Combinatorial bZIP dimers define complex DNA-binding specificity landscapes

José A. Rodríguez-Martínez[1†], Aaron W. Reinke[2†], Devesh Bhimsaria[1, 4†], Amy E. Keating[2,3], Aseem Z. Ansari[1,5 *]

[1] Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706

[2] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139

[3] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139

[4] Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI 53706

[5] The Genome Center of Wisconsin, University of Wisconsin-Madison, Madison, WI 53706

[†] These authors contributed equally to this study

Correspondence should be addressed to:

Professor Aseem Z. Ansari
University of Wisconsin–Madison
433 Babcock Drive, Room 315C
Madison, WI 53706

Email: **azansari@wisc.edu**
Telephone: (608) 265-4690

Abstract

How transcription factor dimerization impacts DNA binding specificity is poorly understood. Guided by protein dimerization properties, we examined DNA binding specificities of 270 human bZIP pairs. DNA interactomes of 80 heterodimers and 22 homodimers revealed that 72% of heterodimer motifs correspond to conjoined half-sites preferred by partnering monomers. Remarkably, the remaining motifs are composed of variably-spaced half-sites (12%) or “emergent” sites (16%) that cannot be readily inferred from half-site preferences of partnering monomers. These binding sites were biochemically validated by EMSA-FRET analysis and validated in vivo by ChIP-seq data from human cell lines. Focusing on ATF3, we observed distinct cognate site preferences conferred by different bZIP partners, and demonstrated that genome-wide binding of ATF3 is best explained by considering many dimers in which it participates. Importantly, our compendium of bZIP-DNA interactomes predicted bZIP binding to 156 disease associated SNPs, of which only 20 were previously annotated with known bZIP motifs.

Introduction

Multiple sequence-specific transcription factors (TFs) converge at enhancers and promoters to control the expression of genes (Ptashne and Gann, 2002). Such TF assemblages permit integration of multiple cellular signals to regulate targeted gene networks (Ciofani et al., 2012; Kittler et al., 2013; Xie et al., 2013). The combinatorial use of a limited number of TFs provides the means to finely control the complex cellular transcriptome. The ability of a given TF to interact with different partners expands the DNA targeting repertoire while enhancing specificity by focusing combinations of factors to a more specific set of regulatory sites across the genome. Different partners also alter the regulatory potential of a given TF, at times completely altering its regulatory output such that the factor switches from an activator to a repressor of transcription depending on the binding partners with which it associates (Ptashne and Gann, 2002).

The bZIP class of human TFs is well suited to play a role in signal integration and combinatorial transcriptional control (Lamb and McKnight, 1991; Miller, 2009; Tsukada et al., 2011). bZIPs bind DNA as either homo- or heterodimers; the discovery in 1988 that JUN and FOS could bind to DNA as a heterodimer immediately suggested the potential for combinatorial regulation by this family (Franza et al., 1988; Lamb and McKnight, 1991). Interestingly, the human bZIP networks display greater ability to form heterodimers compared to simpler eukaryotes, suggesting that more complex combinatorial regulation may contribute to organismal complexity (Reinke et al., 2013). bZIP proteins also interact with other classes of TFs to stabilize higher order oligomers

at enhancers (Jain et al., 1992; Murphy et al., 2013; Thanos and Maniatis, 1995). Their role in nucleating such multi-factor complexes is supported by the observation that certain bZIP dimers such as AP-1 (FOS•JUN) and CEBPA can function as “pioneer” factors that bind inaccessible chromatin and enable the assembly of other TFs at regulatory sites (Biddie et al., 2011; Collins et al., 2014). Notably, the choice of dimerizing partner not only impacts DNA recognition properties but can also influence regulatory function of a given bZIP. For example, ATF3 homodimer acts as a repressor whereas ATF3•JUN activates transcription (Hsu et al., 1992). As a class, bZIPs regulate diverse biological phenomena ranging from response to stress at the cellular level, organ development at the tissue level and viral defense, circadian patterning, memory formation, and ageing at an organismal level (Costa et al., 2003; Herdegen and Waetzig, 2001; Jung and Kwak, 2010; Male et al., 2012). Given their central role in various processes, mutations in bZIP proteins are implicated in the etiology of diseases ranging from cancer and diabetes to neuronal malfunction, developmental defects and behavioral dysfunction (Lopez-Bergami et al., 2010; Tsukada et al., 2011).

