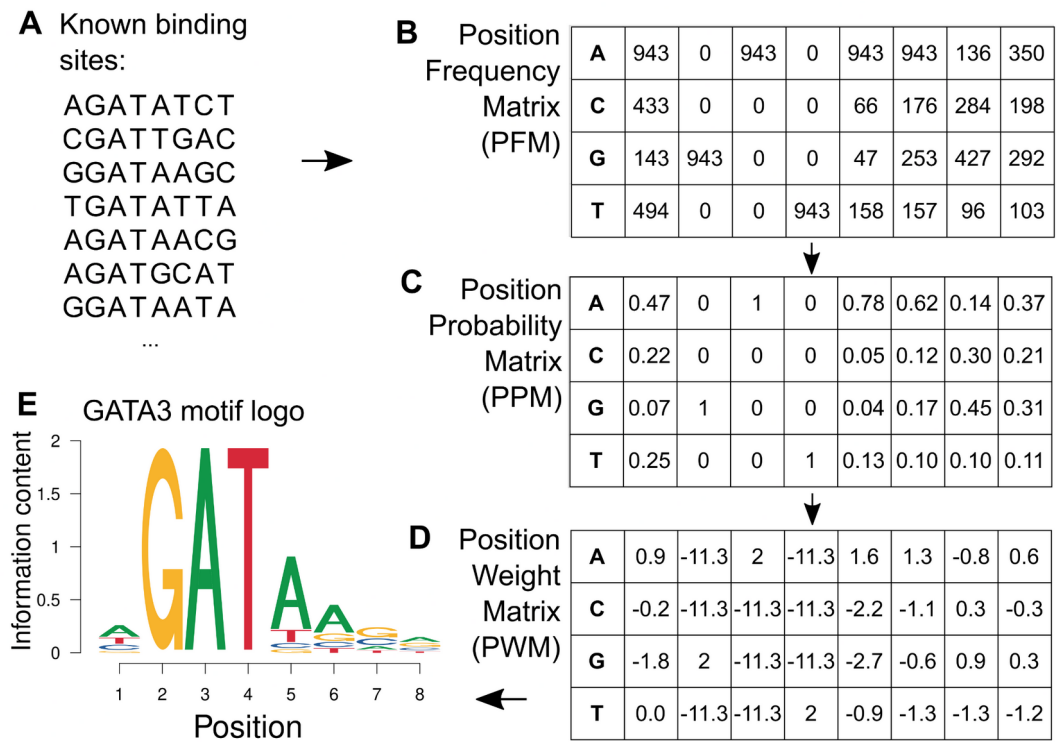


Figure 2: Position weight matrix and motif logo of GATA3. (**A**) The binding sites of the GATA3 TF are identified by high-throughput sequencing techniques. (**B**) A position frequency matrix (PFM) shows the nucleotide counts on each position. (**C**) Normalizing the values (dividing the entries by the total count at each position) produces a position probability matrix (PPM) from a PFM. (**D**) A PWM can be obtained by transforming the entries of a PPM to log likelihoods $M_{k,j}=\log_2\left(\frac{M_{k,j}}{b_k}\right)$, where *b* is a background model ($b=0.25$ if we assume that all nucleotides appear with the same frequency) [21]. In entries where the probability of a nucleotide is zero (due to a small sample size), a pseudocount can be used to avoid undefined values ($M_{k,j}=-\infty$) [23]. (**E**) A sequence logo that visually represents the PWM. Data from Jolma et al. (2013). Cell 152(1): 327-339 (DOI:https://doi.org/10.1016/j.cell.2012.12.009) [24].

1.1.1 DNA binding domains and TF families

TFs are often classified into families based on the type of the DNA binding domain(s) that they use to interact with DNA [18]. The three largest TF families in humans include:

1 *C2H2-ZF*: The Cys_2His_2 zinc-finger protein family is the largest class of TFs in humans with ~750 members. It is also the least understood, since we do not know the motif that is recognized by ~30% of these TFs [18]. Each C2H2-ZFP contains a number of "zinc finger" (ZF) DNA-binding domains, ranging from one to 35 ZFs with an average of ~10 [18]. C2H2-ZFPs usually recognize their target DNA using a subset of their ZFs, with each ZF usually interacting with three to four nucleotides [25].

2 *Homeodomain*: The homeodomain is a protein domain of ~60 amino acids with a structure consisting of three alpha-helices [26]. Homeodomains are encoded by the homeobox genes [27], which are found in animals, fungi, and plants and are particularly highly conserved in vertebrates [28]. Almost all of their motifs are either directly known or can be inferred based on homology [18] TFs of this class are often associated with developmental processes such as differentiation and show highly tissue-specific expression patterns [29].

3 *bHLH*: A basic helix–loop–helix (bHLH) is a protein structural motif that is formed by two regions, one with two α-helices connected by a loop, and a basic region for recognition and binding to DNA [30,31]. Mediated by the HLH motif, the TFs with bHLH often dimerize (either forming homodimers or heterodimers) [32]. They are highly conserved and present in most eukaryotes, including metazoans, plants, and fungi [33].

Beside DBDs, TFs can also contain effector domains; these domains can interact with the basal transcriptional machinery, interact with other TFs, and/or recruit enzymes that modify chromatin state [34]. They can either activate or repress gene expression in a context-dependent manner, which is determined by the local sequences, availability of cofactor, and recruitment of cofactors with opposite effects [35].

1.1.2 ChIP-seq measures TF binding

Various methods exist to determine the sequences that are bound by TFs *in vitro* or *in vivo*. Among the *in vivo* methods, chromatin immunoprecipitation followed by sequencing (ChIP-seq) is by far the most widely used method. ChIP-seq detects binding between DNA and proteins at a genome-wide scale [36], and has been extensively used to detect not only the genomic binding sites of TFs, but also the genome occupancy profiles of RNA polymerase, modified histones, and other targets of interest [37]. ChIP-seq involves crosslinking of DNA-binding proteins and genomic DNA, followed by fragmentation of the genomic DNA (e.g. by sonication). Genomic sites bound by proteins are protected from fragmentation; those bound by the protein of interest can be co-immunoprecipitate using an antibody specific to that protein. Finally, the co-immunoprecipitated DNA goes through library preparation and sequencing [38]. A control experiment is also often performed in parallel to this pull-down experiment, to obtain DNA that was only crosslinked and fragmented without any antibody-based enrichment (input DNA), or DNA that was immunoprecipitated with a non-specific antibody ("IgG" control) [39].

Peak calling is often the first step of downstream ChIP-seq analysis, after mapping and quality control of the sequenced ChIP-seq reads. "Peaks" are regions that are significantly enriched for reads in the pull-down experiment in comparison to the control DNA [37]. As the signal varies across peaks (resulting in strong and weak peaks), the algorithms and tools used for peak calling often calculate p-values and false

discovery rate (FDR) to help identify biologically relevant sites [40,41].

One of the insights obtained from thousands of ChIP-seq experiments across hundreds of TFs is that PWMs are often inadequate models for explaining the *in vivo* specificity of TFs: scanning the genome sequence using PWMs [22] usually results in the identification of tens of thousands of putative TF binding sites (TFBS), most of which are false positives that are not actually functional [15]. While more complicated models such as deep neural networks have had better success [42], *in vivo* TF binding sites remain difficult to predict using DNA sequence alone [15]. This underlines the importance of factors other than sequence that impact TF binding and, consequently, gene regulation, including DNA accessibility and DNA methylation [43,44]. In the next section, we will discuss some of these additional layers that impact TF binding.

1.2 Epigenetic factors that affect TF binding

Various chemical modifications can affect the genome and its associated histones without changing the DNA sequence. The repertoire of these modifications is called the epigenome, and constitutes a key layer of the gene regulation system [45]. The conformation of the epigenome is different across cell types, explaining, to a large extent, why cells with the same genomic sequence have widely different expression patterns and phenotypes [45].

Two of the most important and well-studied components that define the epigenome of a cell are DNA methylation and histone modifications. DNA methylation, which is the addition of methyl groups to the DNA molecule itself, plays a key role in different cellular processes, with its dysregulation associated with various diseases [46]. Histone modification, on the other hand, involves post-translational modification of the N-terminal tail of histone proteins in the nucleosome complex [47]. There are many possible histone modifications that work together to determine chromatin structure, with changes in DNA accessibility being a major consequence of such chromatin

modifications [48]. Particularly, DNA accessibility has been identified as the most important feature, after DNA sequence, to predict the location of TFBSs [43]. Here, we will provide a summary of the mechanisms through which DNA methylation and DNA accessibility impact TF binding, and, consequently, gene regulation.

1.2.1 DNA accessibility

In eukaryotes, nuclear DNA is tightly packaged into chromatin. The basic unit of chromatin is the nucleosome, which is formed by a segment of DNA that wraps around eight histone proteins [47]. Nucleosomes do not occupy the genome in a uniform way, resulting in a range of compactness from regions that are densely packed with nucleosomes to nucleosome-depleted regions that are often found in highly active genomic regions [49]. DNA accessibility refers to the level of possible physical contact between macromolecules and chromatinized DNA—**Fig. 3** shows the continuous nature of DNA accessibility, ranging from nucleosome-packed closed chromatin to nucleosome-depleted accessible DNA [50].

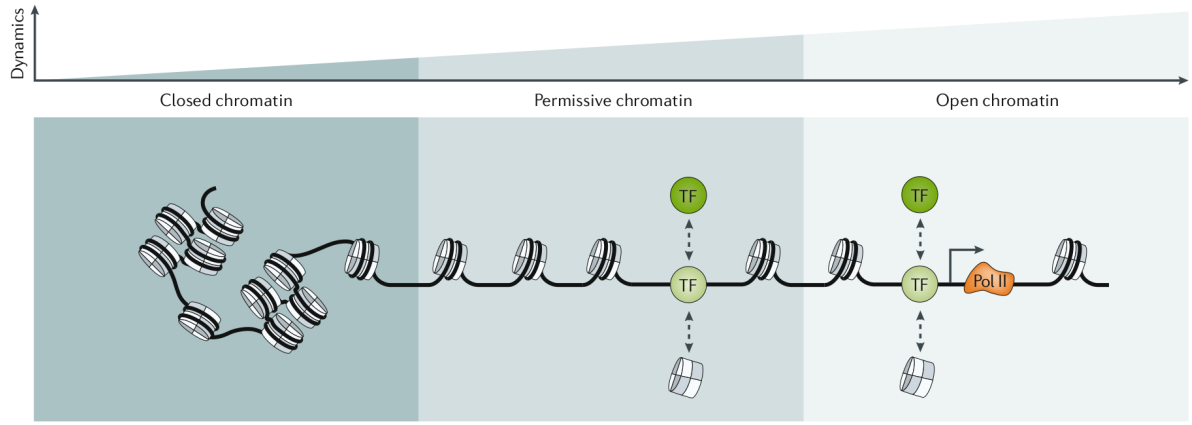


Figure 3: Different continuous chromatin states ranging from close to open chromatin. Adapted from Klemm et al., (2019). Nature Reviews Genetics 20(4): 207-220 with permission (DOI:https://doi.org/10.1038/s41576-018-0089-8) [50].

1.2.1.1 Measuring DNA accessibility using DNase-seq

DNase I hypersensitive site sequencing (DNase-seq) is one of the first methods developed to measure genome-wide DNA accessibility [51] and one of the main methods used for TF footprinting [52]. DNase I is a DNA endonuclease that creates, preferentially, double-stranded breaks in regions where chromatin is not condense and DNA is accessible. In DNase-seq protocols, nuclei are first isolated and permeabilized, then DNA is digested by DNase I into 50-100 bp fragments, followed by library construction and sequencing [53].

1.2.1.2 DNA accessibility and regulatory regions

High levels of accessibility are usually associated with active regulatory loci and transcription. Mapping of highly accessible sites, based on the identification of DNase I hypersensitive sites (DHSs), has shown that these regions encompass ~2-3% of the total genome in any given cell type [8]. The majority of these regions fall within distal enhancers and in lesser amounts within promoters and transcription start site (TSS)-proximal regions [8]. Often, promoter regions are constitutively accessible, but the accessibility of distal enhancers varies by cell type [50]. High accessibility of both promoters and enhancers is correlated with transcriptional activity [8], although regulatory elements that are open but not active are also common **[54].**

1.2.1.3 DNA accessibility and TF binding

TF binding positively correlates with DNA accessibility. Within moderately packed chromatin, a small number of TFs can take advantage of short periods of time where DNA is accessible to recruit cofactors and stabilize chromatin into a more accessible state [8,55]. However, for the great majority of TFs, the existing chromatin state dictates binding [7,8,56]. For example almost all of the binding sites of the glucocorticoid receptor, a TF of the nuclear hormone receptor family, fall within constitutively accessible chromatin , meaning that existing cell type-specific chromatin accessibility landscape determines

occupancy of this TF [7]. Genome wide mapping of DNase hypersensitivity sites of 125 human cell lines by the ENCODE project revealed that the vast majority of sites bound by TFs (90%) fall within open chromatin [8]. This study helped establish DNA accessibility as a useful proxy for TF occupancy [50]. TF binding site prediction methods benefit greatly from DNA accessibility information as it is an important feature to identify functional and non-functional TF binding sites [57,58]. Notably, during the ENCODE-DREAM *in vivo* transcription factor binding site prediction challenge, the methods with the best performance used TF binding motifs and chromatin accessibility information as main sources [43].

1.2.2 DNA methylation

Another key epigenetic layer that affects TF binding and gene regulation is DNA methylation, the addition of a methyl group to a DNA base. Although both adenine and cytosine bases can be methylated, the term DNA methylation is more commonly used to refer to 5-methyl-cytosine modification, the most common DNA methylation. In animals, most 5-methyl-cytosines appear in CpG sites, i.e. a cytosine followed by a guanine in the 5' to 3' direction. This modification is widespread through the genome, with 70-80% of all CpG sites methylated in mammals [59].

CpG methylation is almost always symmetric in somatic cells, meaning that methylation is present on the cytosines of both strands of the self-complementary CpG. This symmetry ensures the preservation of CpG methylation after replication, through recognition and methylation of the resulting hemimethylated DNA by DNMT1 [DNA (cytosine-5)-methyltransferase 1] [60]. Other members of the DNA methyltransferase family also play a fundamental role in *de novo* DNA methylation [61]. Conversely, the ten-eleven translocation (TET) methylcytosine dioxygenases mediate demethylation [62]. Aberrant DNA methylation patterns, caused by malfunction or dysregulation of DNMTs and TET proteins, are linked to cancer and other diseases [46,63], highlighting the

importance of DNA methylation in maintaining proper cell function. Hypo- and hyper-methylation of tumor samples, for example, are a common occurrence in cancer [46,63]. Genome-wide hypo-methylation is associated with increased gene expression and can occur at regulatory elements like promoters and enhancers. Furthermore, transcriptional silencing of tumor suppressor genes by hyper-methylation has been observed in a plethora of cancer types [46,63].

1.2.2.1 WGBS provides base-resolution readout of methylation states

Whole-genome bisulfite sequencing (WGBS) is the most widely used method for genome-wide determination of the methylation status of cytosines at single-base resolution. In this method, sodium bisulfite is used to convert unmethylated cytosines to uracil, while methylated cytosines are protected from this conversion. After PCR amplification and sequencing, the methylated cytosines appear as C in the sequencing reads, while unmethylated cytosines that were converted appear as T [64]. Downstream analysis tools can then use a bisulfite-converted reference genome to align reads and call the fraction of methylated reads at each base [65].

WGBS is experimentally expensive—although it can cover >90% of all genomic CpG sites without considerable bias toward a specific region, it also requires substantial sequencing depth to obtain precise measurement of methylation. Additionally, depending on the biological question and the required downstream analysis, it may be beneficial to have biological replicates [66,67]. Fortunately, consortiums such as the ENCyclopedia Of DNA Elements (ENCODE) and The Cancer Genome Atlas (TCGA) have publicly provided WGBS data for many of the widely used cell lines as well as different cancer types [68,69], enabling studies such as the one described in this thesis.

1.2.2.2 DNA methylation and TF binding

DNA methylation is often associated with gene silencing through both direct and

indirect regulation of gene expression [70]. This gene silencing effect can result from various molecular mechanisms, including the inhibition of the binding of TFs to gene regulatory regions. Some of the main mechanisms through which DNA methylation affects TF binding include:

i *Binding of MBD proteins to methylated DNA*: The mCpG-binding domain (MBD) protein family includes MeCP2, MBD1, MBD2, MBD4, and MBD3, proteins that bind to mCpG in a non-sequence-specific mode [4]. MBD proteins can affect gene expression in two main ways: first, they can outcompete TFs by simply binding to DNA and obstructing TF binding [4]; secondly, by recruiting histone deacetylases (HDAC) [71], MBD proteins can increase chromatin compaction and lead to transcriptional repression [72].

ii *Direct effect of methylated DNA on TF binding*: In TFBSs, the addition of a "bulky" methyl group allows for the formation or loss of possible van der Waals interactions or hydrophobic contacts between DNA and protein side chains [73]. TFBS methylation has been traditionally associated with repression of binding, although it has been found to enhance *in vitro* binding for some TFs [1,74,75].

iii *Methylation-induced structural changes*: The double-helix of methylated DNA has a narrower minor groove compared to unmethylated DNA. Lazarovici et al. used cleavage by DNase I, which is highly shape-sensitive, to probe the DNA shape and found enhanced cleavage adjacent to methylated CpG base pairs [76], suggesting narrowing of the minor groove induced by methylation. These changes in DNA shape can subsequently affect the binding of many of the TFs [77].

1.2.2.3 Relationship between DNA methylation and DNA accessibility

There is a tight relationship between DNA methylation and DNA accessibility, with histone modifications playing an integral part in mediating this association [78]. DNA methylation affects the state of chromatin. Methylated DNA is recognized by MBD proteins, which in turn recruit histone deacetylases (HDACs) [71,79,80] and histone methyltransferases (HMTs) [79]. The effect of both HDACs and HMT results in a more compact chromatin and, consequently, represses transcription [72]. Furthermore, DNA methyltransferases (DNMTs) can directly recruit these histone modifiers. DNMT1 and DNMT3b can interact with HDACs [81,82], and DNMT1 and DNMT3a can bind to the histone methyltransferase SUV39H1 [79]. Conversely, the existing chromatin state can affect the DNA methylation [83]. For instance, the methylation state of histone 3 at lysine 4 (H3K4) is linked to local DNA methylation [83]. Unmethylated H3K4 is recognized by a protein domain found in DNMT3A, DNMT3B and DNMT3L, DNMT3A and DNMT3L can form a tetramer [84]; Interestingly, when this tetramer is modeled into nucleosomal DNA, the active sites of DNMT3A are positioned on adjacent DNA major grooves and can result in de novo CpG methylation [84]. Finally, and consistent with the effect of these histone modifications, next generation sequencing techniques have observed a negative correlation between DNA methylation and DNA accessibility, when measuring them simultaneously over DNase hypersensitive sites [85,86].

1.3 The interplay between TF binding and DNA methylation

The underlying mechanisms that govern TF recognition of DNA methylation are poorly understood. As outlined above, the most widely recognized relationship between TFs and methylated DNA interactions is that DNA methylation prevents TF binding, through competitive binding of proteins containing a methyl-CpG-binding domain (MBD), through promoting closed chromatin, or through direct prevention of TF binding by the methyl moiety [4]. However, there is a growing body of evidence that, TFs can

directly interact with methylated DNA using their DNA binding domains [2,87]. In what follows, we will summarize some of the recent findings that suggest TF-mCpG interactions may be more frequent than currently recognized.

1.3.1 Interactions between methylated DNA and TFs identified by high-throughput *in vitro* methods

Recently, several high-throughput methods that measure the effect of DNA methylation on TF binding have been developed [5,12,13]. Methyl-Spec-seq simultaneously measures the relative affinities of hundreds to thousands of unmethylated and methylated sequences towards TFs [12]. EpiSELEX-seq is another method that quantifies the binding free energy changes in the presence of methylation, by probing the binding affinity of methylated and unmethylated sequences in a single reaction to compare TF occupancy [13]. Both methods use electrophoretic mobility shift assay (EMSA) selection followed by DNA sequencing. Methyl-Spec-seq has been used to probe the *in vitro* methylation preferences of CTCF and ZFP57, while EpiSELEX-seq has been used to study p53 [12,13]. Another method, which is based on systematic evolution of ligands by exponential enrichment (SELEX), was used to carry out a systematic exploration of the effect of methylation on TF binding preferences [5]. Yin et al. analysed the methylation preferences of 519 TFs that recognize CpG-containing sequences, and found that 60% were influenced by mCpG: for 23% of TFs, mCpG inhibited binding, while another 34% of TFs actually preferred binding to methylated CpG-containing sequences, and 5% of TFs showed multiple effects depending on the position of mCpG within the binding site [5]. Interestingly, they found both TF families with homogeneous effects of mCpG in binding as well as TF families with heterogeneous responses to mCpG. For example, in the MAD and CP2 families, mCpG always had a positive effect on binding. In others, like the RUNT and ETS families, mCpG consistently causes a decrease in binding. Notably, the Cys2His2 zinc finger proteins (C2H2-ZFPs) had a heterogeneous response, with some showing increased and some showing decreased binding due to

methylation of CpGs [5].

