Shared Nucleotide Flanks Confer Transcriptional Competency to bZip Core Motifs

Daniel M. Cohen[1,3], Hee-Woong Lim[2,3], Kyoung-Jae Won[2,3] and David J. Steger[1,3,*]

[1]Division of Endocrinology, Diabetes, and Metabolism, Department of Medicine, [2]Department of Genetics, [3]The Institute for Diabetes, Obesity, and Metabolism, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104

*Address Correspondence to:

David J. Steger, Ph.D.

12-103 Smilow Center for Translational Research

3400 Civic Center Boulevard, Bldg 421

Philadelphia, PA 19104-5160

Phone: (215) 746-8520; email: stegerdj@pennmedicine.upenn.edu

ABSTRACT

Sequence-specific DNA binding recruits transcription factors (TFs) to the genome to regulate gene expression. Here, we perform high resolution mapping of CEBP proteins to determine how sequence dictates genomic occupancy. We demonstrate a fundamental difference between the sequence repertoire utilized by CEBPs in vivo versus the palindromic sequence preference reported by classical in vitro models, by identifying a palindromic motif at less than 1% of the genomic binding sites. On the native genome, CEBPs bind a diversity of related 10 bp sequences resulting from the fusion of degenerate and canonical half-sites. Altered DNA specificity of CEBPs in cells occurs through heterodimerization with other bZip TFs, and approximately 40% of CEBP-binding sites in primary human cells harbor motifs characteristic of CEBP heterodimers. In addition, we uncover an important role for sequence bias at core-motif-flanking bases for CEBPs and demonstrate that flanking bases regulate motif function across mammalian bZip TFs. Favorable flanking bases confer efficient TF occupancy and transcriptional activity, and DNA shape may explain how the flanks alter TF binding. Importantly, motif optimization within the 10-mer is strongly correlated with cell-type-independent recruitment of CEBPβ, providing key insight into how sequence sub-optimization affects genomic occupancy of widely expressed CEBPs across cell types.

INTRODUCTION

Sequence-specific DNA binding by transcription factors (TFs) is fundamental to the establishment and maintenance of gene programs that drive cell function in health and disease (1, 2). The genomic distribution of TFs at enhancers and promoters defines the framework by which these proteins orchestrate temporal and spatial regulation of gene expression (3, 4). The genomic landscape of TF-binding sites (TFBSs) is organized by the non-random distribution of DNA recognition sequences, or motifs, that mediate recruitment of their cognate TFs (5). Consequently, defining the motif preferences employed by each TF and mapping the genomic locations of motifs are key to unlocking the basis for gene regulatory networks.

High-throughput approaches have facilitated the identification of TFBSs both in vitro and in vivo (6–8). Protein binding microarrays (PBMs) and high-throughput in vitro selection (HT-SELEX) have determined the specificities of hundreds of isolated TFs from multiple species (9, 10). Alternatively, chromatin immunoprecipitation combined with next generation sequencing (ChIP-seq) has been employed extensively to locate where TFs occupy the native genome and to interrogate motifs from overrepresented sequences in ChIP-seq peaks. In spite of the high information content of consensus sequences, experimentally determined TF motifs have limited ability to predict in vivo binding (11). DNA accessibility (12–14) as well as contextualizing factors including DNA shape (15–19), DNA methylation (20–22), neighboring TF interaction (23, 24) and altered sequence specificity due to heterodimer formation between related TFs (25) constrain and reshape how TFBSs are utilized in native genomes. Collectively, these variables help explain why TFs occupy only a small fraction of candidate motifs in the genome (26).

Contrary to the strong sequence dependence of TF binding in vitro, it has been suggested that TFs are recruited independently of their cognate sequences at many ChIP-seq peaks either through indirect protein-protein interaction (tethering) (24, 27–30) or through recognition of DNA shape (31, 32). Coupled with the observation that motif scores fail to differentiate between bound versus unbound genomic sequences (18, 33, 34), the question of what constitutes the minimal sequence determinants for TFBSs in vivo has become increasingly uncertain. Fortunately, new experimental approaches are providing avenues to address this question. High resolution (20-50 bp) mapping of bound genomic sequences has been facilitated by the development of ChIP with lambda exonuclease digestion and sequencing (ChIP-exo) (35). Close discrimination of bound motifs can revise and improve recognition sequences (36, 37) and resolve dimeric versus monomeric binding (38, 39). In parallel, comparison of bound and unbound motifs in biochemical assays of TF binding to histone-free genomic DNA is providing further insight into the native sites that are sufficient to mediate occupancy

(40–43). Uniting these approaches has the potential to bridge major gaps in our understanding of the relationship between TF sequence specificity, motif occurrence and occupancy at native genomic sites.

CEBP TFs are particularly interesting in terms of how DNA-binding specificity defines genomic occupancy for two key reasons. As lineage determining TFs in several tissues (44–47), CEBPs may function as pioneer factors that overcome the inhibitory effects of chromatin, and thus defining their sequence specificity may be instructive as to whether a relationship exists between binding site affinity and TF occupancy in the genome. In addition, CEBPs can bind DNA as both homodimers and heterodimers (47), and their ability to target different sequence motifs through heterodimerization with other bZip family members (25, 41, 48–50) may enable the utilization of a broad repertoire of motifs to control a variety of gene expression programs (51). Indeed, CEBPs occupy tens of thousands of sites in primary cells and tissues (26, 52–54), however degenerate ChIP-seq motifs obscure the importance of sequence determinants for binding site selection.

Here, we report the high-resolution mapping of CEBP-binding sites in the human and mouse genomes using ChIP-exo. We find that CEBPs occupy a large repertoire of sequences in vivo defined by the fusion of canonical and degenerate CEBP half sites. Positive selection for the nucleotide composition observed within the core motif reflects altered sequence specificity of CEBP homo- versus heterodimers. We demonstrate the importance of the CEBP 10-mer motif by identifying an optimal sequence that is prevalently bound independent of cell type, suggesting that it forms a high-affinity-binding site that overrides chromatin context. Moreover, we reveal a critical role for negative sequence selection, i.e. the exclusion of a particular base, at the first and last position of the 10-mer. Negative selection for specific flanking nucleotide pairs is a general feature shared by multiple bZIP TF motifs, and the distinction between favorable and unfavorable motif contexts correlates with changes in DNA shape. We illustrate the functional importance of flanking base composition by showing that natural genetic variation from single nucleotide polymorphisms (SNPs) that introduce non-permissive flanking bases leads to strain-specific CEBP occupancy in mice. Collectively, these findings provide an expanded motif definition for CEBP that can be generalized to the broader bZip family, and establish important relationships between motif optimization, genomic occupancy and transcriptional activity.

MATERIALS AND METHODS

Animal experiments were reviewed and approved by the Institutional Animal Care and Use Committees of the University of Pennsylvania. Mice were kept under standardized conditions with

water and food *ad libitum* in a pathogen-free animal facility. hMSCs were obtained from Lonza and maintained in low glucose Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum and 2mM glutamine. hMSCs were used at passages 4 through 7. ChIP-exo was performed as described previously (38) with minor modifications explained in the Supplementary Information. Extended and detailed methods for computational analyses of the genomics data are available in the Supplementary Information. Luciferase reporter assay and DNA pyrosequencing are described in the Supplementary Information.

RESULTS

CEBP proteins recognize a diversity of genomic sequences through a degenerate half site.
To identify the genomic sequences targeted by CEBP TFs with high resolution in the native genome, we performed ChIP-exo in primary human mesenchymal stem cells (hMSCs) for CEBPβ as well as in mouse liver tissue for CEBPα and CEBPβ. The approach uses lambda exonuclease to trim ChIP DNA until a bound protein blocks further enzymatic activity (55). This creates 5' borders on both DNA strands that are juxtaposed with the protein, manifested as opposite-stranded peak pairs on a genome browser, and achieves 20-50 bp resolution of DNA binding (35).

Opposite-stranded peak pairs annotate both canonical CEBPβ homodimer motifs and CEBPβ-sequences bound by the ATF4 heterodimer in hMSCs, demonstrating the resolving power of ChIP-exo (Figure 1A). Globally, CEBP peak pairs show an average distance distribution of 15-30 bp, with a predominant distance of 25 bp for CEBPβ and 27 bp for CEBPα (Figure 1B, Supplementary Figure S1A). Motif analysis reveals exclusive enrichment of an 8-mer-core sequence comprised of a degenerate half site (TKnn) fused to a CEBP half site (GCAA) (Figure 1C, Supplementary Figure S1B). Ordered peak pairs flank this motif at a majority of ChIP-seq peaks (Figure 1D, Supplementary Figure S1C), indicating that CEBPs occupy the genome primarily through direct, sequence-specific interaction. Parsing the CEBP cistrome by individual 8-mer variants of the CEBP core motif reveals that the sequence bound most frequently by CEBPβ and CEBPα is TTGTGCAA (Figure 1E), partly due to its high occurrence in the genome (Supplementary Figure S1D). Nevertheless, this sequence accounts for only about 14% of high-confidence ChIP-exo-annotated binding sites. Together with similar ChIP-seq occupancy strengths observed as a function of CEBP 8-mers (Supplementary Figure S1E), these data indicate that no singular sequence explains the majority of CEBP binding. Interestingly, the CEBPβ-ATF4 heterodimer sequence, TGATGCAA, is the second most prevalent CEBP core motif variant, and additional hybrid motifs composed of non-CEBP bZip half sites (TGWN) joined to the CEBP half site are also present within the top-ranked sequences. The tolerance of

substituting G in lieu of the canonical T at the 2nd position of the hybrid core suggests either an intrinsic relaxation of CEBP's sequence specificity in physiological contexts, broadened motif recognition through heterodimerization, or both. As a whole, the ChIP-exo data demonstrate conservation between human and mouse CEBP family members through interaction with a compound motif anchored by a CEBP half site.