Fifty-three bZIPs encoded by the human genome can be grouped into 21 families, and as homodimers they are known to bind at least 6 distinct classes of DNA motifs, including sites labeled as TRE, CRE, CRE-L, CAAT, PAR, and MARE (Figure 1) (Deppmann et al., 2006; Jolma et al., 2013). In 1991, Hai et al. showed that some heterodimers have DNA-binding specificities that are distinct from those of each partnering bZIP. For example, JUN•ATF2 heterodimer binds to a cognate site in the ENK2 promoter that is not bound by either JUN•JUN or ATF2•ATF2 homodimers (Hai and Curran, 1991). However, the past 20 years have provided only a handful of

additional examples of how bZIP heterodimerization influences DNA binding specificity (Cohen et al., 2015; Jolma et al., 2015; Vinson et al., 1993; Yamamoto et al., 2006). Central questions about bZIP transcription factors remain unanswered: What is the influence of protein dimerization on DNA binding? Does DNA stabilize dimer formation? Which protein dimers can bind DNA? Which sequences do they bind? And, how do bZIP binding sites contribute to cellular function and the etiology of various diseases?

Resolving these issues requires systematic examination of the DNA binding specificities of bZIP homo and heterodimers. Fifty-three human bZIP proteins can potentially form as many as 1,431 distinct dimers. Quantitative experiments using fluorescence resonance energy transfer (FRET) in solution indicated that ~30% of all possible bZIP dimers form in the absence of DNA. Most bZIPs can form dimers with different partners, potentially greatly expanding the repertoire of cognate sites that might be targeted by different heterodimers (Reinke et al., 2013). We used this protein-protein interaction dataset to prioritize 270 bZIP dimers for bZIP-DNA interactome studies and to apply FRET-based methods to distinguish DNA bound heterodimers from homodimers.

Insights that emerged from our compendium of bZIP-DNA interactomes include: (i) identification of new bZIP cognate sites, (ii) evidence for three classes of heterodimer binding sites (conjoined half-sites, variably-spaced half-sites, and unpredicted emergent cognate sites), (iii) ability of individual bZIP heterodimers to target a range of binding sites, (iv) evidence for varying heterodimer selectivity between distinct sequences currently classified as a single consensus motif, (v) improved ability to account for in

vivo genome-wide occupancy of heterodimers, and (vi) identification of bZIP cognate sites at 156 SNPs linked to human diseases and quantitative traits. DNA sequence preferences of bZIP heterodimers reported here serve as a valuable resource for many purposes including, but not limited to, evaluating potential bZIP dimer binding at genomic binding sites, providing hypotheses about mechanisms underlying the etiology of disease-linked SNPs, and predicting binding specificities of heterodimeric bZIPs from other species.

Results

Comprehensive bZIP-DNA interactomes

To elucidate DNA sequence-recognition properties of TFs that form obligate dimers, we examined 270 pairs of purified human bZIP proteins. These pairs were composed from 36 bZIP proteins representing 21 bZIP families and encompassing the diversity of all 53 bZIPs encoded in the human genome (Supplementary file 1A). Given the 666 potential dimeric pairs that can be formed with 36 bZIPs, we used biophysically measured protein-protein interactions to prioritize the dimers that were examined (Reinke et al., 2013). We selected 126 pairs (97 hetero- and 29 homodimer) that form stable dimers with protein-protein interaction (PPI) dissociation constants (K_d) less than 1 μM at 21 °C in the absence of DNA. In addition, we tested 144 (137 hetero- and 7 homo-) dimer combinations that do not stably associate in solution in the absence of DNA (K_d > 1 μM at 21 °C). For most TFs, including the bZIP class, the DNA specificity of the isolated DNA binding domain (DBD) is typically indistinguishable from the full-length factor (Jolma et al., 2013). Therefore, we focused our efforts on the bZIP domain, which comprises the basic region that binds DNA and the leucine zipper dimerization module that forms a coiled-coil. Recombinant proteins, overexpressed in bacteria, were purified to homogeneity. Two versions of each protein were made, one conjugated to biotin at the carboxyl terminal and the other without. This set of highly purified DNA binding proteins enabled examination of the innate DNA-binding sequence specificity of the set of 36 representative human bZIPs. Individual bZIP partners were mixed in 3:1 molar ratios with the biotinylated partner at the lower concentration; affinity purification of the

protein-DNA complexes using the less abundant biotinylated partner enriched for heterodimers. To additionally favor examination of the heterodimer, whenever possible, the interaction partner with weaker homodimer was biotinylated and used for isolating protein-DNA complexes. Each protein dimer is denoted by a dot between each monomer –for example, JUN•ATF3. The DNA binding sites are indicated either by a specific sequence or their classical designations, such as CRE or CAAT, or by half-sites connected by a hyphen, such as CRE-CAAT.

