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. Surprisingly, the sequence determinants for CEBPs diverge from classical models. In vivo, CEBPs recognize the fusion of a degenerate and canonical half site, which is atypical for CEBP homodimers and implies altered DNA specificity through heterodimerization. Furthermore, the minimum sequence determinants for CEBP binding are encoded by a 10-mer motif rather than the commonly annotated 8-bp sequence. This extended motif definition is broadly important. First, motif optimization within the 10-mer is strongly correlated with cell-type-independent recruitment of CEBPβ. Second, selection bias at core-motif-flanking nucleotides occurs for multiple bZip proteins. This study sheds new light on DNA-sequence specificity for bZip proteins, and provides key insights into how sequence sub-optimization affects genomic occupancy of 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 (Deplancke, Alpern, & Gardeux, 2016; Vockley, Barrera, & Reddy, 2017). The genomic distribution of TFs at enhancers and promoters defines the framework by which these proteins orchestrate temporal and spatial regulation of gene expression (Shlyueva, Stampfel, & Stark, 2014; Spitz & Furlong, 2012). 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. 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.

Sequence preferences of TFs have been interrogated systematically using high-throughput approaches designed to select TFBSs both in vitro and in vivo (Jolma & Taipale, 2011; Odom, 2011; Stormo & Zhao, 2010). Protein binding microarrays (PBMs) and high-throughput in vitro selection (HT-SELEX) have determined the specificities of hundreds of isolated TFs from multiple species (Jolma et al., 2013; Weirauch et al., 2014). 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, in vitro motif logos have limited ability to predict in vivo binding (Orenstein & Shamir, 2017), perhaps due to relaxed or altered sequence specificity of TFs in their native environments. While this discrepancy can be ascribed in part to the limited, 200 base pair (bp) resolution of ChIP-seq that restricts mapping of TFBSs to short motifs, emerging evidence suggests that altered sequence specificity in cells may be biologically important in TFBS selection. Recent investigations of contextualizing factors such as DNA structure and protein-protein interactions have revealed that DNA methylation (Mann et al., 2013; Yin et al., 2017), neighboring TF interaction (Jolma et al., 2015), and heterodimer formation between related TFs (Rodríguez-Martínez, Reinke, Bhimsaria, Keating, & Ansari, 2017) can change sequence specificity.

Despite these important advances, a causal relationship between the locations of recognition motifs and TFBSs has proven elusive on a site-by-site basis in the genome. A bottom-up approach of motif scanning throughout the genome is daunting given that any particular TF binds only a small fraction of its candidate motifs (Wang et al., 2012), presumably due to the fact that chromatin renders inaccessible the vast majority of sites. A top-down approach of motif scanning at ChIP-seq peaks is often confounded by either the absence of a good match to a consensus sequence or the presence of multiple motif matches, especially for TFs that recognize low-complexity motifs. Fortunately, experimental advances are beginning to resolve these ambiguities. Identification of DNA sequences mediating TF recruitment in the genome has been facilitated by the development of ChIP with lambda exonuclease digestion and sequencing (ChIP-exo) that achieves 20-50 bp resolution of bound sites (Rhee & Pugh, 2011). Close discrimination of bound motifs can revise and improve recognition sequences (Iwata et al., 2017; Luna-Zurita et al., 2016) and resolve dimeric versus monomeric binding (Lim et al., 2015; Starick et al., 2015). 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 (Bartlett et al., 2017; Cohen et al., 2015; Gossett & Lieb, 2008; Guertin, Martins, Siepel, & Lis, 2012). Uniting these approaches has the potential to bridge major gaps in our understanding of the relationship between TF sequence specificity, motif occurrence and occupancy of 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 (Costa, Kalinichenko, Holterman, & Wang, 2003; Friedman, 2002; Rosen et al., 2002; Tsukada, Yoshida, Kominato, & Auron, 2011), 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, and their ability to target different sequence motifs through heterodimerization with other bZip family members (Cohen et al., 2015; Han et al., 2013;

Reinke, Baek, Ashenberg, & Keating, 2013; Rodríguez-Martínez et al., 2017) may enable the utilization of a broad repertoire of motifs to control a variety of gene expression programs. Indeed, CEBPs occupy tens of thousands of sites in primary cells and tissues (Everett et al., 2013; Heinz et al., 2010; Schmidt et al., 2010; Wang et al., 2012), 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 that is anchored by a CEBP half site. The base composition of the degenerate half site at any particular locus determines homo- versus heterodimer occupancy. While positive selection is important, it is striking that, the flanking bases directly abutting the core motif also affect occupancy through negative selection, i.e. the absence of a particular base is more important for binding than the presence of another. 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, natural genetic variation from single nucleotide polymorphisms (SNPs) that introduce non-permissive flanking bases leads to strain-specific CEBP occupancy in mice. Intriguingly, the expanded motif definition for CEBP can be generalized to the broader bZip family, revealing that conserved favorable and unfavorable bases flanking preferred cores determine genomic occupancy and transcriptional activity.

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 (Mymryk & Archer, 1994). 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 (Rhee & Pugh, 2011).

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, Figure 1-figure supplement 1A). 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, Figure 1-figure supplement 1B). Ordered peak pairs flank this motif at a majority of ChIP-seq peaks (Figure 1D, Figure 1-figure supplement 1C), 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 (Figure 1-figure supplement 1D). Nevertheless, this sequence accounts for only about 14% of high-confidence ChIP-exo-annotated binding sites. Taken together with similar ChIP-seq occupancy strengths observed as a function of CEBP 8-mers (Figure 1-figure supplement 1E), 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 (Agre, Johnson, & McKnight, 1989) and more recent systematic biochemical approaches (Isakova et al., 2017; Jolma et al., 2013; Weirauch et al., 2014) report that the CEBP homodimer binds a palindromic motif formed by the fusion of two CEBP half sites (TTGCGCAA). Whether expanded sequence recognition by heterodimers (Cohen et al., 2015; Han et al., 2013; Rodríguez-Martínez et al., 2017) is sufficient to explain the discrepancies between homodimeric versus endogenous CEBP binding is unclear given that in vitro (PBM, HT-SELEX, SMiLE-Seq) and in vivo (ChIP-exo) assays have technical differences. To remove bias introduced 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 (Weirauch et al., 2014), HT-SELEX (Jolma et al., 2013) and SMiLE-seq (Isakova et al., 2017). 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) (Cohen et al., 2015). Sequence-specific interaction by the CEBPβ homo- and heterodimer is indicated by the emergence of peak pairs with fixed spacing that flank both motifs (Figure 1-figure supplement 1F,G). Thus, in vitro ChIP-exo corroborates the DNA sequence specificity reported by established biochemical approaches, and confirms that heterodimer formation with ATF4 alters the specificity of CEBPβ. More broadly, the data illustrate a fundamental, assay-independent difference between the DNA-binding specificity of the CEBPβ homodimer versus CEBPβ in cells. Given the relatively high frequency of CEBPβ occupancy at hybrid sequences comprised of AP-1 or ATF-like half sites fused to a CEBP half site, it seems likely

that heterodimerization with other bZip family members is a major contributor to the broadened motif repertoire recognized by CEBPs in vivo.

Sequence optimization regulates cell-type-specific binding by CEBPβ. The observation that CEBPβ utilizes a similar sequence repertoire in hMSCs and mouse liver raises the question of whether particular 8-mers affect cell-type-specific recruitment. To gain further insight into a relationship between DNA sequence and cell-type-specific binding by CEBPβ, we classified high-confidence, ChIP-exo-annotated CEBPβ sites according to their co-occupancy in 6 different cell lines profiled for CEBPβ binding by the ENCODE consortium (Figure 2A). Consistent with frequency measurements for shared versus unique binding sites for a TF in different cell types (Gertz et al., 2013; John et al., 2011), approximately 10-15% of bound CEBP motifs map to either hMSCs-specific or cell-type-independent peaks, while the remaining 70-80% fall between these extremes. The hMSC-specific and cell-type-independent sites exhibit distinct gene ontologies, suggesting separate biological functions. Genes associated with biological processes characteristic of hMSC differentiation and mesenchymal traits are enriched near hMSC-specific sites, whereas genes with more general roles in the control of transcription and translation reside nearby cell-type-independent sites. Interestingly, CEBPβ-binding strength in hMSCs scales in concert with the number of cell types associated with a bound site (Figure 2B), revealing that the cell-type-independent sites are associated with a genomic context that is more favorable for CEBPβ occupancy. To test whether CEBP sequence varies with cell-type-dependent binding, we generated consensus position weight matrices based on our ChIP-exo data for hMSC-specific or cell-type-independent sites (Figure 2C). For both homodimer and heterodimer-like consensus motifs, we observe a pronounced enrichment for C in the 5^{th} position of the motif at cell-type-independent sites versus hMSC-specific sites. This C adheres to the canonical CEBP half site, revealing that cell-type-independent sites have optimized 8-mers relative to their counterparts at hMSC-specific sites. Since both homodimer and heterodimer-like