The CEBP motif identified in primary cells and tissue differs strikingly from the optimal sequence observed for homodimers in vitro. Both early studies (56–59) and more recent systematic biochemical approaches (9, 10, 60) report that the CEBP homodimer binds a palindromic motif formed by the fusion of two CEBP half sites (ATTGCGCAAT), yet this sequence only represents only a small fraction ($\leq$ 1%) of the genomic sites occupied by CEBPs in vivo. To exclude the possibility that this discrepancy is caused by assay-dependent effects, we performed ChIP-exo utilizing recombinant CEBPβ homodimer or ATF4-CEBPβ heterodimer and protein-free genomic DNA. A sequence resembling the palindromic CEBP motif is enriched at peak pairs for the CEBPβ homodimer (Figure 1F), consistent with findings from PBMs (10), HT-SELEX (9) and SMiLE-seq (60). Yet, this motif is distinct from the consensus motif for endogenous CEBP. In contrast, in vitro ChIP-exo for the CEBPβ-ATF4 heterodimer yields a motif that is very similar to that reported for ATF4 in hMSCs (Figure 1F) (41). Sequence-specific interaction by the CEBPβ homo- and heterodimer is indicated by the emergence of peak pairs with fixed spacing that flank both motifs (Supplementary Figures S1F and S1G). Thus, in vitro ChIP-exo corroborates the DNA sequence specificity reported by established biochemical approaches, and illustrates a fundamental difference between the DNA-binding specificity of the CEBPβ homodimer versus CEBPβ in cells. Although a few thousand heterodimeric sites of ATF4 with various CEBPs have been mapped in vivo (41, 48, 50), they represent only 2-5% of the CEBPβ cistrome in hMSCs. Thus, the observation that roughly 40% of CEBPβ binding in hMSCs occurs at hybrid sequences comprised of AP-1 or ATF-like half sites fused to a CEBP half site suggests that heterodimerization with other bZip family members may occur on a much broader scale than previously envisioned.

Sequence optimization regulates cell-type-specific binding by CEBPβ.

Despite the strong preference of CEBP homodimers for the palindromic sequence, TTGCGCAA, it accounts for less than 3% of all CEBPβ binding sites. However, correcting for the fact that the CEBP palindromic 8-mer occurs rarely in the human genome, it exhibits the highest fraction of genomic occupancy of any CEBP motif (Supplementary Figure S1D). This suggests a relationship between sequence optimization and probability of genomic occupancy. To test whether motif optimization is

correlated with the likelihood of binding and/or transcriptional regulation, we examined the relationships between CEBPβ-binding sites, CEBP motifs, and gene transcription across multiple human cell lines (hMSCs, Helas3, HepG2, K562, IMR90, A549). Consistent with frequency measurements for shared versus unique binding sites for a TF in different cell types (27, 61), and CEBPβ specifically (62), approximately 20% of CEBP sites map to either a single cell type (cell-type specific) or all cell types (cell-type independent), respectively, while the remaining 60% fall between these extremes (Figure 2A). Functional CEBPβ sites show enrichment for RNA Polymerase II (RNAPII) (62), and using gene body RNAPII occupancy as a surrogate for transcription, we observed higher transcriptional activity at genes near cell-type-independent sites compared to genes near cell-type-specific sites (Figure 2B). Moreover, distinct gene ontologies were observed, with genes near cell-type-independent sites enriched for general processes such as mRNA metabolism and translation, whereas cell-type-specific site-gene pairs associate with specialized pathways such as adipocytokine signaling and lipoprotein metabolism. Intriguingly, the genomic distribution of cell-type-independent sites is biased towards transcription start sites (TSSs), whereas cell-type-specific sites display a gene-distal distribution characteristic of the overall CEBPβ cistrome (Supplementary Figure S2A).

Motif quality has been reported to correlate with shared occupancy across cell types for some nuclear receptors (27, 63), but prior studies of cell-type-dependent binding for CEBPβ focused on the role of collaborating TFs and chromatin (62). Interestingly, de novo motif analyses revealed a depletion of C in the 5th position of the CEBP motif at cell-type-specific, but not cell-type-independent, sites (Figure 2C). This C conforms to the canonical CEBP half site, suggesting that the probability of CEBPβ occupancy is correlated with motif optimization. To further elucidate this relationship, we surveyed the utilization of the top motif variants in both classes of sites. Preservation of C at the 4th position of the core 8-mer (5th position of the 10-mer) was strongly correlated with increased probability of binding in multiple cell types, with 90% co-occupancy of the CEBP palindrome in 4 or more cell types (Figure 2D). Conversely, cell-type-specific sites are enriched for sub-optimized motifs, harboring substitutions in the 4th and 6th positions of the core 8-mer. Consistent with the hypothesis that cell-type-independent binding sites are privileged for both high-affinity motifs and highly accessible chromatin, average CEBPβ-binding strength is positively correlated with co-occupancy status (Supplementary Figure S2B). These data demonstrate that motif optimization within the core 8-mer increases the probability that any given CEBP-binding site will be shared across cell types, and also suggest a relationship between optimized CEBP motifs and increased transcriptional activity. Though rare genome-wide, optimized CEBP motifs may serve as elite sequences that enable CEBPβ

to overcome the repressive effects of chromatin to coordinate a limited program of constitutive gene expression.

Bases directly abutting the core CEBP motif impact occupancy.

We would predict occupancy at that nearly every genomic instance of the palindromic CEBP core 8-mer if this motif constitutively recruits CEBPs, and yet most are unoccupied in hMSCs (Supplementary Figure S1D). To investigate potential distinguishing features for bound versus unbound CEBP palindromes, we profiled CEBPβ occupancy at all 3121 palindromic 8-mers (excluding unplaced contigs) in the human genome using our in vitro ChIP-exo data derived from recombinant CEBPβ homodimer bound to histone-free genomic DNA. While the vast majority (84%) of CEBP palindromes showed binding, a subset failed to recruit CEBPβ. Sequence alignments of these unbound sites revealed a pronounced difference in the base composition at positions immediately flanking the core motif (Figure 3A). In contrast, neighboring positions beyond these flanks have random sequence variation (Supplementary Figure S3A). Strikingly, the occurrence of T at the 5' flank or A at the 3' flank is negatively correlated with CEBPβ occupancy. Moreover, while C is also disfavored at the 5' flank, its ability to cripple the functionality of the palindromic 8-mer is most pronounced when paired with A at the 3' flank. Likewise, T-G dinucleotide flanks appear highly deleterious to CEBPβ binding (note that **C**TTGCGCAA**A** and **T**TTGCGCAA**G** are reverse-complementary 10-mer sequences). Importantly, sequence preferences at these flanking positions are implicit within the motif logos from our ChIP-exo experiments (Figure 1C, Supplementary Figure S1B). Careful inspection of these logos reveals the exclusion of T at the 5' flank and A at the 3' flank. However, compared to the core 8-mer, the relatively low information content encoded in these flanks de-emphasizes their importance, and fails to underscore how specific base pairings at the 5' and 3' flanks (T-A, C-A, T-G) can override the ability of an optimized 8-mer core sequence to recruit CEBPβ.

The apparent importance of the flanking bases indicates that the minimal sequence determinant for CEBP binding to the genome is encoded by a 10-mer sequence, corroborating earlier biochemical and structural studies that identified a 5-mer half-site (56–59). To further investigate the relationship between sequence variation at core-motif flanking positions and motif binding strength, we revisited our analysis of cell-type-dependent occupancy across all 3121 CEBP palindromic motifs, parsing by 10-mer variants (Figure 3B). Remarkably, 88% of all genomic instances of the top-6 performing 10-mers are co-occupied by CEBPβ in 5 or more cell types (red bars), and $\geq$ 95% of these sequences are bound in at least one cell type (sum of red and green bars). Conversely, 10-mers comprised of a palindromic 8-mer nested within unfavorable flanks are rarely occupied. These data indicate that flanking nucleotides play a critical role in CEBP motif recognition, and reveal that when

annotated as a 10-mer sequence, the CEBP palindrome functions to recruit CEBPβ independently of cell type or genomic neighborhood. In contrast, 8-mer palindromic sequences with neutral flanking dinucleotide pairs (combinations of one favorable and one unfavorable flanking nucleotide at the 5' and 3' positions) are enriched for cell-type-specific CEBPβ binding, suggesting that while these sequences have the potential to recruit CEBPs, they are more sensitive to cell-type-specific differences in chromatin structure. Reduced affinity for these sequences could explain this behavior, which is indicated by weaker ChIP-exo signal at sites with neutral flanking bases relative to those with favorable flanks (Supplementary Figure S3B).