1.3.2 *In vivo* effects of DNA methylation in TF binding

Disentangling the effect of DNA methylation on TF binding from other factors *in vivo* is experimentally demanding. As a consequence, there are few studies that have probed *in vivo* methyl-CpG preferences of TFs. The following studies either account for changes in DNA accessibility or offer supporting *in vitro* evidence in their analysis.

Maurano et al. found that, although the majority of in vivo CTCF binding sites are unaltered by DNA methylation changes, there is a subset of CTCF sites that is sensitive to DNA methylation, with a diverse occupancy of CTCF across cell types [88]. Importantly, these sites showed no co-binding by other TFs, were methylated when unbound, and were enriched for CpG sites that start at the 2nd and 12th positions [88].

Domcke et al. found that NRF1 is sensitive to methylation of its binding sites. They studied TF binding in wild type and Dnmt1 knockout murine embryonic stem cells, and found that NRF1 gained several thousand binding sites in the unmethylated condition among genomic regions that were accessible in both conditions. Furthermore, when methylation was restored in Dnmt1 knockout cells, NRF1 binding was diminished by methylation of the binding sites [10].

In another study, Mann et al. used protein binding microarrays to evaluate the effect of CpG methylation on B-ZIP transcription factors. They found enhanced binding of CEBPB to DNA when CpGs in array probe sequences were methylated [2]. Another *in vitro* study corroborated that CpG methylation in the 6th position of the CEBPB motif increased binding [5]. Furthermore, a large number of *in vivo* CEBPB binding sites have been reported to be highly methylated [3,4].

Finally, Cusack et al. examined occupancy of five TFs in conditions with

contrasting DNA methylation (wild type and DNMT knockdowns) and chromatin states (normal and treated with HDAC inhibitor) [9]. Pairwise comparisons of these four conditions showed that MAX and NRF1 TFs preferentially bind to unmethylated DNA. Importantly, this preferential binding is observed even after considering changes in DNA accessibility due to recruitment of HDACs to methylated CpGs by MBD proteins [9].

1.4 Computational methods that study TF methyl-binding preferences

As discussed in the previous section, there are only a few cases in which the *in vivo* effect of methylation on TF binding has been tested. As a result, most of our understanding of how DNA methylation affects TF binding comes from *in vitro* studies, with only a few exceptions [2,9,10,13]. While in principle it should be possible to use *in vivo* data of TF binding, DNA methylation, and other confounding factors to fill this gap, available computational studies for performing such analysis are limited in both scope and methodology. Here, we will summarize the available methods for studying TF-mCpG interactions *in vivo*.

One of the methods that consider DNA methylation to model TF binding is Methylphet, which predicts TF binding using a machine-learning approach called random forests (RF) [44]. For a given TF, the RF model of Methylphet incorporates the methylation score, motif score (obtained with a known PWM), and an array of genomic features (distance to TSS, sequence conservation, etc) in order to perform a binary classification of any given genomic region (TF-bound vs. unbound). Methylphet uses ChIP-seq peaks of a TF in order to train the RF model. Xu et al. showed that Methylphet performs better than the motif score alone, or motif score combined with other genomic features (using CENTIPEDE) [44,89]. However, Methylphet uses the average methylation score over the putative TFBS as well as the average of methylation in 30bp bins around the site; therefore it is not able to identify the specific CpG where methylation impacts binding. Furthermore, it is not able to deconvolve the effect of methylation from

correlated confounding variables such as DNA accessibility. Finally, it does not provide an explicit model of how methylation affects TF binding; instead it only predicts whether the TF binds to a specific genomic region.

Other methods have been developed as methylation-aware tools to improve *de novo* motif discovery. Viner et al. and Ngo et al. describe tools that expand the ATGC alphabet by adding symbols for methylated cytosines [90,91]. Specifically, Viner et al. present an extension of MEME, a widely used *de novo* motif-finding algorithm, while Ngo et al. present an algorithm called mEpigram that identifies enriched k-mers in TF peaks and uses them to identify enriched motifs. These methods use WGBS data to create a methylation-aware genome and use it (in addition to a normal genome) to identify methylated motifs (m-motifs) and regular motifs. However, both approaches only consider the sequence and methylation at the potential TFBSs, again leading to the inability to deconvolve the effect of methylation at specific CpG sites from confounding factors such as DNA accessibility and local (regional) methylation level.

This shortcoming is also present in MEDEMO [92], another tool for de novo motif discovery with a similar approach to Viner et al. and Ngo et al. [90,91]. MEDEMO, however, reports the largest number of predicted TF methyl-preferences than any of the previous methods: application of MEDEMO to 335 TFs identified 32 cases where inclusion of methylation in the model improved the prediction of TFBSs, 14 of which represented potential new findings [92]. However, similar to Methylphet, MEDEMO only reports whether considering methylation improves TFBS prediction, without delineating the positive or negative effect of methylation. Furthermore, among the 32 TFs for which methylation was deemed important, only one TF is known to prefer methylated DNA. This might point to a blind spot of this method, considering that 33% of TFs exhibit preferential binding to mCpGs *in vitro* [5]. Furthermore, the TFs with well-established *in vivo* methyl-binding preferences were also missing from MEDEMO's findings. For example, MEDEMO was not able to find any methylation preference for CEBPB, a TF

with well characterized affinity for methylated DNA [2].

Lastly, another relevant method is TFregulomeR, which uses a compendium of ChIP-seq and WGBS datasets to characterize binding partners and cell-specific bindings sites of a TF. Importantly, it allows to investigate the TF function at different DNA methylation levels and with different co-factors. However, this method only allows for a representation of the DNA methylation levels in different contexts (binding partners) and does not provide a quantitative meassurement of its effect on TF binding [93].

Given these shortcomings, there is still a gap in computational methods that can delineate the direct effect of CpG methylation on TF binding *in vivo*, in order to perform a systematic investigation of TF methylation preferences. This project aims to quantitatively model the relationship between DNA methylation and DNA binding preferences of TFs *in vivo*, using a model that is robust to experimental noises and biases inherent to ChIP-seq, able to account for confounding factors that may mask the true methylation-TF interactions, and applicable to a large number of human TFs.

1.5 Hypothesis

We hypothesize that DNA methylation is a major determinant of *in vivo* DNA binding by TFs, and that we can model TF methyl-preferences using genome-wide binding profiles of TFs.

1.6 Objectives

- To develop a method for quantitative modeling of TF binding *in vivo*.
- To use this method for systematic characterization of *in vivo* TF methylation preferences.

CHAPTER 2. MATERIALS AND METHODS

In order to understand the relationship between DNA methylation and TF binding, we began by retrieving and analyzing WGBS, ChIP-seq, and DNase-seq data from different TFs in several cell lines (**Section 2.1**). We developed a method to jointly model these data sets to predict TF-specific binding (**Section 2.2**), and benchmarked it on CTCF ChIP-seq data in HEK293 cells. We expanded our CTCF studies by obtaining differential binding sites of CTCF between different cell lines (**Section 2.3**), and examined whether, using our method, we can predict differential binding that was caused by DNA methylation changes. Finally, we applied our method to a large number of TFs to systematically study the *in vivo* effect of DNA methylation on TF binding (**Section 2.4**). In this chapter, I will describe the methods used to obtain the data and perform these analyses.

2.1 Preprocessing of genomic data

2.1.1 ChIP-seq data processing, peak calling, and peak signal quantification

We limited our analysis to ChIP-seq experiments for TFs done in HepG2, K562, HEK293, GM12878 and HeLa-S3 cell lines, given the availability of WGBS and DNase-seq data for these cell lines. ChIP-seq and ChIP-exo raw reads were retrieved from four main sources: ENCODE [68,94], Najafabadi et al. [95], Schmitges et al. [96], and Imbeault et al. [97]. ENCODE data were downloaded from ENCODE project website (https://www.encodeproject.org/experiments/), while the other data were downloaded from GEO (accession numbers GSE58341, GSE76494, and GSE78099). A total of 2677 ChIP-seq experiments were analyzed, covering 421 TFs and 5 cell lines.

Raw reads were aligned to the human reference genome (GRCh38) with *bowtie2* (version 2.3.4.1) using the “very sensitive local” mode. Mapped reads with mapping quality score smaller than 30 were removed using *Samtools* (version 1.9). ChIP-seq peaks were called using *MACS* (version 1.4) with a permissive p-value threshold of

0.01. We used this permissive p-value to obtain a range of TF binding signals, which our method uses to quantitatively model TF binding strength. We also included negative peaks, i.e. peaks obtained by swapping the treatment with the control experiments, to enable proper modeling of the background signal. In the end, for each ChIP-seq experiment, this process resulted in a list of peaks covering a wide range of pulldown or control (background) signal strengths, along with their associated read counts.

2.1.2 WGBS data processing and DNase-seq data retrieval

Raw reads from Whole-Genome Bisulfite Sequencing (WGBS) of six cell lines were retrieved from ENCODE and GEO (see **Supplementary Table 1** for accession numbers). Raw reads were trimmed based on their quality (phred33 >= 20) with *TrimGalore* (version 0.6.4) [98]. Paired reads were aligned to the human reference genome hg38 [99] using *bismark* (*bowtie2* mode, version 0.22.2), allowing one mismatch during alignment. Reads were deduplicated by removing those that aligned to the same genomic position (*bismark:deduplicate_bismark*). Methylation calls were then extracted, ignoring the first 2 bps from the 5' end of read 2 (*bismark:bismark_methylation_extractor*). A genome wide coverage report with methylated and unmethylated read counts was then generated (*bismark:coverage2cytosine*). Finally, a bigwig file was generated for unmethylated and methylated counts (*bedGraphToBigWig*) [100].

For DNase-seq data, read depth-normalized bigwig files representing DNase-seq signal were retrieved from ENCODE (see **Supplementary Table 1** for accession numbers).

2.2 Joint Accessibility-Methylation-Sequence (JAMS) models

2.2.1 Formatting and preprocessing of data for JAMS

To retrieve the sequence, DNA accessibility, and DNA methylation to train our model we focused on the positive and negative ChIP-seq peak regions that did not fall

within endogenous repeat elements, since the sequence homology of repeat elements can confound the modeling of ChIP-seq data based on sequence [95]. This was done by removing peaks that overlapped any repeat regions, as defined by RepeatMasker [99,101] .

To model the effect of sequence and epigenetic factors on TF binding using our method, it is necessary to align the peaks based on the position of the most likely TF binding site. To do so, we used the known motif of each TF, in the form of position frequency matrices (PFMs), to search for the most likely TFBS within the 100 bp range of the peak summit. PFMs were obtained from CIS-BP [102], and were augmented by *de novo* motifs identified by RCADE2 [103,104] for the C2H2-ZF family of TFs as described later in **Section 2.4**. CISP-BP contains more than one PFMs per TF, as they are derived from different experimental techniques. We selected PFMs exclusively derived from *in vitro* experiments, in order to avoid the confounding effects present *in vivo*. We prioritized, in descending order, PFMs from SELEX, Selective microfluidics-based ligand enrichment followed by sequencing (SMiLE-seq), and Protein-Binding Microarrays (PBM). We used *AffiMx* [105] to identify the best motif match in each peak sequence. This process was uniformly applied to all peaks, including the negative ChIP-seq peak set.

Once the best motif hit in each peak was identified, we extracted the sequence and nucleotide-resolution methylation profile at the motif hit as well as the flanking regions (20 bp) around the motif hit. Sequences were retrieved from the reference genome hg38 using *bedtools:getfasta* [99,106]. Methylated and unmethylated read counts at each position were retrieved from the WGBS bigwig files using *bwtool* [107].

Similarly, normalized DNA accessibility was extracted from the motif hit region and 500 bp upstream and downstream of the motif hit from the DNase-seq bigwig files. ChIP-seq read counts were extracted from the control and pull-down experiments for the +/- 400bp region surrounding the motif match using *bedtools:multicov* (MAPQ score

> 30). (**Fig. 4C, bottom**) [106].

2.2.2 Implementation of JAMS

Our method creates a joint accessibility-methylation-sequence model (JAMS model) for each ChIP-seq experiment, in which the ChIP-seq signal of each peak is explained as a function of accessibility, methylation, and sequence at that peak. Consider the *k*×*m* matrix $\boldsymbol{X}$, which represents the value of *m* predictive features at *k* genomic positions (i.e. peaks). These *m* features include those related to accessibility (A), methylation (M), and sequence (S):

$$\boldsymbol{X} = \left[\boldsymbol{X}_A \boldsymbol{X}_M \boldsymbol{X}_S\right]$$

JAMS models the logarithm of TF binding strength at each of the *k* peaks as a linear function of the matrix $\boldsymbol{X}$:

$$\log \boldsymbol{\mu}_f = \boldsymbol{X} \times \boldsymbol{\beta}_f$$

Here, $\boldsymbol{\mu}_f$ is the vector of the binding strength for transcription factor *f* across *k* peaks, $\boldsymbol{X}$ is the *k*×*m* feature matrix described above, and $\boldsymbol{\beta}_f$ is the vector of *m* coefficients that describe the effect of each of the *m* features on the TF binding strength (matrices are denoted with bold capital letters, and vectors with bold lower-case letters).

Similarly, the background ChIP-seq signal across the peaks is also modeled as a function of $\boldsymbol{X}$:

$$\log \boldsymbol{\mu}_b = \boldsymbol{X} \times \boldsymbol{\beta}_b$$

Here, $\boldsymbol{\mu}_b$ represents the background signal strength across *k* peaks, and $\boldsymbol{\beta}_b$ is the vector of *m* coefficients that describe the effect of each of the *m* features on the background signal.

In a ChIP-seq experiment, the expected control (background) read counts at each peak is simply a function of the background signal multiplied by the library size. Therefore, the logarithm of control reads can be modeled as:

$$\log \boldsymbol{\lambda}_c = \log \boldsymbol{\mu}_b + s_c = \boldsymbol{X} \times \boldsymbol{\beta}_b + s_c$$

Here, $\boldsymbol{\lambda}_c$ is the vector of expected (average) control read counts across the *k* peaks, and s_c is an experiment-specific size factor that can be interpreted as the logarithm of sequencing depth for the control library.

The expected pull-down read counts in a ChIP-seq experiment, however, are a

function of both the background signal and the TF binding strength, multiplied by the library size. Therefore:

$$\log \boldsymbol{\lambda}_p = \log \boldsymbol{\mu}_b + \log \boldsymbol{\mu}_f + s_p = \boldsymbol{X} \times \boldsymbol{\beta}_b + \boldsymbol{X} \times \boldsymbol{\beta}_f + s_p$$

Here, $\boldsymbol{\lambda}_p$ is the vector of expected pulldown read counts across the k peaks, and s_p can be interpreted as the logarithm of sequencing depth for the pulldown library.

While these equations describe the expected control and pulldown read counts, the actual observed read counts are probabilistic observations that may deviate from these expected values. Here, we model the read counts as observations from negative binomial distributions [108] whose mean is given by the equations above, with a shared dispersion parameter across the peaks:

$$\boldsymbol{n}_c = NB\left(\boldsymbol{\lambda}_c, \varphi\right)$$
$$\boldsymbol{n}_p = NB\left(\boldsymbol{\lambda}_p, \varphi\right)$$

Here, $\boldsymbol{n}_c$ and $\boldsymbol{n}_p$ are the vectors of observed control and pulldown read counts across the k peaks, respectively, and φ is the dispersion parameter. The equations above allow us to jointly model the control and pulldown experiments as a function of $\boldsymbol{X}$. We use the glm.nb function in R for this purpose and fit a model of the form *n~XX+t+XX:t*, where *n* is an R vector that concatenates the observed control and pulldown read counts (with length 2*k*), *XX* is the result of duplicating matrix $\boldsymbol{X}$, i.e. *XX=rbind(X,X)*, and *t* is a binary vector of length 2*k* indicating whether the observed read count comes from the control experiment (0) or from the pulldown experiment (1). The coefficients returned by the glm.nb function for *XX* correspond to $\boldsymbol{\beta}_b$ in the equations above, and the coefficients for *XX:t* correspond to $\boldsymbol{\beta}_f$. The glm.nb also returns the standard error of mean and a p-value for each of these coefficients, which we use to determine the statistical significance.

Constructing the matrix $\boldsymbol{X}$: Sequence, DNA methylation and DNA accessibility are used as the predictor variables, which are included in the matrix $\boldsymbol{X}$. We used one-hot encoding for the sequence over the TFBS. Methylated and unmethylated read counts over the motif were used to calculate the methylation percentage at each position. If the average coverage of methylation and unmethylated reads over the motif

is less than 10 counts, the peak is removed. Average DNA accessibility was calculated for bins of 100 bp (10 bins) plus one bin for the TFBS region itself, and then logarithm of DNA accessibility was calculated; a pseudocount equivalent of 1% of the smallest value was used to allow for log transformation of the data. Average methylation percentage and sequence composition of the flanking regions were also used as predictors.

2.3 Differential binding analysis

To calculate differential TF binding between cell lines, we first identified CTCF ChIP-seq experiments from ENCODE that had at least two biological replicates per cell line (**Supplementary Table 1**), and retrieved the pull-down and control experiment data. After aligning and peak calling (**Section 2.1.1**), we defined a unified list of peaks that were present in at least one sample. Peaks that were present in more than one sample and had summits within 100 bp of each other were merged, as they likely represent the same CTCF binding site. Then, the best motif match within 100 bp of each summit was identified [105]. We extracted ChIP-seq read counts present within a 400bp range from the motif hit in the pull-down and control experiments and created a count matrix.

We used the count matrix and a custom model matrix (**Supplementary Table 3**) to compare count ratios (of pulldown and control reads) between pairs of cell lines. The DESeqDataSetFromMatrix function from DESeq2 was used to create a DESeqDataSet object (parameters: countData = count matrix and colData/design = custom matrix) followed by fitting of a negative binomial GLM (function DESeq, parameters: *full* = custom matrix, betaPrior = FALSE), and computing of log2 fold changes and p-values [109]. Significant differentially bound peaks (FDR < 0.1) were identified for every pair of cell lines, excluding cell line pairs whose ChIP-seq experiments were done in different laboratories. The pair of cell lines (GM12878 and HeLa-S3) with the highest number of significantly bound peaks were selected for further analysis.

2.4 Inference of PFMs for C2H2-ZF proteins using RCADE2

We inferred position frequency matrices (PFMs) for canonical C2H2 zinc finger proteins using RCADE2 [103,104]. RCADE2 uses the protein sequence, the DNA sequence of the ChIP-seq peaks, and a previously computed machine learning-based recognition code to predict the DNA-binding preferences of C2H2-ZFPs. The protein sequences for these TFs were retrieved from UNIPROT [110]. We focused on the top 500 ChIP-seq peaks (sorted by p-value) that did not fall within endogenous repeat elements (EREs) [99,101]. The DNA sequence of the +/- 250 region around the peak summits for the top 500 non-ERE peaks along with the protein sequence was provided as input to RCADE2, and the optimized motif was used to augment the CIS-BP motifs.

CHAPTER 3. RESULTS

3.1 Quantitative modeling of ChIP-seq data to infer *the in vivo* methyl-binding preferences of TFs

3.1.1 Modeling the joint effect of accessibility, methylation and sequence on TF binding

Several factors work together to determine the TF binding strength, as measured by ChIP-seq, toward a specific binding site. First, the sequence of the binding site determines the TF affinity, given that the majority of TFs are sequence-specific. Secondly, for most TFs, the existing level of DNA accessibility heavily influences TF binding [7,8]. Finally, regional methylation outside the TFBS may affect the TF binding strength, for example by recruiting Methyl-CpG-binding domain (MBD) proteins, which in turn recruit chromatin remodelers [6]. Therefore, in order to examine the specific effect of methylation of the TFBS on TF binding affinity, we need to jointly model it together with these confounding factors.

