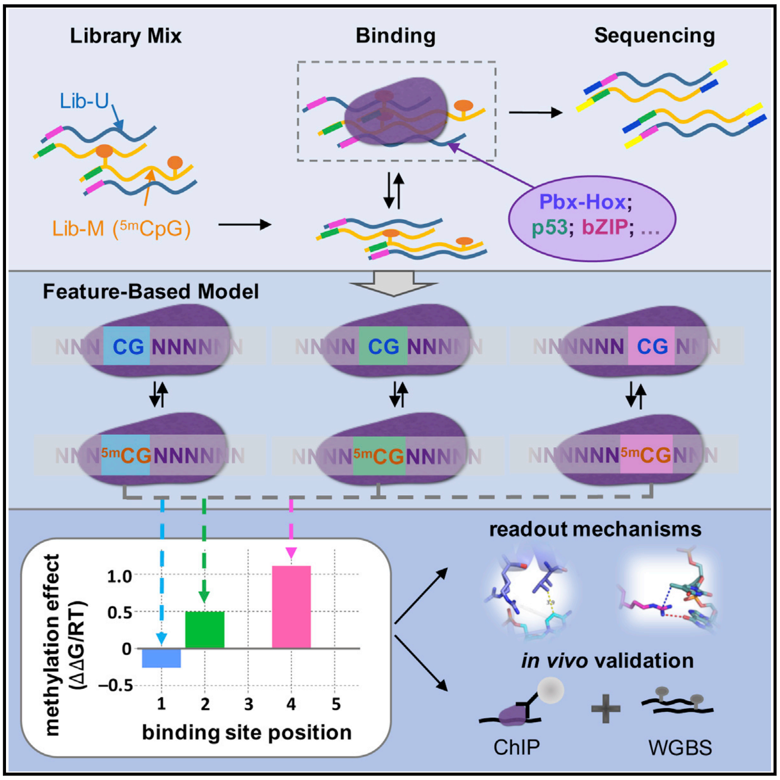
Cell Reports

Article

Quantitative Analysis of the DNA Methylation Sensitivity of Transcription Factor Complexes

Graphical Abstract



Highlights

- Single-round SELEX systematically maps the sensitivity of TF binding to DNA modification
- Feature-based models pinpoint position-specific ^{5}mCpG effects within binding sites
- Human bZIP and Pbx-Hox complexes show paralog-specific ^{5}mCpG sensitivity patterns
- Cytosine methylation stabilizes binding by p53 in vitro and in vivo

Authors

Judith F. Kribelbauer, Oleg Laptenko, Siying Chen, ..., Carol Prives, Richard S. Mann, Harmen J. Bussemaker

Correspondence

rsm10@cumc.columbia.edu (R.S.M.), hjb2004@columbia.edu (H.J.B.)

In Brief

Kribelbauer et al. present a high-throughput method for quantifying the in vitro sensitivity of TF binding to epigenetic DNA modifications. Application to p53 tetramers shows that cytosine methylation can either increase or decrease binding affinity, depending on where the modification occurs within the binding interface. These effects persist in vivo.

Accession Numbers

GSE98652

Kribelbauer et al., 2017, Cell Reports *19*, 2383–2395
June 13, 2017 © 2017 The Author(s).
http://dx.doi.org/10.1016/j.celrep.2017.05.069

OPEN
ACCESS
CellPress

Quantitative Analysis of the DNA Methylation Sensitivity of Transcription Factor Complexes

Judith F. Kribelbauer,[1,2] Oleg Laptenko,[1] Siying Chen,[1,2] Gabriella D. Martini,[1] William A. Freed-Pastor,[1,4] Carol Prives,[1] Richard S. Mann,[2,3,*] and Harmen J. Bussemaker[1,2,5,*]
[1]Department of Biological Sciences, Columbia University, New York, NY 10027, USA
[2]Department of Systems Biology, Columbia University Medical Center, New York, NY 10032, USA
[3]Department of Biochemistry and Molecular Biophysics, Columbia University Medical Center, New York, NY 10032, USA
[4]David H. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02142, USA
[5]Lead Contact
*Correspondence: rsm10@cumc.columbia.edu (R.S.M.), hjb2004@columbia.edu (H.J.B.)

http://dx.doi.org/10.1016/j.celrep.2017.05.069

SUMMARY

Although DNA modifications play an important role in gene regulation, the underlying mechanisms remain elusive. We developed EpiSELEX-seq to probe the sensitivity of transcription factor binding to DNA modification in vitro using massively parallel sequencing. Feature-based modeling quantifies the effect of cytosine methylation (^{5}mC) on binding free energy in a position-specific manner. Application to the human bZIP proteins ATF4 and C/EBPβ and three different Pbx-Hox complexes shows that ^{5}mCpG can both increase and decrease affinity, depending on where the modification occurs within the protein-DNA interface. The TF paralogs tested vary in their methylation sensitivity, for which we provide a structural rationale. We show that ^{5}mCpG can also enhance in vitro p53 binding and provide evidence for increased in vivo p53 occupancy at methylated binding sites, correlating with primed enhancer histone marks. Our results establish a powerful strategy for dissecting the epigenomic modulation of protein-DNA interactions and their role in gene regulation.

INTRODUCTION

High-throughput profiling of in vitro transcription factor (TF) binding specificities is a powerful approach for obtaining sequence motifs for a variety of TF families and in several different organisms (Badis et al., 2009; Jolma et al., 2013; Weirauch et al., 2014). However, despite the growing number of known TF motifs, accurate prediction of in vivo TF binding and its effect on target gene expression has remained surprisingly difficult. One of the complications is that protein-protein interactions can modify the DNA binding specificities of transcription factors (Jolma et al., 2015; Miller, 2009; Slattery et al., 2011). Another potential complication is the existence of covalent modifications of DNA, particularly cytosine methylation (^{5}mCpG), which is widespread in vertebrates. Because of their potential to alter the chromatin state (Hashimshony et al., 2003) or DNA shape (Lazarovici et al., 2013), an important and hotly debated question is to what extent DNA modifications can influence TF binding and, thereby, contribute to changes in the epigenetic landscape and gene regulation. Such a regulatory mechanism is conceptually compelling because DNA modifications could provide an additional layer of temporal and spatial control to fine-tune gene expression.

^{5}mCpG has been shown to be important in gene silencing in normal and cancer cells (Jones and Baylin, 2007; Stein et al., 1982), gene imprinting (Razin and Cedar, 1994), and X chromosome inactivation (Hellman and Chess, 2007; Tribioli et al., 1992). In spite of this progress, there is no general mechanism explaining the effect of DNA methylation on gene expression (Dantas Machado et al., 2015). Several studies have found that, despite the overall association between promoter methylation and transcriptional silencing, some promoters can simultaneously be methylated and transcriptionally active (Gutierrez-Arcelus et al., 2013). In addition, systematic studies with cancer cell lines have found that aberrant methylation, such as hypermethylation of specific CpG islands, is a hallmark of cancer progression (Baylin and Jones, 2011; Paz et al., 2003). Recent studies have identified additional modifications, such as 5-hydroxy-methylcytosine (5hmC) and N^6-methyladenine (6mA), in mammalian genomes, raising the possibility that these also influence gene regulation (Fu et al., 2015; Greer et al., 2015; Zhang et al., 2015). To identify the causal determinants of in vivo TF binding among all of these correlated variables, detailed quantitative characterization of the effect of DNA modification on in vitro transcription factor binding is a prerequisite.

On a limited scale, the in vitro platform of protein binding microarrays (PBMs) has been used to probe TF binding to methylated DNA probes (Hu et al., 2013; Mann et al., 2013). These studies demonstrated that ^{5}mCpGs can have both positive and negative effects on affinity. However, they were limited by the fact that the DNA arrays contained either fully methylated or fully un-methylated sequences (Mann et al., 2013), but not both in competition, or they only considered a select subset of sequences (Hu et al., 2013). In addition, the data analysis in these studies was restricted to oligomer-based methods, which makes it difficult to identify position-specific effects, especially for lower-affinity binding sites that deviate from the consensus motif. To study the effect of cytosine methylation on TF binding at high resolution, a quantitative assay is required that allows simultaneous probing of methylated and unmethylated DNA probes across all possible sequence contexts.

To address these issues, we developed EpiSELEX-seq, a method that uses a single round of gel electrophoresis to simultaneously assess binding to methylated and unmethylated DNA fragments, thus allowing methylation sensitivity to be analyzed for any TF or TF complex. We apply EpiSELEX-seq to human basic leucine zipper (bZIP) and Hox complexes as well as tetramers of the tumor suppressor protein p53. Using a feature-based Poisson regression model, we quantify position-specific methylation effects on in vitro binding in the low-affinity range. For p53, by jointly analyzing whole-genome bisulfite sequencing and in vivo binding (chromatin immunoprecipitation sequencing [ChIP-seq]) data, we provide evidence that the increased in vitro affinity for specific DNA sequences because of methylation leads to enhanced occupancy in vivo. These sites of increased binding have a histone modification pattern associated with primed enhancers, supporting a role for p53 as a pioneer factor that can access methylated DNA sites.

RESULTS

Affinity-Based Selection from Mixed Pools of Methylated and Unmethylated DNA Ligands

To quantitatively assess the effects of DNA methylation on TF binding, we developed a method in which a methylated library (Lib-M) and an unmethylated library (Lib-U) containing a randomized region of a desired length (16 bp or 26 bp) were first separately synthesized, each distinguished by a unique 4-bp bar code located near the variable region (Figure 1A). After treatment of Lib-M with a DNA methyltransferase, both libraries were mixed in equal proportions, incubated with a TF of interest, and subjected to a single round of electrophoretic mobility shift assay (EMSA) selection. Sequencing libraries were prepared from the library mix both before (R0) and after (R1) affinity-based selection (Figure 1B; Figures S1A and S1B). For each sequenced DNA ligand, the bar code allows us to reconstruct the methylation status at the time of TF binding.