These findings enhance our understanding of the optimal CEBP palindromic motif, yet an important question is whether core-motif flanking nucleotides are also important in the context of a more degenerate CEBP motif that is representative of the broader CEBP cistrome. However, the very nature of these degenerate sequences as sub-optimized binding sites presents a challenge in interpreting whether unbound genomic sequences fail to recruit CEBPβ due to chromatin effects, unfavorable flanks, or both. Importantly, the effects of chromatin can be excluded by testing how CEBP 8-mers are populated within known CEBPβ ChIP-seq peaks, which by definition reside within accessible chromatin. While ChIP-exo has the resolution to resolve multiple binding events within single ChIP-seq peaks for CEBPβ, our analysis pipeline picks the strongest peak pair with a characteristic spacing per ChIP-seq peak, and thus does not explicitly identify every instance of a bound motif. As a result, we re-interrogated our ChIP-exo data to discover additional, weaker CEBPβ-binding events within CEBPβ ChIP-seq peaks, and compared their motifs to co-localized 8-mer sequences that failed to exhibit a characteristic ChIP-exo peak pair. A total of 3686 CEBP motif candidates were mapped in the vicinity of a CEBP-bound motif, and classified into bound versus unbound sequences based on their ChIP-exo signature (Figure 3C, Supplementary Figure S3C). Within the set of candidate secondary motifs, 61% lacked appropriately spaced opposite-stranded peak pairs, indicating little or no occupancy. We then examined the frequency of dinucleotide flanks abutting the bound primary and secondary CEBP 8-mers as well as the unbound secondary 8-mers (Figure 3D). Consistent with the behavior of the flanks surrounding the palindromic 8-mer, the flanks present at primary (strong) ChIP-exo bound motifs are enriched for A-T, A-C, A-G, G-T, G-C, and C-T dinucleotide flanks, whereas the unfavorable T-A, C-A, and T-G flanks are essentially absent. This frequency distribution differs from the background rate for CEBP 8-mers across the human genome (Supplementary Figure S3D), such that favorable flanks occur less frequently than expected by chance. Secondary bound CEBP motifs show a preference for favorable flanks that is similar to primary sites, albeit at lower frequencies. In contrast, unbound 8-mers are enriched for unfavorable and neutral flanks. These data demonstrate that flanking bases play a role in discriminating which

candidate CEBP 8-mers are bound within regions of open chromatin. Combined with the earlier analyses, they reveal that a 10-mer sequence enables discrimination of real versus decoy CEBP motifs in the genome.

Bases directly abutting core bZip motifs affect DNA shape.

Crystallography studies of CEBPα and several other bZip proteins complexed with DNA have modeled contacts to the core bases (64–68), whereas interaction with the DNA backbone may take place at the flanks (67). These observations challenge the notion that the flanks interact with CEBPs via base readout, and led us to consider whether the flanking bases could impact genomic occupancy through alteration of DNA shape. To address DNA shape in an unbiased manner, we examined the CEBP motifs residing in accessible chromatin of hMSCs yet differing in their ability to recruit CEBPβ (see Supplementary Figure S3C). Comparison of intrinsic DNA shape features revealed significant differences between bound versus unbound motifs (Figure 4A). The shape changes occurred at or near the flanking bases, consistent with an intimate association between sequence and shape, and suggesting that CEBPβ prefers to bind motifs surrounded by more positive roll, less helical twist, wider minor groove width, and less negative propeller twist.

More broadly, we sought to determine whether the DNA shape features associated with CEBPβ occupancy apply to bZip TFs in general. To examine a potential relationship between DNA shape and bZip TF occupancy, we identified weakly and strongly bound motifs from published PBMs for CREB, NFIL3 and JUND (10). Comparison of differentially bound motifs for each TF revealed significant alterations for various shape features (Figure 4B). Interestingly, the shape changes resembled those at the CEBP motif, suggesting that CREB, NFIL3 and JUND prefer to bind motifs surrounded by more positive roll, less helical twist, wider minor groove width, and less negative propeller twist. Given that DNA sequence determines shape, we examined the motifs enriched in previously published ChIP-seq datasets for CREB, NFIL3 and AP-1 (69). Similar to the CEBP motif, we found a clear exclusion of T and A in the 5' and 3' flanking positions, respectively (Figure 4C). Thus, the motifs for multiple bZip proteins share similar negative selection for T-A pairs at the flanks, indicating that an expanded motif definition including flanking bases is broadly important across mammalian bZips.

Bases directly abutting core bZip motifs affect transcriptional activity.

To test whether a change in the flanks is sufficient to convert a functional CEBP-binding site into a crippled site, we performed ChIP for CEBPs in liver tissue isolated from C57BL/6J (B6) and 129S1/SvImJ (129) mice, and examined occupancy at sites carrying SNPs that introduce unfavorable

flanks into CEBP-binding sites when comparing B6 to 129. While only two sites exist meeting the criteria for type of nucleotide substitution of interest and the absence of a neighboring CEBP-binding site, both show diminished occupancy of CEBPα and CEBPβ in 129 mice relative to B6 (Figure 5A). Consistent with these results, B6x129 F1 mice showed significantly skewed binding of CEBPs to the B6 alleles (Figure 5B). Because the B6 and 129 alleles reside in the same nuclei of F1 mice and are thus exposed to the same trans-acting factors, these data demonstrate that cis effects determine differential binding of CEBPs at these loci. Specifically, the introduction of unfavorable flanking nucleotides may be sufficient to impair CEBP binding independently of the core 8-mer.

Our data establish genome-wide trends between bZip TF occupancy and core-motif flanking bases. To test this relationship and extend its relevance to transcriptional activity, we examined the activity of synthetic luciferase reporters containing multimerized core motifs for distinct bZip TFs flanked by either favorable or unfavorable bases. Replacement of favorable with unfavorable flanks decreased luciferase activity across all bZip reporter constructs tested, with reductions ranging from 6-fold for the CEBP motif to $\geq$ 10-fold for the remaining motifs (Figure 5C). Thus, the data reveal a shared requirement across the bZIP family for favorable motif flanks that confer binding and transcriptional competency to their cognate core recognition sequences.

DISCUSSION

We have used ChIP-exo to perform a genome-wide cataloging of motif utilization within the CEBP cistromes of primary human cells and mouse liver tissue. We demonstrate species-conserved sequence requirements for the recruitment of CEBP proteins to the native genome that are fundamentally different from the sequence preferences of CEBP homodimers in vitro. Pioneering in vitro studies (56–59) and more recent systematic biochemical approaches (9, 10, 60) report that the CEBP homodimer prefers to bind a palindromic motif formed by the fusion of two CEBP half sites (ATTGCGCAAT). Our data reveal that this motif captures less than 1% of the binding sites occurring in cells. On the native genome, CEBPs bind a diversity of related sequences resulting from the fusion of degenerate and canonical CEBP half sites that yields a 10-mer-consensus of the form VTKNNGCAAB. A large majority of binding sites, 70-90% depending on threshold cutoffs, contains bound CEBP motifs. This suggests that CEBPs primarily occupy the genome through direct, sequence-specific interaction, whereas binding to motifs with atypical spacing between half sites (25) and to other DNA-bound TFs through tethering contribute minimally to the genomic recruitment of CEBPs.

It is noteworthy that roughly 40% of CEBP-binding sites contain a G at the third position of the 10-mer, creating a preferred half-site motif for the ATF, AP-1, and CREB families of bZip TFs. Evidence has been found for the heterodimerization and altered sequence specificity of CEBP-ATF complexes compared to their homodimer counterparts (25, 41, 48–50), yet none of these studies addresses the extent to which heterodimerization drives genomic occupancy in vivo. For example, the identification of approximately 1600 ATF4-CEBPβ heterodimer sites in hMSCs represents only 2-5% of the total genomic CEBPβ sites (41). Our data indicate heterodimer binding greatly exceeding that with ATF4 such that CEBP occupancy of hybrid motifs represents a large fraction of the cistrome in vivo. This widespread occupancy is unprecedented and highly impactful for understanding the function of bZip TFs, and may help to explain why CEBPs populate large cistromes comprised of tens of thousands of binding sites in mammalian tissues and primary cells (26, 52–54).

Genome-wide cataloging of motifs within CEBP cistromes preserves genomic information that affords comparisons between bound and unbound sites. Direct examination of optimal TTGCGCAA 8-mers unoccupied both in vitro and in vivo identified T-A flanking bases that are disfavored for CEBPβ binding and transcriptional activation. Reminiscent of an early in vitro study of CEBPs (58), our finding impacts the understanding of DNA-binding specificity for bZip proteins in general, as the same flanking bases also cripple the transcriptional activity of the core motifs for ATF4, AP-1, CREB and PAR TFs. Negative selection against unfavorable flanks suggests that these positions contribute to motif recognition by modulating DNA-binding affinity. Structural studies of bZip proteins bound to DNA show interaction with the DNA backbone (67) but not the bases at the flanking positions (64–68), suggesting that the flanks impact bZip affinity through DNA shape affects. Contrasting the shape of bound and unbound sites for multiple TFs has led to the notion that motif flanks can regulate genomic occupancy through alteration of DNA shape (15, 18, 33, 70). Consistent with this, DNA shape features differ in similar ways between high- and low-affinity binding sites for CEBPβ, CREB, NFIL3 and JUND. The shared sequence bias at the flanks across distinct bZIP motifs, coupled with the fact that mono and di-nucleotide sequences account for more than 90% of the variance in commonly interrogated shape features (71), explains how similar DNA shape features can persist at the motif periphery even while divergent sequences dominate at the core.

Elite CEBP 10-mer motifs comprised of RTTRCGCAAY recruit CEBPβ in a cell-type-independent manner and are associated with higher levels of gene expression relative to cell-type-specific sites. Moreover, for optimized CEBP 10-mers containing a palindromic core, approximately 80% of genomic instances are bound by CEBPβ. Thus, highly optimized CEBP motifs are sufficient to recruit CEBPβ regardless of the genomic context, implying that CEBPs can overcome chromatin-mediated repression. Neutral flanks pair a favorable and unfavorable base at the first and last position

of the 10mer, and they are correlated with a progressive loss of palindromic occupancy across cell types and weaker binding strengths in vitro. Importantly, these relationships between flanking sequence and motif occupancy can be generalized to the more degenerate CEBP motif, suggesting that CEBPs can populate lower-affinity sequences that are readily accessible in open chromatin.