For this purpose, we developed Joint Accessibility-Methylation-Sequence (JAMS) models, which quantitatively explain both the pull-down and background signal in ChIP-seq experiments (https://github.com/csglab/JAMS). The JAMS model for each ChIP-seq experiment considers the pull-down read density as a combination of a background signal and a TF-specific signal. On the other hand, the read count profiles obtained from control experiments (e.g. input DNA) purely reflect the background signal (**Fig. 4A**). Each of the background and TF-specific signals, in turn, is modeled as a function of the peak sequence, chromatin accessibility profile along the peak, and regional as well as base-resolution methylation pattern of the peak (**Fig. 4B-C**). JAMS converts these associations into a generalized linear model whose parameters can then be inferred jointly from pull-down and control experiments, with an appropriate error model that connects the expected (predicted) signal at each peak to the observed read counts—we use negative binomial with a log-link function in this work (**Fig. 4D**).

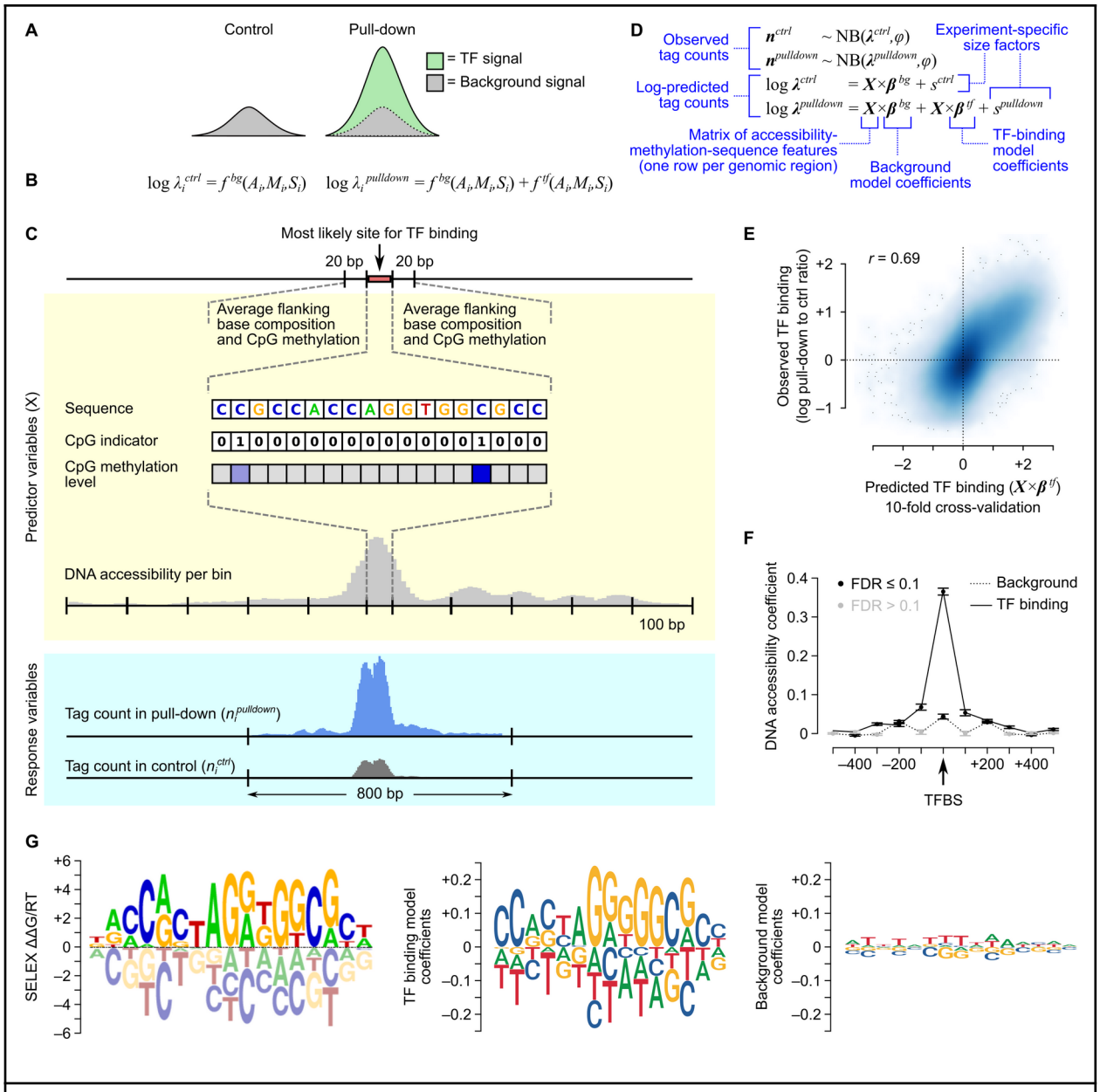


Figure 4: Overview of JAMS model. (**A**) At each genomic region *i*, the JAMS model considers the control tag count (left) or the pull-down tag count (right) as a combination of background and/or TF-binding signals at that position. (**B**) Each of these signals are then modeled as a function of accessibility (A_i), methylation (M_i), and sequence (S_i) at each region *i*. (**C**) Schematic summary of the predictor features extracted for each genomic location and the outcome variables. (**D**) The specifications of the generalized linear model used by JAMS. (**E**) Comparison between the observed

and predicted CTCF binding signal in HEK293 cells [96]. (**F**) DNA accessibility coefficients learned by the CTCF JAMS model; each dot corresponds to the effect of accessibility at a 100bp-bin. (**G**) Sequence motif logos representing the known CTCF binding preference (based on SELEX [24] (left), the TF binding specificity learned by JAMS (middle), and the effect of sequence on the background signal (right). JAMS motif logos are plotted using ggseqLogo [111] , with letter heights representing model coefficients; SELEX motif logo was obtained from the CIS-BP database [102].

To fit the parameters of its model, JAMS assumes that TF binding occurs at a fixed position and orientation in each of the provided peaks. To satisfy this assumption, we use existing position frequency matrices (PFMs) of each TF to identify the most likely TF binding site within each peak, and use that position as the reference for extracting the accessibility-methylation-sequence features at, and around, the binding site (see **Method, Section 2.2** for details). Also, to ensure that JAMS can correctly learn the features associated with both TF-specific and background signals, we include not only the peaks that have significantly high pull-down signal, but also peaks with low pull-down signal as well as genomic locations that have significantly high control signal (**Method, Section 2.1.1**).

In order to examine the ability of JAMS models to recover the *in vivo* binding preferences of TFs, we first applied it to ChIP-seq data from CTCF, a widely studied TF that is constitutively expressed across cell lines and tissues [112,113] and has a long residence time on DNA [114]. We initially focused on the cell line HEK293, and generated a JAMS model of CTCF binding in this cell line using previously published ChIP-seq [96], WGBS [115], and chromatin accessibility data [94] (**Methods, Section 2.1.1**). To evaluate the performance of the JAMS model, we used 10-fold cross-validation, and examined the correlation between the predicted TF-specific signal and the observed pulldown-to-control signal ratio across the peak regions. As **Fig. 4E** shows, the JAMS model predictions correlate strongly with the pulldown-to-control signal ratio (Pearson r=0.69),

suggesting that accessibility-methylation-sequence features can quantitatively predict the CTCF-binding strength.

Examining the coefficients of the fitted JAMS model, we observed that DNA accessibility, especially at the peak center, has a strong effect on the TF-specific signal (which only affects the pull-down read count), but limited effect on the background ChIP-seq signal (which affects both the control and pull-down read counts; **Fig. 4F**). Nonetheless, the effect on background signal was still statistically significant, consistent with previously observed bias of DNA sonication toward accessible chromatin regions [116]. Importantly, sequence features at the TF binding site are strongly predictive of the CTCF binding strength, while they have limited and diffuse effect on the background signal (**Fig. 4G**). Furthermore, the sequence model learned by JAMS is highly correlated with the known motif for CTCF (r=0.86, **Fig. 4G**), suggesting that JAMS models can recapitulate the underlying biology of TF binding. We emphasize that while the known CTCF motif is used initially to identify an offset for each peak and align the peak regions, this process is not expected to confound the sequence features learned by JAMS, since it is uniformly applied to all peaks regardless of the signal strength.

3.1.2 JAMS models reveal the contribution of CpG methylation to TF binding

By jointly considering the contribution of accessibility, methylation and sequence to TF binding, JAMS models should be able to deconvolve the specific effect of methylation from the confounding effect of other variables. To begin to explore this possibility, we examined the JAMS model of CTCF. For this purpose, in addition to sequence motif logos, we developed “dot plot logos” to enable easier visual inspection of JAMS coefficients that correspond to sequence and methylation effects. As **Fig. 5A** shows, the JAMS model of CTCF binding in HEK293 cells suggests that CpG methylation in the 2nd and 12th positions of the binding site has a significantly negative effect on CTCF binding (but not on the background signal; **Supplementary Fig. 1**). In

other words, while a large fraction of CTCF binding sites have CpGs at those two positions, CTCF preferentially binds when these CpGs are not methylated.

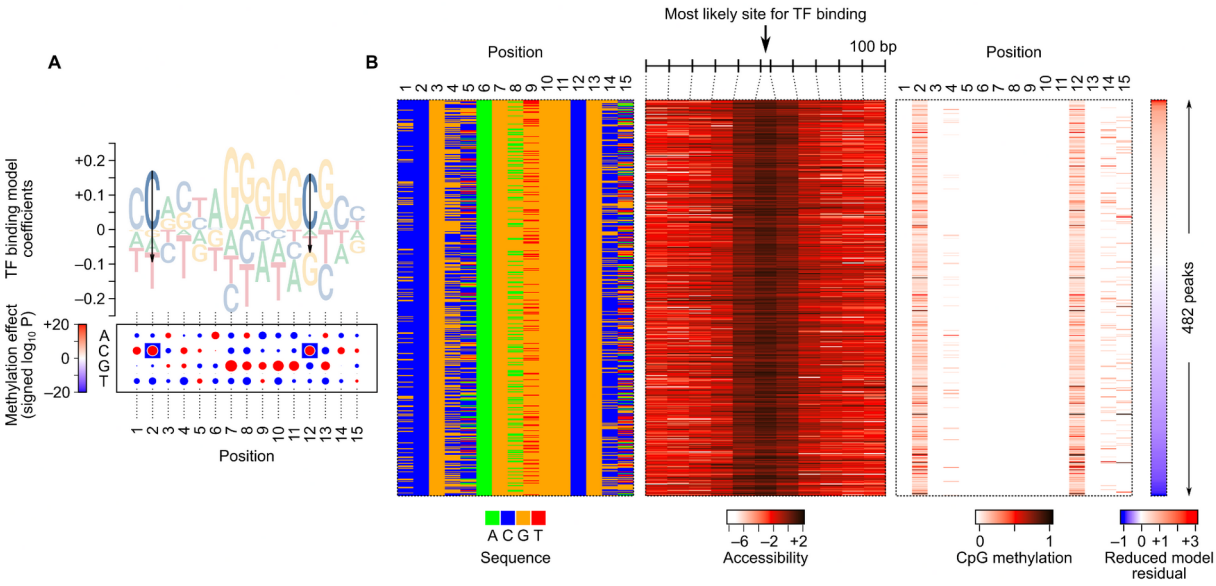


Figure 5: CpG methylation preference of CTCF in HEK293 cells. (**A**) Motif logo and dot plot representations of the sequence/methylation preference of CTCF. The logo (top) shows methylation coefficients as arrows, with the arrow length proportional to the mean estimate of methylation effect. The heatmap (bottom) shows the magnitude of the preference for each nucleotide at each position using the size of the dots, with red and blue representing positive and negative coefficients, respectively. The signed logarithm of P-value of the methylation coefficient is shown using the color of the squares around the dots, with red and blue corresponding to increased or decreased binding to methylated C, respectively (only significant methylation coefficients at FDR<1×10^{-5} are shown). (**B**) Heatmap representation of the sequence, accessibility, and CpG methylation, for a subset of CTCF peaks that have high DNA accessibility, a close sequence match to the initial CTCF motif, and CpGs at positions 2 and 12. Peaks are sorted by the residual of a reduced JAMS model that does not use the methylation level of C2 and C12 for predicting the CTCF binding signal.

To ensure that this observation is not confounded by other variables such as accessibility and the average local methylation level, we also trained a JAMS model

with all the variables except the CpG methylation level at each binding site position; we then compared these reduced models to the full model using a likelihood ratio test. This analysis revealed that removing the CpG methylation levels at positions 2 or 12 of the binding site significantly reduces the fit of the model to the observed data (**Supplementary Fig. 2**). Therefore, the CpG methylation level in these positions is informative about CTCF binding signal even after considering the effect of other confounding variables such as sequence, accessibility, and the average methylation of flanking regions.

The independent effect of CpG methylation on CTCF binding can also be observed after stratification of CTCF peaks based on the confounding variables. Specifically, we repeated the JAMS modeling after removing the variables that represent the TF-specific contribution of methylation at positions 2 and 12, sorted the peaks by the residual of this model (i.e. by the ChIP-seq signal that could not be explained by the reduced model), and visualized the methylation pattern of the peaks, limiting to the peaks that (a) had a sequence similar to the CTCF-preferred binding site, (b) had CpGs at positions 2 and 12, and (c) had high DNA accessibility. As **Fig. 5B** shows, even if we focus on the peaks with similar sequence and accessibility, the residual of the reduced model still correlates negatively with CpG methylation at positions 2 and 12. In other words, peaks whose signal is smaller than what the reduced model predicts have higher CpG methylation, supporting the negative effect of CpG methylation on CTCF binding.

Similar JAMS models can be obtained using CTCF ChIP-seq, WGBS, and accessibility data from several other cell lines (**Supplementary Fig. 3**), highlighting the reproducibility of these results across different contexts. Importantly, our observation that CpG methylation at positions 2 and 12 negatively affects CTCF binding is consistent with previous reports on CTCF methylation preferences *in vivo* and *in vitro* [12,88]. These results overall suggest that JAMS models have the potential to faithfully recapitulate the methylation preferences of TFs using ChIP-seq data.

3.2 Prediction of cell type specific TF binding using JAMS models

3.2.1 CTCF JAMS models are transferable across cell types

A JAMS model that encodes the intrinsic binding preference of a TF should be able to predict the ChIP-seq signal of that TF in new contexts, such as in previously unseen cell types that were not used in model training. We began to examine this possibility by investigating the transferability of the CTCF model that was learned in HEK293 cells to other cell types. We used DNase-seq and WGBS data (**Methods, Section 2.1.1** and **Supplementary Table 1**) from six cell lines (H1, GM12878, HeLa-S3, HepG2, and K562) to predict the CTCF binding signal (using the HEK293-trained JAMS model), and compared the predictions to experimental CTCF ChIP-seq data obtained for each cell type (**Supplementary Table 1**). We observed that the CTCF JAMS model that was trained on HEK293 data could successfully predict the ChIP-seq pulldown-to-control ratio in other cell types, with a performance comparable to JAMS models that were specifically trained on the data from each type (**Table 1**). These results support the transferability of JAMS models across cell types.

Table 1: Pearson correlation between observed and predicted CTCF-binding across cell types. The third column shows the *r* between observed and predicted signal for JAMS models that were trained on each individual cell type. The fourth column shows the *r* between the predictions of the JAMS model that was trained on HEK293 and the observed ChIP-seq data in other cell lines.

Cell line	ChIP-seq peaks	10-fold CV	HEK293-trained *r*
HEK293	135,717	0.69	-
H1	128,123	0.72	0.62
GM12878	39,535	0.69	0.54
HeLa-S3	65,865	0.72	0.60
HepG2	81,188	0.73	0.64
K562	85,122	0.74	0.68

3.2.2 Differential binding between cell lines is captured by JAMS models

The analyses described in the previous section show that the JAMS models learned from one cell type can be transferred to another cell type. However, a considerable proportion of CTCF binding sites are shared across these cell types [88];

therefore, it is not immediately clear to what extent this transferability corresponds to cell-invariant features of the JAMS model (sequence) as opposed to potentially cell type-specific features (methylation and accessibility). In fact, one of the most challenging aspects of modeling TF binding is the ability to identify TF binding sites that are differentially occupied across cell types [43]. To understand the extent to which differential accessibility and methylation of DNA drives differential CTCF binding, and the extent to which these effects can be captured by JAMS, we decided to use the JAMS model learned from HEK293 cells to predict differential binding of CTCF in other cell lines. We started by identification of differentially bound CTCF peaks in pairwise comparisons of cell lines listed in **Table 1**. For any given two cell lines, we used the log-fold change (logFC) in the pulldown-to-control ratio as the measure of differential binding (**Fig. 6A**). The mean and standard error of mean (SEM) of this metric was calculated using a statistical model that assumes a negative binomial distribution for the tag counts, which also allows us to calculate a P-value for the null hypothesis that logFC is equal to zero (see **Methods, Section 2.3**).

Application of this method to all pairwise cell comparisons revealed the largest number of differentially bound CTCF peaks between GM12878 and HeLa-S3 cells (**Fig. 6B**); therefore, we focused on prediction of the differential peaks between these two cell lines using the HEK293 JAMS model of CTCF. Specifically, we used the JAMS model to predict the CTCF binding signal in each of the GM12878 and HeLa-S3 cell lines (based on the accessibility and methylation data of each cell line), and then calculated the logFC of the JAMS predictions between the two cells. As shown in **Fig. 6C**, the JAMS-predicted changes in CTCF binding are strongly correlated with the experimental logFC values (r=0.40 across peaks with logFC SEM<1.28; see **Supplementary Fig. 4** for details on the choice of SEM cutoff). These results suggest that the CTCF JAMS model can quantitatively predict the change in CTCF binding strength based on differential accessibility and methylation. Importantly, for the set of peaks that pass the statistical significance threshold for differential binding between the two cell lines (FDR<0.1), the

correlation between JAMS predictions and experimental logFC reaches as high as 0.84 (**Fig. 6C**), with JAMS being able to distinguish GM12878-specific from HeLa-S3-specific binding events with 95% accuracy.

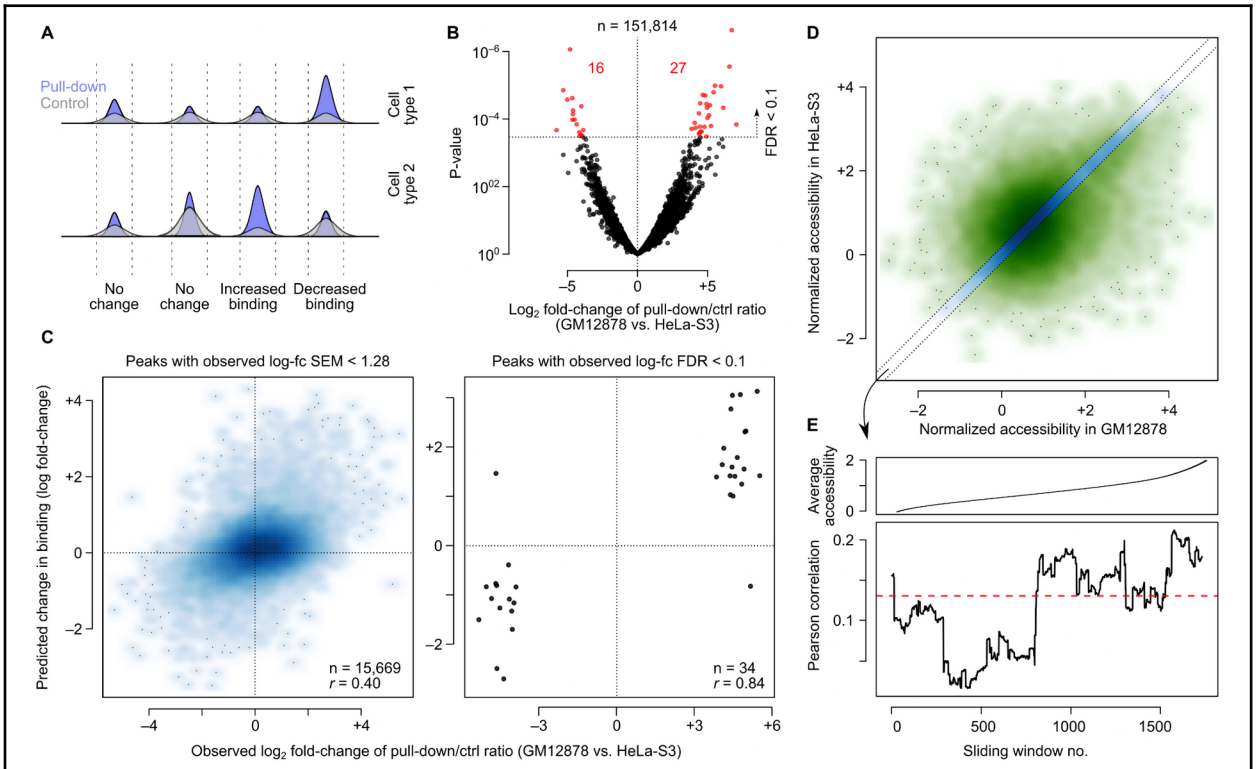


Figure 6: Prediction of differentially bound CTCF peaks using JAMS. (**A**) Schematic representation of identifying differentially bound peaks based on the combination of pulldown and control signal in two cell lines. See **Methods** for details. (**B**) Volcano plot showing differential binding of ChIP-seq peaks between GM12878 and HeLa-S3. Significant peaks at FDR < 0.1 are shown in red. (**C**) Left: Scatter plot of JAMS-predicted changes in CTCF binding and observed differential binding between GM12878 and HeLa-S3 cells. Peaks with observed logFC SEM <1.3 are included. Right: Limited to peaks that pass FDR<0.1 for differential binding of CTCF. (**D**) Comparison of the accessibility of putative CTCF peaks between two cell lines. The diagonal band in the middle (blue) shows the region that was selected as no-change in accessibility (difference in accessibility < 0.2). (**E**) Predicting differential CTCF binding for peaks with no change in accessibility. Peaks were ranked by accessibility, and the correlation between predicted and observed logFC of CTCF

binding was calculated for sliding windows of 500 peaks (bottom). The average accessibility for each sliding window is shown on top.