182 motifs share a bias for C at position 5, we infer that this sequence preference is conserved between
183 CEBPβ and its heterodimer partners. Parsing ubiquitous and selective CEBPβ sites by individual core
184 8-mers illustrates the association of C within the first half site of cell-type-independent sites, and
185 reveals the specific sequences that are widely bound across cell types (Figure 2D). The motif-centric
186 perspective also highlights the elite behavior of the CEBP palindrome, which exhibits the strongest
187 predilection for cell-type-specific binding such that 60% or more of its bound locations are shared
188 across all cells and 90% or more are bound in at least 5 cell types. These data demonstrate that high-
189 affinity motifs can recruit CEBP homodimers and heterodimers independent of cell type or chromatin
190 structure. Substitution of other bases in lieu of C at the 5th position of the CEBP motif is tolerated, but
191 it yields a suboptimal motif that is bound in a contingent manner, likely attributable to the inherent
192 differences in chromatin structure across cell types (Yue et al., 2014). Together, these data suggest
193 that CEBPβ may play a pioneering role by overcoming chromatin-mediated repression at high-affinity
194 motifs but not at suboptimal sequences.
195
196 **Bases directly abutting the core CEBP motif impact occupancy**. The palindromic motif is superior
197 to all other CEBP motifs for recruitment of CEBPs to the genome. Palindromic CEBP motifs are bound
198 at the highest rate relative to genomic frequency (Figure 1-figure supplement 1D), and sites that are
199 competent for binding have the highest probability of being occupied in multiple cell types (Figure 2D).
200 And yet, most palindromic 8-mers are unoccupied in hMSCs (Figure 1-figure supplement 1D),
201 suggesting that two CEBP half sites comprising a high-affinity sequence for CEBP homodimers is not
202 sufficient to mediate occupancy throughout the genome. To test this, we profiled CEBPβ occupancy at
203 every palindromic motif in the human genome, excluding unplaced contigs, using our in vitro ChIP-exo
204 data generated with purified CEBPβ and histone-free genomic DNA. While the vast majority (84%) of
205 CEBP palindromes showed binding, a subset failed to recruit CEBPβ. Sequence alignments of these
206 unbound regions revealed a pronounced difference in the nucleotide composition of positions

immediately flanking the palindromic 8-mer (Figure 3A). In contrast, neighboring bases extending beyond these flanks are randomly associated (Figure 3-figure supplement 1A). 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). Indeed, the sequence biases at the flanking positions are implicit within the motif logos from our ChIP-exo experiments (Figure 1C, Figure 1-figure supplement 1B), which, if inspected carefully, reveal the exclusion of T at the 5' flank and A at the 3' flank. However, compared to the core 8-mer, the relatively lower information content encoded in these flanks de-emphasizes their importance in the motif logos, and fails to underscore how specific base pairings at the 5' and 3' flanks (T-A, C-A, T-G) can override recruitment of CEBPβ to an optimized 8-mer core sequence.

This observation that flanking nucleotides affect CEBPβ occupancy suggests that the minimal CEBP motif is functionally a 10-mer sequence. Furthermore, it follows that the added specificity constraints encoded by these flanks would influence CEBP occupancy in vivo. To explore this possibility, we examined the occupancy of all 3121 palindromic sites as measured by ChIP-seq for CEBPβ across 6 different cell lines profiled by the ENCODE consortium in combination with our hMSC dataset (Figure 3B). Remarkably, we observe that for the top-5 ranked 10-mer sequences, CEBPβ is commonly bound across 5 or more cell types at the vast majority of sites (63-83%, red bars), and occupancy is observed at $\geq$95% of sites in at least one cell type (sum of red and green bars). Conversely, for 10-mers that contain unfavorable flanks, occupancy is rarely observed in cells, and those sites that are occupied are not shared amongst multiple cell types. These data indicate that flanking nucleotides influence binding to the canonical palindromic CEBP motif in the native genome. Moreover, they reveal that when defined correctly as a 10-mer, CEBP motif candidates predict

genomic occupancy with reasonable accuracy without any prior knowledge of cell type or chromatin structure. 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-dependent 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 (Figure 3-figure supplement 1B).

While these observations enhance our understanding of the optimal CEBP homodimer motif, an important question is whether flanking nucleotides influence the broader CEBP cistrome by affecting binding at core motifs that diverge from the canonical CEBP palindrome. To address this question, we mined our CEBPβ hMSC ChIP-exo dataset to see if the flanks could correctly predict real (ChIP-exo annotated) versus decoy (unbound) candidates for CEBP-binding sites. Specifically, we mapped frequently bound 8-mers (see Figure 1E) that reside in the open chromatin of high-confidence, ChIP-exo-annotated CEBPβ-binding sites (Figure 3C). This analysis detected 3686 candidate 8-mers, or secondary motifs, in the vicinity of a known primary CEBP-binding site. These sites were then classified into bound and unbound based on enrichment for ChIP-exo reads (Figure 3-figure supplement 1C). Within the set of candidate secondary motifs, 61% lacked prominent spaced opposite-stranded peak pairs, indicating little or no occupancy. We then examined the frequency of dinucleotide flanks abutting the primary CEBP 8-mer as well as the bound and unbound secondary 8-mers (Figure 3D). Consistent with our ChIP-exo motif logos and analyses of the palindromic 8-mer, the primary sites 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 (Figure 3-figure supplement 1D), such that favorable flanks occur less frequently than expected by chance. Remarkably, weakly bound

secondary motifs show a preference for favorable flanks that is similar to primary sites, albeit at lower frequencies. In contrast, the 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 a 10-mer sequence as the minimal CEBP recognition motif.

Bases directly abutting core bZip motifs affect transcriptional activity. The aforementioned ChIP-exo studies establish genome-wide trends between CEBP occupancy and the flanking bases for the CEBP core 8-mer motif. To directly address whether a change in the flanks is sufficient to render a change in CEBP binding, 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 constraints of SNP location relative to the CEBP core 8-mer, the 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 4A). Consistent with these results, B6x129 F1 mice showed significantly skewed binding of CEBPs to the B6 alleles (Figure 4B). 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.

More broadly, we sought to determine whether the influence of flanking bases is specific to CEBPs or a more general feature of the bZip family of TFs. A general role for flanking bases in sequence recognition by bZip TFs is supported by the motifs enriched in previously published ChIP-seq datasets for AP-1, CREB and NFIL3 (Mei et al., 2017). Each reveals a clear exclusion of T and A in the 5' and 3' flanking positions, respectively (Figure 4C). 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 ≥ 10-fold for the remaining motifs (Figure 4D). 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

Our ChIP-exo data resolve species-conserved sequence requirements for the recruitment of CEBP proteins to the native genome. Consistent with in vitro data obtained from PBMs (Weirauch et al., 2014), HT-SELEX (Jolma et al., 2013) and SMiLE-seq (Isakova et al., 2017), the CEBP homodimer exhibits a strong preference for palindromic or near-palindromic motifs, comprised of two abutting half sites of VTTRC, that occur rarely in mammalian genomes. In contrast, CEBPs employ an expanded repertoire of recognition sequences in vivo characterized by the fusion of degenerate and canonical CEBP half sites to yield a 10-mer-consensus motif of VTKNNGCAAB. This enables CEBPs to populate large cistromes comprised of tens of thousands of binding sites in mammalian tissues and primary cells (Everett et al., 2013; Heinz et al., 2010; Schmidt et al., 2010; Wang et al., 2012).

It is noteworthy that roughly a third 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 TFs. Extensive evidence has been found for the heterodimerization and altered sequence specificity of CEBP-ATF complexes compared to their homodimer counterparts (Cohen et al., 2015; Han et al., 2013; Reinke et al., 2013; Rodríguez-Martínez et al., 2017). An intriguing question is whether interfamily bZip heterodimerization is broadly important for directing CEBPs to genomic loci with VTGNNGCAAB motifs. Though our current study can only infer that heterodimerization contributes to the genomic recruitment of CEBPs, it does unambiguously demonstrate that 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. Moreover, binding to motifs with atypical spacing between half sites (Rodríguez-Martínez et al., 2017) or to other DNA-bound TFs through tethering contribute minimally to the genomic recruitment of CEBPs.

Rather than simply examining enrichment of overrepresented sequences within CEBP-binding sites, ChIP-exo enables a genome-wide cataloging of motif utilization within the CEBP cistrome. Moreover, in contrast to in vitro methods interrogating sequence-specific binding, ChIP-exo preserves genomic information that affords important comparisons between bound and unbound sites within and across cell types. Harnessing these strengths, we demonstrate that elite CEBP 10-mer motifs comprised of RTTRCGCAAY recruit CEBPβ in a cell-type-independent manner. 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 to access high-affinity sites. In parallel, direct examination of unbound TTGCGCAA motifs reveal flanking bases that are disfavored for CEBPβ binding. Pronounced negative selection against unfavorable flanks suggests that these positions contribute to motif recognition by modulating DNA-binding affinity.

Comparison of bound versus unbound motifs is often excluded from analyses of high-throughput assays interrogating motif preferences of TFs, yet it provides a powerful lens to uncover additional determinants of DNA-binding specificity. By anchoring on CEBP palindromes and examining genomic occupancy as a function of cell type, we further classified a third group of flanks as neutral. 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. Comparison of bound versus unbound sequences at secondary CEBP motifs that occur in the vicinity of strong primary-binding sites reveals enrichment for neutral flanks at bound secondary motifs relative to primary sites.

This suggests that CEBPs can populate low-affinity sequences that reside in open chromatin. In contrast, secondary motifs that have unfavorable flanks are rarely bound.