Unlike high throughput assays that select for optimal TF-binding sequences, analysis of bound TF sequences suggests that optimized motifs play limited biological roles in genomic recruitment of TFs. A relationship between sub-optimized motifs and cell-type-dependent binding has been documented for ERα (27, 63), yet whether deviations from consensus motifs are biological drivers of differential TF occupancy is unknown, especially given the dominant effect of chromatin structure on the accessibility of DNA motifs. Intriguingly, motif sub-optimization through somatic mutation of the central CG dinucleotide of the CEBP motif has been reported in human cancers (72), suggesting an evolutionary pressure selecting against optimized CEBP motifs that mirrors the overall sparsity of these motifs in the human genome. Placed in the context of our work, perhaps the rarity of fully optimized TF motifs in eukaryotic genomes serves to limit constitutive genomic recruitment, suppressing the potential for TFs to trigger unregulated gene expression with regard to tissue or cell type. Conversely, the majority of CEBP-bound motifs are sub-optimized and occupied in a cell-type-specific manner. This observation fits with an emerging paradigm whereby tissue-specific gene expression is mediated by composite enhancers (24, 73–75) that recruit multiple TFs through sub-optimized motifs (73, 74). Rather than a fortuitous event, sub-optimization may be biologically favorable to impart a dependency of TF occupancy on chromatin environment, and render enhancers readily amenable to evolutionary turnover (76–78).

ACCESSION NUMBERS

High throughput sequencing data have been deposited at GEO under accession number GSE111515.

ACKNOWLEDGMENTS

We are grateful to Raymond Soccio for guidance on the experimental strategy interrogating strain-specific TF binding, and for providing liver tissue from 129S1/SvImJ and 129S1/SvImJxC57BL/6J F1 mice. We also thank Chris Krapp and Marisa Bartolomei for help with pyrosequencing and generously providing access to their sequencer. We are indebted to members of the Lazar laboratory for insightful discussions, and also thank the Functional Genomics Core of the Penn Diabetes Center (DK19525) for deep sequencing.

FUNDING

This work was supported by National Institutes of Health grants R01 DK106027 (to K-JW) and R01 DK098542 (to DJS). Funding for open access charge: National Institutes of Health.

CONFLICT OF INTEREST

We have no conflicts of interest to report.

REFERENCES

1. Deplancke,B., Alpern,D. and Gardeux,V. (2016) The Genetics of Transcription Factor DNA Binding Variation. *Cell*, **166**, 538–554.

2. Vockley,C.M., Barrera,A. and Reddy,T.E. (2017) Decoding the role of regulatory element polymorphisms in complex disease. *Curr. Opin. Genet. Dev.*, **43**, 38–45.

3. Shlyueva,D., Stampfel,G. and Stark,A. (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.*, **15**, 272–286.

4. Spitz,F. and Furlong,E.E.M. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, **13**, 613–626.

5. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The Human Transcription Factors. *Cell*, **172**, 650–665.

6. Jolma,A. and Taipale,J. (2011) Methods for Analysis of Transcription Factor DNA-Binding Specificity In Vitro. *Subcell. Biochem.*, **52**, 155–173.

7. Odom,D.T. (2011) Identification of Transcription Factor-DNA Interactions In Vivo. *Subcell. Biochem.*, **52**, 175–191.

8. Stormo,G.D. and Zhao,Y. (2010) Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.*, **11**, 751–760.

9. 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.

10. 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.

11. Orenstein,Y. and Shamir,R. (2017) Modeling protein-DNA binding via high-throughput in vitro technologies. *Brief. Funct. Genomics*, **16**, 171–180.

12. Workman,J.L. and Kingston,R.E. (1998) Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu. Rev. Biochem.*, **67**, 545–579.

13. Arvey,A., Agius,P., Noble,W.S. and Leslie,C. (2012) Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res.*, **22**, 1723–1734.

14. Kaplan,T., Li,X.-Y., Sabo,P.J., Thomas,S., Stamatoyannopoulos,J.A., Biggin,M.D. and Eisen,M.B. (2011) Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. *PLoS Genet.*, **7**, e1001290.

15. Zhou,T., Yang,L., Lu,Y., Dror,I., Dantas Machado,A.C., Ghane,T., Di Felice,R. and Rohs,R. (2013) DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res.*, **41**, W56-62.

16. Mathelier,A., Xin,B., Chiu,T.-P., Yang,L., Rohs,R. and Wasserman,W.W. (2016) DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Syst.*, **3**, 278-286.e4.

17. Dror,I., Golan,T., Levy,C., Rohs,R. and Mandel-Gutfreund,Y. (2015) A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res.*, **25**, 1268–1280.

18. Gordân,R., Shen,N., Dror,I., Zhou,T., Horton,J., Rohs,R. and Bulyk,M.L. (2013) Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.*, **3**, 1093–1104.

19. Levo,M., Zalckvar,E., Sharon,E., Dantas Machado,A.C., Kalma,Y., Lotam-Pompan,M., Weinberger,A., Yakhini,Z., Rohs,R. and Segal,E. (2015) Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res.*, **25**, 1018–1029.

20. 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.

21. 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**.

22. Maurano,M.T., Wang,H., John,S., Shafer,A., Canfield,T., Lee,K. and Stamatoyannopoulos,J.A. (2015) Role of DNA Methylation in Modulating Transcription Factor Occupancy. *Cell Rep.*, **12**, 1184–1195.

23. 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.

24. Wang,J., Zhuang,J., Iyer,S., Lin,X., Whitfield,T.W., Greven,M.C., Pierce,B.G., Dong,X., Kundaje,A., Cheng,Y., *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798–1812.

25. Rodríguez-Martínez,J.A., Reinke,A.W., Bhimsaria,D., Keating,A.E. and Ansari,A.Z. (2017) Combinatorial bZIP dimers display complex DNA-binding specificity landscapes. *eLife*, **6**.

26. Wang,J., Zhuang,J., Iyer,S., Lin,X., Whitfield,T.W., Greven,M.C., Pierce,B.G., Dong,X., Kundaje,A., Cheng,Y., *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798–1812.

27. Gertz,J., Savic,D., Varley,K.E., Partridge,E.C., Safi,A., Jain,P., Cooper,G.M., Reddy,T.E., Crawford,G.E. and Myers,R.M. (2013) Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol. Cell*, **52**, 25–36.

28. Glass,C.K. and Saijo,K. (2010) Nuclear receptor transrepression pathways that regulate inflammation in macrophages and T cells. *Nat. Rev. Immunol.*, **10**, 365–376.

29. Langlais,D., Couture,C., Balsalobre,A. and Drouin,J. (2012) The Stat3/GR Interaction Code: Predictive Value of Direct/Indirect DNA Recruitment for Transcription Outcome. *Mol. Cell*, **47**, 38–49.

30. Zhang,Y., Fang,B., Emmett,M.J., Damle,M., Sun,Z., Feng,D., Armour,S.M., Remsberg,J.R., Jager,J., Soccio,R.E., *et al.* (2015) Discrete functions of nuclear receptor Rev-erbα couple metabolism to the clock. *Science*, **348**, 1488–1492.

31. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.*, **79**, 233–269.

32. Rohs,R., West,S.M., Sosinsky,A., Liu,P., Mann,R.S. and Honig,B. (2009) The role of DNA shape in protein-DNA recognition. *Nature*, **461**, 1248–1253.

33. Rossi,M.J., Lai,W.K.M. and Pugh,B.F. (2018) Genome-wide determinants of sequence-specific DNA binding of general regulatory factors. *Genome Res.*, **28**, 497–508.

34. Love,M.I., Huska,M.R., Jurk,M., Schöpflin,R., Starick,S.R., Schwahn,K., Cooper,S.B., Yamamoto,K.R., Thomas-Chollier,M., Vingron,M., *et al.* (2017) Role of the chromatin landscape and sequence in determining cell type-specific genomic glucocorticoid receptor binding and gene regulation. *Nucleic Acids Res.*, **45**, 1805–1819.

35. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408–1419.

36. Iwata,A., Durai,V., Tussiwand,R., Briseño,C.G., Wu,X., Grajales-Reyes,G.E., Egawa,T., Murphy,T.L. and Murphy,K.M. (2017) Quality of TCR signaling determined by differential affinities of enhancers for the composite BATF-IRF4 transcription factor complex. *Nat. Immunol.*, **18**, 563–572.

37. Luna-Zurita,L., Stirnimann,C.U., Glatt,S., Kaynak,B.L., Thomas,S., Baudin,F., Samee,M.A.H., He,D., Small,E.M., Mileikovsky,M., *et al.* (2016) Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and Coordinates Cardiogenesis. *Cell*, **164**, 999–1014.

38. Lim,H.-W., Uhlenhaut,N.H., Rauch,A., Weiner,J., Hübner,S., Hübner,N., Won,K.-J., Lazar,M.A., Tuckermann,J. and Steger,D.J. (2015) Genomic redistribution of GR monomers and dimers

mediates transcriptional response to exogenous glucocorticoid in vivo. *Genome Res.*, **25**, 836–844.

39. Starick,S.R., Ibn-Salem,J., Jurk,M., Hernandez,C., Love,M.I., Chung,H.-R., Vingron,M., Thomas-Chollier,M. and Meijsing,S.H. (2015) ChIP-exo signal associated with DNA-binding motifs provides insight into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. *Genome Res.*, **25**, 825–835.

40. Bartlett,A., O'Malley,R.C., Huang,S.-S.C., Galli,M., Nery,J.R., Gallavotti,A. and Ecker,J.R. (2017) Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat. Protoc.*, **12**, 1659–1672.

41. Cohen,D.M., Won,K.-J., Nguyen,N., Lazar,M.A., Chen,C.S. and Steger,D.J. (2015) ATF4 licenses C/EBPβ activity in human mesenchymal stem cells primed for adipogenesis. *eLife*, **4**, e06821.