We note that many of the CTCF binding sites are differentially accessible between GM12878 and HeLa-S3 (**Fig. 6D**). The above analysis cannot rule out the possibility that differential accessibility is responsible for the differential CTCF binding between these two cell lines. To specifically examine the role of differential methylation in driving cell type-specific CTCF binding, we further limited our analysis to the set of peaks that had similar accessibility in both cell lines (**Fig. 6D**), and also removed all the JAMS predictor variables corresponding to accessibility. We observed that this reduced JAMS model can still predict differential CTCF binding among the peaks that are not differentially accessible (r=0.14 between predicted and observed logFC across n=2232 peaks; **Fig. 6E**). This correlation increases to 0.22 for the set of peaks that have high accessibility in both cell lines (**Fig. 6E**), suggesting that the effect of CpG methylation is most noticeable when the putative CTCF binding site is accessible.

Overall, these analyses suggest that JAMS models can accurately predict differential TF binding across cell types, including differential TF binding events that are driven by differential methylation of the putative binding sites. The ability of JAMS models to predict cell type-specific TF binding events further highlight their reliability in capturing the biochemical determinants of TF binding using ChIP-seq data.

3.3 JAMS models reveal the landscape of TF methyl-binding preferences

3.3.1 A high-confidence compendium of JAMS models for 260 TFs

A recent large-scale *in vitro* study has revealed that methyl-binding preferences are heterogeneous across TFs, and vary even within TF families [5]. While this *in vitro* study provided a first global picture of TF methyl-preferences, it is not clear to what extent its conclusions can be extended to *in vivo* TF function. However, establishing the relationship between TF binding and CpG methylation *in vivo* is experimentally taxing and time consuming [9,10,14]. Therefore, we decided to apply JAMS to a comprehensive

compendium of ChIP-seq data in order to identify TFs whose *in vivo* binding is positively or negatively affected by methylation of CpGs at their binding sites.

We collected and uniformly processed data from 2368 ChIP-seq and ChIP-exo experiments [94,96,97], covering the *in vivo* binding profiles of 260 TFs in six cell lines, along with the WGBS and DNase-seq assays in those cell lines (see **Supplementary Table 1** for accession numbers). On average, we identified ~60k peaks per ChIP-seq experiment using the permissive P-value threshold of 0.01 (**Fig. 7A**). We then used the peak tag counts to fit a JAMS model to each ChIP-seq experiment. We noticed that the quality of the JAMS models, measured by the Pearson correlation between the predicted and observed TF-specific signal, varied substantially across the experiments, with correlations ranging from 0 to 0.8 (median 0.48, **Fig. 7B**). This variation may reflect a multitude of factors, including the ChIP-seq data quality as well as the extent to which the TF signal can be explained by our model specifications. We therefore decided to keep only a subset of high-confidence models. Specifically, we selected at most one representative model per TF based on the following criteria: (i) the model should have used at least 10,000 peaks for training, (ii) Pearson correlation >0.2 between the predicted and observed TF-specific signal after cross-validation, (iii) Pearson correlation >0.3 between the known and JAMS-inferred sequence motif, (iv) and low contribution of the sequence to the background signal compared to the TF-specific signal (control-to-pulldown ratio of the sequence coefficients mean < 0.4). As an example, in **Fig. 7C** we show two JAMS models for BHE40, obtained from two different ChIP-seq experiments, only one of which passes all the criteria mentioned above. Overall, we obtained high-confidence JAMS models for 260 TFs, spanning a range of TF families (**Fig. 7D**).

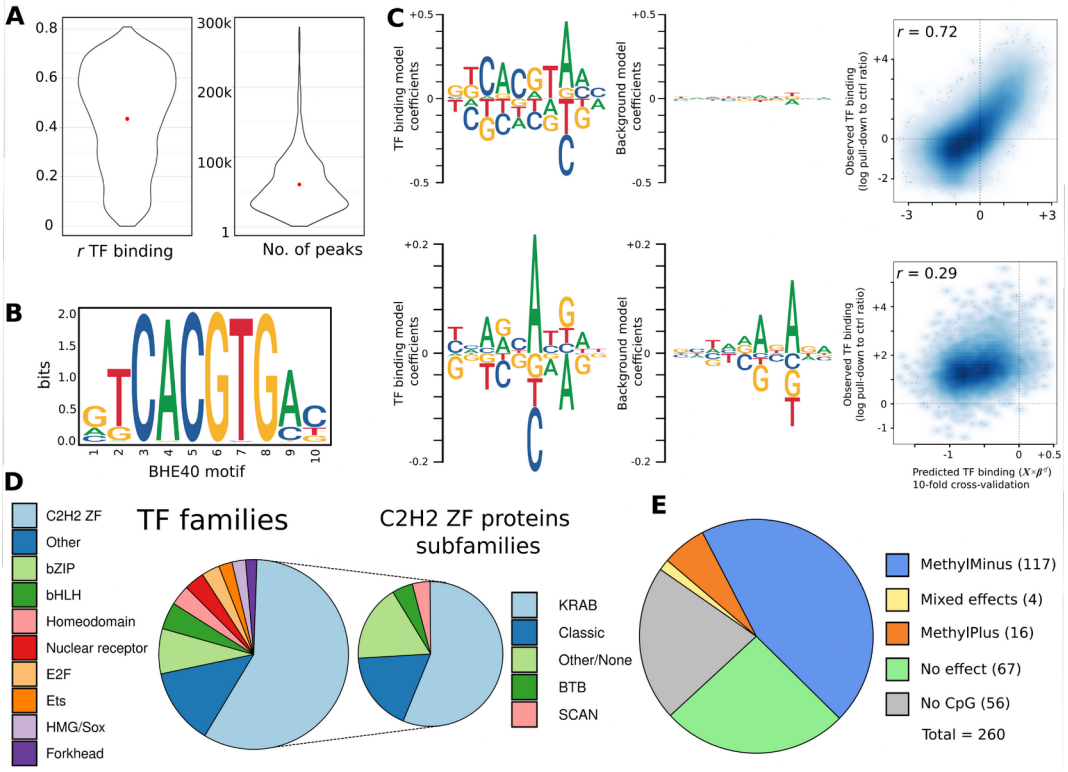


Figure 7: Systematic application of JAMS. (**A**) Left: Violin plot showing the distribution of Pearson correlation between the observed and predicted TF binding signal. Right: Distribution of the number of peaks used to create JAMS models. The violin plots represent a total of 2368 ChIP-seq experiments that were analyzed by JAMS. (**B**) Known BHE40 motif obtained from the CIS-BP database, shown as an example [102]. (**C**) Results from a high-quality (top) and a low-quality (bottom) JAMS model for BHE40. Inferred sequence coefficients for TF binding (left) and background (middle), as well as the predicted vs. observed TF binding signal (right) are shown. (**D**) Pie charts of the main TF families (left) and C2H2 ZF proteins subfamilies (right) for TFs with at least one high-quality JAMS model. (**E**) Pie chart of the methyl-binding preferences of TFs with at least one high quality JAMS model. We obtained high-quality models for a total of 260 TFs.

3.3.2 Systematic inference of the *in vivo* TF methyl-binding preferences

After selecting one JAMS model per TF, we used the JAMS-inferred effects of methylation to classify the TFs according to their inferred methyl-binding preferences. We use a notation similar to Yin et al. [5]. Specifically, we classified a TF as (a) MethylMinus if its JAMS model included at least one significantly negative mCpG effect (FDR<1×10^{-5}), (b) MethylPlus if the model included at least one significantly positive mCpG effect, (c) mixed-effect if the model included both significantly positive and negative mCpG effects, (d) and no-effect if the motif included a CpG but its methylation level did not have a significant effect. Overall, we found 117 MethylMinus TFs, 16 MethylPlus TFs, four mixed-effect TFs, and 67 TFs with no significant mCpG effects; we also identified a set of 56 TFs without a CpG site in their binding site (**Fig. 7E**).

To understand whether our JAMS-based classification captures known methyl-binding preferences of TFs, we started by examining a few TFs whose methyl-binding preferences have been extensively studied *in vitro* and *in vivo*, including CEBPB and NRF1. Using protein-binding microarrays (PBMs), Mann et al. have previously reported enhanced binding of CEBPB to its CpG-containing target sequence when the array probes were methylated [2]. Consistent with this observation, a large number of the genomic binding sites of CEBPB is highly methylated *in vivo* [4]. The JAMS model for CEBPB (**Fig. 8, top**) is concordant with these previous reports, showing that CpG methylation at the 6th position of CEBPB target sequence has a positive effect on its binding strength (**Fig. 8C**). This effect is in fact highly reproducible, and is present in three out of four JAMS models that we obtained using different CEBPB ChIP-seq experiments.

Another well studied TF is NRF1, which has been found to be sensitive to CpG methylation of DNase-I-hypersensitive sites in murine stem cells [10]. Moreover, Cusack

et al. found that NRF1 preferentially binds to unmethylated DNA even after accounting for changes in DNA accessibility caused by the recruitment of HDACs to methylated CpGs by MBD proteins [9]. Consistent with these reports, we found that CpG methylation of the 3rd and 9th positions of the NFR1 target sequence has a negative effect on its binding (**Fig 8G**); these effects were consistent across all the cell lines we analyzed.

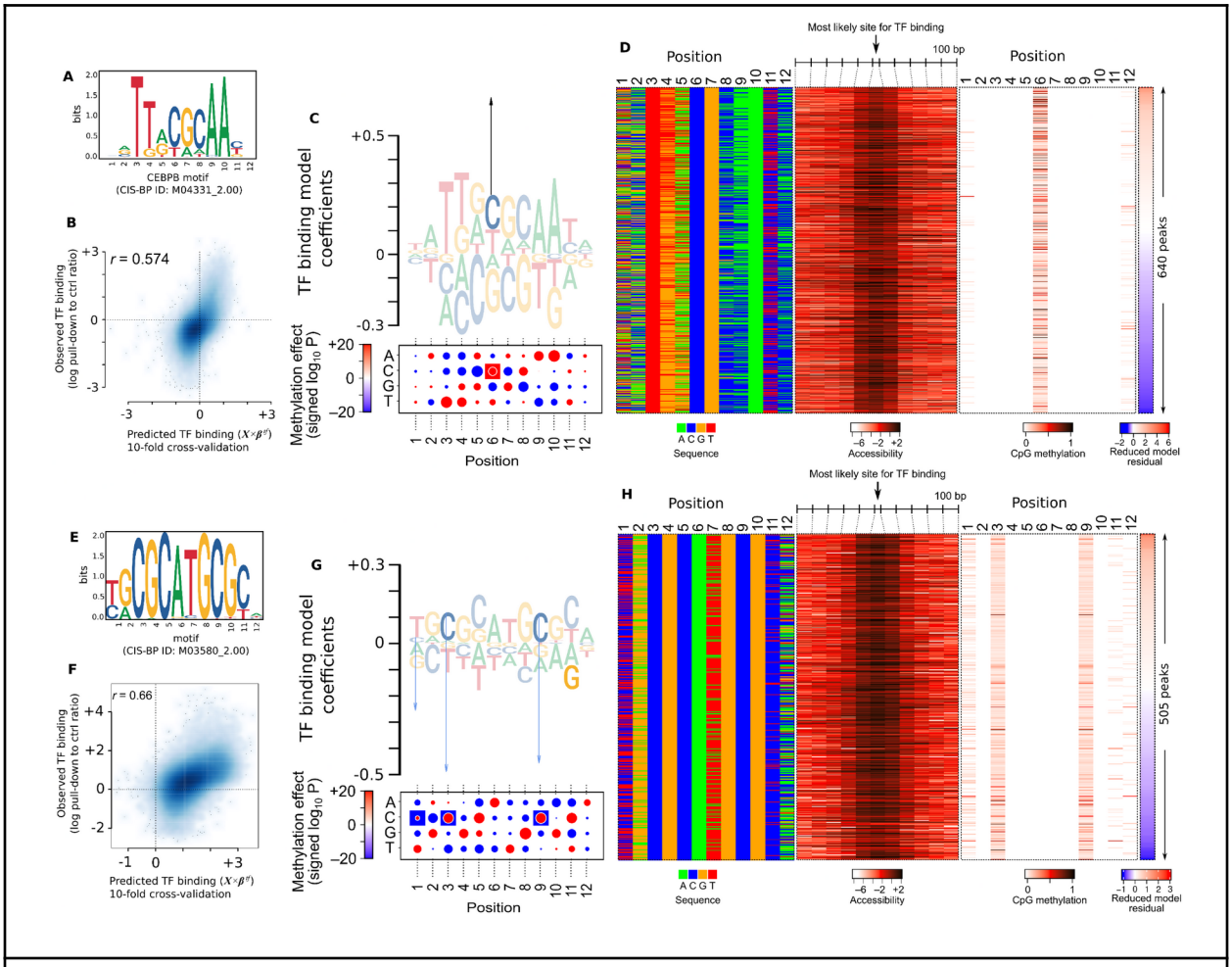


Figure 8: Examples of known TF methyl-binding preferences that were also captured by JAMS. Panels **A-D** correspond to CEBPB, a known methyl-plus TF. Panels **E-H** correspond to NRF1, a known TF whose binding is inhibited by methylation. (**A**) Known motif for CEBPB. (**B**) Scatter plot of JAMS-predicted vs. observed TF binding signal for CEBPB. (**C**) Motif logo and dot plot representations of the sequence/methylation preference of CEBPB as inferred by JAMS (see **Figure 5** for how these representations should be interpreted). (**D**) Heatmap representation of

the sequence, accessibility, and CpG methylation, for a subset of CEBPB peaks that have high DNA accessibility. Peaks are sorted by the residual of a reduced JAMS model that does not use the methylation level for predicting the TF binding signal. (**E-H**) Similar to panels **A**-**D**, but for NFR1.

The above examples suggest that JAMS models are consistent with previously reported methylation preferences of TFs. However, there are only a handful of TFs whose methylation preferences have been validated *in vivo*. Therefore, to systematically evaluate our JAMS-based classification of TFs, we compared our inferred methyl-binding preferences with those obtained from methylation-sensitive SELEX (bisulfite-SELEX) by Yin et al. [5]. Overall, 76 out of the 260 TFs that we studied here were also included in the Yin et al. study (**Table 2**). These included 44 TFs that we classified as MethylMinus based on *in vivo* data; 29 of these TFs (~66%) were also identified as MethylMinus by bisulfite-SELEX, and another 7 TFs (16%) were identified as mixed-effect. This suggests that our approach has ~82% precision for identification of TFs that are negatively affected by CpG methylation in at least one position in their target sequence. On the other hand, out of 39 MethylMinus TFs found by bisulfite-SELEX, 31 were also classified as either MethylMinus or mixed-effect by JAMS, suggesting that ~79% of *in vitro*-observed MethylMinus effects can be captured using *in vivo* data.

Table 2: Contingency table of TF classifications by JAMS (rows) and bisulfite-SELEX [5] (columns).

JAMS \ bisulfite-SELEX	MethylMinus	MethylPlus	MixedEffects	Little effect	Not studied
MethylMinus	29	4	7	4	73
MethylPlus	1	4	0	0	11
MixedEffects	2	1	0	0	1
NoCpG	0	0	0	0	56
NoEffect	7	11	4	2	43

Similarly, out of five JAMS-based MethylPlus TFs that were also studied by Yin et al., four were classified as MethylPlus based on SELEX, suggesting a precision of ~80% [5]. However, despite this high precision, analysis of *in vivo* data appears to have low sensitivity in detecting MethyPlus events, with only 5 out of 20 SELEX-based MethylPlus TFs being identified as either MethylPlus or mixed-effect by JAMS (~25% sensitivity). This observation might reflect the difficulty of modeling MethylPlus effects using *in vivo* data. Nonetheless, we found 11 MethylPlus TFs that were previously unclassified—this is in addition to 73 previously unclassified MethylMinus and one novel mixed-effect TF, highlighting the utility of JAMS models in revealing novel TF methyl preferences (**Table 2**). For example, we show a novel TF methyl preferences for ZKSC1 (**Fig. 9**). The TF-specific logos for all the MethylPlus and MixedEffects inferred by JAMS are shown in **Supplementary figure 6**.

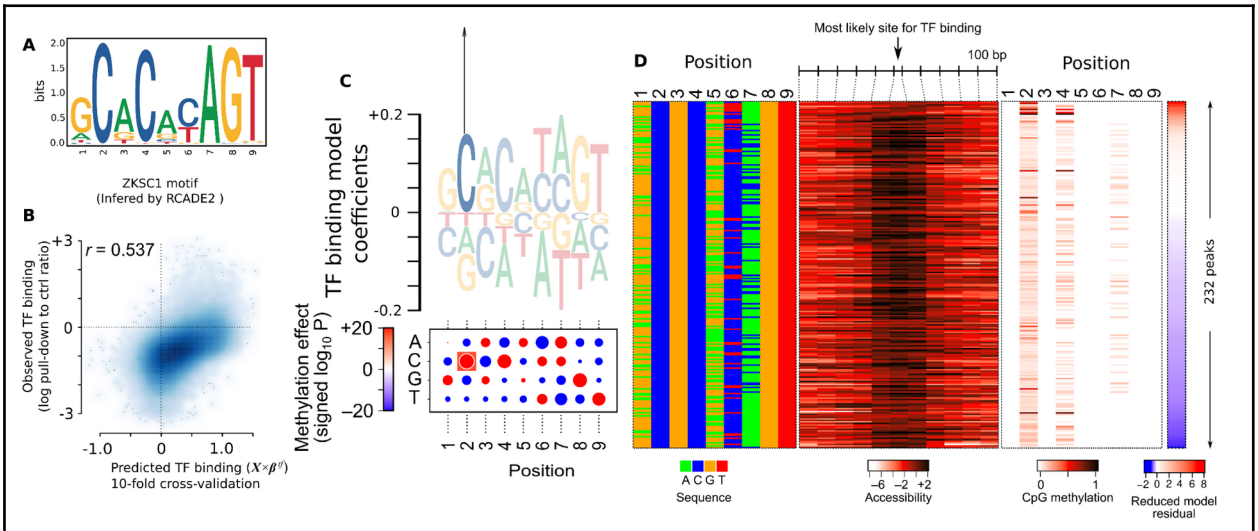


Figure 9: Example of a novel TF methyl-binding preference found by JAMS. (**A**) Known motif for ZKSC1, a C2H2 zinc finger transcription factor (motif inferred by RCADE2 [103,104]). (**B**) Scatter Plot of JAMS-predicted vs. observed TF binding signal. (**C**) Motif logo and dot plot representations of the sequence/methylation preference, as inferred by JAMS. (**D**) Heatmap representation of the sequence, accessibility, and CpG methylation, for a subset of TF peaks that have high DNA accessibility. Peaks are sorted by the residual of a reduced JAMS model that does not use the methylation level on position 2 for predicting the TF binding signal. Note the high level of methylation at position 2 among the peaks that have an excess binding

signal that cannot be explained by this reduced model.