For accurate affinity estimation, it is important that the two cytosines in each CpG base-pair step in Lib-M be fully methylated because incomplete methylation would lead to underestimation of the effect of ^{5}mCpG on TF binding. We employed two separate tests to confirm full methylation: methylation, bisulfite treatment, and sub-cloning of a test sequence containing four CpGs and high-throughput sequencing followed by dinucleotide analysis of a Lib-M that was either treated or not treated with bisulfite. In the first test, we determined that optimal methylation efficiency is achieved after two successive rounds of methylation with $\leq$ 250 ng of input DNA per reaction (Table S1). Using larger amounts of DNA (e.g., the recommended 1 μg) resulted in incomplete methylation of the test probes. The short size of our probes ($\sim$50 bp) compared with typical genomic fragments (>1 kb) might be the source of this discrepancy because suboptimal conditions typically resulted in the methylation of either all four CpGs or none, arguing for a processive nature of the DNA methyltransferase. In the second test, bisulfite treatment of a Lib-U of random 16-mers showed depletion of all CpN dinucleotides, as expected (Figure 1C). In contrast, under optimal methylation conditions, bisulfite treatment of a Lib-M showed depletion of all CpN dinucleotides except CpG, which was recovered at levels identical to those observed in non-bisulfite-treated, methylated libraries (Figure 1D).

EpiSELEX-Seq Identifies Differences in Methylation Sensitivity within the bZIP Family

To benchmark our method, we considered the bZIP transcription factors ATF4 and C/EBPβ, previously reported to be sensitive to DNA methylation (Mann et al., 2013). Many bZIP homo- or heterodimers preferentially bind to the cyclic AMP (cAMP) response element (CRE) TGACGTCA and/or the C/EBP consensus TTGCGCAA (Figures 2A and 2B). These palindromic sequences both contain a central CpG dinucleotide, creating the potential for methylation-sensitive DNA binding. For ATF4 homodimers, as expected, the relative enrichment of 10-bp sequences (encompassing the suspected TF footprint) that do not contain any CpG dinucleotides is similar between Lib-U and Lib-M (Figure 2C). However, sequences that contain at least one CpG fall into distinct groups, each with a different ratio between Lib-M and Lib-U, indicative of a sensitivity to cytosine methylation that depends on the position of the CpG dinucleotide within the binding site (Figure 2C; Figure S2A). When a CpG base pair step is present at the center of the ATF4 binding site, methylation of both cytosines leads to a decrease in affinity. By contrast, sequences that contain a CpG in the flank of the motif (at positions –3/–4 or +3/+4) are bound much more strongly when methylated, leading to an alternative optimal left half-site, (^{5}mC)GAT. Interestingly, these methylation sensitivities are not observed for C/EBPβ (Figure 2D), consistent with a previous observation that in vivo binding by this factor tolerates CpG methylation (Zhu et al., 2016). The methylation sensitivity for ATF4 is also reflected in the energy logos (Foat et al., 2006) that can be derived from the oligomer enrichment tables by considering all possible point mutations away from the optimal sequence (Experimental Procedures). The logo derived from Lib-M, compared with its equivalent for Lib-U, no longer has a central CpG as the most preferred sequence and shows an increased preference for a CpG at position –4/–3 (Figures 2E and 2F). Together, these findings demonstrate that sensitivity to DNA methylation can differ between paralogs from the same structural family.

Feature-Based Modeling Quantifies Position-Specific Methylation Effects

To systematically analyze the quantitative effect of cytosine methylation on binding affinity, we developed a feature-based generalized linear model to estimate the change in binding free energy associated with cytosine modification at any particular offset within the binding site. The frequency of DNA ligand S after one round (R1) of affinity-based selection, F_1, is proportional to the frequency of the same probe in the initial (R0) pool, F_0, as well as to the relative affinity of the interaction:

$$F_1\,(S) \propto F_0\,(S)\,\times\,exp\left[-\frac{\Delta\Delta G\,(S)}{RT}\right].$$

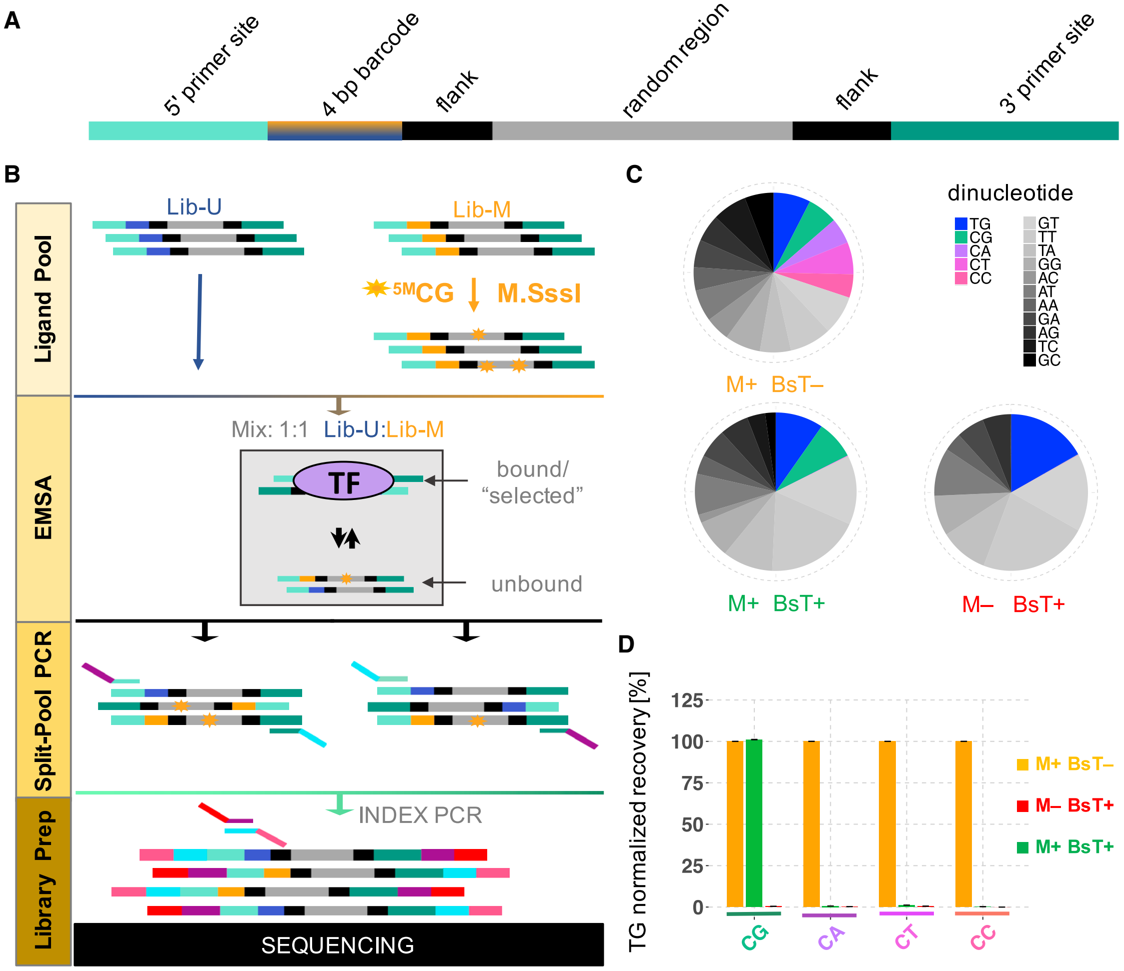


Figure 1. Overview and Validation of the EpiSELEX-Seq Design

(A) Library design. 4-bp bar codes distinguish unmodified (Lib-U) and modified (Lib-M) DNA ligands. All libraries share a random region, reverse complement symmetric flanks, and a pair of 5′ and 3′ primer sites.

(B) EpiSELEX-seq workflow. Lib-M is methylated and mixed with Lib-U. The mixed pool is incubated with a TF of interest, and the bound fraction is separated by an EMSA, purified, split, and amplified using two sets of primers. Unique Illumina bar codes are added for multiplexing.

(C) Validation of the methylation protocol. Shown are dinucleotide frequencies in Lib-M after various combinations of optional methylation (M+/M–) and bisulfite treatment (BsT+/BsT–), determined by Illumina sequencing. The four CpN dinucleotides for which the methylation status of the cytosine is unambiguous are highlighted, as is TpG, which serves as a reference for CpN dinucleotides.

(D) TpG-normalized recovery of the four CpN dinucleotides. Only the CpGs protected by methylation are retained after bisulfite conversion. Error bars are based on counting statistics and error propagation.

We model $\Delta\Delta G(S)$, the difference in binding free energy between ligand S, and the optimal ligand S_{opt} as a sum of contributions because of the specific (binary) features ϕ associated with S:

$$\frac{\Delta\Delta G(S)}{RT} \equiv \frac{\Delta G(S) - \Delta G(S_{opt})}{RT} = \sum_{\phi} \beta_{\phi} X_{\phi}(S).$$

Some features indicate the presence (X = 1) or absence (X = 0) of a specific base at a given position within the binding site, whereas others indicate the methylation status of a particular CpG dinucleotide. We estimate the corresponding coefficients β_{ϕ} from the data by fitting a generalized linear model based on counting statistics to the read counts in R1 while accounting for biases in R0 (see Experimental Procedures for details). To validate this modeling approach, we first inferred free energy effects for the three possible substitutions of the optimal base A_{-5} using the ATF4 homodimer data. Good agreement is observed with the results obtained using oligomer enrichment (Figure 2G). Next we used an extended model that included features indicating methylation status. The coefficients from this fit indicate that methylation of $C_{-1}G_{+1}$ represses binding ($\Delta\Delta G/RT$ = 1.5, corresponding to 0.9 kcal/mol or equivalently, a 4.5-fold reduction in affinity), consistent with the changes in oligomer enrichment between Lib-U and Lib-M (Figure 2H). The coefficients for the equivalent flanking positions $C_{-4}G_{-3}$ and $C_{+3}G_{+4}$ are