42. Gossett,A.J. and Lieb,J.D. (2008) DNA Immunoprecipitation (DIP) for the Determination of DNA-Binding Specificity. *CSH Protoc.*, **2008**, pdb.prot4972.

43. Guertin,M.J., Martins,A.L., Siepel,A. and Lis,J.T. (2012) Accurate prediction of inducible transcription factor binding intensities in vivo. *PLoS Genet.*, **8**, e1002610.

44. Costa,R.H., Kalinichenko,V.V., Holterman,A.-X.L. and Wang,X. (2003) Transcription factors in liver development, differentiation, and regeneration. *Hepatol. Baltim. Md*, **38**, 1331–1347.

45. Friedman,A.D. (2002) Transcriptional regulation of granulocyte and monocyte development. *Oncogene*, **21**, 3377–3390.

46. Rosen,E.D., Hsu,C.-H., Wang,X., Sakai,S., Freeman,M.W., Gonzalez,F.J. and Spiegelman,B.M. (2002) C/EBPalpha induces adipogenesis through PPARgamma: a unified pathway. *Genes Dev.*, **16**, 22–26.

47. Tsukada,J., Yoshida,Y., Kominato,Y. and Auron,P.E. (2011) The CCAAT/enhancer (C/EBP) family of basic-leucine zipper (bZIP) transcription factors is a multifaceted highly-regulated system for gene regulation. *Cytokine*, **54**, 6–19.

48. Han,J., Back,S.H., Hur,J., Lin,Y.-H., Gildersleeve,R., Shan,J., Yuan,C.L., Krokowski,D., Wang,S., Hatzoglou,M., *et al.* (2013) ER-stress-induced transcriptional regulation increases protein synthesis leading to cell death. *Nat. Cell Biol.*, **15**, 481–490.

49. Reinke,A.W., Baek,J., Ashenberg,O. and Keating,A.E. (2013) Networks of bZIP protein-protein interactions diversified over a billion years of evolution. *Science*, **340**, 730–734.

50. Huggins,C.J., Mayekar,M.K., Martin,N., Saylor,K.L., Gonit,M., Jailwala,P., Kasoji,M., Haines,D.C., Quiñones,O.A. and Johnson,P.F. (2015) C/EBPγ Is a Critical Regulator of Cellular Stress Response Networks through Heterodimerization with ATF4. *Mol. Cell. Biol.*, **36**, 693–713.

51. Kilberg,M.S., Shan,J. and Su,N. (2009) ATF4-dependent transcription mediates signaling of amino acid limitation. *Trends Endocrinol. Metab. TEM*, **20**, 436–443.

52. Everett,L.J., Le Lay,J., Lukovac,S., Bernstein,D., Steger,D.J., Lazar,M.A. and Kaestner,K.H. (2013) Integrative genomic analysis of CREB defines a critical role for transcription factor networks in mediating the fed/fasted switch in liver. *BMC Genomics*, **14**, 337.

53. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.

54. Schmidt,D., Wilson,M.D., Ballester,B., Schwalie,P.C., Brown,G.D., Marshall,A., Kutter,C., Watt,S., Martinez-Jimenez,C.P., Mackay,S., *et al.* (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, **328**, 1036–1040.

55. Mymryk,J.S. and Archer,T.K. (1994) Detection of transcription factor binding in vivo using lambda exonuclease. *Nucleic Acids Res.*, **22**, 4344–4345.

56. Agre,P., Johnson,P.F. and McKnight,S.L. (1989) Cognate DNA binding specificity retained after leucine zipper exchange between GCN4 and C/EBP. *Science*, **246**, 922–926.

57. Johnson,P.F. (1993) Identification of C/EBP basic region residues involved in DNA sequence recognition and half-site spacing preference. *Mol. Cell. Biol.*, **13**, 6919–6930.

58. Osada,S., Yamamoto,H., Nishihara,T. and Imagawa,M. (1996) DNA binding specificity of the CCAAT/enhancer-binding protein transcription factor family. *J. Biol. Chem.*, **271**, 3891–3896.

59. Vinson,C.R., Sigler,P.B. and McKnight,S.L. (1989) Scissors-grip model for DNA recognition by a family of leucine zipper proteins. *Science*, **246**, 911–916.

60. Isakova,A., Groux,R., Imbeault,M., Rainer,P., Alpern,D., Dainese,R., Ambrosini,G., Trono,D., Bucher,P. and Deplancke,B. (2017) SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nat. Methods*, **14**, 316–322.

61. John,S., Sabo,P.J., Thurman,R.E., Sung,M.-H., Biddie,S.C., Johnson,T.A., Hager,G.L. and Stamatoyannopoulos,J.A. (2011) Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.*, **43**, 264–268.

62. Savic,D., Roberts,B.S., Carleton,J.B., Partridge,E.C., White,M.A., Cohen,B.A., Cooper,G.M., Gertz,J. and Myers,R.M. (2015) Promoter-distal RNA polymerase II binding discriminates active from inactive CCAAT/ enhancer-binding protein beta binding sites. *Genome Res.*, **25**, 1791–1800.

63. Joseph,R., Orlov,Y.L., Huss,M., Sun,W., Kong,S.L., Ukil,L., Pan,Y.F., Li,G., Lim,M., Thomsen,J.S., *et al.* (2010) Integrative model of genomic factors for determining binding site selection by estrogen receptor-α. *Mol. Syst. Biol.*, **6**, 456.

64. Ellenberger,T.E., Brandl,C.J., Struhl,K. and Harrison,S.C. (1992) The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted alpha helices: crystal structure of the protein-DNA complex. *Cell*, **71**, 1223–1237.

65. Fujii,Y., Shimizu,T., Toda,T., Yanagida,M. and Hakoshima,T. (2000) Structural basis for the diversity of DNA recognition by bZIP transcription factors. *Nat. Struct. Biol.*, **7**, 889–893.

66. Glover,J.N. and Harrison,S.C. (1995) Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature*, **373**, 257–261.

67. Miller,M., Shuman,J.D., Sebastian,T., Dauter,Z. and Johnson,P.F. (2003) Structural basis for DNA recognition by the basic region leucine zipper transcription factor CCAAT/enhancer-binding protein alpha. *J. Biol. Chem.*, **278**, 15178–15184.

68. Schumacher,M.A., Goodman,R.H. and Brennan,R.G. (2000) The structure of a CREB bZIP.somatostatin CRE complex reveals the basis for selective dimerization and divalent cation-enhanced DNA binding. *J. Biol. Chem.*, **275**, 35242–35247.

69. Mei,S., Qin,Q., Wu,Q., Sun,H., Zheng,R., Zang,C., Zhu,M., Wu,J., Shi,X., Taing,L., *et al.* (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.*, **45**, D658–D662.

70. Joshi,R., Passner,J.M., Rohs,R., Jain,R., Sosinsky,A., Crickmore,M.A., Jacob,V., Aggarwal,A.K., Honig,B. and Mann,R.S. (2007) Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell*, **131**, 530–543.

71. Rube,H.T., Rastogi,C., Kribelbauer,J.F. and Bussemaker,H.J. (2018) A unified approach for quantifying and interpreting DNA shape readout by transcription factors. *Mol. Syst. Biol.*, **14**, e7902.

72. Melton,C., Reuter,J.A., Spacek,D.V. and Snyder,M. (2015) Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat. Genet.*, **47**, 710–716.

73. Ballester,B., Medina-Rivera,A., Schmidt,D., Gonzàlez-Porta,M., Carlucci,M., Chen,X., Chessman,K., Faure,A.J., Funnell,A.P.W., Goncalves,A., *et al.* (2014) Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. *eLife*, **3**, e02626.

74. Farley,E.K., Olson,K.M. and Levine,M.S. (2015) Regulatory Principles Governing Tissue Specificity of Developmental Enhancers. *Cold Spring Harb. Symp. Quant. Biol.*, **80**, 27–32.

75. Siersbæk,R., Rabiee,A., Nielsen,R., Sidoli,S., Traynor,S., Loft,A., La Cour Poulsen,L., Rogowska-Wrzesinska,A., Jensen,O.N. and Mandrup,S. (2014) Transcription factor cooperativity in early adipogenic hotspots and super-enhancers. *Cell Rep.*, **7**, 1443–1455.

76. Moses,A.M., Pollard,D.A., Nix,D.A., Iyer,V.N., Li,X.-Y., Biggin,M.D. and Eisen,M.B. (2006) Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput. Biol.*, **2**, e130.

77. Prescott,S.L., Srinivasan,R., Marchetto,M.C., Grishina,I., Narvaiza,I., Selleri,L., Gage,F.H., Swigut,T. and Wysocka,J. (2015) Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. *Cell*, **163**, 68–83.

78. Stergachis,A.B., Neph,S., Sandstrom,R., Haugen,E., Reynolds,A.P., Zhang,M., Byron,R., Canfield,T., Stelhing-Sun,S., Lee,K., *et al.* (2014) Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature*, **515**, 365–370.