The deeper understanding of the CEBP motif as a 10-mer sequence comprised of a core 8-mer with tunable binding strength controlled by the nucleotide flanks has profound implications for understanding the relationships between motif occurrence, DNA-binding affinity, and genomic occupancy. Optimal flanks are observed at only 25-30% of candidate VTKNNGCAAB 10-mers in the human genome, compared to 37.5% if all flanks have equal probability of occurrence. Optimized sequences incorporating a palindromic core are extremely rare, totaling 934 sites or 0.2% of candidate 10-mers with optimized flanks. Neutral flanks are present at approximately 50% of the candidate CEBP 10-mers across the genome, revealing a large sequence space for regulated binding by tissue-specific DNA accessibility. Finally, unfavorable flanks could explain why many computationally predicted motifs are rarely bound in any cell type or tissue (Wang et al., 2012).

It is striking that the majority of the CEBP cistrome is populated by motifs that are sub-optimized in the core 8-mer, the flanking bases or both, and that these motifs display narrow occupancy across cell types. This relationship is reminiscent of the sub-optimization of motifs reported for cell-type-dependent binding of ERα (Gertz et al., 2013; Joseph et al., 2010). Intriguingly, high-throughput transcriptional screens in Drosophila have demonstrated a genetic bias towards motif sub-optimization at developmentally-regulated enhancers (Farley et al., 2015). 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.

Moreover, our findings may have broad impact on the understanding of DNA-binding specificity for bZip proteins in general. Core motifs for ATF4, AP-1, CREB and PAR are also de-enriched for T-A flanks that cripple their ability to promote transcription. While negative selection at the flanks appears to be a conserved feature of bZip motifs, a thermodynamic basis for this observation is currently lacking. Crystallography studies of several bZip-DNA complexes, including the

CEBPα DNA-binding domain (Ellenberger, Brandl, Struhl, & Harrison, 1992; Fujii, Shimizu, Toda, Yanagida, & Hakoshima, 2000; Glover & Harrison, 1995; Miller, Shuman, Sebastian, Dauter, & Johnson, 2003; Schumacher, Goodman, & Brennan, 2000), have modeled binding to short DNA duplexes of about 20 bp. Given that the apparent footprint of CEBP by ChIP-exo is approximately 25 bp, it is possible that protein-DNA interactions outside of the core 8-mer may have been missed due to the short and flexible nature of oligonucleotide duplexes. Unlike positive selection within the core 8-mer, negative selection at the flanks may be detected by contacts between bZip proteins and the DNA backbone at these positions. Consistent with this hypothesis, it has been suggested that motif flanks regulate genomic occupancy of E-box TFs through alteration of DNA shape (Gordân et al., 2013). However, given that shape parameters are computed from the underlying DNA sequence, it is difficult to experimentally uncouple DNA sequence from DNA shape at the resolution of a single position in the context of binding assays such as ChIP-seq or ChIP-exo. Rather, structural analyses of bZip-DNA complexes using longer DNA templates and comparing motifs with optimized versus sub-optimized flanks will likely be necessary to explain how flanking nucleotides influence motif recognition by bZip TFs.

MATERIALS AND METHODS

Animal Care and Cell Culture: 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 and ChIP-exo: The following antibodies were used: CEBPα (sc-61, Santa Cruz), CEBPβ (sc-150, Santa Cruz), ATF4 (sc-200, Santa Cruz) and normal IgG (2729, Cell Signaling). ChIP in mouse liver was performed with a minimum of 3 individual mice per genotype. Primers for human and mouse ChIP are reported in Table S1.

An Illumina-based ChIP-exo method (Serandour, Brown, Cohen, & Carroll, 2013) was performed with mouse liver and hMSCs. ChIP-exo in hMSCs was performed using approximately 7 million cells. ChIP-exo performed in vitro used binding conditions described previously for an in vitro cistromics assay modeled after ChIP-seq (Cohen et al., 2015). Binding reactions (100 μl) were treated with 1% formaldehyde for 1 min at room temperature, quenched with 125 mM glycine for 5 min, and brought to 1 ml with binding buffer lacking DTT for immunoprecipitation. ATF4 and CEBPβ were expressed in BL21(DE3) *Escherichia coli* from pET30a vectors carrying an N-terminal His Tag and full-length cDNAs. Proteins were purified by Co^{2+} affinity chromatography and quantified by comparison with a BSA standard after SDS-PAGE. Approximately 25 μM ATF4 and/or 600 nM CEBPβ were present in the initial binding reaction of 100 μl. Library preparation was performed similarly to ChIP-exo with cells or tissue.

ChIP-exo data processing including initial peak calling of putative flanking borders and cross-correlation analysis of opposite-stranded peak pairs was performed as described previously (Lim et al., 2015). Biological replicates were performed for each liver and hMCS experiment, and peak pairs were scored as positive if called in each replicate, while downstream analyses utilized pooled data from both replicates. De novo motif search was performed within 50 bp windows from the top-1000 sites showing 15-30 bp spacing between opposite-stranded peak pairs using MEME-ChIP (Machanick & Bailey, 2011). Motif visualization emphasizing de-enriched bases was performed with REDUCE (Roven & Bussemaker, 2003). ChIP-exo signal was visualized after scanning anchoring regions with the de novo motif PWM and identifying the site with maximum score in each region. Peaks from matching ChIP-seq data (Bauer et al., 2015; Cohen et al., 2015) served as anchoring regions for

ChIP-exo experiments performed with tissue or cells. For in vitro ChIP-exo studies, opposite-stranded peak pairs of 15-30 bp were pooled, merged and extended to a 50 bp window.

Co-occupancy in ENCODE datasets: ENCODE peak calls (narrowPeak) for CEBPβ binding from 6 cell lines (A549, H1esc, HelaS3, HepG2, IMR90, K562) were pooled with 300bp ChIP-seq calls for CEBPβ in hMSCs stimulated with DMI for 24 hours. Co-bound CEBPβ peaks were defined as having at least 50bp overlap across cell types and each ChIP-seq peak scored for cell-type specific binding using Galaxy Tools suite available at cistrome.org/ap. High confidence ChIP exo-bound motifs were assigned the cell-type specific binding score of their corresponding ChIP-seq peak. Co-occupancy of CEBPβ binding sites with other TFs was assessed using the peak calls in the wgEncodeRegTfbsClusteredV3 supertrack, filtered to exclude contributions to co-occupancy from CEBPβ, CEBPδ, RNA polymerase II, or RNA polymerase III.

Analysis of secondary candidate CEBP motifs: We mapped candidate 8-mers that matched the sequence any of the top 13 occupied CEBP sequences (see Figure 1E) in the hg19 genome using HOMER. Candidate 8-mers that occurred within an 11-200bp window adjacent to a high confidence CEBP ChIP exo-bound (primary) site were included in the analyses in Figure 3C/D. ChIP exo-bound secondary sites were defined as having either a positive or negative peak pair exceeding 0.1 RPM read count and a summed RPM score of 0.15 RPM or greater. Heatmaps in Figure S3C were ordered based on decreasing RPM exo strength (sum of positive and negative exo peak pair) at the secondary 8-mer sequence. Heatmaps visualizations were generated in Treeview. Polar bar graphs of flanking nucleotide pair frequency were generated using R.

Luciferase Reporter Assays: Synethic bZip reporters were comprised of four repeats of each motif, separated by identical 18 bp spacer sequences. Unique flanking 5' and 3' extensions were added as PCR anchor points. Reporter cassettes totaled 163 bp or 159 bp for 10 bp or 9 bp bZip motifs, respectively. DNA cassettes were purchased as Ultramers from IDT, PCR amplified, and subcloned

435 into the pGL4.24 vector using XhoI and BglII sites. Sequence for each multimerized motif is available
436 in Table S1. Sanger sequencing verified all plasmids. HEK293T were co-transfected with 320 ng of
437 luciferase reporter construct, 40 ng of CMV-Renilla plasmid, and 40 ng of pCDNA3.1 or pCDNA3.1-
438 Cebpb, using 1.2 μl of Lipofectamine 2000 (Life Technologies) per well, in a reverse transfection
439 protocol using 24 well plates. Cells were lysed in passive lysis buffer either 24 hours (CRE reporter) or
440 48 hours (all other reporters) post-transfection. Relative luminescence (Firefly to Renilla ratio) was
441 determined on a Biotek Synergy H1 microplate reader. Samples were assayed in a minimum of
442 triplicates per biological condition.
443
444 **Pyrosequencing**: DNA (5% of ChIP samples or 0.5% of input DNA) was amplified by PCR using
445 Phusion polymerase (NEB), a biotinylated forward primer designed using the PyroMark Assay Design
446 software (Qiagen), and the reverse primer employed for ChIP-qPCR (see Table S1). A total of 37
447 amplification cycles was used. PCR amplified ChIP DNA was gel isolated and purified using ChIP
448 DNA Clean & Concentrate columns and eluted in 15uL of water. 3uL of biotinylated PCR products
449 was used per pyrosequencing reaction. Pyrosequencing was performed on a PyroMark Q96 MD
450 instrument using the PyroMark Gold reagents per the manufacturer's instructions (Qiagen).
451
452 **ACKNOWLEDGMENTS**

453 We are grateful to Raymond Soccio for guidance on the experimental strategy interrogating
454 strain-specific TF binding, and for providing liver tissue from 129S1/SvImJ and
455 129S1/SvImJxC57BL/6J F1 mice. We also thank Chris Krapp and Marisa Bartolomei for help with
456 pyrosequencing and generously providing access to their sequencer. We are indebted to members of
457 the Lazar laboratory for insightful discussions, and also thank the Functional Genomics Core of the
458 Penn Diabetes Center (DK19525) for deep sequencing. This work was supported by NIH grants R01
459 DK106027 (to K-JW) and R01 DK098542 (to DJS).
460

COMPETING INTERESTS

We have none to report.