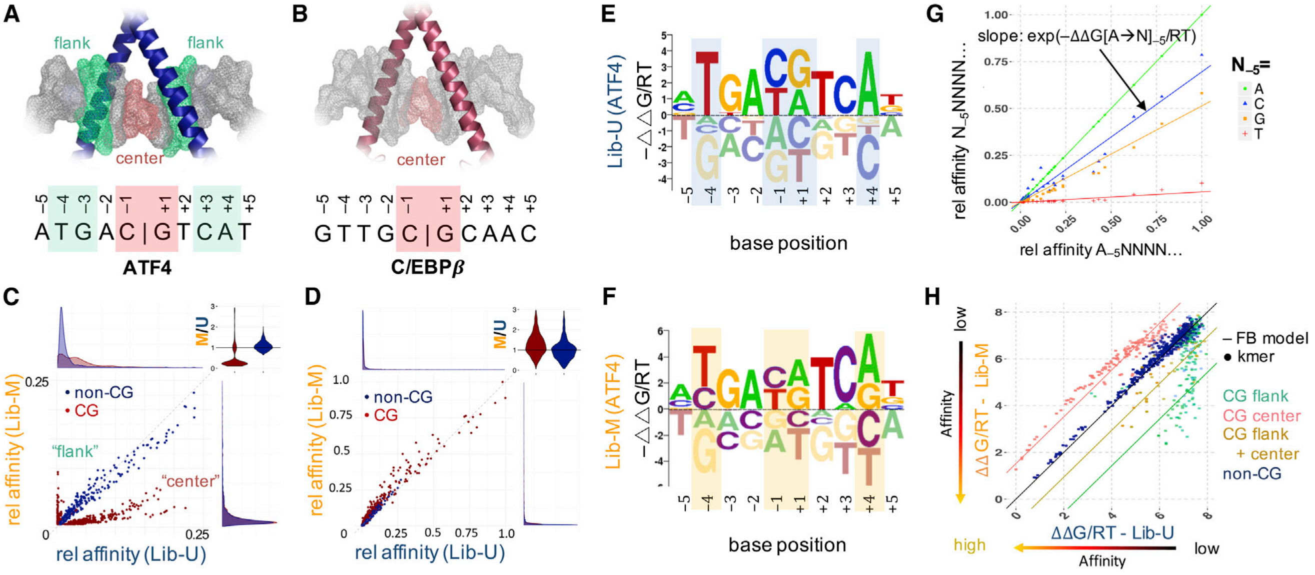


Figure 2. Probing Methylation Sensitivity for ATF4 and C/EBPβ

(A and B) Crystal structure (PDB: 1GTW) for the human bZIP homodimer C/EBPβ along with the symmetric consensus motifs for ATF4 (A) or C/EBPβ (B) and the definition of "flank" (green) and "center" (pink) positions in the binding sites.

(C) Enlargement of low-affinity range comparing the relative enrichment of 10-bp oligonucleotides between Lib-M versus Lib-U for ATF4. Non-CpG sequences (blue) show similar enrichment in both libraries, whereas distinct subsets of the CpG-containing sequences (red) are either preferred in Lib-U (center) or in Lib-M (flank).

(D) As in (C), but for C/EBPβ homodimers. Non-CpG- and CpG-containing sequences show similar enrichments in both libraries across the entire sequence range. The insets in (C) and (D) show the marginal distributions and the distribution of methylated/unmethylated ratio for all oligomers with a relative enrichment above 10^{-3}.

(E and F) Energy logo for ATF4 derived from Lib-U (E) and Lib-M (F). The central CpG is no longer the top choice in the methylated library. ^{5}mCpGs at the equivalent positions –4/–3 and +3/+4 appear as a new sequence feature in Lib-M.

(G) Relative affinities for ATF4 in Lib-U. Each point represents a 10-bp oligomer containing either an A (reference base) or a point mutation (C, T, or G) at position –5. The slope of the lines represents the exponentiated value of -ΔΔG/RT associated with each point mutation, as estimated from the Lib-U read counts using a feature-based model.

(H) Comparison of binding free energies between Lib-M and Lib-U. Each point denotes a unique 10-mer. Vertical line offsets correspond to the position-dependent methylation effects (ΔΔG/RT) estimated using a feature-based (FB) model.

almost identical, as expected based on symmetry, and indicate a strong increase in binding because of methylation (ΔΔG/RT = –2.6). Our model also predicts the combined effect of methylating both $C_{-1}G_{+1}$ and $C_{+3}G_{+4}$ (or $C_{-3}G_{-4}$) by simply adding up the respective free energy coefficients (Figure 2H).

Explaining the Effect of Cytosine Methylation by "Thymine Mimicry"

Although it has distinct base-pairing preferences, ^{5}mC is chemically similar to thymine in that both have a methyl group at the carbon 5 position of the pyrimidine ring (Figure 3A). Therefore, the total effect of a C-to-T transition on protein-DNA binding free energy, $\Delta\Delta G[\mathrm{C}\rightarrow\mathrm{T}]$, can be separated into the effect of the methyl group alone, $\Delta\Delta G[\mathrm{C}\rightarrow 5\mathrm{mC}]$, and changes in charge and base pair interactions, $\Delta\Delta G[5\mathrm{mC}\rightarrow\mathrm{T}]$ (Figure 3A). Following this logic, the value of $\Delta\Delta G[\mathrm{C}\rightarrow\mathrm{T}]$ and $\Delta\Delta G[5\mathrm{mC}\rightarrow\mathrm{T}]$, as estimated using Lib-U and Lib-M, respectively, can be subtracted from each other to obtain an estimate of the effect because of methylation $\Delta\Delta G[\mathrm{C}\rightarrow 5\mathrm{mC}]$. This approach was successful when applied to ATF4 to predict the effect of methylating the CpG dinucleotide, both at the central (–1/+1) and the flanking (–4/–3) positions (Figures 3B–3D). In agreement with these observations, many bZIP proteins contain two conserved hydrophobic amino acids that, in crystal structures, make van der Waals (VdW) contacts with the carbon 5 methyl group of thymidine at position –4 in the binding site (Figure S2B). C/EBPβ, but not ATF4, has a valine instead of an alanine at one of these positions, providing a possible mechanistic explanation for the increased preference of ATF4 for ^{5}mC over C, where the gain of a methyl group on the base may compensate for the loss of a methyl group in alanine compared with valine.

Deciphering the DNA Binding Specificity of Human Pbx-Hox Complexes

An important aspect of gene regulation is the capacity of TFs to form complexes with cofactors. A prominent example of such cooperative binding is that of Hox proteins and their three amino acid loop extension (TALE) cofactors, which play a crucial role in animal development (Merabet and Mann, 2016). As monomers, Hox family members bind to similar DNA sequences in vitro but have distinct functions in vivo. Previously, we used SELEX-seq to capture the latent binding specificity of all eight *Drosophila* Hox proteins with their TALE cofactors Extradenticle (Exd) and the HM isoform (HM) of Homothorax (Hth), which is required

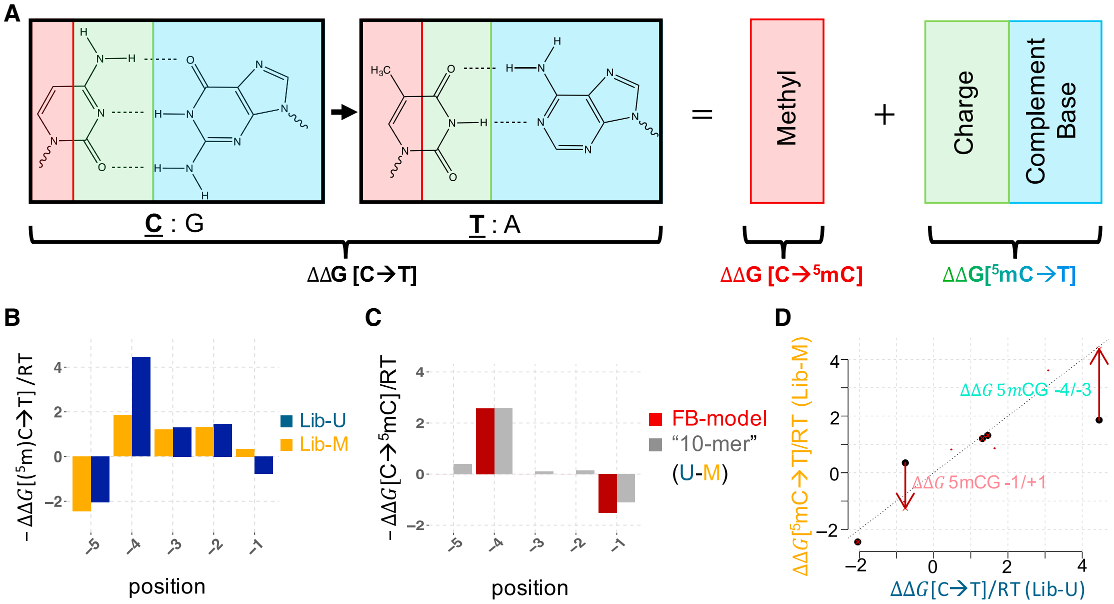


Figure 3. Deconvolving the Methylation Sensitivity for ATF4

(A) Decomposition of the position-specific DNA-protein binding free energy change associated with a C → T transition. The C → T change is the sum of C → ^{5}mC and ^{5}mC → T, allowing an interpretation of methylation sensitivity in terms of thymine mimicry.

(B) Change in binding free energy associated with C → T transition in each library as derived from an oligomer-based PSAM.

(C) Position-specific methylation effect on binding free energy as estimated based on either the oligomer-enrichment-based approach (as in B, gray) or the feature-based-modeling approach (red).

(D) The methylation effect, as estimated using the feature-based model (red arrows), explains the differences in the C → T transition effect observed for Lib-U and Lib-M.

for optimal Exd-Hox interaction (Slattery et al., 2011). In mammals, where the Hox cluster has been duplicated several times in the genome, multiple cofactors from the PBC and MEIS class of TALE factors as well as epigenetic DNA modifications all have the potential to modulate DNA binding.