FIGURE LEGENDS

Figure 1. CEBP proteins occupy multiple sequence motifs on the native genome. (**A**) Comparison of ChIP-exo and ChIP-seq results for CEBPβ in hMSCs. Left, an opposite-stranded peak pair from ChIP-exo resides near the center of the ChIP-seq peak for either a homodimer-binding site (top) or a heterodimer site with ATF4 (bottom). Right, closer inspection reveals canonical DNA motifs for CEBPβ (green) or CEBPβ-ATF4 (green-orange) between the ChIP-exo peak pairs. Red and blue indicate the 5' ends of the forward- and reverse-stranded sequence tags, respectively. (**B**) Distance distributions for the spacing between opposite-stranded peak pairs. Predominant distances are indicated. (**C**) MEME de novo motif analyses of the 1000-top-ranked ChIP-exo peak pairs spaced 15-30 bp apart. (**D**) Average profiles (top) and density heat maps (bottom) of the ChIP-exo sequence tags at CEBP-binding sites in hMSCs or liver. (**E**) Top-ranked core motifs at CEBPβ peak pairs in hMSCs compared with CEBPs in liver. The CEBP half site, GCAA, is uncolored; degenerate half site is green (CEBP related) or orange (bZip related). (**F**) MEME de novo motif analyses of the 1000-top-ranked peak pairs spaced 10-30 bp apart are shown for the CEBPβ homodimer and CEBPβ-ATF4 heterodimer. Motif analysis from ATF4 ChIP-seq in hMSCs is shown for comparison.

Figure 2. Selective versus widespread occupancy across cell types for distinct CEBPβ motifs. (**A**) Pie chart comparing CEBPβ occupancy across 6 human cell types (hMSC + 6 ENCODE cell lines). sites unique to one cell type (cell-type-specific) or shared between 2-3, 4-5, and all 6 cell types. (**B**) Box plot interrogating RNAPII occupancy (ChIP-seq reads per million, RPM) at expressed genes within 100 kb of cell-type-independent or cell-type-specific binding sites for CEBPβ. Wilcoxon rank sum test used to compare adjacent classes. Highly ranked gene ontology (GO) terms are shown. (**C**) De novo motif analyses showing de-enriched bases that differ between cell-type-independent and cell-type-specific sites. (**D**) Individual 8-mers enriched at cell-type-independent or cell-type-specific sites for CEBPβ were examined to display occupancy across the 6 human cell types. Base positions of interest within the first and second half sites are highlighted.

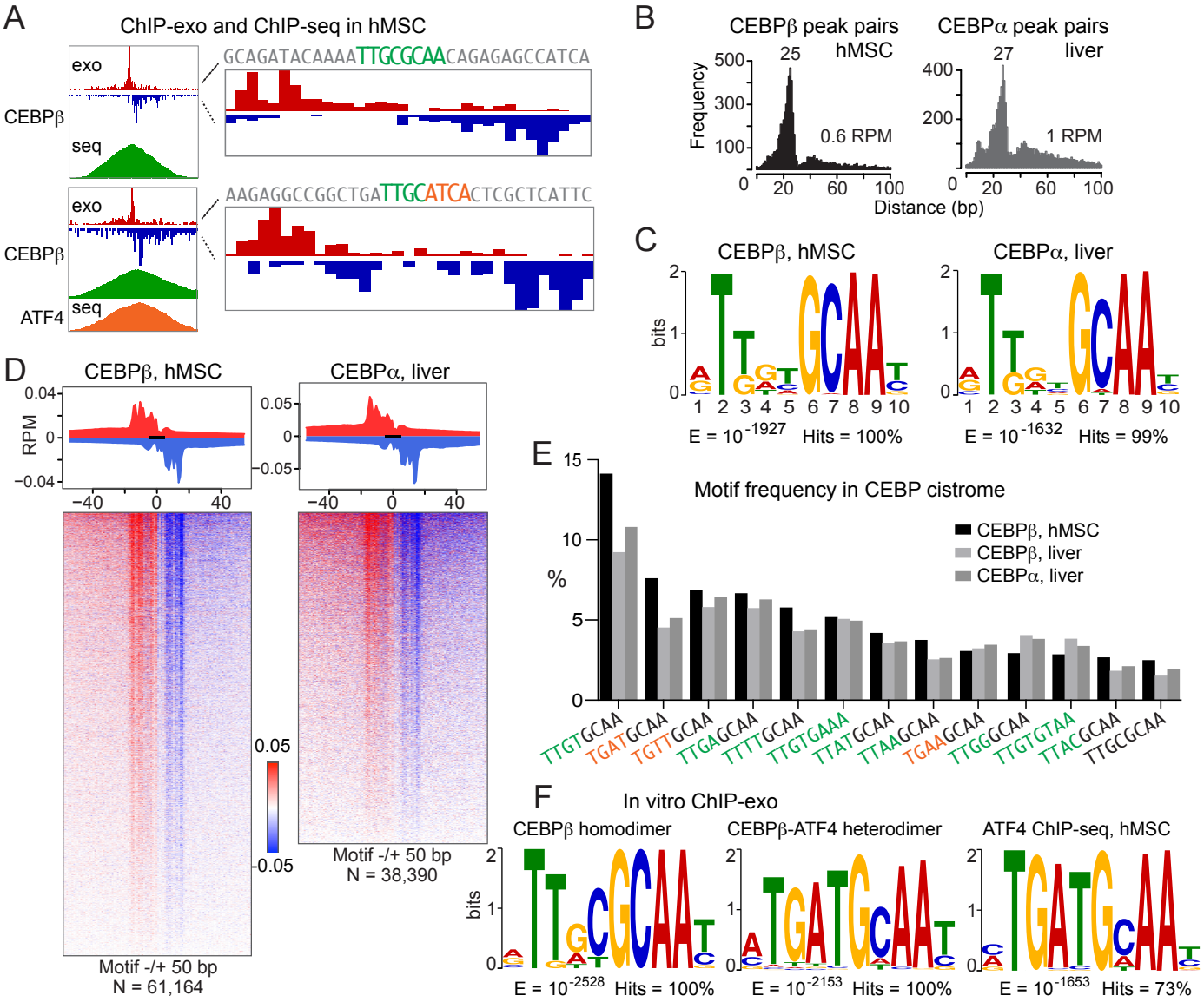
Figure 3. Bases directly flanking the core CEBP 8-mer affect occupancy. (**A**) Density heat map of the in vitro ChIP-exo reads for the CEBPβ homodimer at all canonical palindromic 8-mers with mappable sequence. Binding strength is ordered from top to bottom. Color charts show the base identity at the first position next to the 8-mer on the 5' and 3' ends. Grey boxes indicate sites without

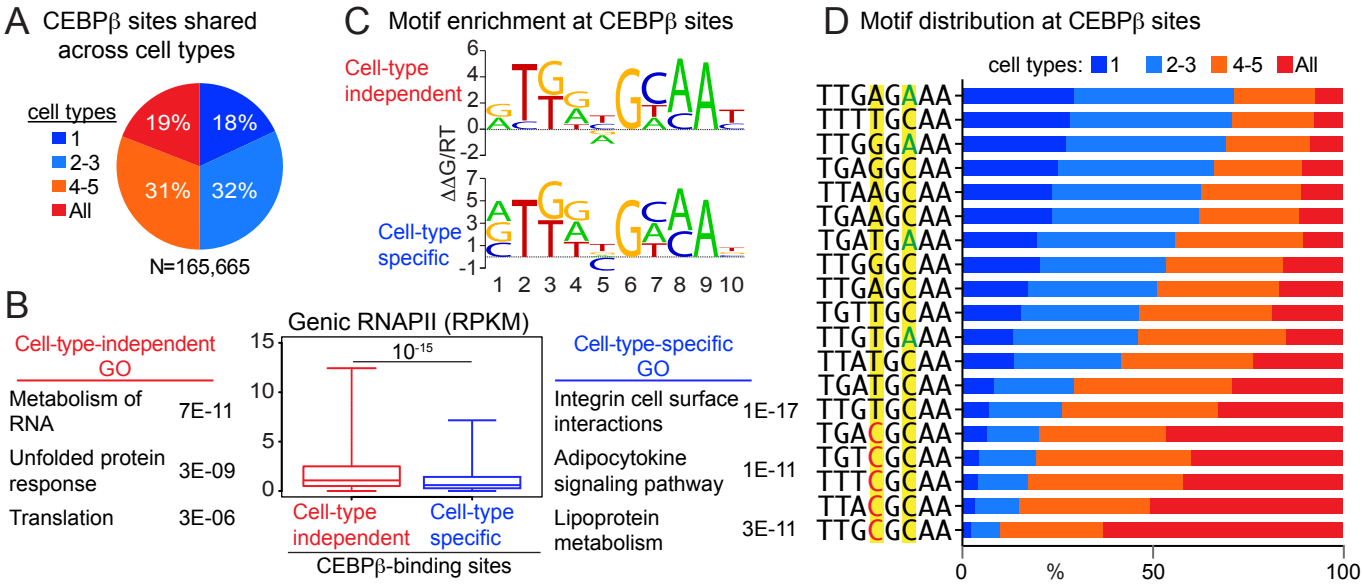
detectable ChIP-exo reads. (**B**) Histogram interrogating relationship between flanking bases and CEBPβ occupancy at all canonical palindromic 8-mers across 7 cell types. Equivalent flanking pairs (5'-3') are grouped together. ENCODE ChIP-seq peak calls are plotted. Favorable flanks associate with occupancy in most cell types at most locations. Unfavorable flanks are not bound in any cell type at most locations. (**C**) CEBPβ ChIP-exo reads at a ChIP-seq peak (insert) from hMSCs with 3 CEBP motifs of the form TKnnGCAA. Purple shading indicates motif locations. Binding to the primary motif (right) is indicated by co-localization with an opposite-stranded peak pair. The secondary motifs co-localize with either a weaker peak pair having too few reads to meet binding cutoffs (left) or background reads (center). Flanking bases (larger font) are indicated as favorable (red) or unfavorable (blue) based on the findings from the palindromic 8-mer. (**D**). Polar bar graphs indicating the frequency of each pair of bases (5' and 3') flanking a generic CEBP motif at primary and secondary motifs. Comparison of the frequencies for favorable, intermediate and unfavorable flanks between the secondary weakly bound versus unbound motifs are highly statistically significant (p < E-13 by hypergeometic distribution).