REFERENCES

Agre, P., Johnson, P. F., & McKnight, S. L. (1989). Cognate DNA binding specificity retained after leucine zipper exchange between GCN4 and C/EBP. *Science (New York, N.Y.)*, *246*(4932), 922–926.

Bartlett, A., O'Malley, R. C., Huang, S.-S. C., Galli, M., Nery, J. R., Gallavotti, A., & Ecker, J. R. (2017). Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nature Protocols*, *12*(8), 1659–1672. https://doi.org/10.1038/nprot.2017.055

Bauer, R. C., Sasaki, M., Cohen, D. M., Cui, J., Smith, M. A., Yenilmez, B. O., … Rader, D. J. (2015). Tribbles-1 regulates hepatic lipogenesis through posttranscriptional regulation of C/EBPα. *The Journal of Clinical Investigation*, *125*(10), 3809–3818. https://doi.org/10.1172/JCI77095

Cohen, D. M., Won, K.-J., Nguyen, N., Lazar, M. A., Chen, C. S., & Steger, D. J. (2015). ATF4 licenses C/EBPβ activity in human mesenchymal stem cells primed for adipogenesis. *ELife*, *4*, e06821. https://doi.org/10.7554/eLife.06821

Costa, R. H., Kalinichenko, V. V., Holterman, A.-X. L., & Wang, X. (2003). Transcription factors in liver development, differentiation, and regeneration. *Hepatology (Baltimore, Md.)*, *38*(6), 1331–1347. https://doi.org/10.1016/j.hep.2003.09.034

Deplancke, B., Alpern, D., & Gardeux, V. (2016). The Genetics of Transcription Factor DNA Binding Variation. *Cell*, *166*(3), 538–554. https://doi.org/10.1016/j.cell.2016.07.012

Ellenberger, T. E., Brandl, C. J., Struhl, K., & 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*(7), 1223–1237.

Everett, L. J., Le Lay, J., Lukovac, S., Bernstein, D., Steger, D. J., Lazar, M. A., & 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. https://doi.org/10.1186/1471-2164-14-337

Farley, E. K., Olson, K. M., Zhang, W., Brandt, A. J., Rokhsar, D. S., & Levine, M. S. (2015). Suboptimization of developmental enhancers. *Science (New York, N.Y.)*, *350*(6258), 325–328. https://doi.org/10.1126/science.aac6948

Friedman, A. D. (2002). Transcriptional regulation of granulocyte and monocyte development. *Oncogene*, *21*(21), 3377–3390. https://doi.org/10.1038/sj.onc.1205324

Fujii, Y., Shimizu, T., Toda, T., Yanagida, M., & Hakoshima, T. (2000). Structural basis for the diversity of DNA recognition by bZIP transcription factors. *Nature Structural Biology*, *7*(10), 889–893. https://doi.org/10.1038/82822

Gertz, J., Savic, D., Varley, K. E., Partridge, E. C., Safi, A., Jain, P., … Myers, R. M. (2013). Distinct properties of cell-type-specific and shared transcription factor binding sites. *Molecular Cell*, *52*(1), 25–36. https://doi.org/10.1016/j.molcel.2013.08.037

Glover, J. N., & Harrison, S. C. (1995). Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature*, *373*(6511), 257–261. https://doi.org/10.1038/373257a0

Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., & Bulyk, M. L. (2013). Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Reports*, *3*(4), 1093–1104. https://doi.org/10.1016/j.celrep.2013.03.014

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

Guertin, M. J., Martins, A. L., Siepel, A., & Lis, J. T. (2012). Accurate prediction of inducible transcription factor binding intensities in vivo. *PLoS Genetics*, *8*(3), e1002610. https://doi.org/10.1371/journal.pgen.1002610

511 Han, J., Back, S. H., Hur, J., Lin, Y.-H., Gildersleeve, R., Shan, J., … Kaufman, R. J. (2013). ER-
512 stress-induced transcriptional regulation increases protein synthesis leading to cell death.
513 *Nature Cell Biology*, *15*(5), 481–490. https://doi.org/10.1038/ncb2738
514 Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., … Glass, C. K. (2010). Simple
515 combinations of lineage-determining transcription factors prime cis-regulatory elements
516 required for macrophage and B cell identities. *Molecular Cell*, *38*(4), 576–589.
517 https://doi.org/10.1016/j.molcel.2010.05.004
518 Isakova, A., Groux, R., Imbeault, M., Rainer, P., Alpern, D., Dainese, R., … Deplancke, B. (2017).
519 SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nature*
520 *Methods*, *14*(3), 316–322. https://doi.org/10.1038/nmeth.4143
521 Iwata, A., Durai, V., Tussiwand, R., Briseño, C. G., Wu, X., Grajales-Reyes, G. E., … Murphy, K. M.
522 (2017). Quality of TCR signaling determined by differential affinities of enhancers for the
523 composite BATF-IRF4 transcription factor complex. *Nature Immunology*, *18*(5), 563–572.
524 https://doi.org/10.1038/ni.3714
525 John, S., Sabo, P. J., Thurman, R. E., Sung, M.-H., Biddie, S. C., Johnson, T. A., …
526 Stamatoyannopoulos, J. A. (2011). Chromatin accessibility pre-determines glucocorticoid
527 receptor binding patterns. *Nature Genetics*, *43*(3), 264–268. https://doi.org/10.1038/ng.759
528 Jolma, A., & Taipale, J. (2011). Methods for Analysis of Transcription Factor DNA-Binding Specificity
529 In Vitro. *Sub-Cellular Biochemistry*, *52*, 155–173. https://doi.org/10.1007/978-90-481-9069-0_7
530 Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., … Taipale, J. (2013). DNA-
531 binding specificities of human transcription factors. *Cell*, *152*(1–2), 327–339.
532 https://doi.org/10.1016/j.cell.2012.12.009
533 Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., … Taipale, J. (2015). DNA-
534 dependent formation of transcription factor pairs alters their binding specificity. *Nature*,
535 *527*(7578), 384–388. https://doi.org/10.1038/nature15518

536 Joseph, R., Orlov, Y. L., Huss, M., Sun, W., Kong, S. L., Ukil, L., … Liu, E. T. (2010). Integrative
537 model of genomic factors for determining binding site selection by estrogen receptor-α.
538 *Molecular Systems Biology*, *6*, 456. https://doi.org/10.1038/msb.2010.109
539 Lim, H.-W., Uhlenhaut, N. H., Rauch, A., Weiner, J., Hübner, S., Hübner, N., … Steger, D. J. (2015).
540 Genomic redistribution of GR monomers and dimers mediates transcriptional response to
541 exogenous glucocorticoid in vivo. *Genome Research*, *25*(6), 836–844.
542 https://doi.org/10.1101/gr.188581.114
543 Luna-Zurita, L., Stirnimann, C. U., Glatt, S., Kaynak, B. L., Thomas, S., Baudin, F., … Bruneau, B. G.
544 (2016). Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and
545 Coordinates Cardiogenesis. *Cell*, *164*(5), 999–1014. https://doi.org/10.1016/j.cell.2016.01.004
546 Machanick, P., & Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets.
547 *Bioinformatics (Oxford, England)*, *27*(12), 1696–1697.
548 https://doi.org/10.1093/bioinformatics/btr189
549 Mann, I. K., Chatterjee, R., Zhao, J., He, X., Weirauch, M. T., Hughes, T. R., & Vinson, C. (2013). CG
550 methylated microarrays identify a novel methylated sequence bound by the CEBPB|ATF4
551 heterodimer that is active in vivo. *Genome Research*, *23*(6), 988–997.
552 https://doi.org/10.1101/gr.146654.112
553 Mei, S., Qin, Q., Wu, Q., Sun, H., Zheng, R., Zang, C., … Liu, X. S. (2017). Cistrome Data Browser: a
554 data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids*
555 *Research*, *45*(D1), D658–D662. https://doi.org/10.1093/nar/gkw983
556 Miller, M., Shuman, J. D., Sebastian, T., Dauter, Z., & Johnson, P. F. (2003). Structural basis for DNA
557 recognition by the basic region leucine zipper transcription factor CCAAT/enhancer-binding
558 protein alpha. *The Journal of Biological Chemistry*, *278*(17), 15178–15184.
559 https://doi.org/10.1074/jbc.M300417200
560 Mymryk, J. S., & Archer, T. K. (1994). Detection of transcription factor binding in vivo using lambda
561 exonuclease. *Nucleic Acids Research*, *22*(20), 4344–4345.