Here we used EpiSELEX-seq to characterize the binding of human heterodimeric Pbx-Hox complexes to DNA (Figure 4A). To cover the three Hox subclasses defined in Slattery et al. (2011), we performed these experiments using HoxA1, HoxA5, and HoxA9, each in complex with the cofactor PBX1, which was purified together with the HM domain of MEIS1. Comparing the pattern of 12-bp oligomer enrichment from R0 to R1 for each complex, we found similar cofactor-dependent differences in binding specificity between these Hox proteins, as previously observed for their *D. melanogaster* orthologs (Slattery et al., 2011; Figure 4B; Figure S4A): the preferred central dinucleotide spacer (underlined) in the binding site consensus NTGAYNNAYNNN (where Y denotes C or T) is TG for the anterior (class I) factor HoxA1, TA for the central (class II) factor HoxA5, and TT for the posterior (class III) factor HoxA9 (Figure 4D).

Human Pbx-Hox Dimers Show Position-Specific Methylation Sensitivity

The EpiSELEX-seq protocol allows us to assess the three human Pbx-Hox complexes for sensitivity to cytosine methylation. We first constructed separate energy logos for Lib-U and Lib-M by considering all possible point mutations from the most enriched 12-bp sequence (Figure 4D). Although paralog-dependent differences in the central spacer (shaded area) are readily apparent, the logos for the Lib-U and Lib-M human libraries are otherwise highly similar to each other and to those of their fly orthologs. However, this oligomer enrichment-based approach is unable to detect methylation sensitivity for any cytosine that does not occur in a CpG context in the optimal sequence (Figure S3). For ATF4, both cytosine positions at which methylation sensitivity was observed (–4 and −1) were fortuitously followed by a guanine (cf. Figure 2A), but this is not the case for Pbx-Hox. Indeed, when we used our feature-based Poisson regression model to jointly analyze Lib-U and Lib-M to quantify the effect of ^{5}mCpG on binding, all three Hox proteins and Pbx showed significant methylation sensitivity at various positions throughout the binding interface (Figures 4C and 4E; Figure S4B). The direction and amplitude of the methylation effect are highly position-dependent: methylation of CpG dinucleotides that start at positions 5 or 9 (underlined in the consensus sequence NTGAYNNAYNNN) enhance binding by severalfold. In contrast, methylation of CpGs shifted by one position (positions 6 or 10, underlined in NTGAYNNAYNNN) decreases binding by up to 7-fold (Figure 4E). This is reflected in both the energy coefficients (lines in Figure 4E) and in the relative enrichment of 12-mers (points in Figure 4E and Figure S4C). We tested these predictions using competition DNA binding experiments. Consistent with our EpiSELEX-seq analysis, using binding sites that contain a CpG at position 9/10 revealed that a

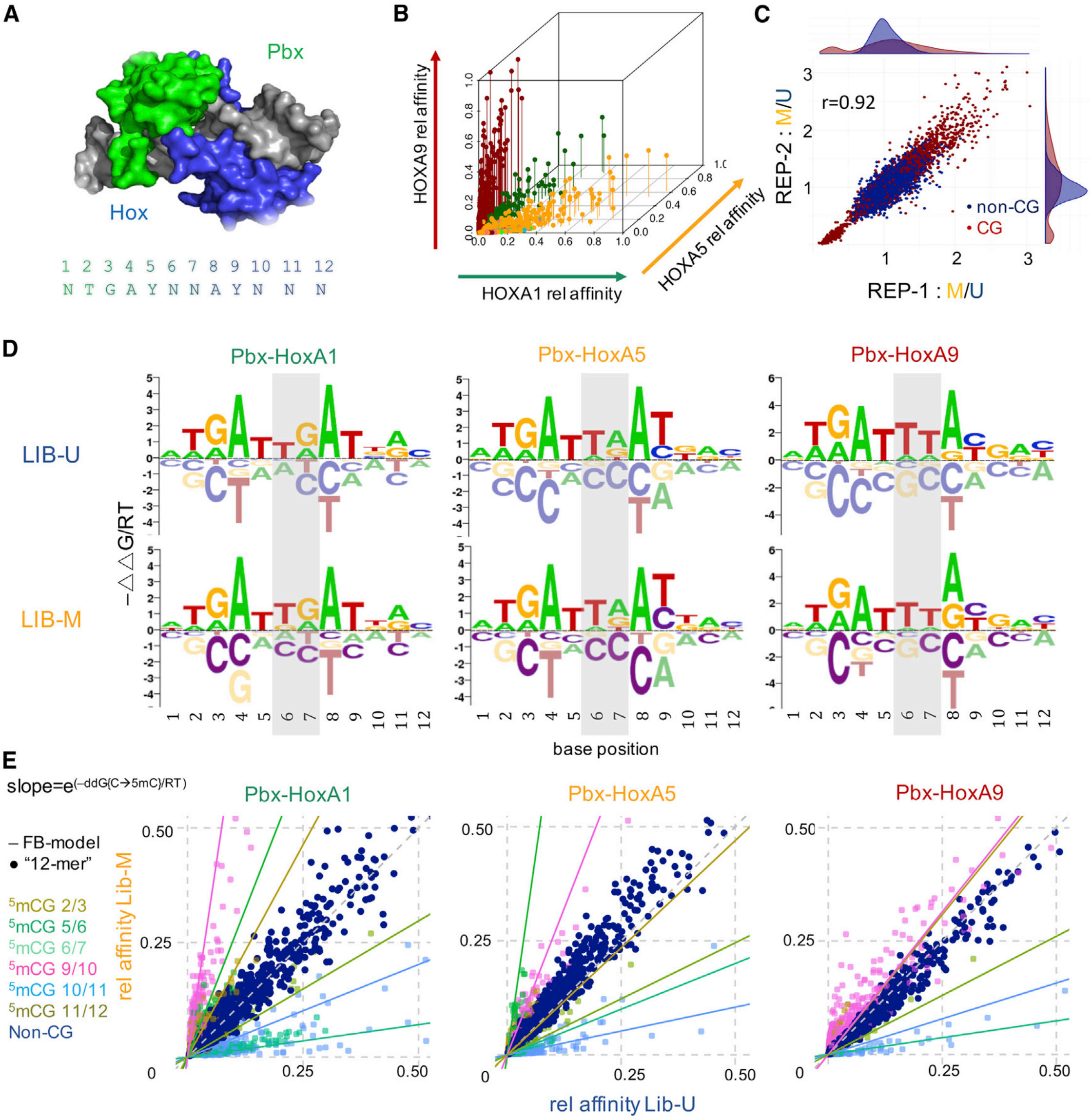


Figure 4. Methylation Sensitivity of Human Pbx-Hox Complexes

(A) Crystal structure (PDB: 1PUF) of human Pbx-HoxA9, with Hox shown in blue and Pbx in green. The consensus sequence with position labels is shown as a reference.

(B) Relative affinity comparison of Pbx1 plus HoxA1, HoxA5, or HoxA9 (green, orange, and red, respectively). Each Hox prefers distinct sets of 12-mers. Preferred central spacers (positions 6 and 7) are TG, TA, and TT for HoxA1, HoxA5, and HoxA9, respectively.

(C) Replicate agreement for EpiSELEX-seq of Pbx1-HoxA9. Methylated/unmethylated (M/U) ratios for 12-mers are shown for one replicate versus the other. Sequences with or without CpGs are shown in red or dark blue, respectively (Pearson correlation of 0.92). Staggered density plots show a narrow distribution of non-CpG 12-mers around 1 but a much broader and bimodal distribution for CpG 12-mers.

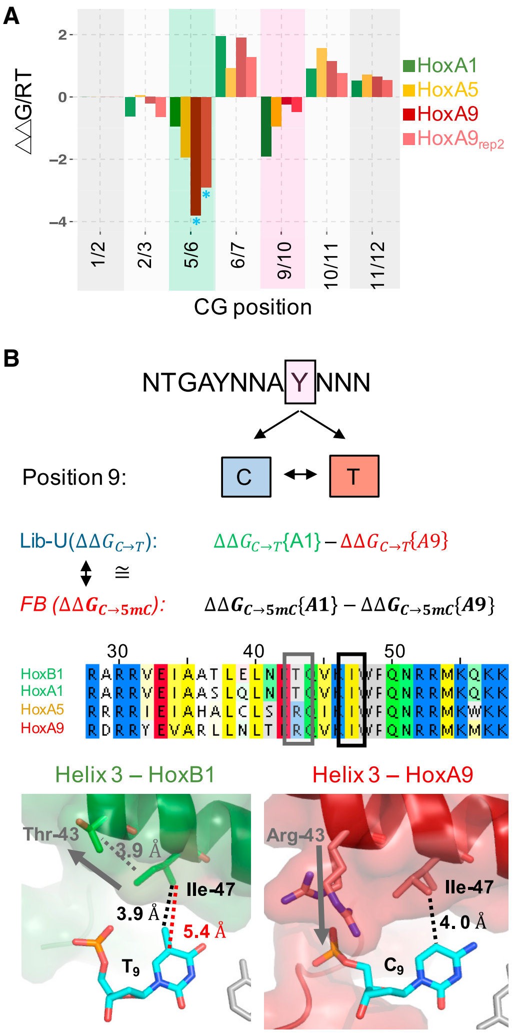
(D) Oligomer-based energy logos for all three Pbx-Hox complexes for Lib-U and Lib-M. No obvious differences between the methylated and unmethylated libraries are observed. The central spacer is shaded in gray.

(E) Lib-M versus Lib-U relative affinity plots for all three complexes. Points are colored based on the position of the CpG dinucleotide (dark blue for non-CpG sequences). The slopes of the lines represent the exponentiated free energy coefficient for the methylation effect in the feature-based (FB) model.

higher concentration was required for unmethylated (inhibitor concentration at which the response is reduced by half [IC_{50}] = 45.5 ± 14.7) than for methylated (IC_{50} = 20.3 ± 2.6) binding sites to compete with a radioactively labeled consensus probe for Pbx-HoxA1 binding (Figure S5).