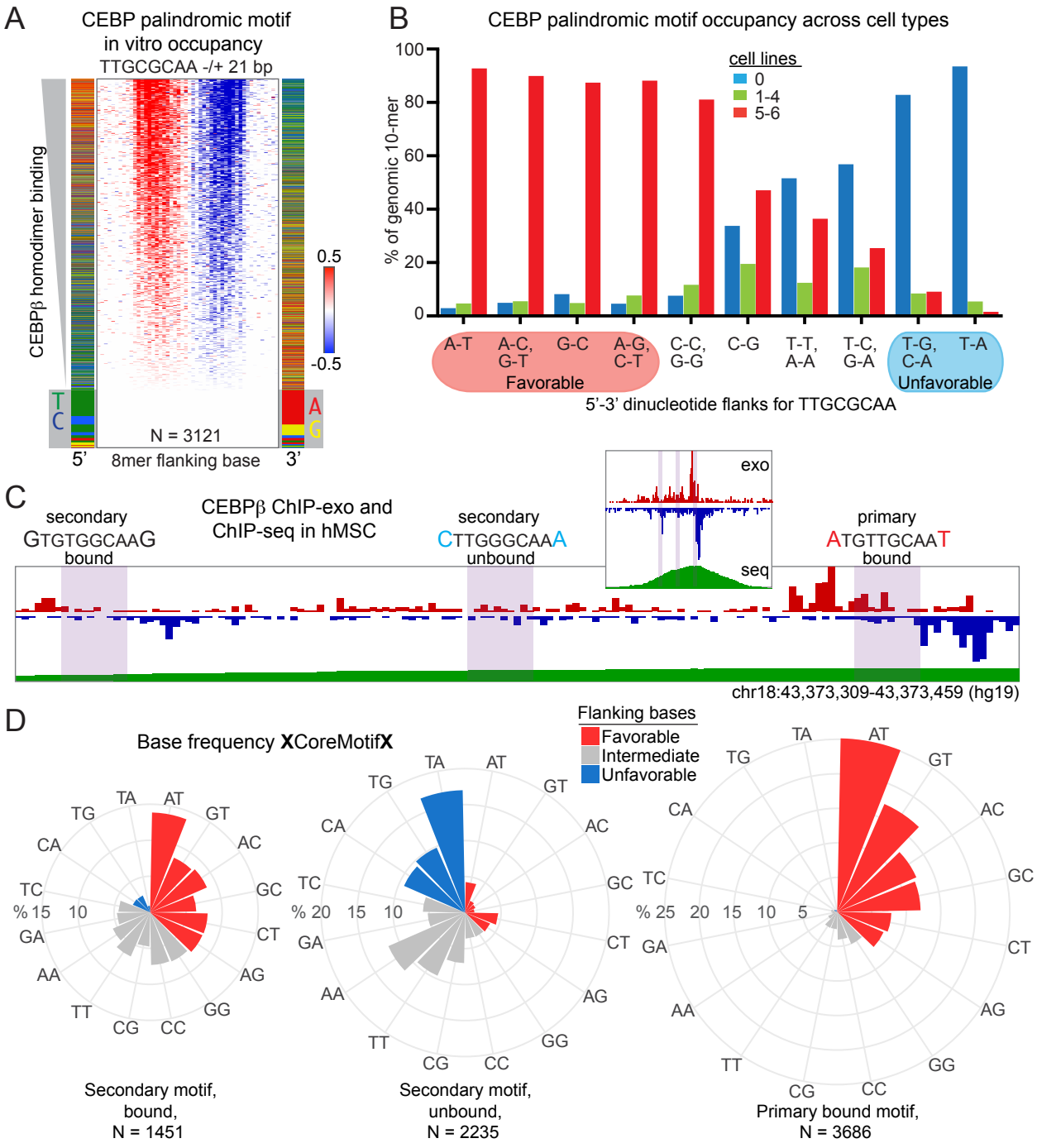
Figure 4. Bases directly flanking core bZip motifs influence DNA shape. (**A**) Density heat maps of DNA shape features (roll, helical twist, minor groove width and propeller twist) at occupied and unoccupied CEBP motifs within CEBPβ ChIP-seq peaks from hMSCs. CEBPβ ChIP-exo signal sets the ordering for all heat maps. F indicates the positions flanking the core 8 bp motif. * denotes p < E-50, Wilcoxon rank sum test. (**B**) Average profile plots of DNA shape parameters at weakly and strongly bound motifs for CREB, NFIL3 and JUND. Differential binding was determined by protein binding microarrays (10). Core motifs and flanking positions (f) are shown. Dashed line indicates the genome-wide average for each shape feature. *, denotes p < 0.01, Wilcoxon rank sum test. (**C**) Motif analyses of CREB1, NFIL3 and AP-1 ChIP-seq data. Top-ranked motif is shown for each emphasizing de-enriched bases in the 5' and 3' positions flanking the core sequences. Negative selection of bases within the cores is not shown.

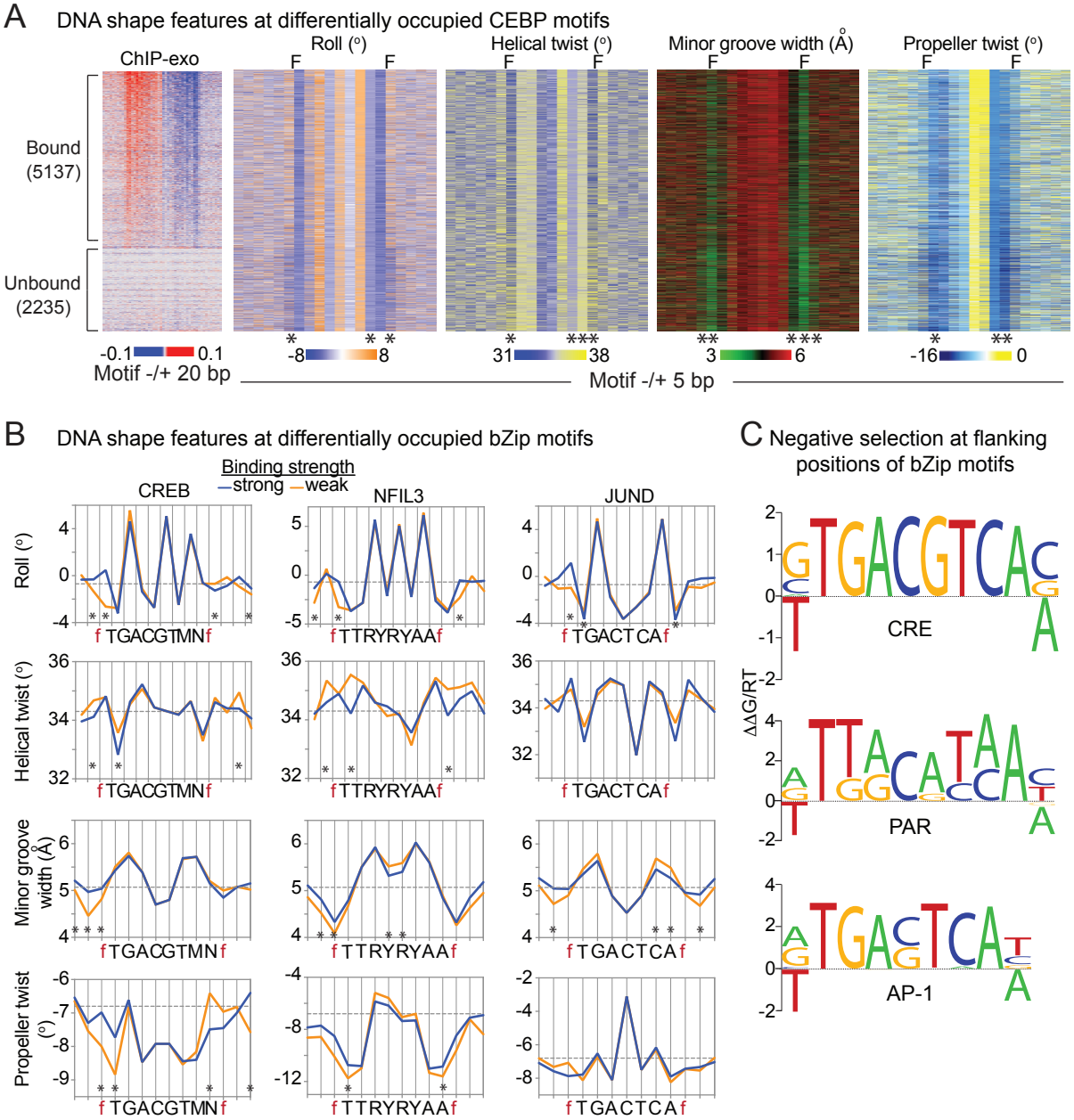
Figure 5. Bases directly flanking core bZip motifs regulate function. (**A**) CEBP ChIP in liver tissue isolated from B6 and 129 mice interrogating binding sites with and without SNPs in the bases flanking the core CEBP 8-mer. ChIP-exo tracks (bottom) show location of SNP relative to core 8-mer and opposite-stranded peak pairs. Ins1, non-specific control site. Error bars depict SEM from 5 biological replicates. *, denotes p < 0.05, Student's *t*-test comparison of B6 with 129. (**B**) Pyrosequencing of CEBPα, CEBPβ and IgG ChIP DNA prepared from liver tissue of B6x129 F1 mice. Chromatograms show raw data for SNP2. Note that these data report the opposite DNA strand shown