562 Odom, D. T. (2011). Identification of Transcription Factor-DNA Interactions In Vivo. *Sub-Cellular*
563 *Biochemistry*, *52*, 175–191. https://doi.org/10.1007/978-90-481-9069-0_8
564 Orenstein, Y., & Shamir, R. (2017). Modeling protein-DNA binding via high-throughput in vitro
565 technologies. *Briefings in Functional Genomics*, *16*(3), 171–180.
566 https://doi.org/10.1093/bfgp/elw030
567 Reinke, A. W., Baek, J., Ashenberg, O., & Keating, A. E. (2013). Networks of bZIP protein-protein
568 interactions diversified over a billion years of evolution. *Science (New York, N.Y.)*, *340*(6133),
569 730–734. https://doi.org/10.1126/science.1233465
570 Rhee, H. S., & Pugh, B. F. (2011). Comprehensive genome-wide protein-DNA interactions detected at
571 single-nucleotide resolution. *Cell*, *147*(6), 1408–1419.
572 https://doi.org/10.1016/j.cell.2011.11.013
573 Rodríguez-Martínez, J. A., Reinke, A. W., Bhimsaria, D., Keating, A. E., & Ansari, A. Z. (2017).
574 Combinatorial bZIP dimers display complex DNA-binding specificity landscapes. *ELife*, *6*.
575 https://doi.org/10.7554/eLife.19272
576 Rosen, E. D., Hsu, C.-H., Wang, X., Sakai, S., Freeman, M. W., Gonzalez, F. J., & Spiegelman, B. M.
577 (2002). C/EBPalpha induces adipogenesis through PPARgamma: a unified pathway. *Genes &*
578 *Development*, *16*(1), 22–26. https://doi.org/10.1101/gad.948702
579 Roven, C., & Bussemaker, H. J. (2003). REDUCE: An online tool for inferring cis-regulatory elements
580 and transcriptional module activities from microarray data. *Nucleic Acids Research*, *31*(13),
581 3487–3490.
582 Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., … Odom, D. T.
583 (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor
584 binding. *Science (New York, N.Y.)*, *328*(5981), 1036–1040.
585 https://doi.org/10.1126/science.1186176
586 Schumacher, M. A., Goodman, R. H., & Brennan, R. G. (2000). The structure of a CREB
587 bZIP.somatostatin CRE complex reveals the basis for selective dimerization and divalent

588 cation-enhanced DNA binding. *The Journal of Biological Chemistry*, *275*(45), 35242–35247.
589 https://doi.org/10.1074/jbc.M007293200
590 Serandour, A. A., Brown, G. D., Cohen, J. D., & Carroll, J. S. (2013). Development of an Illumina-
591 based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties.
592 *Genome Biology*, *14*(12), R147. https://doi.org/10.1186/gb-2013-14-12-r147
593 Shlyueva, D., Stampfel, G., & Stark, A. (2014). Transcriptional enhancers: from properties to genome-
594 wide predictions. *Nature Reviews. Genetics*, *15*(4), 272–286. https://doi.org/10.1038/nrg3682
595 Spitz, F., & Furlong, E. E. M. (2012). Transcription factors: from enhancer binding to developmental
596 control. *Nature Reviews. Genetics*, *13*(9), 613–626. https://doi.org/10.1038/nrg3207
597 Starick, S. R., Ibn-Salem, J., Jurk, M., Hernandez, C., Love, M. I., Chung, H.-R., … Meijsing, S. H.
598 (2015). ChIP-exo signal associated with DNA-binding motifs provides insight into the genomic
599 binding of the glucocorticoid receptor and cooperating transcription factors. *Genome*
600 *Research*, *25*(6), 825–835. https://doi.org/10.1101/gr.185157.114
601 Stormo, G. D., & Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. *Nature*
602 *Reviews. Genetics*, *11*(11), 751–760. https://doi.org/10.1038/nrg2845
603 Tsukada, J., Yoshida, Y., Kominato, Y., & Auron, P. E. (2011). The CCAAT/enhancer (C/EBP) family
604 of basic-leucine zipper (bZIP) transcription factors is a multifaceted highly-regulated system for
605 gene regulation. *Cytokine*, *54*(1), 6–19. https://doi.org/10.1016/j.cyto.2010.12.019
606 Vockley, C. M., Barrera, A., & Reddy, T. E. (2017). Decoding the role of regulatory element
607 polymorphisms in complex disease. *Current Opinion in Genetics & Development*, *43*, 38–45.
608 https://doi.org/10.1016/j.gde.2016.10.007
609 Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., … Weng, Z. (2012). Sequence
610 features and chromatin structure around the genomic regions bound by 119 human
611 transcription factors. *Genome Research*, *22*(9), 1798–1812.
612 https://doi.org/10.1101/gr.139105.112

613 Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., … Hughes, T.
614 R. (2014). Determination and inference of eukaryotic transcription factor sequence specificity.
615 *Cell*, *158*(6), 1431–1443. https://doi.org/10.1016/j.cell.2014.08.009
616 Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., … Taipale, J. (2017).
617 Impact of cytosine methylation on DNA binding specificities of human transcription factors.
618 *Science (New York, N.Y.)*, *356*(6337). https://doi.org/10.1126/science.aaj2239
619 Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., … Mouse ENCODE Consortium.
620 (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature*,
621 *515*(7527), 355–364. https://doi.org/10.1038/nature13992
622
623

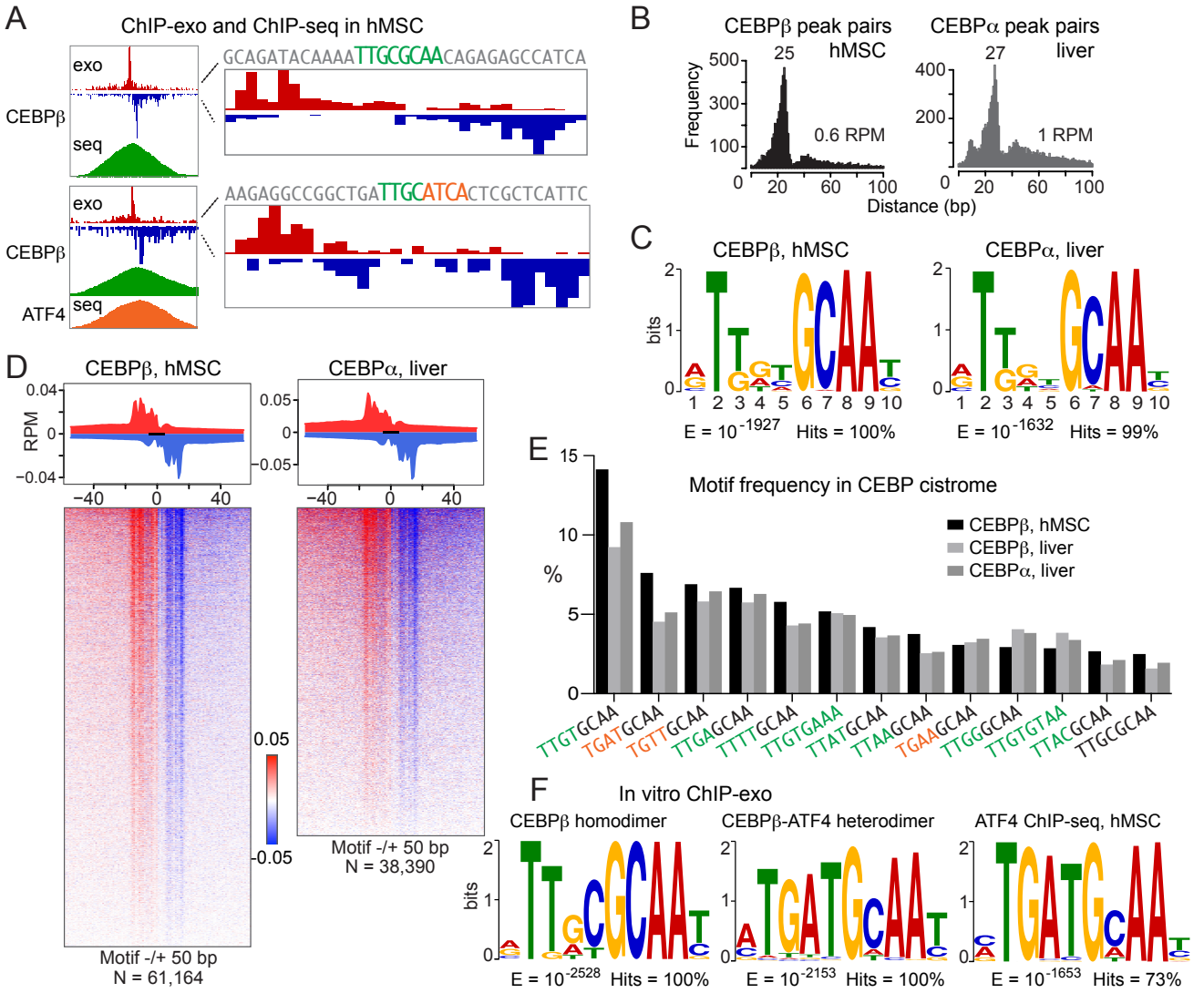

Figure 1. CEBP proteins occupy multiple sequence motifs on the native genome. (**A**) Comparison of ChIP-exo and ChIP-seq results for CEBPβ in hMSCs. Left, an opposite-stranded peak pair from ChIP-exo resides near the center of the ChIP-seq peak for either a homodimer-binding site (top) or a heterodimer 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.