Fig. 10 shows the distribution of different methyl-preferences across main TF families. We noticed that a disproportionately large number of MethylPlus TFs appears to belong to the C2H2-ZF family (also shown in **Table 3**). Specifically, among KRAB-ZF TFs whose binding is significantly affected by methylation, ~24% preferentially bind to methylated CpGs, compared to only ~12% of non-KRAB TFs (Fisher's exact test P<0.009, **Supplementary Table 2**). This is an intriguing observation, given that a majority of KRAB-ZF proteins evolved to specifically bind and repress transposable elements, which largely reside in highly methylated genomic regions [117]. Our observation suggests that many of these proteins preferentially bind to methylated instances of their target sequence, potentially allowing them to distinguish the transposable elements from other genomic regions that contain their preferred binding sequence. In fact, ~56% of all MethylPlus TFs that we identified are KRAB-ZF proteins, suggesting that recognition of methylated transposable elements might have been a primary force in the evolution of methyl-binding TFs.

Overall, our results demonstrate that the methylation preferences of TFs can be reliably inferred from their *in vivo* binding profiles, and provide a comprehensive resource for classification of TF methyl-preferences.

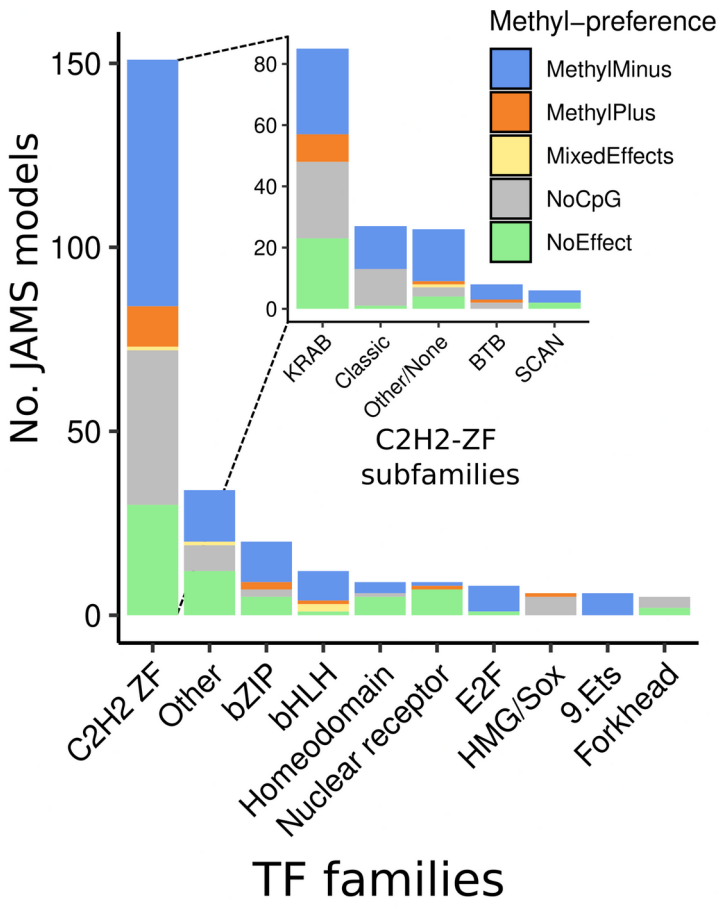


Figure 10: Methylation preferences per TF family. Stacked bar plots showing the distribution of TF methylation preferences inferred with JAMS, grouped by TF families. The inset shows the distribution of methylation preferences for C2H2-ZFP subfamilies.

Table 3: TFs with MethylPlus and MixedEffects methyl-binding preferences, as inferred by JAMS using *in vivo* data. For MixedEffect TFs, both the position at which a positive methylation effect was observed as well as the position with a negative methylation effect are indicated.

TF name (Uniprot entry name)	Gene name (HGNC symbol)	Family	JAMS classification	Effect of methylation by position		Classification by Yin et al., (2017)
				Positive	Negative	
ZN793_HUMAN	ZNF793	C2H2 ZF (KRAB)	MethylPlus	7		
ZKSC1_HUMAN	ZKSCAN1	C2H2 ZF (KRAB+SCAN)	MethylPlus	2		
CEBPB_HUMAN	CEBPB	bZIP	MethylPlus	6		MethylPlus
ZN141_HUMAN	ZNF141	C2H2 ZF (KRAB)	MethylPlus	17		
ZN320_HUMAN	ZNF320	C2H2 ZF (KRAB)	MethylPlus	17		
ZN605_HUMAN	ZNF605	C2H2 ZF (KRAB)	MethylPlus	15		
COT2_HUMAN	NR2F2	Nuclear receptor	MethylPlus	5, 8		
ZN479_HUMAN	ZNF479	C2H2 ZF (KRAB)	MethylPlus	11		
SP1_HUMAN	SP1	C2H2 ZF	MixedEffects	5	8	MethylPlus
ZN490_HUMAN	ZNF490	C2H2 ZF (KRAB)	MethylPlus	7		
ZN506_HUMAN	ZNF506	C2H2 ZF (KRAB)	MethylPlus	5		
ZN417_HUMAN	ZNF417	C2H2 ZF (KRAB)	MethylPlus	16		
USF1_HUMAN	USF1	bHLH	MixedEffects	7	5	MethylMinus
USF2_HUMAN	USF2	bHLH	MixedEffects	7	5	MethylMinus
TCF7_HUMAN	TCF7	HMG/Sox	MethylPlus	2		MethylMinus
KAISO_HUMAN	ZBTB33	C2H2 ZF (BTB)	MethylPlus	5, 7		MethylPlus
TFAP4_HUMAN	TFAP4	bHLH	MethylPlus	7		
NFYB_HUMAN	NFYB	NFYB/HAP3	MixedEffects	9	13	
SCRT1_HUMAN	SCRT1	C2H2 ZF	MethylPlus	3		MethylPlus
CEBPG_HUMAN	CEBPG	bZIP	MethylPlus	6		MethylPlus

CHAPTER 4. DISCUSSION

In this study, we built Joint Accessibility-Methylation-Sequence (JAMS) models to capture the relationship between TF binding and DNA methylation *in vivo*. This method models TF binding as a function of DNA accessibility, sequence and methylation at and around TF binding sites, while separating the background from TF-specific signals. We started by applying this method to CTCF, which revealed that CpG methylation at the 2nd and 12th positions of the CTCF motif is associated with decreased TF binding. This methylation sensitivity is reproduced in multiple cell lines, can be observed even among highly accessible genomic regions, and can explain differential CTCF binding between different cell lines.

As mentioned in the previous chapter, methylation-sensitivity of CTCF has been previously reported [88]. An intriguing observation in this regard was made by Zuo et al., who used a high-throughput *in vitro* method to quantify the effect of CpG methylation on CTCF binding: they found a substantial negative effect of the CpG methylation at the 2nd position of the motif [12], which is also one of the CpG sites we identified. However, we also identified a second CpG site at the 12th position whose methylation reduces CTCF binding, which was not reported by Zuo et al. [12]. Using a likelihood ratio test we showed that the observed effect of methylation at this position cannot be simply explained by its correlation with the first CpG site (**Supplementary Fig. 2**), suggesting that we may have identified a novel CpG methylation effect.

One possible explanation as to why the methylation effect at position 12th could not be observed *in vitro* is that it may reflect the direct competition between CTCF and MBD proteins, with the latter not included in the *in vitro* assay. While JAMS is able to capture the effect of changes in DNA accessibility that result from chromatin remodelling factors recruited by MBD proteins, it currently does not model the direct competition of TFs and MBD proteins. This undetected direct competition between MBD proteins and TFs for the binding sites could affect the interpretation of our model parameters: methylation coefficients obtained by JAMS models should be more

accurately interpreted as the affinity of a TF toward mCpG sites relative to the affinity of MDB proteins.

Accordingly, a positive methylation coefficient means that the TF binds more strongly to the mCpG than MDB proteins do, therefore outcompeting them. This interpretation may in fact explain why a large number of *in vitro*-detected MethylPlus TFs [5] could not be identified by JAMS: even though these TFs can bind to mCpGs *in vitro*, competition with MDB proteins might attenuate this effect *in vivo*. On the other hand, a negative JAMS methylation coefficient could mean that the MDB proteins outcompete the TF *in vivo*, or that the TF simply does not bind to mCpGs even without considering the effect of MBDs. Since the majority of MethylMinus TFs that we identified match *in vitro* observations [5], the latter scenario is likely more prevalent, with most negative coefficients reflecting the TF preference for unmethylated CpG even without considering competition with MBDs. We would like to note the possibility of directly deconvolving these scenarios (i.e. intrinsic preference for unmethylated CpGs vs. competition with MBDs) by including the MBD protein occupancy profiles as additional variables in future versions of the JAMS model.

Another aspect of JAMS models is that it can explain differential TF binding that is associated with changes in DNA methylation, even after controlling for differences in DNA accessibility and sequence. Specifically, we showed that while the large majority of CTCF binding sites are constant across cell types, there are a limited set of methylation-sensitive sites that are highly variable across cell types [88] that can be explained based on our JAMS model of CTCF binding. It is relevant to mention that modelling, and predicting, cell type-specific TF binding is a challenging task and an actively researched problem [43]—Our results support the notion that using CpG methylation data in these methods is an important consideration in order to improve TF binding modelling.

One potential limitation of inference of methyl-preferences of TFs from *in vivo*

data, e.g. using our JAMS models, is that it is difficult to establish the direction of causality: while it is likely that the observed associations reflect the effect of methylation on TF binding, it could also be that they reflect the effect of TF binding on the DNA methylation level [118]. However, TFs that influence DNA methylation most often have an effect on the local neighborhood of their binding sites, which often spans tens of nucleotides [119–121]. JAMS takes into account the neighboring levels of methylation, and tries to identify the site-specific methylation effects within the motif that cannot be explained by (or are independent of) the flanking methylation levels. Furthermore, most known associations between DNA methylation and TF binding entail DNA methylation effect on TF binding rather than the reverse direction [14]; this viewpoint is also supported by high-throughput *in vitro* studies, in which binding site methylation levels are established before measuring TF binding [5,12,13]. We note that the majority of our MethylMinus and MethylPlus findings (>80%) match *in vitro* observations when available [5]; therefore, it is likely that we are observing the effect that CpG methylation has on TF binding, rather than the effect of TF on CpG methylation

Our results represent, to our knowledge, the largest resource for exploring the *in vivo* effect of methylation on TF binding: only a handful of studies have previously investigated methylation preferences of a limited number of TFs *in vivo* while accounting for changes in DNA accessibility. Our results match what has been reported in these studies, e.g. for NRF1, MAX, CEBPB, and KAISO [2,3,9,10,87], but also reveals a substantial number of novel TFs that are affected by CpG methylation Of particular interest, our study revealed a significant number of MethylPlus TFs consistent with *in vitro* studies [5], in stark contrast to previous methods that have attempted to infer the effect of DNA methylation on TF binding [44,90–92].

Finally, we found that a large proportion of the TFs that we identified as MethylPlus belonged to the C2H2-ZF family. C2H2-ZF proteins recognize DNA with an array of zinc fingers (ZFs) [122], with each ZF three or four nucleotides using its base-

contacting residues [123] (the base-contacting residues occupy specific positions in the ZF domain [124]). Identifying the methylation preferences of C2H2-ZF proteins opens the possibility of associating the identity of base-contacting residues to mCpG binding: with a sufficiently large number of methyl-binding ZFs, we could potentially infer a “grammar” of mCpG binding recognition, similar to previous studies that have identified the grammar that connects the identity of base-contacting residues to that of the nucleotides bound by each zinc finger.

CHAPTER 5. CONCLUSIONS AND FUTURE DIRECTIONS

Our study presents, to our knowledge, the first method that quantitatively models the strength of TF-binding as well as background signal using ChIP-seq data, with the goal of deconvolving the effects of sequence, DNA accessibility, and CpG methylation on TF-DNA interactions. We showed that this method captures intra-motif methylation effects, explains differential TF binding between cell lines driven by changes in methylation, and produces TF binding models that are largely consistent with *in vitro* data as well as existing literature. By systematically applying this method to a compendium of 2368 ChIP-seq experiments, we were able to obtain high-confidence models for 260 TFs, representing the largest resource for exploring *in vivo* TF methylation preferences to date.

Future work could focus on using JAMS to understand the processes in which epigenetic changes affects TF binding, especially in conditions where epigenetic remodelling is widespread and frequent, as it is the case in cancer and cellular differentiation. Furthermore, we could associate mCpG recognition, infered by JAMS, to TF features, namely the structure and residue sequence of their DNA binding domains. The C2H2-ZF family is of special interest as they bind to DNA in a modular manner. Finally, albeit our results match current literature, we could improve our method by integrating other relevant genomic features, for example MBD protein profiles.

JAMS is available as a GitHub repository (https://github.com/csglab/JAMS) that includes the computational tools described here as well as the comprehensive dataset of TF methyl-preferences that we have inferred.

CHAPTER 6. REFERENCES

1. 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 Dev.* **2**, 1136–1143 (1988).

2. Mann, I. K. *et al.* CG methylated microarrays identify a novel methylated sequence bound by the CEBPB|ATF4 heterodimer that is active in vivo. *Genome Res.* **23**, 988–997 (2013).

3. Lin, Q. X. X., Rebbani, K., Jha, S. & Benoukraf, T. ZBTB33 (Kaiso) methylated binding sites are associated with primed heterochromatin. *bioRxiv* 585653 (2019) doi:10.1101/585653.

4. Zhu, H., Wang, G. & Qian, J. Transcription factors as readers and effectors of DNA methylation. *Nat. Rev. Genet.* **17**, 551–565 (2016).

5. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017).

6. Du, Q., Luu, P.-L., Stirzaker, C. & Clark, S. J. Methyl-CpG-binding domain proteins: readers of the epigenome. *Epigenomics* **7**, 1051–1073 (2015).

7. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43**, 264–268 (2011).

8. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).

9. Cusack, M. *et al.* Distinct contributions of DNA methylation and histone acetylation to the genomic occupancy of transcription factors. *Genome Res.* **30**, 1393–1406

(2020).

10. Domcke, S. *et al.* Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528**, 575–579 (2015).

11. Wan, J. *et al.* Methylated cis-regulatory elements mediate KLF4-dependent gene transactivation and cell migration. *eLife* **6**, e20068 (2017).

12. Zuo, Z., Roy, B., Chang, Y. K., Granas, D. & Stormo, G. D. Measuring quantitative effects of methylation on transcription factor–DNA binding affinity. *Sci. Adv.* **3**, eaao1799 (2017).

13. Kribelbauer, J. F. *et al.* Quantitative Analysis of the DNA Methylation Sensitivity of Transcription Factor Complexes. *Cell Rep.* **19**, 2383–2395 (2017).

14. Kribelbauer, J. F., Lu, X.-J., Rohs, R., Mann, R. S. & Bussemaker, H. J. Toward a Mechanistic Understanding of DNA Methylation Readout by Transcription Factors. *J. Mol. Biol.* **432**, 1801–1815 (2020).

15. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10**, 252–263 (2009).

16. Akerblom, I. E., Slater, E. P., Beato, M., Baxter, J. D. & Mellon, P. L. Negative regulation by glucocorticoids through interference with a cAMP responsive enhancer. *Science* **241**, 350–353 (1988).

17. Reiter, F., Wienerroither, S. & Stark, A. Combinatorial function of transcription factors and cofactors. *Curr. Opin. Genet. Dev.* **43**, 73–81 (2017).

18. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).

19. Geertz, M., Shore, D. & Maerkl, S. J. Massively parallel measurements of molecular interaction kinetics on a microfluidic platform. *Proc. Natl. Acad. Sci.* **109**, 16540–16545 (2012).

20. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).

21. Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23 (2000).

22. Bailey, T. L. *et al.* MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).

23. Nishida, K., Frith, M. C. & Nakai, K. Pseudocounts for transcription factor binding sites. *Nucleic Acids Res.* **37**, 939–944 (2009).

24. Jolma, A. *et al.* DNA-Binding Specificities of Human Transcription Factors. *Cell* **152**, 327–339 (2013).

25. Barazandeh, M., Lambert, S. A., Albu, M. & Hughes, T. R. Comparison of ChIP-Seq Data and a Reference Motif Set for Human KRAB C2H2 Zinc Finger Proteins. *G3 Genes Genomes Genet.* **8**, 219–229 (2018).

26. Bürglin, T. R. & Affolter, M. Homeodomain proteins: an update. *Chromosoma* **125**, 497–521 (2016).

27. Bürglin, T. R. Homeodomain Subtypes and Functional Diversity. in *A Handbook of Transcription Factors* (ed. Hughes, T. R.) 95–122 (Springer Netherlands, 2011). doi:10.1007/978-90-481-9069-0_5.

28. McGinnis, W., Garber, R. L., Wit-z, J., Kuroiwa, A. & Gehring, W. J. A homologous

protein-coding sequence in drosophila homeotic genes and its conservation in other metazoans. *Cell* **37**, 403–408 (1984).

29. Dunwell, T. L. & Holland, P. W. H. Diversity of human and mouse homeobox gene expression in development and adult tissues. *BMC Dev. Biol.* **16**, 40 (2016).

30. Massari, M. & Murre, C. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Mol. Cell. Biol.* **20**, 429–440 (2000).

31. Murre, C. *et al.* Structure and function of helix-loop-helix proteins. *Biochim. Biophys. Acta BBA - Gene Struct. Expr.* **1218**, 129–135 (1994).

32. Jones, S. An overview of the basic helix-loop-helix proteins. *Genome Biol.* **5**, 226 (2004).

33. Sebé-Pedrós, A., de Mendoza, A., Lang, B. F., Degnan, B. M. & Ruiz-Trillo, I. Unexpected Repertoire of Metazoan Transcription Factors in the Unicellular Holozoan Capsaspora owczarzaki. *Mol. Biol. Evol.* **28**, 1241–1254 (2011).

34. Frietze, S. & Farnham, P. J. Transcription Factor Effector Domains. in *A Handbook of Transcription Factors* (ed. Hughes, T. R.) 261–277 (Springer Netherlands, 2011). doi:10.1007/978-90-481-9069-0_12.

35. Rosenfeld, M. G., Lunyak, V. V. & Glass, C. K. Sensors and signals: a coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response. *Genes Dev.* **20**, 1405–1428 (2006).

36. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* **316**, 1497–1502 (2007).

37. Nakato, R. & Shirahige, K. Recent advances in ChIP-seq analysis: from quality

management to whole-genome annotation. *Brief. Bioinform.* **18**, 279–290 (2017).

38. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**, 669–680 (2009).

39. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).

40. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).

41. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, 1–9 (2008).

42. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).

43. Keilwagen, J., Posch, S. & Grau, J. Accurate prediction of cell type-specific transcription factor binding. *Genome Biol.* **20**, 9 (2019).

44. Xu, T. *et al.* Base-resolution methylation patterns accurately predict transcription factor bindings in vivo. *Nucleic Acids Res.* **43**, 2757–2766 (2015).

45. Dupont, C., Armant, D. R. & Brenner, C. A. Epigenetics: Definition, Mechanisms and Clinical Perspective. *Semin. Reprod. Med.* **27**, 351–357 (2009).

46. Baylin, S. B. & Jones, P. A. Epigenetic determinants of cancer. *Cold Spring Harb. Perspect. Biol.* **8**, a019505 (2016).

47. Campos, E. I. & Reinberg, D. Histones: Annotating Chromatin. *Annu. Rev. Genet.*

43, 559–599 (2009).

48. Buschbeck, M. & Hake, S. B. Variants of core histones and their roles in cell fate decisions, development and cancer. *Nat. Rev. Mol. Cell Biol.* **18**, 299–314 (2017).

49. Lee, C.-K., Shibata, Y., Rao, B., Strahl, B. D. & Lieb, J. D. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat. Genet.* **36**, 900–905 (2004).

50. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019).

51. Boyle, A. P. *et al.* High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* **132**, 311–322 (2008).

52. Sung, M.-H., Baek, S. & Hager, G. L. Genome-wide footprinting: ready for prime time? *Nat. Methods* **13**, 222–228 (2016).

53. Minnoye, L. *et al.* Chromatin accessibility profiling methods. *Nat. Rev. Methods Primer* **1**, 1–24 (2021).

54. Dogan, N. *et al.* Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics Chromatin* **8**, 16 (2015).

55. Krebs, A. R. *et al.* Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Mol. Cell* **67**, 411-422.e4 (2017).

56. Di Stefano, B. *et al.* C/EBPα creates elite cells for iPSC reprogramming by upregulating Klf4 and increasing the levels of Lsd1 and Brd4. *Nat. Cell Biol.* **18**, 371–381 (2016).