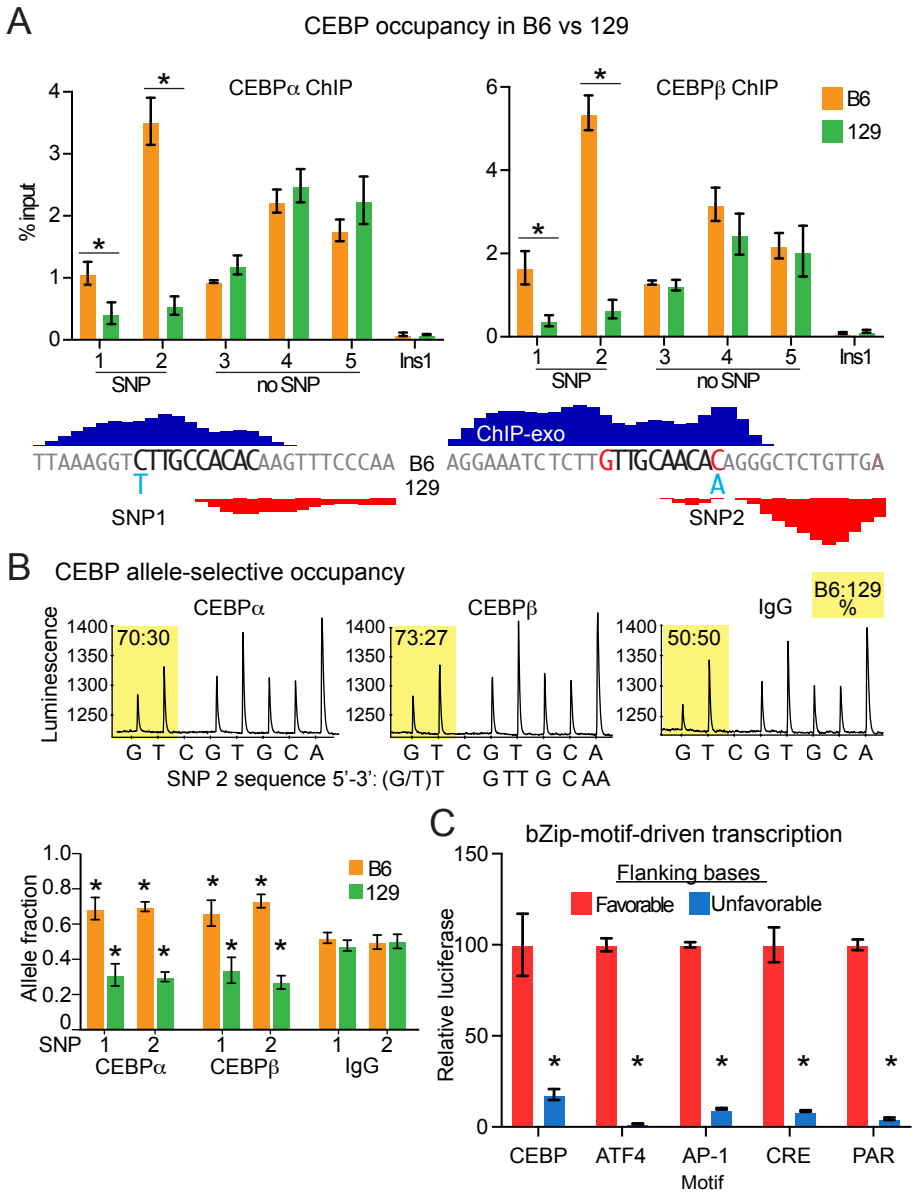
in A. Bar plot (lower left) reports results for SNPs 1 and 2 with error bars depicting SEM from 5 biological replicates. *, denotes $p < 0.05$, Student's *t*-test comparison of CEBPα or CEBPβ with IgG. (**C**) Core bZip motifs were assembled into repeats of four and assayed by a luciferase reporter in HEK293T cells. Flanking bases (X-X) for the CEBP (XTTGTGCAAX), ATF4 (XTGATGCAAX), AP-1 (XTGACTCAX), CRE (XTGACGTCAX) and PAR (XTTACGTAAX) motifs were either favorable (A-T) or unfavorable (T-A, excepting the ATF4 motif, where a T-T pair was used to selectively target the flank of the ATF4 half site) for TF occupancy. Error bars depict SEM from 3 replicates. *, denotes $p < 0.01$, Student's *t*-test comparison of favorable with unfavorable flanks for each motif.


A
ChIP-exo and ChIP-seq in hMSC
exo
CEBPβ
seq
GCAGATACAAAATTGCGCAACAGAGAGCCATCA
exo
CEBPβ
ATF4
seq
AAGAGGCCGGCTGATTGCATCACTCGCTCATTC
B
CEBPβ peak pairs
hMSC
CEBPα peak pairs
liver
Frequency
25
0.6 RPM
27
1 RPM
Distance (bp)
C
CEBPβ, hMSC
CEBPα, liver
bits
E = 10^-1927
Hits = 100%
E = 10^-1632
Hits = 99%
D
CEBPβ, hMSC
CEBPα, liver
RPM
Motif -/+ 50 bp
N = 61,164
Motif -/+ 50 bp
N = 38,390
E
Motif frequency in CEBP cistrome
CEBPβ, hMSC
CEBPβ, liver
CEBPα, liver
%
TTGTGCAA
TGATGCAA
TGTTGCAA
TTGAGCAA
TTTTGCAA
TTGTGAAA
TTATGCAA
TTAAGCAA
TGAAGCAA
TTGGGCAA
TTGTGTAA
TTACGCAA
TTGCGCAA
F
In vitro ChIP-exo
CEBPβ homodimer
CEBPβ-ATF4 heterodimer
ATF4 ChIP-seq, hMSC
bits
E = 10^-2528
Hits = 100%
E = 10^-2153
Hits = 100%
E = 10^-1653
Hits = 73%


A
CEBPβ sites shared across cell types
cell types
1
2-3
4-5
All
19%
18%
31%
32%
N=165,665
B
Cell-type-independent GO
Metabolism of RNA 7E-11
Unfolded protein response 3E-09
Translation 3E-06
Genic RNAPII (RPKM)
10-15
Cell-type independent
Cell-type specific
CEBPβ-binding sites
Cell-type-specific GO
Integrin cell surface interactions 1E-17
Adipocytokine signaling pathway 1E-11
Lipoprotein metabolism 3E-11
C
Motif enrichment at CEBPβ sites
Cell-type independent
Cell-type specific
ΔΔG/RT
1 2 3 4 5 6 7 8 9 10
D
Motif distribution at CEBPβ sites
cell types: 1 2-3 4-5 All
TTGAGAAA
TTTTGCAA
TTGGGAAA
TGAGGCAA
TTAAGCAA
TGAAGCAA
TGATGAAA
TTGGGCAA
TTGAGCAA
TGTTGCAA
TTGTGAAA
TTATGCAA
TGATGCAA
TTGTGCAA
TGACGCAA
TGTCGCAA
TTTCGCAA
TTACGCAA
TTGCGCAA
0
%
50
100


A
CEBP palindromic motif
in vitro occupancy
TTGCGCAA -/+ 21 bp
CEBPβ homodimer binding
0.5
-0.5
T
C
A
G
N = 3121
5'
8mer flanking base
3'
B
CEBP palindromic motif occupancy across cell types
% of genomic 10-mer
100
80
60
40
20
0
cell lines
0
1-4
5-6
A-T
A-C, G-T
G-C
A-G, C-T
C-C, G-G
C-G
T-T, A-A
T-C, G-A
T-G, C-A
T-A
Favorable
Unfavorable
5'-3' dinucleotide flanks for TTGCGCAA
C
CEBPβ ChIP-exo and
ChIP-seq in hMSC
secondary
GTGTGGCAAG
bound
secondary
CTTGGGCAAA
unbound
primary
ATGTTGCAAT
bound
exo
seq
chr18:43,373,309-43,373,459 (hg19)
D
Base frequency XCoreMotifX
Flanking bases
Favorable
Intermediate
Unfavorable
TA
AT
GT
AC
GC
CT
AG
GG
CC
CG
TT
AA
GA
TC
CA
TG
% 15
10
% 20
15
10
% 25
20
15
10
5
Secondary motif,
bound,
N = 1451
Secondary motif,
unbound,
N = 2235
Primary bound motif,
N = 3686


A
DNA shape features at differentially occupied CEBP motifs
ChIP-exo
Roll (°)
Helical twist (°)
Minor groove width (Å)
Propeller twist (°)
Bound (5137)
Unbound (2235)
-0.1
0.1
-8
8
31
38
3
6
-16
0
Motif -/+ 20 bp
Motif -/+ 5 bp
B
DNA shape features at differentially occupied bZip motifs
Binding strength
strong
weak
CREB
NFIL3
JUND
Roll (°)
Helical twist (°)
Minor groove width (Å)
Propeller twist (°)
f TGACGTMN f
f TTRYRYAA f
f TGACTCA f
C
Negative selection at flanking positions of bZip motifs
ΔΔG/RT
CRE
PAR
AP-1


A
CEBP occupancy in B6 vs 129
CEBPα ChIP
CEBPβ ChIP
B6
129
% input
SNP
no SNP
Ins1
ChIP-exo
TTAAAGGTCTTGCCACACAAGTTTCCCAA
B6
129
AGGAAATCTCTTGTTGCAACACAGGGCTCTGTTGA
SNP1
SNP2
B
CEBP allele-selective occupancy
CEBPα
CEBPβ
IgG
B6:129 %
70:30
73:27
50:50
Luminescence
G T C G T G C A
SNP 2 sequence 5'-3': (G/T)T G TT G C AA
Allele fraction
SNP
CEBPα
CEBPβ
IgG
C
bZip-motif-driven transcription
Flanking bases
Favorable
Unfavorable
Relative luciferase
CEBP
ATF4
AP-1
CRE
PAR
Motif