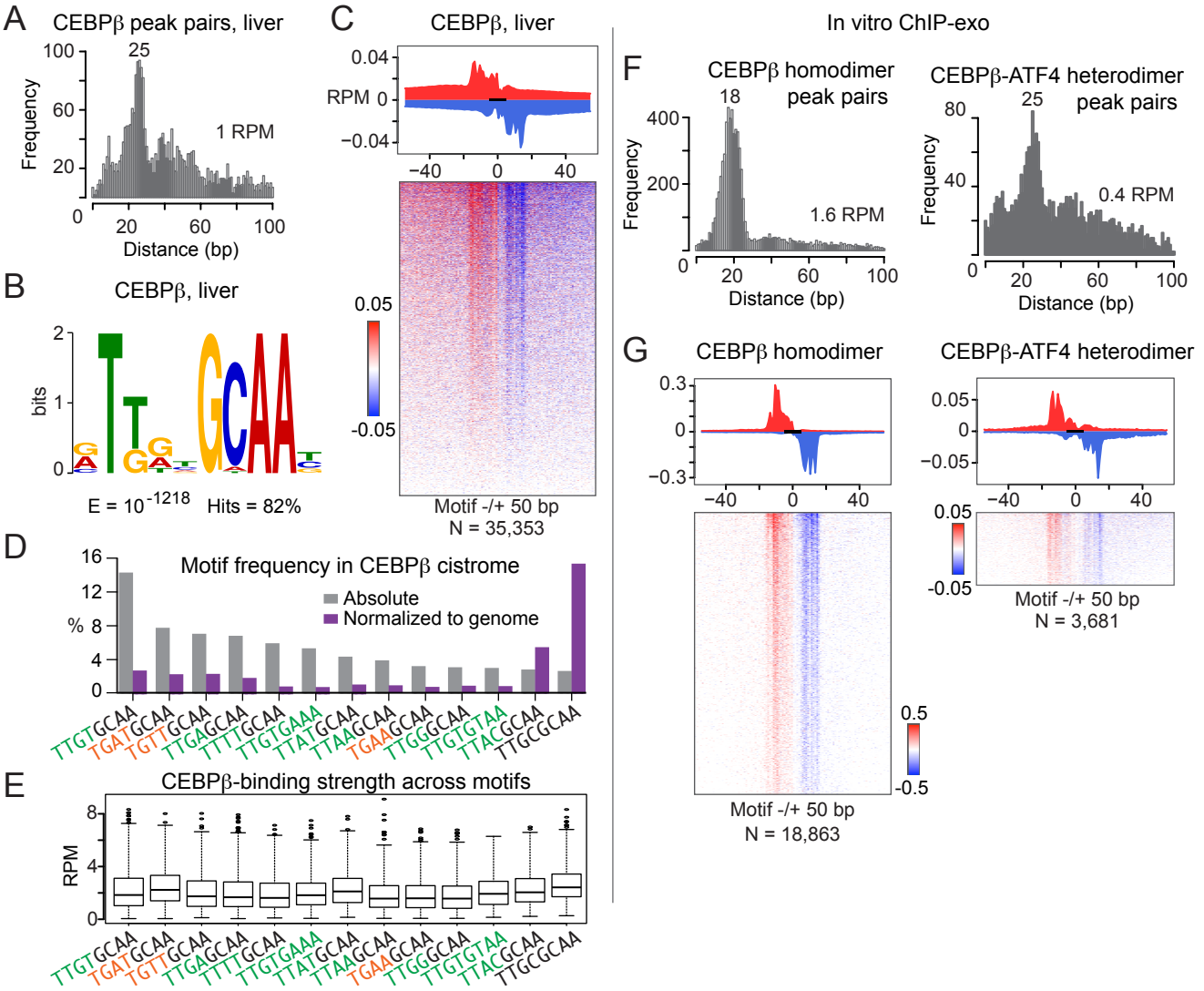

Figure 1-figure supplement 1. In vivo and in vitro ChIP-exo analyses of CEBPβ. (**A**) Distance distribution for the spacing between opposite-stranded peak pairs for CEBPβ in liver. The predominant distance is indicated. (**B**) MEME de novo motif analysis of the 1000-top-ranked peak pairs spaced 15-30 bp apart. (**C**) Average profile (top) and density heat map (bottom) of the ChIP-exo sequence tags at CEBPβ-binding sites in liver. (**D**) Frequencies of the top-ranked core motifs for CEBPβ in hMSCs with and without normalization to the genomic frequency for each motif. The CEBP half site, GCAA, is uncolored; degenerate half site is green (CEBP related) or orange (bZip related). (**E**) Box plots showing CEBPβ-binding strength (ChIP-seq reads per million, RPM) in hMSCs at the top-ranked core motifs. (**F**) Distance distributions for the spacing between opposite-stranded peak pairs from an in vitro cistromics assay modeled after ChIP-exo. The CEBPβ homodimer and CEBPβ-ATF4 heterodimer were immunoprecipitated by CEBPβ and ATF4 antibodies, respectively. Predominant distances are indicated. (**G**) Average profiles (top) and density heat maps (bottom) of the in vitro ChIP-exo sequence tags at binding sites for the CEBPβ homodimer and heterodimer.

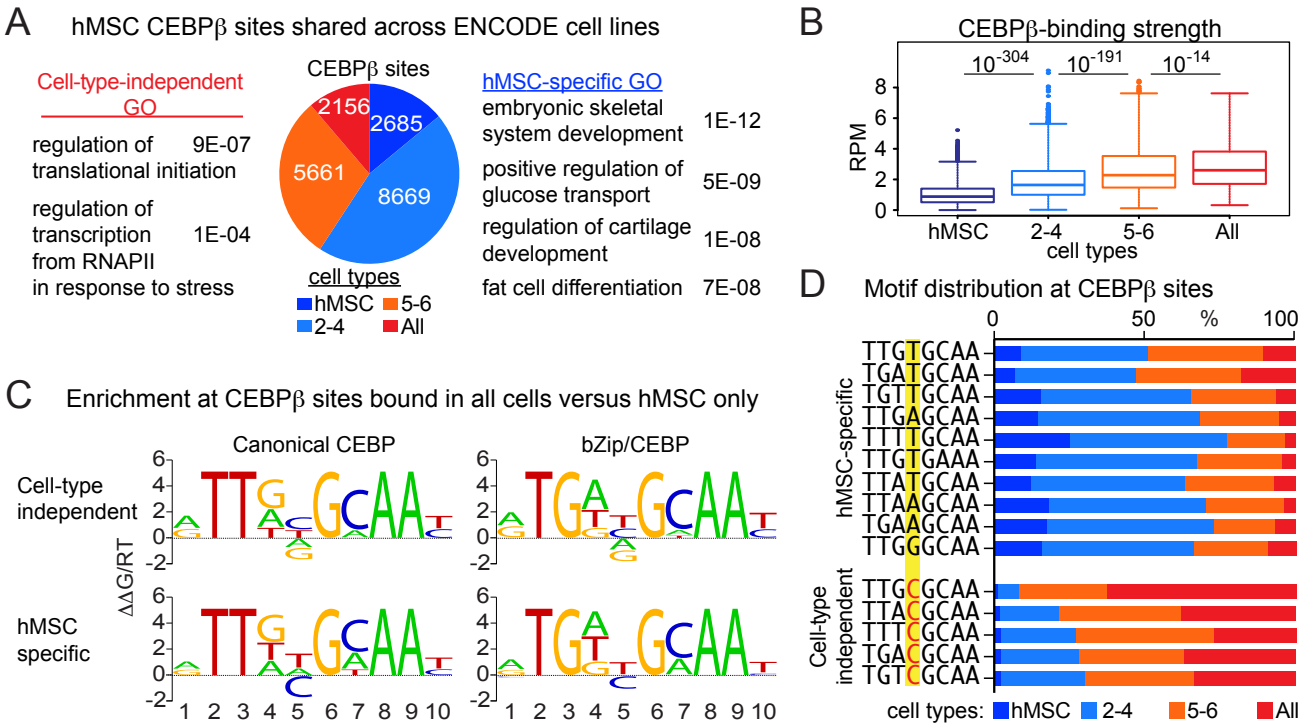

Figure 2. Selective versus widespread occupancy across cell types for distinct CEBPβ motifs. (**A**) Pie chart comparing CEBPβ occupancy across 7 human cell types (hMSC + 6 ENCODE cell lines) at 19,171 ChIP-exo-annotated CEBPβ sites from hMSCs. ENCODE ChIP-seq peak calls determined occupancy in cells other than hMSCs. Top-ranked gene ontology (GO) terms for sites bound in all cells (cell-type independent) or in hMSCs only (hMSC specific) are shown. (**B**) Box plots interrogating CEBPβ-binding strength (ChIP-seq reads per million, RPM) in hMSCs at the classes of sites defined in A. Wilcoxon rank sum test used to compare adjacent classes. (**C**) De novo motif analyses showing de-enriched bases that differ between cell-type-independent and hMSC-specific sites. Top-ranked sequence was subdivided into the canonical and hybrid CEBP motifs. (**D**) Individual 8-mers enriched at cell-type-independent or hMSC-specific sites for CEBPβ were examined to display occupancy across the 7 human cell types. Base position differing between the two classes of sites is highlighted.