57. Galas, D. J. & Schmitz, A. DNAase footprinting a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.* **5**, 3157–3170 (1978).

58. Srivastava, D., Aydin, B., Mazzoni, E. O. & Mahony, S. An interpretable bimodal neural network characterizes the sequence and preexisting chromatin predictors of induced TF binding. *bioRxiv* 672790 (2020) doi:10.1101/672790.

59. Jabbari, K. & Bernardi, G. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene* **333**, 143–149 (2004).

60. Holliday, R. & Pugh, J. E. DNA modification mechanisms and gene activity during development. *Science* **187**, 226–232 (1975).

61. Lyko, F. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nat. Rev. Genet.* **19**, 81–92 (2018).

62. Wu, X. & Zhang, Y. TET-mediated active DNA demethylation: mechanism, function and beyond. *Nat. Rev. Genet.* **18**, 517–534 (2017).

63. Pfister, S. X. & Ashworth, A. Marked for death: targeting epigenetic changes in cancer. *Nat. Rev. Drug Discov.* **16**, 241–263 (2017).

64. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).

65. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinforma. Oxf. Engl.* **27**, 1571–1572 (2011).

66. Wreczycka, K. *et al.* Strategies for analyzing bisulfite sequencing data. *J. Biotechnol.* **261**, 105–115 (2017).

67. Ziller, M. J., Hansen, K. D., Meissner, A. & Aryee, M. J. Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nat. Methods* **12**, 230–232 (2015).

68. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).

69. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

70. Newell-Price, J., Clark, A. J. L. & King, P. DNA Methylation and Silencing of Gene Expression. *Trends Endocrinol. Metab.* **11**, 142–148 (2000).

71. Nan, X. *et al.* Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393**, 386–389 (1998).

72. Haberland, M., Montgomery, R. L. & Olson, E. N. The many roles of histone deacetylases in development and physiology: implications for disease and therapy. *Nat. Rev. Genet.* **10**, 32–42 (2009).

73. Liu, Y., Zhang, X., Blumenthal, R. M. & Cheng, X. A Common Mode of Recognition for Methylated CpG. *Trends Biochem. Sci.* **38**, 177–183 (2013).

74. Hu, S. *et al.* DNA methylation presents distinct binding sites for human transcription factors. *eLife* **2**, e00726 (2013).

75. Iguchi-Ariga, S. M. & Schaffner, W. CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. *Genes Dev.* **3**, 612–619 (1989).

76. Lazarovici, A. *et al.* Probing DNA shape and methylation state on a genomic scale

with DNase I. *Proc. Natl. Acad. Sci.* **110**, 6376–6381 (2013).

77. Rohs, R. *et al.* The role of DNA shape in protein–DNA recognition. *Nature* **461**, 1248–1253 (2009).

78. Moore, L. D., Le, T. & Fan, G. DNA Methylation and Its Basic Function. *Neuropsychopharmacology* **38**, 23–38 (2013).

79. Fuks, F., Hurd, P. J., Deplus, R. & Kouzarides, T. The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase. *Nucleic Acids Res.* **31**, 2305–2312 (2003).

80. Jones, P. L. *et al.* Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat. Genet.* **19**, 187–191 (1998).

81. Fuks, F., Burgers, W. A., Brehm, A., Hughes-Davies, L. & Kouzarides, T. DNA methyltransferase Dnmt1 associates with histone deacetylase activity. *Nat. Genet.* **24**, 88–91 (2000).

82. Geiman, T. M. *et al.* DNMT3B interacts with hSNF2H chromatin remodeling enzyme, HDACs 1 and 2, and components of the histone methylation system. *Biochem. Biophys. Res. Commun.* **318**, 544–555 (2004).

83. Du, J., Johnson, L. M., Jacobsen, S. E. & Patel, D. J. DNA methylation pathways and their crosstalk with histone methylation. *Nat. Rev. Mol. Cell Biol.* **16**, 519–532 (2015).

84. Jia, D., Jurkowska, R. Z., Zhang, X., Jeltsch, A. & Cheng, X. Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature* **449**, 248–251 (2007).

85. Clark, S. J. *et al.* scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).

86. Kelly, T. K. *et al.* Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.* **22**, 2497–2506 (2012).

87. Prokhortchouk, A. *et al.* The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. *Genes Dev.* **15**, 1613–1618 (2001).

88. Maurano, M. T. *et al.* Role of DNA Methylation in Modulating Transcription Factor Occupancy. *Cell Rep.* **12**, 1184–1195 (2015).

89. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21**, 447–455 (2011).

90. Ngo, V., Wang, M. & Wang, W. Finding de novo methylated DNA motifs. *Bioinforma. Oxf. Engl.* **35**, 3287–3293 (2019).

91. Viner, C. *et al.* Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet. *bioRxiv* 043794 (2016) doi:10.1101/043794.

92. Grau, J., Schmidt, F. & Schulz, M. H. Widespread effects of DNA methylation and intra-motif dependencies revealed by novel transcription factor binding models. *bioRxiv* 2020.10.21.348193 (2020) doi:10.1101/2020.10.21.348193.

93. Lin, Q. X. X., Thieffry, D., Jha, S. & Benoukraf, T. TFregulomeR reveals transcription factors' context-specific features and functions. *Nucleic Acids Res.* **48**, e10 (2020).

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

95. Najafabadi, H. S., Albu, M. & Hughes, T. R. Identification of C2H2-ZF binding preferences from ChIP-seq data using RCADE. *Bioinformatics* **31**, 2879–2881 (2015).

96. Schmitges, F. W. *et al.* Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res.* **26**, 1742–1752 (2016).

97. Imbeault, M., Helleboid, P.-Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017).

98. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

99. Navarro Gonzalez, J. *et al.* The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res.* **49**, D1046–D1057 (2021).

100. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).

101. Smit, A., Hubley, R. & Green, P. RepeatMasker. http://www.repeatmasker.org/ (2013).

102. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).

103. Dogan, B., Kailasam, S., Corchado, A. H., Nikpoor, N. & Najafabadi, H. S. A domain-resolution map of in vivo DNA binding reveals the regulatory consequences of somatic mutations in zinc finger transcription factors. *bioRxiv* 630756 (2020) doi:10.1101/630756.

104. Dogan, B. & Najafabadi, H. S. Computational Methods for Analysis of the DNA-Binding Preferences of Cys2His2 Zinc-Finger Proteins. *Methods Mol. Biol. Clifton NJ* **1867**, 15–28 (2018).

105. Lambert, S. A., Albu, M., Hughes, T. R. & Najafabadi, H. S. Motif comparison based on similarity of binding affinity profiles. *Bioinformatics* btw489 (2016) doi:10.1093/bioinformatics/btw489.

106. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinforma.* **47**, (2014).

107. Pohl, A. & Beato, M. bwtool: a tool for bigWig files. *Bioinformatics* **30**, 1618–1619 (2014).

108. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*. (Springer-Verlag, 2002). doi:10.1007/978-0-387-21706-2.

109. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

110. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

111. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645–3647 (2017).

112. Filippova, G. N. *et al.* An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol. Cell. Biol.* **16**, 2802–2813 (1996).

113. Holwerda, S. J. B. & de Laat, W. CTCF: the protein, the binding partners, the binding sites and their chromatin loops. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **368**, 20120369 (2013).

114. Hansen, A. S., Pustova, I., Cattoglio, C., Tjian, R. & Darzacq, X. CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *eLife* **6**, (2017).

115. Lakisic, G. *et al.* Role of the BAHD1 Chromatin-Repressive Complex in Placental Development and Regulation of Steroid Metabolism. *PLOS Genet.* **12**, e1005898 (2016).

116. Auerbach, R. K. *et al.* Mapping accessible chromatin regions using Sono-Seq. *Proc. Natl. Acad. Sci.* **106**, 14926–14931 (2009).

117. Thomas, J. H. & Schneider, S. Coevolution of retroelements and tandem zinc finger genes. *Genome Res.* **21**, 1800–1812 (2011).

118. Ambrosi, C., Manzo, M. & Baubec, T. Dynamics and Context-Dependent Roles of DNA Methylation. *J. Mol. Biol.* **429**, 1459–1475 (2017).

119. de la Rica, L. *et al.* PU.1 target genes undergo Tet2-coupled demethylation and DNMT3b-mediated methylation in monocyte-to-osteoclast differentiation. *Genome Biol.* **14**, 1–21 (2013).

120. Guilhamon, P. *et al.* Meta-analysis of IDH-mutant cancers identifies EBF1 as an interaction partner for TET2. *Nat. Commun.* **4**, 2166 (2013).

121. Suzuki, T. *et al.* RUNX1 regulates site specificity of DNA demethylation by recruitment of DNA demethylation machineries in hematopoietic cells. *Blood Adv.* **1**, 1699–1711 (2017).

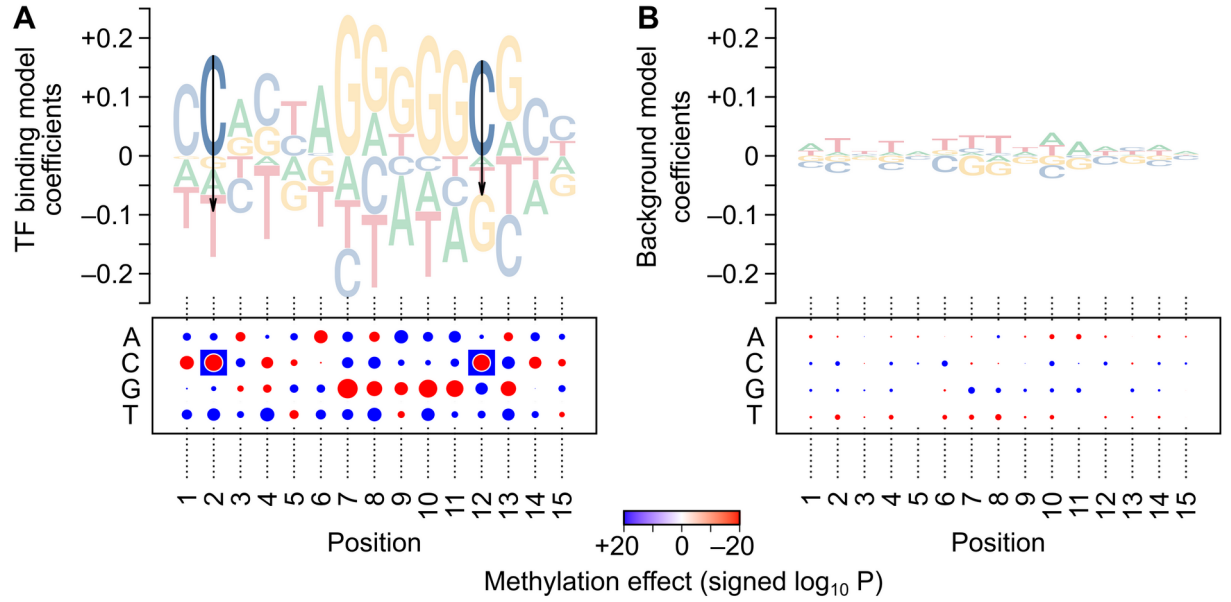
122. Garton, M. *et al.* A structural approach reveals how neighbouring C2H2 zinc fingers influence DNA binding specificity. *Nucleic Acids Res.* **43**, 9147–9157 (2015).

123. Pavletich, N. P. & Pabo, C. O. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science* **252**, 809–817 (1991).

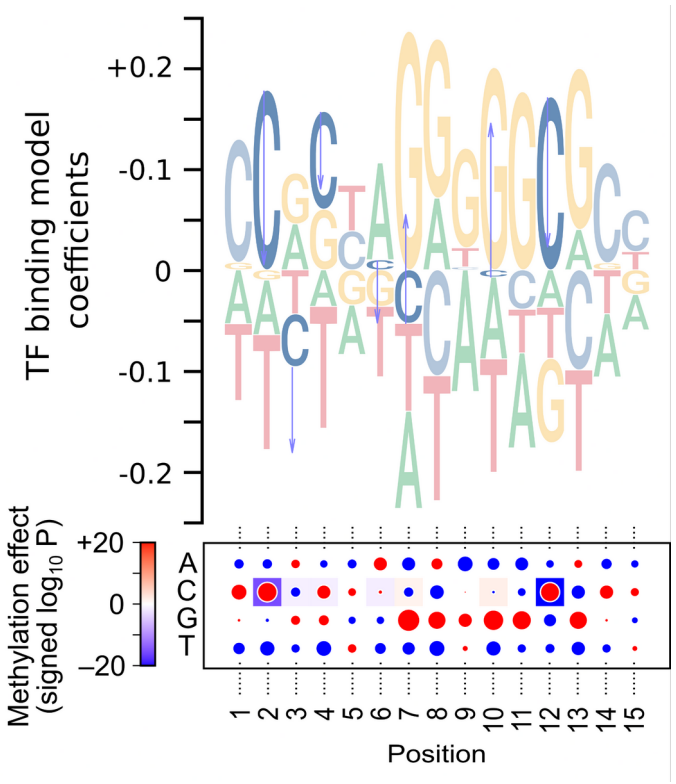
124. Wolfe, S. A., Nekludova, L. & Pabo, C. O. DNA Recognition by Cys2His2 Zinc Finger Proteins. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 183–212 (2000).

APPENDICES

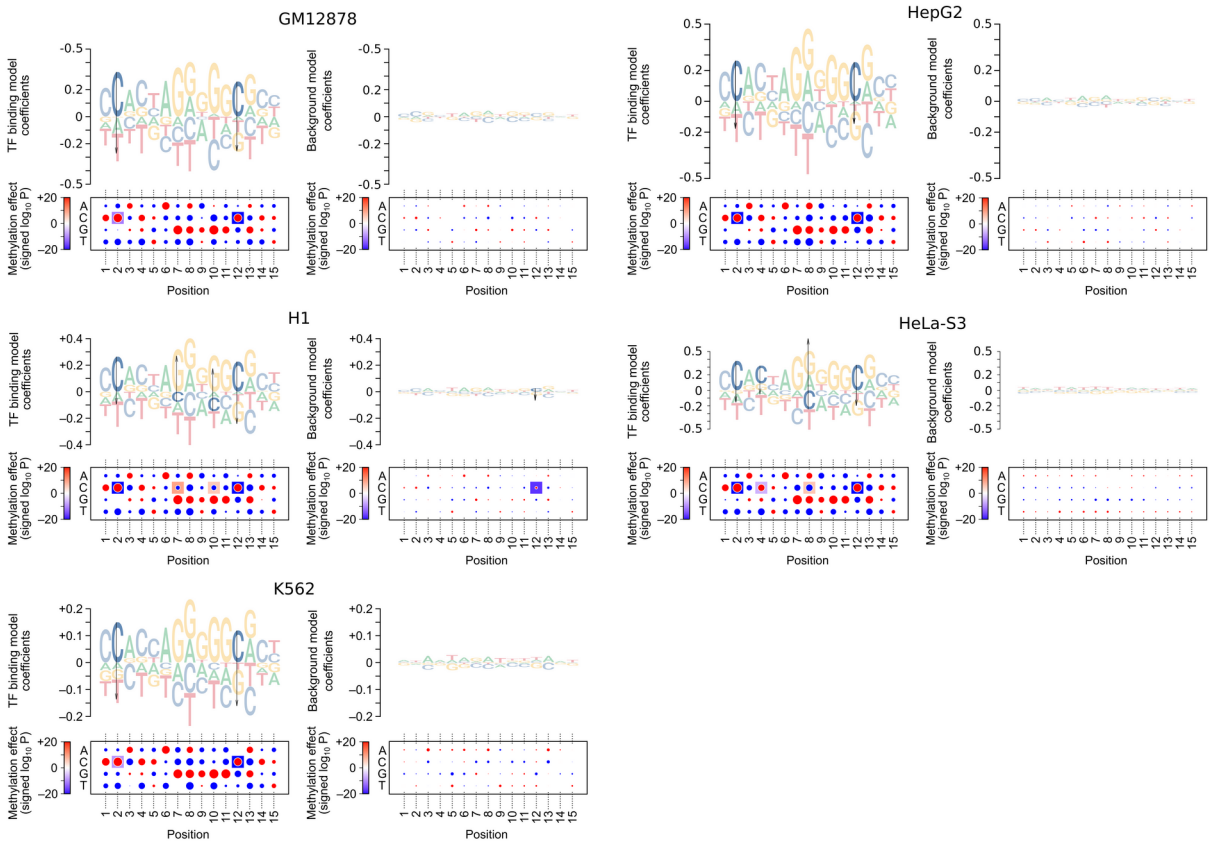
Supplementary figures



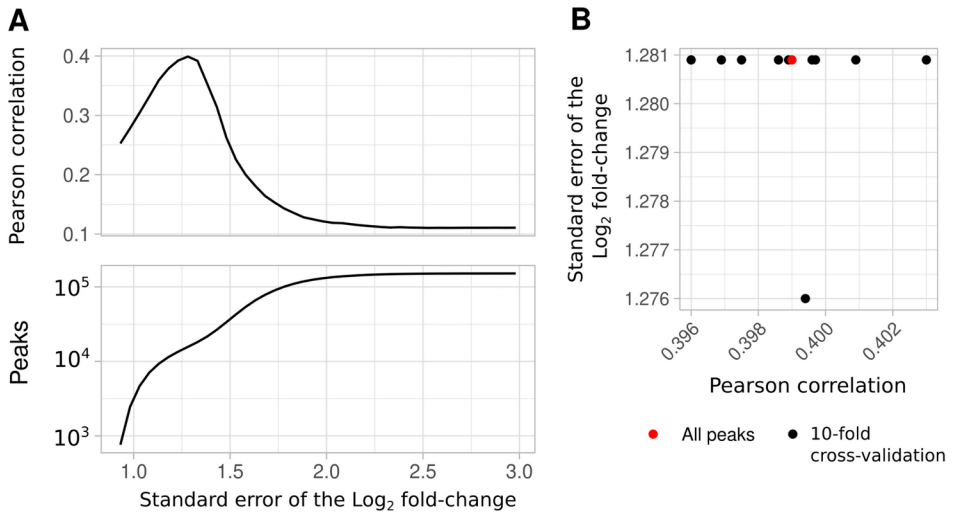
Supplementary Figure 1: Background coefficients for CTCF in HEK293 cells. (**A**) Motif logo and dot plot representations of the sequence/methylation preference of the TF-specific (**A**) and background (**B**) signals. The logo (top) shows methylation coefficients as arrows, with the arrow length proportional to the mean estimate of methylation effect. The heatmap (bottom) shows the magnitude of the preference for each nucleotide at each position using the size of the dots, with red and blue representing positive and negative coefficients, respectively. The signed logarithm of P-value of the methylation coefficient is shown using the color of the squares around the dots, with red and blue corresponding to increased or decreased binding to methylated C, respectively (only significant methylation coefficients at FDR<1×10–5 are shown).