Thymine Mimicry Explains Variation in Methylation Sensitivity among Hox Paralogs

The effect of methylation on binding not only depends on the position of the CpG dinucleotide within the protein-DNA interface but also differs between Hox paralogs (Figure 5A). At



ΔΔG/RT	HoxA1	HoxA9	"Difference" ΔΔG/RT
C→T	–2.0	+0.0	–2.0
C→^{5}mC	–2.0	–0.3	–1.7

Figure 5. Collinearity of Methylation Sensitivity Explained by Structural Differences

(A) Comparison of the methylation effect for all three Pbx-Hox complexes. The two A9 replicates are shown in different shades of red and have good agreement (blue asterisks indicate that coefficients were fit at sub-optimal affinity thresholds because of low counts). Position 9/10 shows large paralog-dependent differences, with HoxA1 having high, HoxA5 medium, and HoxA9 almost no methylation sensitivity; position 5/6 shows the opposite trend.

(B) Comparison of Hox-specific C or T readout for position 9. HoxA1 prefers a T over a C, whereas HoxA9 has equal preference. The observed difference in binding free energy associated with a C→T transition should equal the methylation sensitivity difference between HoxA1 and HoxA9. Alignment of helix3 of several Hox TFs (B1, A1, A5, A9) reveals conservation of Ile47 for the Hox family but polymorphism at residue 43. Ile47 interacts with the pyrimidine at position 9 in both the HoxB1 and the HoxA9 structures. The distance to the aromatic carbon 5 is 5.4 Å for HoxB1 but only 3.9 Å for HoxA9. Addition of a methyl group in HoxB1 reduces the distance to 4.0 Å, allowing for the same VdW interaction as seen in HoxA9. Arg43 (A9) aids in bringing Ile47 closer to the DNA by interacting with the phosphate backbone at nucleotide C9, whereas Thr43 (B1/A1) does not interact with the backbone but, rather, pulls Ile47 away from T9. The C→T energy difference between HoxA1 and HoxA9 is most likely driven by the methyl readout. The table shows that the C→T free energy difference is comparable with the difference in methylation sensitivity (feature-based model) between the two paralogs.

dinucleotide positions 5/6 and 9/10, the strength of methylation sensitivity is collinear with the Hox expression domain along the anterior-posterior axis (HoxA1-HoxA5-HoxA9), similar to other aspects of Hox function (Slattery et al., 2011). To gain more insight into the structural mechanisms underlying these differences in binding, we compared HoxA1 and HoxA9, which show distinct differences in methylation preference at position 9: Pbx-HoxA1 strongly prefers T over C ($\Delta\Delta G[\mathrm{C}\rightarrow\mathrm{T}]/RT \approx -2.0$), whereas Pbx-HoxA9 shows no such preference (Figure 5B). Close examination of a Pbx-HoxB1 (a proxy for HoxA1) crystal structure reveals that isoleucine at position 47 (Ile47) within the homeodomain has a VdW interaction with the carbon 5 methyl group on base T9 of the forward DNA strand (Figure 5B). In contrast, in a Pbx-HoxA9 crystal structure, Ile47 is closer to and interacts with the C9 base even without this methyl group. Accordingly, we would predict that HoxB1/A1 should benefit from the methylation of a C9, whereas HoxA9 should be indifferent to methylation. Indeed, $\Delta\Delta G[\mathrm{C9}\rightarrow\mathrm{T9}]/RT$ is similar to $(\Delta\Delta G[\mathrm{C9}\rightarrow 5\mathrm{mC9}]/RT)$ for HoxA1, whereas, for HoxA9, $\Delta\Delta G[\mathrm{C9}\rightarrow 5\mathrm{mC9}]/RT$ is close to zero (Figure 5B). Because the crystal structures show no further base-specific interactions at position 9, these differences can be fully accounted for by the relative benefit of gaining a methyl group for each paralog.

EpiSELEX-Seq Identifies Non-consensus p53 Binding Sequences Whose Affinity Is Increased upon Methylation

Because altered methylation patterns are observed in many cancers, we tested whether binding by the human tumor suppressor protein p53 might be methylation-sensitive. In vivo, p53 is thought to bind as a tetramer to two dimer sites, RRRCWWGYYY (which we will refer to as CWWG), separated by a spacer of 0–13 bp (el-Deiry et al., 1992; Funk et al., 1992; Figure 6A). Consistently, the palindromic sequence GGACATGTCC site independently emerged from our data as the most enriched 10-mer in both Lib-M and Lib-U (Figure S6A). Comparing Lib-M and Lib-U directly reveals that there are three different classes of CpG-containing sequences that show altered p53 binding upon methylation (Figure 6A). Methylation of a CpG occurring at the 3′ end of the half site (RRRCATGY**CG**, which we will refer to as $C_{+4}G_{+5}$, relative to the motif center) decreases binding by ~20%, whereas methylation at a CpG shifted 1 bp to the left (RRRCATG**CG**Y or $C_{+3}G_{+4}$) increases binding by ~50%. The largest effect, an ~250% increase in binding affinity, was observed when the CpG was in the core of the binding site (RRRCA**CG**YYY or $C_{+1}G_{+2}$). Analysis of a p53 crystal structure (3Q06; Petty et al., 2011) reveals that the methyl group at carbon 5 of the T_{+1} base pyrimidine ring in the CATG core is stacked above the polar guanidinium plane of p53 amino acid R280. The latter is crucial for p53 binding because it forms hydrogen bonds with the G_{+2} base (Figure 6B). The thymine

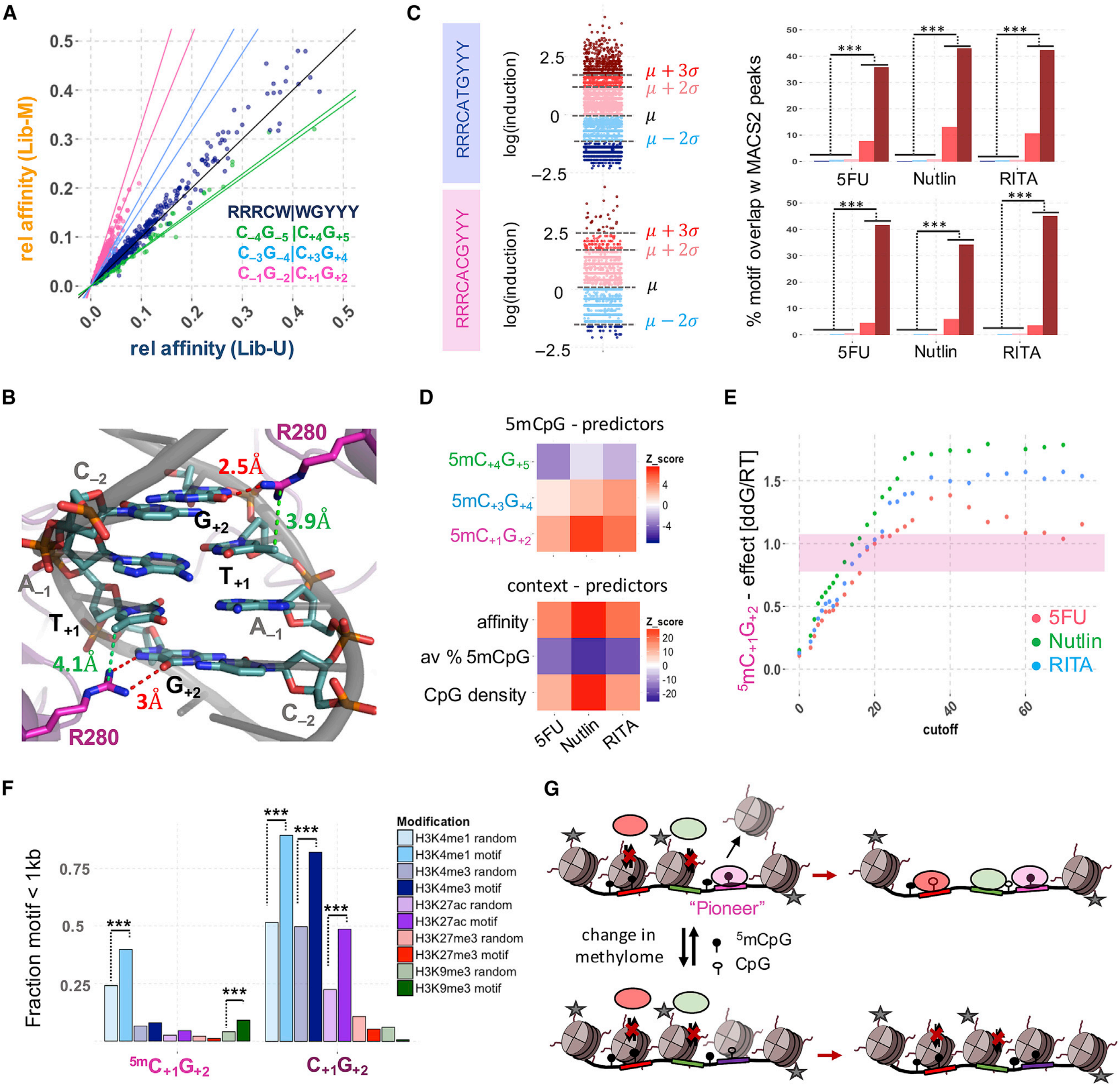


Figure 6. p53 Differentially Binds Methylated Motifs In Vivo in Distinct Chromatin Modification States

(A) EpiSELEX-seq 10-mer relative affinity plot showing the consensus motif (RRRCWWGYYY, blue) and three classes of CpG-containing motifs. CpG motifs are differentially bound upon methylation, with methylation of $C_{4+}G_{5+}$ (green) half-sites reducing binding about 20%, whereas methylation of $C_{3+}G_{4+}$ (cyan) and $C_{1+}G_{2+}$ (pink) sites increases binding ~1.5-fold and ~2- to 3-fold, respectively. Non-CpG consensus sites, as expected, show no difference between Lib-U and Lib-M. The slope of the lines represents the value of ΔΔG associated with methylation at each of the identified CpG positions using the feature-based model; methylation effects related by reverse-complement symmetry, estimated independently, are shown as separate lines.