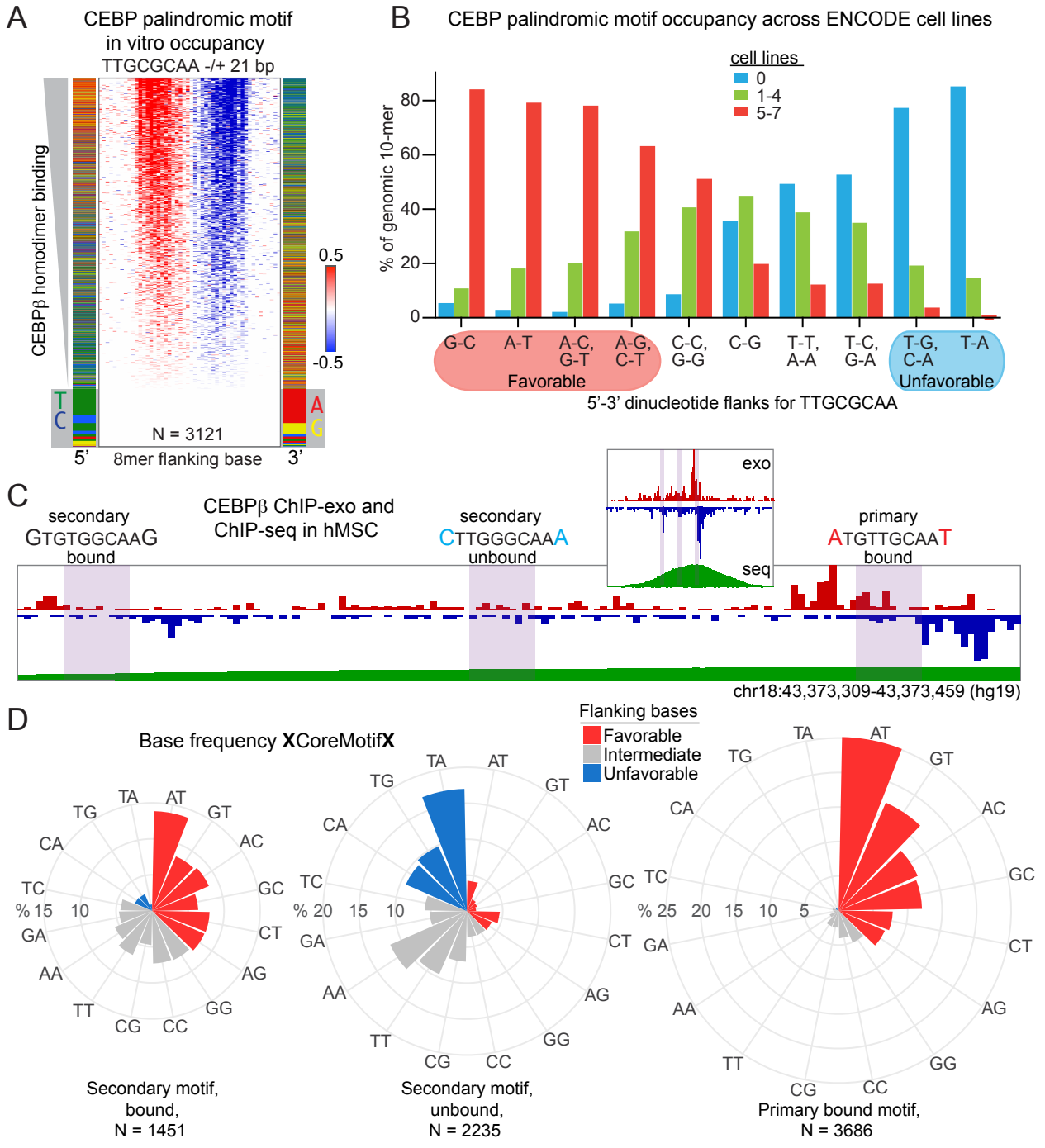

Figure 3. Bases directly flanking the core CEBP 8-mer affect occupancy. (**A**) Density heat map of the in vitro ChIP-exo reads for the CEBPβ homodimer at all canonical palindromic 8-mers with mappable sequence. Binding strength is ordered from top to bottom. Color charts show the base identity at the 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

676 at most locations. (**C**) CEBPβ ChIP-exo reads at a ChIP-seq peak (insert) from hMSCs with 3 CEBP
677 motifs of the form TKnnGCAA. Purple shading indicates motif locations. Binding to the primary motif
678 (right) is indicated by co-localization with an opposite-stranded peak pair. The secondary motifs co-
679 localize with either a weaker peak pair having too few reads to meet binding cutoffs (left) or
680 background reads (center). Flanking bases (larger font) are indicated as favorable (red) or
681 unfavorable (blue) based on the findings from the palindromic 8-mer. (**D**). Polar bar graphs indicating
682 the frequency of each pair of bases (5' and 3') flanking a generic CEBP motif at primary and
683 secondary motifs. Comparison of the frequencies for favorable, intermediate and unfavorable flanks
684 between the secondary bound versus unbound motifs are highly statistically significant (p < E-13 by
685 hypergeometic distribution).

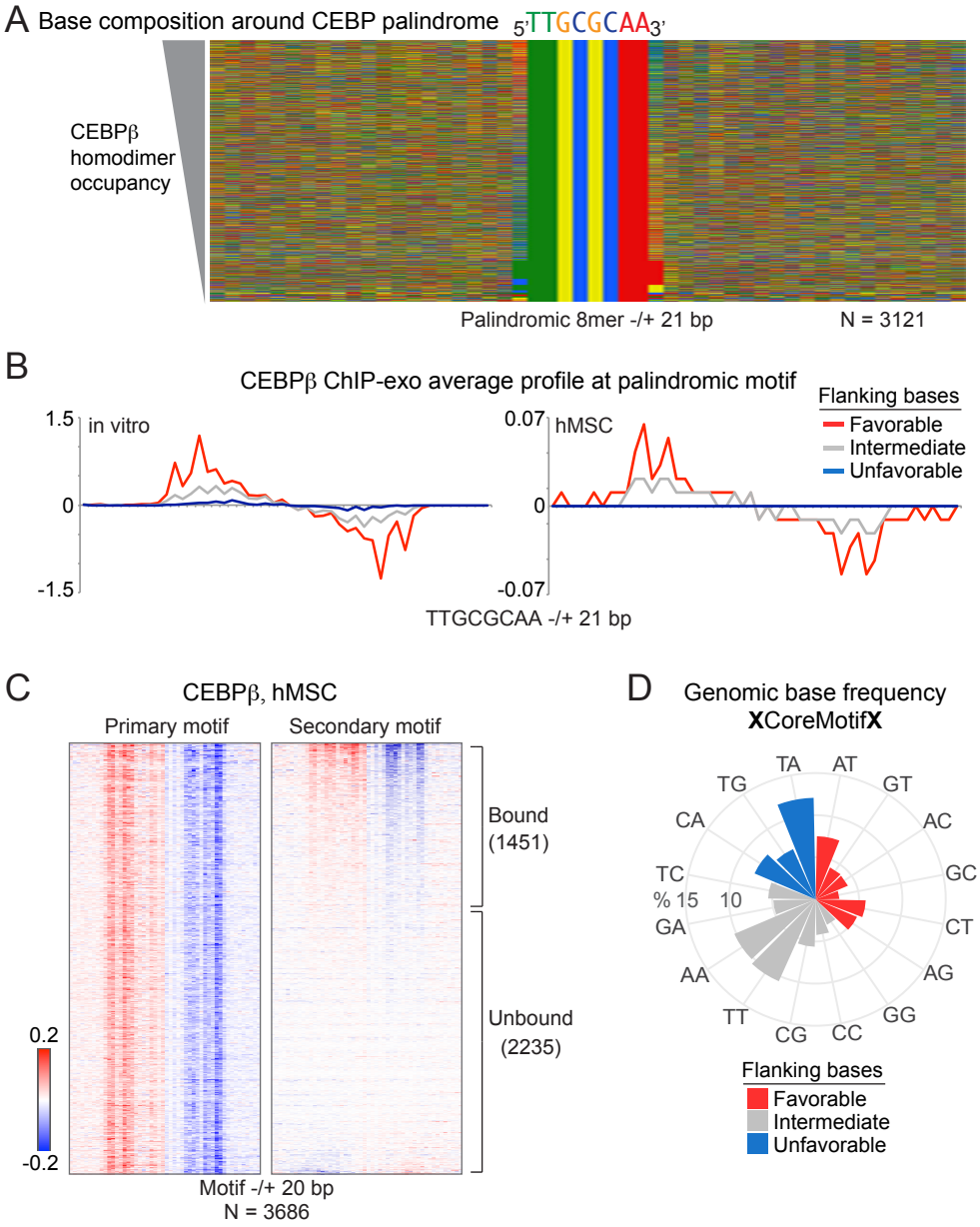

Figure 3-figure supplement 1. Flanking bases for the core CEBP 8-mer regulate TF binding. (**A**) Sequence enrichment beyond the palindromic 8-mer is restricted to a single base on either end of the motif. Color chart corresponding to the heat map of figure 3A showing the base composition of the surrounding region for all canonical palindromic 8-mers with mappable sequence. (**B**) Average profiles of the CEBPβ ChIP-exo sequence tags at the canonical palindromic motif generated from in vitro (left) and hMSC (right) studies. (**C**) Density heat maps of the CEBPβ ChIP-exo reads at the primary and secondary motifs for hMSC ChIP-seq peaks that carry more than one instance of the frequently occupied CEBP core 8-mers. Ranking of motifs based on ChIP-exo signal shows 557 sites with opposite-stranded peak pairs that define secondary bound motifs. Reads for the corresponding primary motif within each ChIP-seq peak are shown at the left. (**D**) Polar bar graph indicating the frequency of each pair of bases (5' and 3') flanking the top CEBP motif variants at all sites in the human genome.