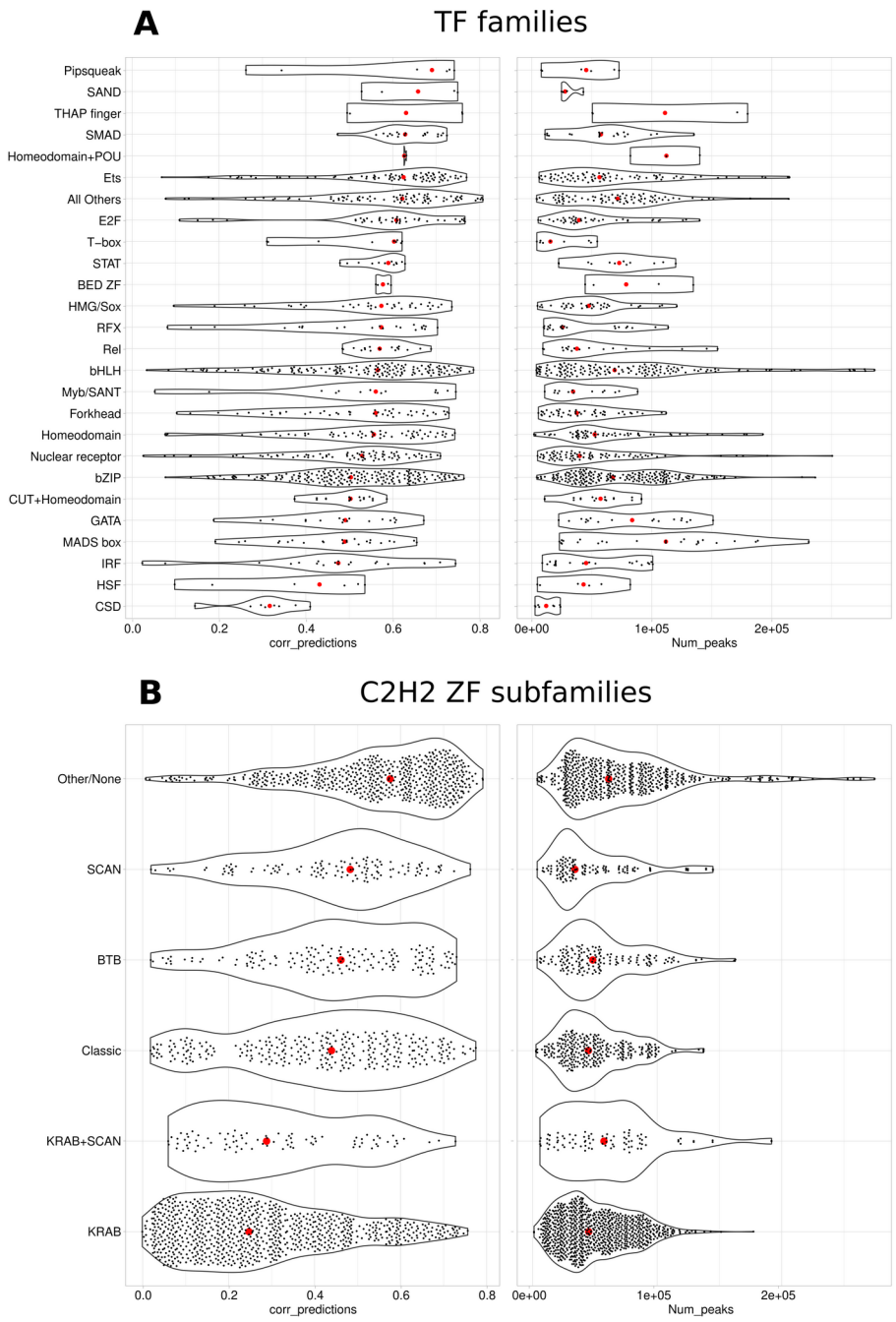
Supplementary Figure 2: Likelihood ratio test per position to identify CTCF binding site positions with significant methylation effects. For each position of the binding site, a reduced model was trained, each exclusing methylation of that position from the predictive variables. Then, each of these reduced models were compared to the whole CTCF JAMS model using a likelihood ratio test (LRT). The p-values obtained from the LRT are shown as the color of the squares. The effect sizes for the bases and methylation are obtained from the full CTCF JAMS model. Significant LRT p-values indicate that removing the methylation of the corresponding position from the model reduces the goodness of fit.



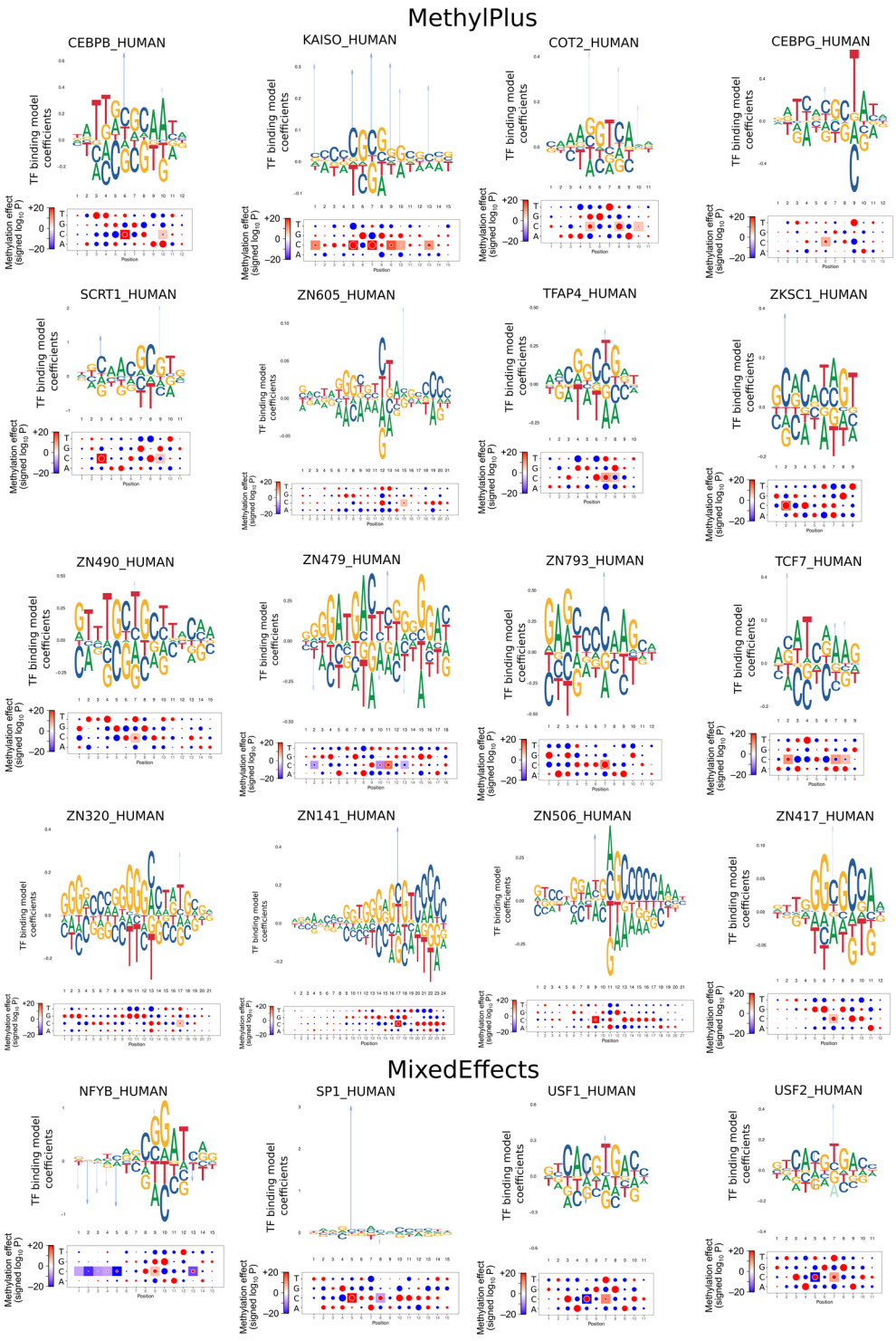
Supplementary Figure 3: JAMS coefficients for CTCF across different cell lines. Motif logs and dot plot representations follow the same format as described in **Fig. 5**. TF-specific and background model coefficients are shown side-by-side for each of the six cell lines (GM12878, HepG2, H1, HeLa-S3, and K562).



Supplementary Figure 4: Calculating LFC threshold. (**A**) Pearson correlation between the predicted and observed change in CTCF binding, after filtering the CTCF peaks based on different cutoffs for standard errors of the LFC of pull-down/control ratio. An optimal threshold is observed at LFC SE = 1.28. (**B**) To discard the possibility of overfitting of the threshold, different optimal thresholds were calculated using a 10-fold cross-validation approach. The LFC SE threshold obtained by using all peaks is similar to the threshold obtained with cross-validation and leads to a similar correlation between the predicted and observed change in CTCF binding.



Supplementary figure 5: MethylChIP results by TF families. Violin plots showing the Pearson correlation between observed and predicted pull-down tag density (left) and number of peaks used to train the GLM (right), shown separately for each TF families (**A**) and C2H2-ZF subfamilies (**B**). The red dots represent the median.



Supplementary figure 6: TF-specific motif logos for MethylPlus and MixedEffects TFs identified by JAMS. Motif logo and dot plot representations follow the same formatting as **Fig. 5**.

Supplementary tables

Supplementary Table 1: GEO and ENCODE FASTQ identification numbers per cell lines for the data that were used to train JAMS CTCF models

Cell line	CTCF ChIP-seq	Input ChIP-seq	Pull-down and Input ChIP-seq lab source	WGBS	DNase-seq
HEK293	GSM2026781	HEK293_input_Hughes_7	Hughes lab, Toronto	GSM1254259	ENCFF148BGE
H1	ENCFF000ONR ENCFF000OOF	ENCFF000OSS ENCFF000OSP	Richard Myers, HAIB	ENCFF677BSB ENCFF800KIP ENCFF311PSV ENCFF335TUD	ENCFF131HMO
GM12878	ENCFF000ARP ENCFF000ARV	ENCFF621LOE ENCFF904LCW	Bradley Bernstein, Broad	ENCFF585BXF, ENCFF851HAT ENCFF798RSS ENCFF113KRQ	ENCFF743ULW
HeLa-S3	ENCFF000BAS ENCFF000BAT	ENCFF000BAO ENCFF000BAU	Bradley Bernstein, Broad	ENCFF953ELH ENCFF718LOZ ENCFF751KHK ENCFF192ITK	ENCFF256QQH
K562	ENCFF000PYD ENCFF000PYJ	ENCFF000QFI ENCFF000QFJ	Richard Myers, HAIB	ENCFF413KHN ENCFF567DAI ENCFF336KJH ENCFF585HYM	ENCFF413AHU
HepG2	ENCFF000PHE ENCFF000PHG	ENCFF000POU ENCFF000POV	Richard Myers, HAIB	ENCFF406GDR ENCFF508BUS ENCFF706BRZ ENCFF220NMH	ENCFF577SOF

Supplementary Table 2: TFs with high quality JAMS model, stratified by methyl-binding preference.

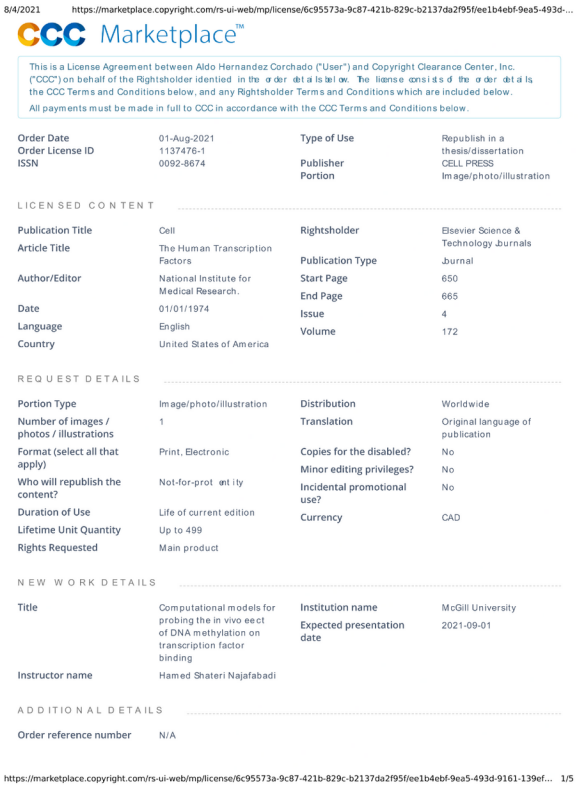
JAMS-inferred methyl preference	TFs in the ZF-KRAB family	Total TFs
No effect of CpG methylation or no CpG	48	123
MethylPlus	9	16
MethylMinus	28	117
MixedEffects	0	4

Supplementary Table 3: Model matrix used to compare count ratios with DESeq2.

	(Intercept)	sample_2	sample_3	sample_4	read_type	cell_line_1
cell_line_2_replicate_1.pulldown	1	0	0	0	0	0
cell_line_2_replicate_2.pulldown	1	1	0	0	0	0
cell_line_1_replicate_1.pulldown	1	0	1	0	0	0
cell_line_1_replicate_2.pulldown	1	0	0	1	0	0
cell_line_2_replicate_1.control	1	0	0	0	1	0
cell_line_2_replicate_2.control	1	1	0	0	1	0
cell_line_1_replicate_1.control	1	0	1	0	1	1
cell_line_1_replicate_2.control	1	0	0	1	1	1

Copyright clearance

Copyright Clearance Center's License agreement for Figure 1 and 3.



8/4/2021 https://marketplace.copyright.com/rs-ui-web/mp/license/6c95573a-9c87-421b-829c-b2137da2f95f/ee1b4ebf-9ea5-493d-...

CCC Marketplace™

This is a License Agreement between Aldo Hernandez Corchado ("User") and Copyright Clearance Center, Inc. ("CCC") on behalf of the Rightsholder identied in the order details below. The license consists of the order details, the CCC Terms and Conditions below, and any Rightsholder Terms and Conditions which are included below.

All payments must be made in full to CCC in accordance with the CCC Terms and Conditions below.

Order Date	01-Aug-2021	Type of Use	Republish in a thesis/dissertation
Order License ID	1137476-1		
ISSN	0092-8674	Publisher	CELL PRESS
		Portion	Image/photo/illustration

LICENSED CONTENT

Publication Title	Cell	Rightsholder	Elsevier Science & Technology Journals
Article Title	The Human Transcription Factors	Publication Type	Journal
Author/Editor	National Institute for Medical Research.	Start Page	650
		End Page	665
Date	01/01/1974	Issue	4
Language	English	Volume	172
Country	United States of America		

REQUEST DETAILS

Portion Type	Image/photo/illustration	Distribution	Worldwide
Number of images / photos / illustrations	1	Translation	Original language of publication
Format (select all that apply)	Print, Electronic	Copies for the disabled?	No
		Minor editing privileges?	No
Who will republish the content?	Not-for-prot ent ity	Incidental promotional use?	No
Duration of Use	Life of current edition	Currency	CAD
Lifetime Unit Quantity	Up to 499		
Rights Requested	Main product		

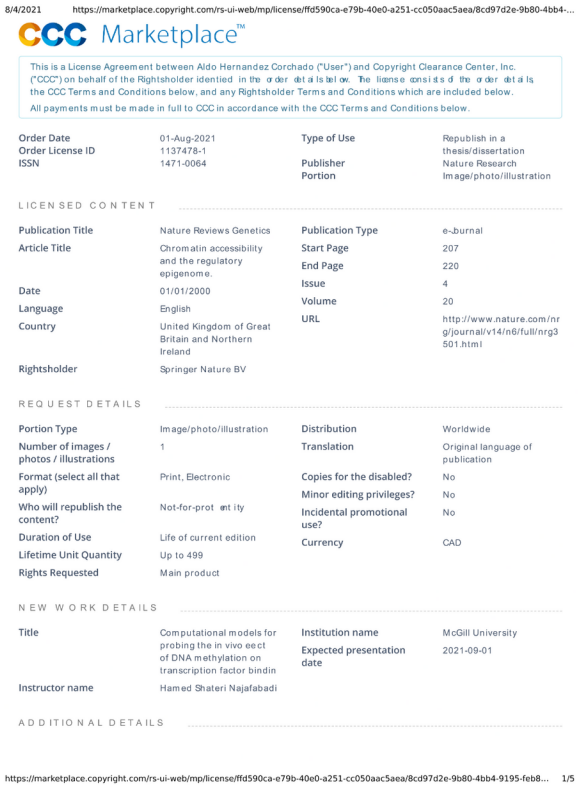
NEW WORK DETAILS

Title	Computational models for probing the in vivo eect of DNA methylation on transcription factor binding	Institution name	McGill University
		Expected presentation date	2021-09-01
Instructor name	Hamed Shateri Najafabadi		

ADDITIONAL DETAILS

Order reference number	N/A

https://marketplace.copyright.com/rs-ui-web/mp/license/6c95573a-9c87-421b-829c-b2137da2f95f/ee1b4ebf-9ea5-493d-9161-139ef... 1/5



8/4/2021 https://marketplace.copyright.com/rs-ui-web/mp/license/ffd590ca-e79b-40e0-a251-cc050aac5aea/8cd97d2e-9b80-4bb4-...

CCC Marketplace™

This is a License Agreement between Aldo Hernandez Corchado ("User") and Copyright Clearance Center, Inc. ("CCC") on behalf of the Rightsholder identied in the order details below. The license consists of the order details, the CCC Terms and Conditions below, and any Rightsholder Terms and Conditions which are included below.

All payments must be made in full to CCC in accordance with the CCC Terms and Conditions below.

Order Date	01-Aug-2021	Type of Use	Republish in a thesis/dissertation
Order License ID	1137478-1		
ISSN	1471-0064	Publisher	Nature Research
		Portion	Image/photo/illustration

LICENSED CONTENT

Publication Title	Nature Reviews Genetics	Publication Type	e-Journal
Article Title	Chromatin accessibility and the regulatory epigenome.	Start Page	207
		End Page	220
		Issue	4
Date	01/01/2000	Volume	20
Language	English	URL	http://www.nature.com/nrg/journal/v14/n6/full/nrg3501.html
Country	United Kingdom of Great Britain and Northern Ireland		
Rightsholder	Springer Nature BV		

REQUEST DETAILS

Portion Type	Image/photo/illustration	Distribution	Worldwide
Number of images / photos / illustrations	1	Translation	Original language of publication
Format (select all that apply)	Print, Electronic	Copies for the disabled?	No
		Minor editing privileges?	No
Who will republish the content?	Not-for-prot ent ity	Incidental promotional use?	No
Duration of Use	Life of current edition	Currency	CAD
Lifetime Unit Quantity	Up to 499		
Rights Requested	Main product		

NEW WORK DETAILS

Title	Computational models for probing the in vivo eect of DNA methylation on transcription factor bindin	Institution name	McGill University
		Expected presentation date	2021-09-01
Instructor name	Hamed Shateri Najafabadi		

ADDITIONAL DETAILS

https://marketplace.copyright.com/rs-ui-web/mp/license/ffd590ca-e79b-40e0-a251-cc050aac5aea/8cd97d2e-9b80-4bb4-9195-feb8... 1/5

The requesting person / organization to appear on the license	Aldo Hernandez Corchado

REUSE CONTENT DETAILS

Title, description or numeric reference of the portion(s)	Figure 1 A. The Human Transcription Factor Repertoire	Title of the article/chapter the portion is from	The Human Transcription Factors
Editor of portion(s)	Weirauch, Matthew T.; Hughes, Timothy R.; Taipale, Jussi; Chen, Xiaoting; Albu, Mihai; Yin, Yimeng; Das, Pratyush K.; Campitelli, Laura F.; Jolma, Arttu; Lambert, Samuel A.	Author of portion(s)	Weirauch, Matthew T.; Hughes, Timothy R.; Taipale, Jussi; Chen, Xiaoting; Albu, Mihai; Yin, Yimeng; Das, Pratyush K.; Campitelli, Laura F.; Jolma, Arttu; Lambert, Samuel A.
Volume of serial or monograph	172	Issue, if republishing an article from a serial	4
Page or page range of portion	651-651	Publication date of portion	2018-02-08

RIGHTSHOLDER TERMS AND CONDITIONS

Elsevier publishes Open Access articles in both its Open Access journals and via its Open Access articles option in subscription journals, for which an author selects a user license permitting certain types of reuse without permission. Before proceeding please check if the article is Open Access on **http://www.sciencedirect.com** and refer to the user license for the individual article. Any reuse not included in the user license terms will require permission. You must always fully and appropriately credit the author and source. If any part of the material to be used (for example, figures) has appeared in the Elsevier publication for which you are seeking permission, with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder. Please contact **permissions@elsevier.com** with any queries.

CCC Terms and Conditions

1. Description of Service; Defined Terms. This Republication License enables the User to obtain licenses for republication of one or more copyrighted works as described in detail on the relevant Order Confirmation (the "Work(s)"). Copyright Clearance Center, Inc. ("CCC") grants licenses through the Service on behalf of the rightsholder identified on the Order Confirmation (the "Rightsholder"). "Republication", as used herein, generally means the inclusion of a Work, in whole or in part, in a new work or works, also as described on the Order Confirmation. "User", as used herein, means the person or entity making such republication.

2. The terms set forth in the relevant Order Confirmation, and any terms set by the Rightsholder with respect to a particular Work, govern the terms of use of Works in connection with the Service. By using the Service, the person transacting for a republication license on behalf of the User represents and warrants that he/she/it (a) has been duly authorized by the User to accept, and hereby does accept, all such terms and conditions on behalf of User, and (b) shall inform User of all such terms and conditions. In the event such person is a "freelancer" or other third party independent of User and CCC, such party shall be deemed jointly a "User" for purposes of these terms and conditions. In any event, User shall be deemed to have accepted and agreed to all such terms and conditions if User republishes the Work in any fashion.