(B) p53 structure (PDB: 3Q06) showing the DNA interface of a p53 dimer with the *RRR*C<u>A</u>|<u>T</u>G*YYY* core (labeled ± relative to the motif center). The two arginines (R280) form hydrogen bonds with the respective G_{+2} bases of each pentamer half-sites (2.5 and 3.0 Å, red) guided by the methyl groups of the pyrimidine carbon 5 of the T_{+1} base, which stack on top of the polar guanidinium plane (3.9 and 4.0 Å, green), thus constraining the possible orientations of the positive charge in favor of forming hydrogen bonds with G_{+2}. Methylation of a $T_{+1} \rightarrow C_{+1}$ substitution would therefore result in stabilization because of regaining a methyl group at position +1.

(C) Comparison of motif-centric analysis and MACS2 peak calling. Left: distribution of induction levels (defined as the logarithm of the ratio of drug-induced and uninduced IP coverage) for all covered CATG or $C_{1+}G_{2+}$ sites (μ = mean and σ = SD). Right: fraction of decamer sites overlapping with MACS2 peak regions split by their log-transformed induction. For all three drugs and both the consensus CATG and the $C_{1+}G_{2+}$ motifs, there is a highly significant trend between motif-centric induction levels and MACS2 peak calling.

(legend continued on next page)

methyl group might thus direct and constrain R280 toward G_{+2}, which has been proposed to serve as a methylation readout mechanism of zinc-finger proteins (Liu et al., 2013). $T_{+1} \rightarrow C_{+1}$ replacement would thus eliminate the guiding methyl group, providing an explanation for the stabilizing effect of methylation at position C_{+1}.

Evidence for Enhanced p53 Binding to Methylated Sites In Vivo

When unmethylated, sequences of type $C_{+1}G_{+2}$ are bound by p53 at a relative affinity of $<10\%$. However, our analysis shows that binding to these sites is strongly enhanced by cytosine methylation. To test whether this effect on in vitro binding is also observable in vivo, we jointly analyzed whole-genome bisulfite sequencing (ENCODE Project Consortium, 2012) and p53 genomic occupancy data—generated by ChIP-seq both before and after induction of p53—for the cell line MCF7 (Nikulenkov et al., 2012). Using standard peak calling (Zhang et al., 2008) at a false discovery rate of 5%, we detected 40 sites that were both occupied by p53 and had an underlying DNA sequence containing a match to RRRCA**CG**YYY, a sample too small to allow for statistical analysis of the effect of methylation status (Figure S6B). Moreover, the negative effect of methylation on chromatin accessibility in vivo may obscure the positive effect on binding suggested by our SELEX analysis. To address this issue, we developed a motif-centric analysis strategy that avoids peak calling. We started by identifying all individual matches to the most strongly bound RRRCATGYYY sites in the genome and classifying each of these p53 half-sites in terms of the change in the number of ChIPed DNA fragments covering it before and after p53 induction. We observed a strong and statistically significant trend between motif-centric fold induction and the probability of falling within a peak region based on model-based analysis of ChIP-seq (MACS2) (Figure 6C), indicating that this approach captures the underlying p53 binding signature. In addition, this trend was robust for three different inducers of p53 activity and was also observed for the CpG-containing $C_{+1}G_{+2}$ motif (Figure 6C).

Encouraged by this observation, we used a generalized linear model that explains how the number of sequenced immunoprecipitation (IP) fragments covering an individual genomic match to any of the four decamer half-site motif classes (CATG, $C_{+1}G_{+2}$, $C_{+3}G_{+4}$, and $C_{+4}G_{+5}$) is distributed between the uninduced and induced conditions. The CATG motif, which does not match any CpG-containing decamers, serves to estimate the effect of local chromatin context, which is represented by the average methylation level and CpG content of the flanking regions as predictors in the model. To account for variation in binding affinity unrelated to methylation, we also included as a covariate the relative affinity of the 10-bp half-site as derived from the DNA sequence using a scoring matrix derived from our Lib-U data (see Experimental Procedures and Supplemental Experimental Procedures for details). Finally, and most importantly, the coefficients associated with three binary indicators for the presence of a methylated CpG dinucleotide at each offset quantify the effect of cytosine methylation on the responsiveness of in vivo p53 binding.

When the model is fit to ChIP-seq data, the position-dependent effects of cytosine methylation within the binding site identified by our EpiSELEX-seq assay are recapitulated in MCF7 cells, with methylation of $C_{+1}G_{+2}$ having a significant stabilizing effect (Figure 6D). The coefficients for the confounding contributions in the model also behave as expected, with positive effects for CpG density and sequence-derived p53 affinity and a negative effect for regional methylation (Figure 6D). Considering that the in vivo methylation effects should more closely reflect the in vitro effect at higher levels of ChIP enrichment, where the local chromatin context is, presumably, more permissive, we repeated our model fit using increasing cutoffs on the sum of induced and uninduced read counts for all consensus matches in the genome (Figure 6E). The coefficient for $C_{+1}G_{+2}$ behaves as expected and saturates at $\Delta\Delta G/RT = +1.5$, corresponding to an $\sim$4.5-fold increase in binding affinity upon full methylation of the CpG dinucleotide (Figure 6E). Thus, the in vivo methylation effect appears to be even higher than in vitro, which could reflect contributions from additional methylated CpG dinucleotides within the full p53 tetramer binding site or cooperativity with other factors. For the other two motif classes ($C_{+3}G_{+4}$ and $C_{+4}G_{+5}$), the coverage by IP fragments is too sparse to allow quantification, consistent with the weaker in vitro methylation sensitivity observed for these CpG offsets with our EpiSELEX-seq assay.

It has been suggested that p53 can bind to high-nucleosome-occupancy regions and act as a pioneer factor to alter chromatin accessibility (Laptenko et al., 2011; Sammons et al., 2015). We therefore analyzed five histone modifications that, in combination, can be used to classify enhancers or promoters as active, closed, or primed (Calo and Wysocka, 2013). Methylated $C_{+1}G_{+2}$ sites are significantly enriched for H3K9me3 and H3K4me1 but not H3K27ac (associated with active enhancers)

(D) Feature model fits of drug-induced (5FU, Nutlin, and RITA) in vivo P53 ChIP-seq data for MCF7 using Lib-U relative affinities, average methylation levels, and CpG density within a 500-bp region as context-dependent predictors and three position-specific binary methylation indicator features. Datasets were subsampled to 50 sites for each possible methylation-motif combination (see Experimental Procedures for details). Top: the significance of the methylation features, with red signifying positive and blue negative effects on binding. *Z* scores for $C_{1+}G_{2+}$ ranges from 3.0 (5Fu) to 6.3 (Nutlin). Bottom: the scores for the context-dependent, confounding model predictors (highly significant across all drugs).

(E) Methylation coefficients for the $C_{1+}G_{2+}$ sites were computed on the entire dataset using the feature-based model from (D), with increasing cutoffs on the sum of uninduced and drug-induced p53 IP coverage. The pink area shows the expected difference in binding free energy from EpiSELEX-seq results.

(F) Overlap with peaks of histone modifications (<1 kb) for methylated and unmethylated $C_{1+}G_{2+}$ motifs (>2 SD above mean induction, dark shade). Equally sized, methylation-matched random control sets (light shade) show the expected overlap. Primed enhancer (H3K4me1) and heterochromatin (H3K9me3) modifications, but not marks of active transcription, are significantly enriched in methylated $C_{1+}G_{2+}$ sites, whereas unmethylated $C_{1+}G_{2+}$ sites show patterns of active transcription (H3K4me1, H3K4me3, and H3K27ac), perhaps reflecting increased accessibility at active promoters.

(G) Potential mechanism of how aberrant methylation patterns might contribute to altered p53 binding and, thus, potentially contribute to changes in chromatin landscape and gene regulation.

or H3K4me3 (associated with active transcription) compared with a matched control set (see Experimental Procedures for details; Figure 6F). These histone modifications have been suggested to mark either heterochromatin (H3K9me3) (Grewal and Jia, 2007) or enhancers that are primed to become active (H3K4me1) (Calo and Wysocka, 2013). We observed the same pattern for CATG sites within methylated regions (Figure S6C). By contrast, unmethylated $C_{+1}G_{+2}$ sites tend to have a strong signature of H3K4me1 and H3K4me3 or H3K27ac (Figure 6F), arguing that ChIP enrichment at those loci may be due to transcriptional activity rather than specific p53 targeting. This again underscores the need to account for confounding effects when analyzing in vivo binding data.

Interestingly, 67 of 90 (74%, with 44% expected, $p = 3 \times 10^{-10}$) of the methylated $C_{+1}G_{+2}$ sites occur within 3 kb of a protein-coding gene (60 genes total) or long non-coding RNAs (lncRNAs) (20 total) (Figure S6D) annotated in GENCODE (Derrien et al., 2012). The enrichment for sites occurring near lncRNAs (21 of 90 sites or 23%, with 8% expected, $p = 5 \times 10^{-7}$) (Figure S6D) is consistent with previous findings about p53 regulation of lncRNA expression (Léveillé et al., 2015).

DISCUSSION

With EpiSELEX-seq, we have developed a method that can accurately quantify the change in binding free energy associated with the presence of a methylated cytosine at any position within the protein DNA interface. A key aspect of our approach, which allows us to robustly identify methylation sensitivity, is that modified and unmodified DNA ligands are probed simultaneously in a single reaction, ensuring a direct comparison of TF occupancy. One round of selection is sufficient to accurately capture methylation effects, even for lower-affinity sites that deviate from the consensus and, thus, readily escape detection when binding to methylated and unmethylated ligands is assayed separately or over multiple rounds of selection. The context-sensitive nature of our analysis is essential because opposing methylation effects can occur within a single binding site, making it difficult or impossible to detect the effect of methylation using less precise approaches, such as oligomer enrichment only. This point is illustrated by our analysis of human Pbx-Hox heterodimers, whose DNA binding specificity we studied here for the first time at high resolution. The net effect of methylation on binding is close to neutral, but methylation of different CpGs in the binding site can modulate the binding affinity by up to 7-fold in either direction. This also illustrates why it may be difficult to detect methylation sensitivity by looking at motif enrichment in differentially methylated regions (DMRs). Pbx-Hox sequence logos constructed separately for the unmethylated and methylated libraries were nearly indistinguishable and did not reveal significant methylation sensitivity of Pbx-Hox complexes (Figure 4D). Only when we examined the consequences of methylation at specific positions were we able to identify clear effects.