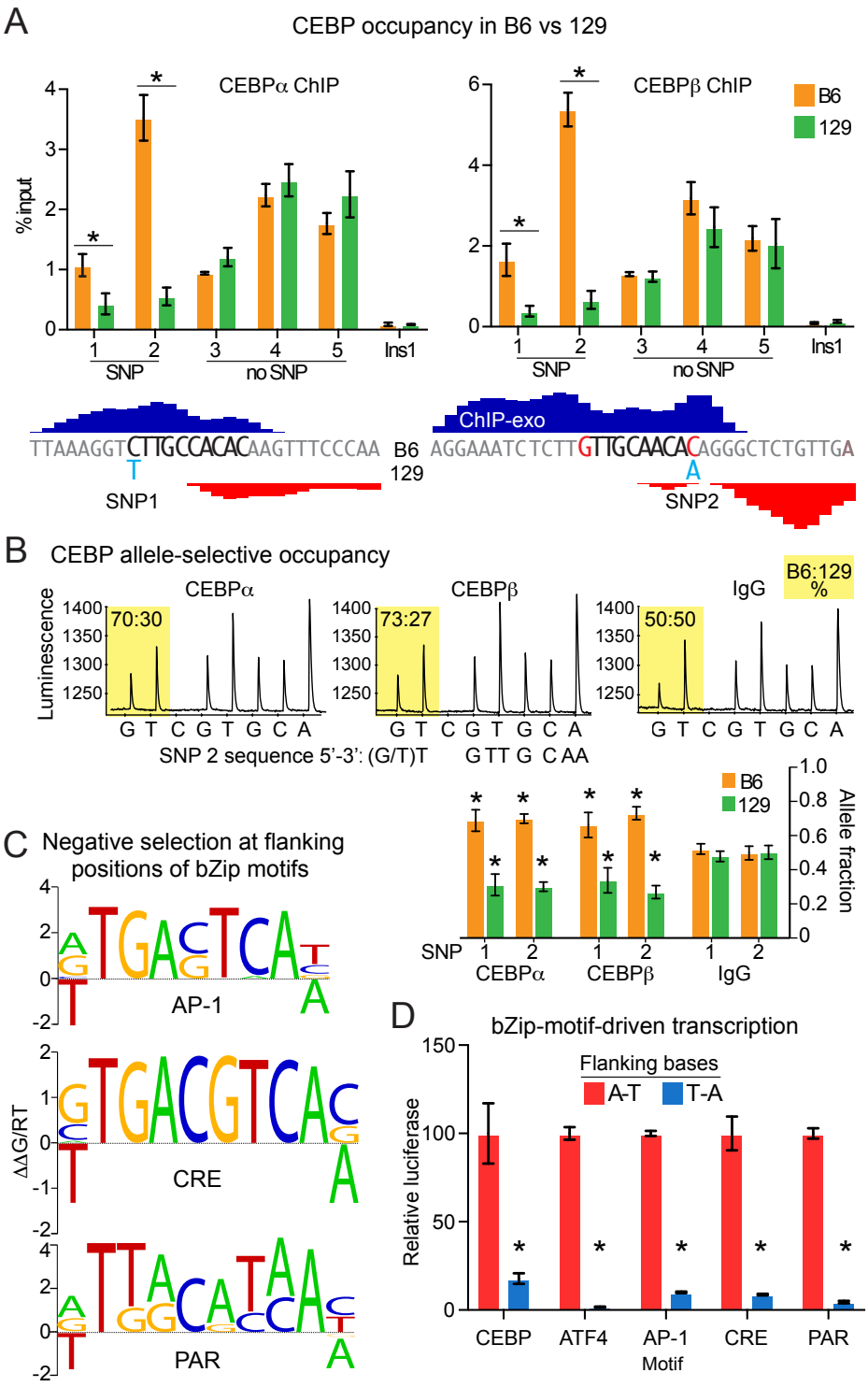

Figure 4. Bases directly flanking core bZip motifs regulate transcription. (**A**) CEBP ChIP in liver tissue isolated from B6 and 129 mice interrogating binding sites with and without SNPs in the bases flanking the core CEBP 8-mer. ChIP-exo tracks (bottom) show location of SNP relative to core 8-mer 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 right) 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**) Motif analyses of AP-1, CREB1 and NFIL3 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. (**D**) 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) 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.

Supplemental Table 1. Oligonucleotides Used in the Study. List contains the templates used to subclone motif multimers for luciferase assays and PCR primers used for ChIP.

Luciferase templates

4xAT AP-1
ctcgagtaatacgactcactatagggtactcgATGACTCATtctagatcactatactcgATGACTCATtctagatcactata
ctcgATGACTCATtctagatcactatactcgATGACTCATtctagatcactactgcagatcgtgagttatgagatct
4xTA AP-1
ctcgagtaatacgactcactatagggtactcgTTGACTCAAtctagatcactatactcgTTGACTCAAtctagatcactata
ctcgTTGACTCAAtctagatcactatactcgTTGACTCAAtctagatcactactgcagatcgtgagttatgagatct
4xAT ATF4
ctcgagtaatacgactcactatagggtactcgATTGCATCATtctagatcactatactcgATTGCATCATtctagatcacta
tactcgATTGCATCATtctagatcactatactcgATTGCATCATtctagatcactactgcagatcgtgagttatgagatct
4xTA ATF4
ctcgagtaatacgactcactatagggtactcgTTTGCATCAAtctagatcactatactcgTTTGCATCAAtctagatcacta
tactcgTTTGCATCAAtctagatcactatactcgTTTGCATCAAtctagatcactactgcagatcgtgagttatgagatct
4xAT CEBP
ctcgagtaatacgactcactatagggtactcgATTGCACAATtctagatcactatactcgATTGCACAATtctagatcacta
tactcgATTGCACAATtctagatcactatactcgATTGCACAATtctagatcactactgcagatcgtgagttatgagatct
4xTA CEBP
ctcgagtaatacgactcactatagggtactcgTTTGCACAAAtctagatcactatactcgTTTGCACAAAtctagatcacta
tactcgTTTGCACAAAtctagatcactatactcgTTTGCACAAAtctagatcactactgcagatcgtgagttatgagatct
4xAT CRE
ctcgagtaatacgactcactatagggtactcgATGACGTCATtctagatcactatactcgATGACGTCATtctagatcacta
tactcgATGACGTCATtctagatcactatactcgATGACGTCATtctagatcactactgcagatcgtgagttatgagatct
4xTA CRE
ctcgagtaatacgactcactatagggtactcgTTGACGTCAAtctagatcactatactcgTTGACGTCAAtctagatcacta
tactcgTTGACGTCAAtctagatcactatactcgTTGACGTCAAtctagatcactactgcagatcgtgagttatgagatct
4xAT PAR
ctcgagtaatacgactcactatagggtactcgATTACGTAATtctagatcactatactcgATTACGTAATtctagatcacta
tactcgATTACGTAATtctagatcactatactcgATTACGTAATtctagatcactactgcagatcgtgagttatgagatct
4xTA PAR
ctcgagtaatacgactcactatagggtactcgTTTACGTAAAtctagatcactatactcgTTTACGTAAAtctagatcacta
tactcgTTTACGTAAAtctagatcactatactcgTTTACGTAAAtctagatcactactgcagatcgtgagttatgagatct

PCR primers

4xmer FOR	agcactcgagtaatacgactcactataggg
4xmer REV	atatagatctcataactcacgatctgcag
INS1 F	ggacccacaagtggaacaac
INS1 R	gtgcagcactgatccacaat
CEBP ChIP site 1 SNP F	catcatcatcaacaacaacaaca
CEBP ChIP site 1 SNP R	gcagagagcaactttgtgga
CEBP ChIP site 2 SNP F	gagtggtgtttccagaggcta
CEBP ChIP site 2 SNP R	tgagccatctctccagcttt
CEBP ChIP site 3 no SNP F	ctctccctctttgtcgcatt
CEBP ChIP site 3 no SNP R	tccgacattttgcagacatc
CEBP ChIP site 4 no SNP F	cccagcttgctcaactaagg
CEBP ChIP site 4 no SNP R	accacatccatggtggagag
CEBP ChIP site 5 no SNP F	tcttccagggaaatgctgag
CEBP ChIP site 5 no SNP R	aggtgattgcaggagattgg
CEBP SNP 1 pyro F-bio	tggcccagaaatatggcttagag
CEBP SNP 2 pyro F-bio	gctgcaggcggtcaaatg
pyro seq SNP1	gaaacttgtgtggcaa
pyro seq SNP2	ggtcaacagagccct