3. Scope of License; Limitations and Obligations.

   3.1. All Works and all rights therein, including copyright rights, remain the sole and exclusive property of the Rightsholder. The license created by the exchange of an Order Confirmation (and/or any invoice) and payment by User of the full amount set forth on that document includes only those rights expressly set

forth in the Order Confirmation and in these terms and conditions, and conveys no other rights in the Work(s) to User. All rights not expressly granted are hereby reserved.

   3.2. General Payment Terms: You may pay by credit card or through an account with us payable at the end of the month. If you and we agree that you may establish a standing account with CCC, then the following terms apply: Remit Payment to: Copyright Clearance Center, 29118 Network Place, Chicago, IL 60673-1291. Payments Due: Invoices are payable upon their delivery to you (or upon our notice to you that they are available to you for downloading). After 30 days, outstanding amounts will be subject to a service charge of 1-1/2% per month or, if less, the maximum rate allowed by applicable law. Unless otherwise specifically set forth in the Order Confirmation or in a separate written agreement signed by CCC, invoices are due and payable on "net 30" terms. While User may exercise the rights licensed immediately upon issuance of the Order Confirmation, the license is automatically revoked and is null and void as if it had never been issued, if complete payment for the license is not received on a timely basis either from User directly or through a payment agent, such as a credit card company.

   3.3. Unless otherwise provided in the Order Confirmation, any grant of rights to User (i) is "one-time" (including the editions and product family specified in the license), (ii) is non-exclusive and non-transferable and (iii) is subject to any and all limitations and restrictions (such as, but not limited to, limitations on duration of use or circulation) included in the Order Confirmation or invoice and/or in these terms and conditions. Upon completion of the licensed use, User shall either secure a new permission for further use of the Work(s) or immediately cease any new use of the Work(s) and shall render inaccessible (such as by deleting or by removing or severing links or other locators) any further copies of the Work (except for copies printed on paper in accordance with this license and still in User's stock at the end of such period).

   3.4. In the event that the material for which a republication license is sought includes third party materials (such as photographs, illustrations, graphs, inserts and similar materials) which are identified in such material as having been used by permission, User is responsible for identifying, and seeking separate licenses (under this Service or otherwise) for, any of such third party materials; without a separate license, such third party materials may not be used.

   3.5. Use of proper copyright notice for a Work is required as a condition of any license granted under the Service. Unless otherwise provided in the Order Confirmation, a proper copyright notice will read substantially as follows: "Republished with permission of [Rightsholder's name], from [Work's title, author, volume, edition number and year of copyright]; permission conveyed through Copyright Clearance Center, Inc." Such notice must be provided in a reasonably legible font size and must be placed either immediately adjacent to the Work as used (for example, as part of a by-line or footnote but not as a separate electronic link) or in the place where substantially all other credits or notices for the new work containing the republished Work are located. Failure to include the required notice results in loss to the Rightsholder and CCC, and the User shall be liable to pay liquidated damages for each such failure equal to twice the use fee specified in the Order Confirmation, in addition to the use fee itself and any other fees and charges specified.

   3.6. User may only make alterations to the Work if and as expressly set forth in the Order Confirmation. No Work may be used in any way that is defamatory, violates the rights of third parties (including such third parties' rights of copyright, privacy, publicity, or other tangible or intangible property), or is otherwise illegal, sexually explicit or obscene. In addition, User may not conjoin a Work with any other material that may result in damage to the reputation of the Rightsholder. User agrees to inform CCC if it becomes aware of any infringement of any rights in a Work and to cooperate with any reasonable request of CCC or the Rightsholder in connection therewith.

4. Indemnity. User hereby indemnifies and agrees to defend the Rightsholder and CCC, and their respective employees and directors, against all claims, liability, damages, costs and expenses, including legal fees and expenses, arising out of any use of a Work beyond the scope of the rights granted herein, or any use of a Work which has been altered in any unauthorized way by User, including claims of defamation or infringement of rights of copyright, publicity, privacy or other tangible or intangible property.

The requesting person / organization to appear on the license	Aldo Hernandez Corchado

REUSE CONTENT DETAILS

Title, description or numeric reference of the portion(s)	Figure 1, A continuum of accessibility states broadly reects the distribution of chromatin dynamics across the genome.	Title of the article/chapter the portion is from	Chromatin accessibility and the regulatory epigenome.
Editor of portion(s)	Greenleaf, William J; Klemm, Sandy L.; Shipony, Zohar	Author of portion(s)	Greenleaf, William J; Klemm, Sandy L.; Shipony, Zohar
Volume of serial or monograph	20	Publication date of portion	2019-04-01
Page or page range of portion	208-208		

RIGHTSHOLDER TERMS AND CONDITIONS

If you are placing a request on behalf of/for a corporate organization, please use RightsLink. For further information visit **http://www.nature.com/reprints/permission-requests.html** and **https://www.springer.com/gp/rights-permissions/obtaining-permissions/882**. If the content you are requesting to reuse is under a CC-BY 4.0 licence (or previous version), you do not need to seek permission from Springer Nature for this reuse as long as you provide appropriate credit to the original publication. https://creativecommons.org/licenses/by/4.0/

CCC Terms and Conditions

1. Description of Service; Defined Terms. This Republication License enables the User to obtain licenses for republication of one or more copyrighted works as described in detail on the relevant Order Confirmation (the "Work(s)"). Copyright Clearance Center, Inc. ("CCC") grants licenses through the Service on behalf of the rightsholder identified on the Order Confirmation (the "Rightsholder"). "Republication", as used herein, generally means the inclusion of a Work, in whole or in part, in a new work or works, also as described on the Order Confirmation. "User", as used herein, means the person or entity making such republication.

2. The terms set forth in the relevant Order Confirmation, and any terms set by the Rightsholder with respect to a particular Work, govern the terms of use of Works in connection with the Service. By using the Service, the person transacting for a republication license on behalf of the User represents and warrants that he/she/it (a) has been duly authorized by the User to accept, and hereby does accept, all such terms and conditions on behalf of User, and (b) shall inform User of all such terms and conditions. In the event such person is a "freelancer" or other third party independent of User and CCC, such party shall be deemed jointly a "User" for purposes of these terms and conditions. In any event, User shall be deemed to have accepted and agreed to all such terms and conditions if User republishes the Work in any fashion.

3. Scope of License; Limitations and Obligations.

   3.1. All Works and all rights therein, including copyright rights, remain the sole and exclusive property of the Rightsholder. The license created by the exchange of an Order Confirmation (and/or any invoice) and payment by User of the full amount set forth on that document includes only those rights expressly set forth in the Order Confirmation and in these terms and conditions, and conveys no other rights in the Work(s) to User. All rights not expressly granted are hereby reserved.

   3.2. General Payment Terms: You may pay by credit card or through an account with us payable at the end of the month. If you and we agree that you may establish a standing account with CCC, then the following

terms apply: Remit Payment to: Copyright Clearance Center, 29118 Network Place, Chicago, IL 60673-1291. Payments Due: Invoices are payable upon their delivery to you (or upon our notice to you that they are available to you for downloading). After 30 days, outstanding amounts will be subject to a service charge of 1-1/2% per month or, if less, the maximum rate allowed by applicable law. Unless otherwise specifically set forth in the Order Confirmation or in a separate written agreement signed by CCC, invoices are due and payable on "net 30" terms. While User may exercise the rights licensed immediately upon issuance of the Order Confirmation, the license is automatically revoked and is null and void as if it had never been issued, if complete payment for the license is not received on a timely basis either from User directly or through a payment agent, such as a credit card company.

   3.3. Unless otherwise provided in the Order Confirmation, any grant of rights to User (i) is "one-time" (including the editions and product family specified in the license), (ii) is non-exclusive and non-transferable and (iii) is subject to any and all limitations and restrictions (such as, but not limited to, limitations on duration of use or circulation) included in the Order Confirmation or invoice and/or in these terms and conditions. Upon completion of the licensed use, User shall either secure a new permission for further use of the Work(s) or immediately cease any new use of the Work(s) and shall render inaccessible (such as by deleting or by removing or severing links or other locators) any further copies of the Work (except for copies printed on paper in accordance with this license and still in User's stock at the end of such period).

   3.4. In the event that the material for which a republication license is sought includes third party materials (such as photographs, illustrations, graphs, inserts and similar materials) which are identified in such material as having been used by permission, User is responsible for identifying, and seeking separate licenses (under this Service or otherwise) for, any of such third party materials; without a separate license, such third party materials may not be used.

   3.5. Use of proper copyright notice for a Work is required as a condition of any license granted under the Service. Unless otherwise provided in the Order Confirmation, a proper copyright notice will read substantially as follows: "Republished with permission of [Rightsholder's name], from [Work's title, author, volume, edition number and year of copyright]; permission conveyed through Copyright Clearance Center, Inc." Such notice must be provided in a reasonably legible font size and must be placed either immediately adjacent to the Work as used (for example, as part of a by-line or footnote but not as a separate electronic link) or in the place where substantially all other credits or notices for the new work containing the republished Work are located. Failure to include the required notice results in loss to the Rightsholder and CCC, and the User shall be liable to pay liquidated damages for each such failure equal to twice the use fee specified in the Order Confirmation, in addition to the use fee itself and any other fees and charges specified.

   3.6. User may only make alterations to the Work if and as expressly set forth in the Order Confirmation. No Work may be used in any way that is defamatory, violates the rights of third parties (including such third parties' rights of copyright, privacy, publicity, or other tangible or intangible property), or is otherwise illegal, sexually explicit or obscene. In addition, User may not conjoin a Work with any other material that may result in damage to the reputation of the Rightsholder. User agrees to inform CCC if it becomes aware of any infringement of any rights in a Work and to cooperate with any reasonable request of CCC or the Rightsholder in connection therewith.

4. Indemnity. User hereby indemnifies and agrees to defend the Rightsholder and CCC, and their respective employees and directors, against all claims, liability, damages, costs and expenses, including legal fees and expenses, arising out of any use of a Work beyond the scope of the rights granted herein, or any use of a Work which has been altered in any unauthorized way by User, including claims of defamation or infringement of rights of copyright, publicity, privacy or other tangible or intangible property.

5. Limitation of Liability. UNDER NO CIRCUMSTANCES WILL CCC OR THE RIGHTSHOLDER BE LIABLE FOR ANY DIRECT, INDIRECT, CONSEQUENTIAL OR INCIDENTAL DAMAGES (INCLUDING WITHOUT LIMITATION DAMAGES FOR LOSS OF BUSINESS PROFITS OR INFORMATION, OR FOR BUSINESS INTERRUPTION) ARISING OUT OF THE USE OR INABILITY TO USE A WORK, EVEN IF ONE OF THEM HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. In any event, the total liability of the Rightsholder and CCC (including their respective employees and directors) shall not exceed

5. Limitation of Liability. UNDER NO CIRCUMSTANCES WILL CCC OR THE RIGHTSHOLDER BE LIABLE FOR ANY DIRECT, INDIRECT, CONSEQUENTIAL OR INCIDENTAL DAMAGES (INCLUDING WITHOUT LIMITATION DAMAGES FOR LOSS OF BUSINESS PROFITS OR INFORMATION, OR FOR BUSINESS INTERRUPTION) ARISING OUT OF THE USE OR INABILITY TO USE A WORK, EVEN IF ONE OF THEM HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. In any event, the total liability of the Rightsholder and CCC (including their respective employees and directors) shall not exceed the total amount actually paid by User for this license. User assumes full liability for the actions and omissions of its principals, employees, agents, affiliates, successors and assigns.

6. Limited Warranties. THE WORK(S) AND RIGHT(S) ARE PROVIDED "AS IS". CCC HAS THE RIGHT TO GRANT TO USER THE RIGHTS GRANTED IN THE ORDER CONFIRMATION DOCUMENT. CCC AND THE RIGHTSHOLDER DISCLAIM ALL OTHER WARRANTIES RELATING TO THE WORK(S) AND RIGHT(S), EITHER EXPRESS OR IMPLIED, INCLUDING WITHOUT LIMITATION IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. ADDITIONAL RIGHTS MAY BE REQUIRED TO USE ILLUSTRATIONS, GRAPHS, PHOTOGRAPHS, ABSTRACTS, INSERTS OR OTHER PORTIONS OF THE WORK (AS OPPOSED TO THE ENTIRE WORK) IN A MANNER CONTEMPLATED BY USER; USER UNDERSTANDS AND AGREES THAT NEITHER CCC NOR THE RIGHTSHOLDER MAY HAVE SUCH ADDITIONAL RIGHTS TO GRANT.

7. Effect of Breach. Any failure by User to pay any amount when due, or any use by User of a Work beyond the scope of the license set forth in the Order Confirmation and/or these terms and conditions, shall be a material breach of the license created by the Order Confirmation and these terms and conditions. Any breach not cured within 30 days of written notice thereof shall result in immediate termination of such license without further notice. Any unauthorized (but licensable) use of a Work that is terminated immediately upon notice thereof may be liquidated by payment of the Rightsholder's ordinary license price therefor; any unauthorized (and unlicensable) use that is not terminated immediately for any reason (including, for example, because materials containing the Work cannot reasonably be recalled) will be subject to all remedies available at law or in equity, but in no event to a payment of less than three times the Rightsholder's ordinary license price for the most closely analogous licensable use plus Rightsholder's and/or CCC's costs and expenses incurred in collecting such payment.

8. Miscellaneous.

8.1. User acknowledges that CCC may, from time to time, make changes or additions to the Service or to these terms and conditions, and CCC reserves the right to send notice to the User by electronic mail or otherwise for the purposes of notifying User of such changes or additions; provided that any such changes or additions shall not apply to permissions already secured and paid for.

8.2. Use of User-related information collected through the Service is governed by CCC's privacy policy, available online here:https://marketplace.copyright.com/rs-ui-web/mp/privacy-policy

8.3. The licensing transaction described in the Order Confirmation is personal to User. Therefore, User may not assign or transfer to any other person (whether a natural person or an organization of any kind) the license created by the Order Confirmation and these terms and conditions or any rights granted hereunder; provided, however, that User may assign such license in its entirety on written notice to CCC in the event of a transfer of all or substantially all of User's rights in the new material which includes the Work(s) licensed under this Service.

8.4. No amendment or waiver of any terms is binding unless set forth in writing and signed by the parties. The Rightsholder and CCC hereby object to any terms contained in any writing prepared by the User or its principals, employees, agents or affiliates and purporting to govern or otherwise relate to the licensing transaction described in the Order Confirmation, which terms are in any way inconsistent with any terms set forth in the Order Confirmation and/or in these terms and conditions or CCC's standard operating procedures, whether such writing is prepared prior to, simultaneously with or subsequent to the Order Confirmation, and whether such writing appears on a copy of the Order Confirmation or in a separate instrument.

8.5. The licensing transaction described in the Order Confirmation document shall be governed by and construed under the law of the State of New York, USA, without regard to the principles thereof of conflicts of law. Any case, controversy, suit, action, or proceeding arising out of, in connection with, or related to such licensing transaction shall be brought, at CCC's sole discretion, in any federal or state court located in the County of New York, State of New York, USA, or in any federal or state court whose geographical jurisdiction covers the location of the Rightsholder set forth in the Order Confirmation. The parties expressly submit to the personal jurisdiction and venue of each such federal or state court. If you have any comments or questions about the Service or Copyright Clearance Center, please contact us at 978-750-8400 or send an e-mail to support@copyright.com.

v 1.1

the total amount actually paid by User for this license. User assumes full liability for the actions and omissions of its principals, employees, agents, affiliates, successors and assigns.

6. Limited Warranties. THE WORK(S) AND RIGHT(S) ARE PROVIDED "AS IS". CCC HAS THE RIGHT TO GRANT TO USER THE RIGHTS GRANTED IN THE ORDER CONFIRMATION DOCUMENT. CCC AND THE RIGHTSHOLDER DISCLAIM ALL OTHER WARRANTIES RELATING TO THE WORK(S) AND RIGHT(S), EITHER EXPRESS OR IMPLIED, INCLUDING WITHOUT LIMITATION IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. ADDITIONAL RIGHTS MAY BE REQUIRED TO USE ILLUSTRATIONS, GRAPHS, PHOTOGRAPHS, ABSTRACTS, INSERTS OR OTHER PORTIONS OF THE WORK (AS OPPOSED TO THE ENTIRE WORK) IN A MANNER CONTEMPLATED BY USER; USER UNDERSTANDS AND AGREES THAT NEITHER CCC NOR THE RIGHTSHOLDER MAY HAVE SUCH ADDITIONAL RIGHTS TO GRANT.

7. Effect of Breach. Any failure by User to pay any amount when due, or any use by User of a Work beyond the scope of the license set forth in the Order Confirmation and/or these terms and conditions, shall be a material breach of the license created by the Order Confirmation and these terms and conditions. Any breach not cured within 30 days of written notice thereof shall result in immediate termination of such license without further notice. Any unauthorized (but licensable) use of a Work that is terminated immediately upon notice thereof may be liquidated by payment of the Rightsholder's ordinary license price therefor; any unauthorized (and unlicensable) use that is not terminated immediately for any reason (including, for example, because materials containing the Work cannot reasonably be recalled) will be subject to all remedies available at law or in equity, but in no event to a payment of less than three times the Rightsholder's ordinary license price for the most closely analogous licensable use plus Rightsholder's and/or CCC's costs and expenses incurred in collecting such payment.

8. Miscellaneous.

8.1. User acknowledges that CCC may, from time to time, make changes or additions to the Service or to these terms and conditions, and CCC reserves the right to send notice to the User by electronic mail or otherwise for the purposes of notifying User of such changes or additions; provided that any such changes or additions shall not apply to permissions already secured and paid for.

8.2. Use of User-related information collected through the Service is governed by CCC's privacy policy, available online here:https://marketplace.copyright.com/rs-ui-web/mp/privacy-policy

8.3. The licensing transaction described in the Order Confirmation is personal to User. Therefore, User may not assign or transfer to any other person (whether a natural person or an organization of any kind) the license created by the Order Confirmation and these terms and conditions or any rights granted hereunder; provided, however, that User may assign such license in its entirety on written notice to CCC in the event of a transfer of all or substantially all of User's rights in the new material which includes the Work(s) licensed under this Service.

8.4. No amendment or waiver of any terms is binding unless set forth in writing and signed by the parties. The Rightsholder and CCC hereby object to any terms contained in any writing prepared by the User or its principals, employees, agents or affiliates and purporting to govern or otherwise relate to the licensing transaction described in the Order Confirmation, which terms are in any way inconsistent with any terms set forth in the Order Confirmation and/or in these terms and conditions or CCC's standard operating procedures, whether such writing is prepared prior to, simultaneously with or subsequent to the Order Confirmation, and whether such writing appears on a copy of the Order Confirmation or in a separate instrument.

8.5. The licensing transaction described in the Order Confirmation document shall be governed by and construed under the law of the State of New York, USA, without regard to the principles thereof of conflicts of law. Any case, controversy, suit, action, or proceeding arising out of, in connection with, or related to such licensing transaction shall be brought, at CCC's sole discretion, in any federal or state court located in the County of New York, State of New York, USA, or in any federal or state court whose geographical jurisdiction covers the location of the Rightsholder set forth in the Order Confirmation. The parties expressly submit to the personal jurisdiction and venue of each such federal or state court. If you have any comments or questions about the Service or Copyright Clearance Center, please contact us at 978-750-8400 or send an e-mail to support@copyright.com.

v 1.1

8/4/2021 https://marketplace.copyright.com/rs-ui-web/mp/license/6c95573a-9c87-421b-829c-b2137da2f95f/ee1b4ebf-9ea5-493d-...

of law. Any case, controversy, suit, action, or proceeding arising out of, in connection with, or related to such licensing transaction shall be brought, at CCC's sole discretion, in any federal or state court located in the County of New York, State of New York, USA, or in any federal or state court whose geographical jurisdiction covers the location of the Rightsholder set forth in the Order Confirmation. The parties expressly submit to the personal jurisdiction and venue of each such federal or state court. If you have any comments or questions about the Service or Copyright Clearance Center, please contact us at 978-750-8400 or send an e-mail to support@copyright.com.

v 1.1

8/4/2021 https://marketplace.copyright.com/rs-ui-web/mp/license/ffd590ca-e79b-40e0-a251-cc050aac5aea/8cd97d2e-9b80-4bb4-...

comments or questions about the Service or Copyright Clearance Center, please contact us at 978-750-8400 or send an e-mail to support@copyright.com.

v 1.1