Despite an ongoing debate concerning to what extent CpG methylation is a driver of gene silencing or the consequence thereof (Ambrosi et al., 2017), the general view is that methylation has a repressive effect on TF binding. For example, in a study that compared binding of TFs between wild-type and Dnmt1 knockout embryonic stem cells (ESCs) (Domcke et al., 2015), the authors showed that removal of methylation marks at specific nuclear respiratory factor 1 (NRF1) binding sites led to increased binding and expression of nearby genes. In addition, experimentally induced methylation reduced NRF1 binding to those sites. Here, and in agreement with recently published data (Yin et al., 2017), we demonstrate that endogenous methylated motifs containing a CpG at specific sites within the protein-DNA interface can also increase binding and that the mechanisms underlying the epigenetic control of TF binding and, thus, gene expression are more nuanced than previously thought.

For p53, despite a general negative effect of regional methylation on genomic occupancy, the increased binding to methylated RRRCACGYYY sites our analysis revealed implies that methylated binding sites are functional and might direct p53 to alter previously inaccessible loci in the genome. This conclusion is supported by our finding that these occupied and methylated binding sites are associated with a histone modification pattern that indicates either compacted chromatin (Grewal and Jia, 2007) or transcriptionally poised enhancers. Additional evidence that p53 can access nucleosomal DNA in vitro and in vivo and, thus, might be a pioneer factor also supports this notion (Laptenko et al., 2011; Sammons et al., 2015). Many diseases, in particular many forms of cancer, are accompanied by aberrant methylation patterns (Kulis and Esteller, 2010), raising the question whether even subtle changes in the methylome could trigger differential TF binding and, thus, contribute to the onset of disease. Interestingly, H3K4me1 has also been shown to be significantly associated with loss of methylation during aging in multiple human cell types (Fernández et al., 2015), providing yet additional support for the functionality and importance of such sites.

EXPERIMENTAL PROCEDURES

Protein Expression and Purification

Human proteins, either full-length or nearly full-length, were affinity-purified using a polyhistidine tag (HIS). p53 protein was purified as described in Laptenko et al. (2015), containing a deletion in the C-terminal basic region to prevent non-specific DNA binding contributions outside of the core DNA-binding domain.

Library Design and Methylation

Full library sequences were as follows: 5′-GGTAGTGGAGG-TGGG-CCTGG-16(26)xN-CCAGG-GAGGTGGAGTAGG-3′ for Lib-U and 5′-GGTAGTG-GAGG-GCAC-CCTGG-16(26)xN-CCAGG-GAGGTGGAGTAGG-3′ for Lib-M. Libraries were double-stranded by annealing and extension using Klenow polymerase (New England Biolabs). Lib-M was methylated with M.SssI (NEB) using only ~250 ng/1× reaction and two subsequent incubation cycles of 2 hr at 37°C each. Up to 400 ng of 1× methylated DNA can be combined in the second step using Oligo-Clean-up columns (ZymoGenetics) for purification.

EpiSELEX-Seq Protocol

EMSAs and extraction of bound DNA were performed as described previously (Slattery et al., 2011) with an equal mix of Lib-U and Lib-M. Purified, bound DNA was split in two and amplified using a 15-cycle PCR protocol with high-fidelity enzymes (Phusion or Q5, NEB), with overhang primers adding TruSeq Illumina adaptor sites in two orientations, respectively, to allow sequencing from both 5′ and 3′ ends. Efficient splitting was analyzed by comparing the number of reads resulting from each set of primers (Figure S1). Specific Illumina bar codes were added by a five-cycle PCR using NEBNext Multiplex Oligos for Illumina sequencing and Phusion/Q5 polymerase. Indexed libraries were gel-purified as described previously (Slattery et al., 2011), pooled, and sequenced using a v2 75-cycle high-output kit on an Illumina NEXTSeq series desktop sequencer. For initial (R0) and enriched (R1) libraries, 5–35 million single-end reads were obtained.

Testing for Methylation Efficiency

A DNA probe containing four CpG sites was methylated, bisulfite-treated, and cloned into the pBlueScript vector; four to eight colonies were picked for sequencing. We assessed methylation efficiency by counting retained CpG dinucleotides. Alternatively, Lib-M was split into three parts (methylated and not treated, methylated and treated, unmethylated and treated with bisulfite) and prepared for sequencing as described above. Dinucleotide frequencies were computed, and CpN ratios were compared across all treatments to assess methylation efficiency using "TpG" as a reference.

EpiSELEX-Seq Data Processing

FASTQ files were pre-processed using the FASTX toolkit (Hannon lab). Files were reverse-complemented, merged, and trimmed to have unidirectional reads starting from the 4-bp bar code site. Each dataset was assigned to either Lib-U or Lib-M. A fifth-order Markov model was generated from the R0 libraries for Lib-U and Lib-M, respectively, using the R package bioconductor.org/packages/SELEX (Riley et al., 2014).

Oligomer Enrichment Analysis

Relative affinities for oligomers of length *k* were estimated as described previously (Slattery et al., 2011). Fold enrichments were normalized based on the most enriched oligomer for each library. Position-specific affinity matrices (PSAMs) (Foat et al., 2006) were generated by considering all 3*k* point mutations away from the most enriched oligomer. Binding free energy differences $\Delta\Delta G/RT$ between the mutated and optimal sequence were computed as the negative logarithm of the relative fold enrichment. The binding free energy change $\Delta\Delta G/RT$ for a $\mathrm{C}\rightarrow\mathrm{T}$ transition was calculated separately for Lib-U and Lib-M and used to estimate the effect of methylation on binding as follows:

$$\Delta\Delta G[\mathrm{C}\rightarrow 5\mathrm{mC}] \approx \Delta\Delta G[\mathrm{C}\rightarrow \mathrm{T}]_{\mathrm{Lib-U}} - \Delta\Delta G[\mathrm{C}\rightarrow \mathrm{T}]_{\mathrm{Lib-M}}.$$

Feature-Based Modeling

A feature-based Poisson regression model was fit to the R1 read counts. First, PSAMs were constructed from oligomer enrichment tables for each sample and used to estimate relative binding affinity in either orientation and at each offset relative to the random region, allowing for up to 2-bp overlap with the constant flanks. Only probes for which a single offset/orientation contributing at least 95% of the total were kept. To avoid bias, R1 reads were randomly split; one half was used to define the set of oligomers that correspond to the rows in the design matrix and the other to obtain R1 counts. Regression models were fit in two ways: (1) using the Lib-U R1 count for a particular motif of length *k*, the Markov model prediction from the corresponding R0 as an offset, and 4*k* base indicator features at each position in the motif as independent variables; (2) the same as before but including Lib-M and using both base and ^{5}mCpG features. As expected, adding $^{5\ m}$CpG features and jointly fitting to Lib-U and Lib-M did not affect the base feature estimates (Figure S1C). For p53, a separate intercept was fit for each library.

Competition Assay for Pbx-HoxA1

Two 12-mer competitor probes with identical sequence—ATGATTGA**CG**AC—but different methylation statuses at position 9 were tested for their capacity to compete with a labeled probe for Pbx-HoxA1 binding in an EMSA. Pbx-HoxA1 and labeled probe concentrations were held constant while increasing the concentrations of the unlabeled competitor DNA over a 1,000-fold range. Experiments were performed in duplicate. IC_{50} values were calculated using ImageJ for quantification and R-package drc to fit a dose-response curve.

Data Processing for In Vivo p53 Binding

We downloaded ChIP-seq data (FASTQ files, Sequence Read Archive accession number: SRP007261) for p53 in MCF7 cells (no-drug control and drugs Nutlin, reactivation of p53 and induction of tumor cell apoptosis [RITA], and fluorouracil [5FU]), and MCF7 whole-genome bisulfite sequencing data (browser extensible data [BED] file, GSM1328112). FASTQ files were aligned to hg19 (Bowtie2) and converted to coverage tracks (deeptools/bamCoverage) after extending reads by 200 nt. MACS2 was run with options –g hs and –q 0.05 using the uninduced p53 IP as a control. BED peak files (hg19) for five histone modifications in MCF7 were downloaded (ENCODE: ENCSR000EWP, ENCSR000EWQ, ENCSR000EWR, ENCSR493NBY, and ENCSR985MIB). General transfer format (GTF) files for the current releases (v25) of human whole-genome annotation and lncRNA-specific annotation data were downloaded from GENCODE (mapped to GRCh37/hg19).

In Vivo Motif-Centric p53 Binding Analysis

Analysis was done in *R*. The hg19 genome was scanned for sites mapping to the consensus RRRCATGYYY or the three CpG-containing motif classes (RRRCATGYCG, RRRCATGCGY, and RRRCACGYYY). CpG sites were intersected with whole genome bisulfite sequencing (WGBS) BED files, and methylation status was assigned based on the percentage methylated ("1" for > 80% and "0" for < 10% methylated), keeping only sites with $\geq$ 10× coverage. Average methylation levels and CpG density of the 500 bp centered around the motif were computed, and the per-motif p53 coverage for uninduced (control) and drug-induced (IP) were obtained, keeping only motifs with $\geq$ 1× coverage in both control and IP. Individual genomic motif occurrences correspond to the rows of the design matrix (X); columns are the in vitro 10-mer affinities (unmethylated EpiSELEX-seq), the three position-specific binary ^{5}mCpG indicators, average methylation level, and CpG density. The glm function with family = "binomial" was used in *R* to fit the following model of the probability of a specific motif being bound:

$$p(bound) \equiv \frac{IP}{IP+CTRL} = \frac{1}{1+e^{-\frac{\Delta\Delta G}{RT}}},$$

where

$$\frac{\Delta\Delta G}{RT} = \sum_{\phi}\beta_{\phi}X_{\phi}.$$

Regression coefficients and *Z* scores quantifying their statistical significance were obtained for each model fit. To avoid bias in the size of motif classes in the training data, sub-sampling (~200 times) was applied in a way that guarantees an equal number of occurrences of methylation status and motif class (for the CATG motif, a 50% regional methylation level was used as a threshold). Models for enriched p53 occupancy were fit by sequentially removing rows for which the sum of drug-induced and uninduced IP fragment counts fell below a certain threshold. Fisher's exact test was used to compute statistical associations between MACS2 peaks (1 kb around the peak summit) and grouped motif-centered log-transformed enrichment values.

Overlap with GENCODE Annotation and Histone Marks

Enrichment for either GENCODE annotations or histone marks was scored by computing the overlap between the motif sets (log induction > 2 SD above the mean) and either the gene annotations (within 3 kb) or the histone peaks (within 1 kb). p Values were computed by generating >100 random methylation-matched sets from the WGBS data and calculating the probability of observing the actual overlap based on the sampling of random overlaps.

ACCESSION NUMBERS

The accession number for the raw sequencing data and tables reported in this paper, including the PSAM-filtered Lib-M and Lib-U relative affinities and respective oligomer sequences for all protein complexes, is NCBI GEO: GSE98652.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2017.05.069.

AUTHOR CONTRIBUTIONS

J.F.K., R.S.M., and H.J.B. conceived the experiments. J.F.K. carried out the experiments with help from O.L. and S.C. J.F.K., G.D.M., and H.J.B. constructed the feature-based model. J.F.K. analyzed the ChIP-seq and WGBS data. O.L., W.A.F.P., and C.P. helped interpret the p53 data. J.F.K., R.S.M., and H.J.B. interpreted the data and wrote the paper. All authors discussed the findings and contributed to the manuscript.

ACKNOWLEDGMENTS

We thank Remo Rohs, Timothy Bestor, and Barry Honig, as well as the members of the Bussemaker, Mann, and Rohs labs for valuable discussions. This work was supported by a Howard Hughes Medical Institute International Student Research Fellowship (to J.F.K.) and NIH grants R01HG003008 (to H.J.B.), R35GM118336 (to R.S.M.), and P01CA087497 and R01CA196234 (to C.P.). Columbia University's Shared Research Computing Facility is supported by NIH grant G20RR030893 and NYSTAR contract C090171.

Received: August 31, 2016
Revised: April 7, 2017
Accepted: May 22, 2017
Published: June 13, 2017

REFERENCES

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

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. Science *324*, 1720–1723.

Baylin, S.B., and Jones, P.A. (2011). A decade of exploring the cancer epigenome - biological and translational implications. Nat. Rev. Cancer *11*, 726–734.

Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? Mol. Cell *49*, 825–837.

Dantas Machado, A.C., Zhou, T., Rao, S., Goel, P., Rastogi, C., Lazarovici, A., Bussemaker, H.J., and Rohs, R. (2015). Evolving insights on how cytosine methylation affects protein-DNA binding. Brief. Funct. Genomics *14*, 61–73.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

Domcke, S., Bardet, A.F., Adrian Ginno, P., Hartl, D., Burger, L., and Schübeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. Nature *528*, 575–579.

el-Deiry, W.S., Kern, S.E., Pietenpol, J.A., Kinzler, K.W., and Vogelstein, B. (1992). Definition of a consensus binding site for p53. Nat. Genet. *1*, 45–49.

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

Fernández, A.F., Bayón, G.F., Urdinguio, R.G., Toraño, E.G., García, M.G., Carella, A., Petrus-Reurer, S., Ferrero, C., Martinez-Camblor, P., Cubillo, I., et al. (2015). H3K4me1 marks DNA regions hypomethylated during aging in human stem and differentiated cells. Genome Res. *25*, 27–40.

Foat, B.C., Morozov, A.V., and Bussemaker, H.J. (2006). Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. Bioinformatics *22*, e141–e149.

Fu, Y., Luo, G.Z., Chen, K., Deng, X., Yu, M., Han, D., Hao, Z., Liu, J., Lu, X., Doré, L.C., et al. (2015). N6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. Cell *161*, 879–892.

Funk, W.D., Pak, D.T., Karas, R.H., Wright, W.E., and Shay, J.W. (1992). A transcriptionally active DNA-binding site for human p53 protein complexes. Mol. Cell. Biol. *12*, 2866–2871.

Greer, E.L., Blanco, M.A., Gu, L., Sendinc, E., Liu, J., Aristizábal-Corrales, D., Hsu, C.H., Aravind, L., He, C., and Shi, Y. (2015). DNA Methylation on N6-Adenine in C. elegans. Cell *161*, 868–878.

Grewal, S.I., and Jia, S. (2007). Heterochromatin revisited. Nat. Rev. Genet. *8*, 35–46.

Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S.B., Buil, A., Ongen, H., Yurovsky, A., Bryois, J., Giger, T., Romano, L., Planchon, A., et al. (2013). Passive and active DNA methylation and the interplay with genetic variation in gene regulation. eLife *2*, e00523.

Hashimshony, T., Zhang, J., Keshet, I., Bustin, M., and Cedar, H. (2003). The role of DNA methylation in setting up chromatin structure during development. Nat. Genet. *34*, 187–192.

Hellman, A., and Chess, A. (2007). Gene body-specific methylation on the active X chromosome. Science *315*, 1141–1143.

Hu, S., Wan, J., Su, Y., Song, Q., Zeng, Y., Nguyen, H.N., Shin, J., Cox, E., Rho, H.S., Woodard, C., et al. (2013). DNA methylation presents distinct binding sites for human transcription factors. eLife *2*, e00726.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327–339.

Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature *527*, 384–388.

Jones, P.A., and Baylin, S.B. (2007). The epigenomics of cancer. Cell *128*, 683–692.

Kulis, M., and Esteller, M. (2010). DNA methylation and cancer. Adv. Genet. *70*, 27–56.

Laptenko, O., Beckerman, R., Freulich, E., and Prives, C. (2011). p53 binding to nucleosomes within the p21 promoter in vivo leads to nucleosome loss and transcriptional activation. Proc. Natl. Acad. Sci. USA *108*, 10385–10390.

Laptenko, O., Shiff, I., Freed-Pastor, W., Zupnick, A., Mattia, M., Freulich, E., Shamir, I., Kadouri, N., Kahan, T., Manfredi, J., et al. (2015). The p53 C terminus controls site-specific DNA binding and promotes structural changes within the central DNA binding domain. Mol. Cell *57*, 1034–1046.

Lazarovici, A., Zhou, T., Shafer, A., Dantas Machado, A.C., Riley, T.R., Sandstrom, R., Sabo, P.J., Lu, Y., Rohs, R., Stamatoyannopoulos, J.A., and Bussemaker, H.J. (2013). Probing DNA shape and methylation state on a genomic scale with DNase I. Proc. Natl. Acad. Sci. USA *110*, 6376–6381.

Léveillé, N., Melo, C.A., Rooijers, K., Díaz-Lagares, A., Melo, S.A., Korkmaz, G., Lopes, R., Akbari Moqadam, F., Maia, A.R., Wijchers, P.J., et al. (2015). Genome-wide profiling of p53-regulated enhancer RNAs uncovers a subset of enhancers controlled by a lncRNA. Nat. Commun. *6*, 6520.

Liu, Y., Zhang, X., Blumenthal, R.M., and Cheng, X. (2013). A common mode of recognition for methylated CpG. Trends Biochem. Sci. *38*, 177–183.

Mann, I.K., Chatterjee, R., Zhao, J., He, X., Weirauch, M.T., Hughes, T.R., and Vinson, C. (2013). CG methylated microarrays identify a novel methylated sequence bound by the CEBPB|ATF4 heterodimer that is active in vivo. Genome Res. *23*, 988–997.

Merabet, S., and Mann, R.S. (2016). To Be Specific or Not: The Critical Relationship Between Hox And TALE Proteins. Trends Genet. *32*, 334–347.

Miller, M. (2009). The importance of being flexible: the case of basic region leucine zipper transcriptional regulators. Curr. Protein Pept. Sci. *10*, 244–269.

Nikulenkov, F., Spinnler, C., Li, H., Tonelli, C., Shi, Y., Turunen, M., Kivioja, T., Ignatiev, I., Kel, A., Taipale, J., and Selivanova, G. (2012). Insights into p53 transcriptional function via genome-wide chromatin occupancy and gene expression analysis. Cell Death Differ. *19*, 1992–2002.

Paz, M.F., Fraga, M.F., Avila, S., Guo, M., Pollan, M., Herman, J.G., and Esteller, M. (2003). A systematic profile of DNA methylation in human cancer cell lines. Cancer Res. *63*, 1114–1121.

Petty, T.J., Emamzadah, S., Costantino, L., Petkova, I., Stavridi, E.S., Saven, J.G., Vauthey, E., and Halazonetis, T.D. (2011). An induced fit mechanism regulates p53 DNA binding kinetics to confer sequence specificity. EMBO J. *30*, 2167–2176.

Razin, A., and Cedar, H. (1994). DNA methylation and genomic imprinting. Cell *77*, 473–476.

Riley, T.R., Slattery, M., Abe, N., Rastogi, C., Liu, D., Mann, R.S., and Bussemaker, H.J. (2014). SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. Methods Mol. Biol. *1196*, 255–278.

Sammons, M.A., Zhu, J., Drake, A.M., and Berger, S.L. (2015). TP53 engagement with the genome occurs in distinct local chromatin environments via pioneer factor activity. Genome Res. *25*, 179–188.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H.J., and Mann, R.S. (2011). Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell *147*, 1270–1282.

Stein, R., Razin, A., and Cedar, H. (1982). In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. Proc. Natl. Acad. Sci. USA *79*, 3418–3422.

Tribioli, C., Tamanini, F., Patrosso, C., Milanesi, L., Villa, A., Pergolizzi, R., Maestrini, E., Rivella, S., Bione, S., Mancini, M., et al. (1992). Methylation and sequence analysis around EagI sites: identification of 28 new CpG islands in XQ24-XQ28. Nucleic Acids Res. *20*, 727–733.

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431–1443.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F., et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science *356*, eaaj2239.

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

Zhang, G., Huang, H., Liu, D., Cheng, Y., Liu, X., Zhang, W., Yin, R., Zhang, D., Zhang, P., Liu, J., et al. (2015). N6-methyladenine DNA modification in Drosophila. Cell *161*, 893–906.

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