DNA-binding specificity of the pairs was queried using systematic evolution of ligands by exponential enrichment coupled to deep sequencing (SELEX-seq) (Figure 1 – figure supplement 1) (Jolma et al., 2010; Slattery et al., 2011; Tietjen et al., 2011; Zhao et al., 2009; Zykovich et al., 2009). In our cognate site identification (CSI) effort using SELEX-seq, a DNA library spanning the entire sequence space of a 20-mer (10^{12} different sequence permutations) was independently incubated with each of the 270 different bZIP pairs (234 hetero- and 36 homodimers), and protein-bound DNA sequences were enriched, amplified, and re-probed for an additional two cycles to further enrich cognate sites over non-cognate sites that comprise the majority of the starting library. The starting DNA library as well as selectively enriched sequences were barcoded and sequenced using massively parallel DNA sequencing methods. The CSI intensity, corresponding to the z-score for the enrichment of a sequence, was computed for each 10-mer as described in the methods. Repeated experiments demonstrated an average Pearson's correlation r = 0.8 ± 0.1 between CSI intensities from replicates (Figure 1 – figure supplement 1). The CSI intensity (z-score) correlates with the binding affinity for a

particular sequence (Figure 1 – figure supplement 2) (Carlson et al., 2010; Puckett et al., 2007; Tietjen et al., 2011). In three cases, reciprocal biotinylation of each partner was performed to ensure that the choice of partner did not skew the results (Pearson's correlation for comparing experiments was r = 0.89 – 0.98; Figure 1 – figure supplement 1C).

Among the 270 pairs tested were 12 homodimers that had been previously examined by other groups (Jolma et al., 2013). Overall, we found excellent agreement between the cognate sites identified using highly purified bZIP modules in our study versus full-length proteins in unpurified cell lysates in other studies, with only a few inconsistent examples that can be seen in Figure 1 – figure supplement 3A (e.g. MAFB, NFE2, ATF4). Interestingly, we found that the previously reported DNA specificity of ATF4 has a higher correlation with the specificity of ATF4•CEBPG heterodimer identified in this report than with the specificity of the ATF4 homodimer (Figure 1 – figure supplement 3B), suggesting that CEBPG possibly formed a complex with ATF4 in the cell lysates used in the prior study (Jolma et al., 2013). This observation highlights the advantage of using highly purified proteins over cell lysates and validates our focus on the bZIP domain to capture the innate specificity of this class of transcription factors that bind DNA as obligate dimers.

Overall, 30 out of the 36 bZIP proteins tested in this study enriched specific DNA sequences as part of at least one dimer. bZIPs that are not known to bind DNA as homodimers did not yield cognate sites in our studies (e.g., JUNB and FOS) (Deng and

194 Karin, 1993; Hai and Curran, 1991). 73 of 126 (58%) of bZIP pairs that dimerize in the
195 absence of DNA yielded specific cognate sites (Figure 1A). Surprisingly, 29 of the 144
196 (20%) bZIP pairs that do not stably associate in the absence of DNA (K_d > 1 μM at 21
197 °C) yielded evidence of sequence-specific binding to DNA, indicating that protein-
198 protein interactions were stabilized by binding to specific DNA sites (Figure 1A;
199 Supplementary file 1C). This finding has important implications, given that the majority
200 of the potential bZIP protein-protein interaction space consists of protein pairs that do
201 not associate strongly in the absence of DNA (Reinke et al., 2013).

202 ***Conjoined, Variably-spaced, and Emergent cognate sites bound by heterodimers***

203

204 For 184 bZIP-DNA interactomes that showed evidence for an enriched motif, we
205 computationally parsed and retained datasets that could be attributed with high
206 confidence to 80 heterodimers and 22 homodimers (Materials and methods). We
207 assigned a specificity profile to a heterodimer when it bound sequences that were
208 significantly different (t-test $p < 0.05$) from the sequences preferred by the homodimer of
209 the biotinylated bZIP (e.g., ATF4 vs. ATF4•CEBPA, r = 0.1; Figure 1 – figure
210 supplement 4) or when the biotinylated bZIP did not bind specific DNA sequences as a
211 homodimer (e.g., FOS and JUNB).

212

213 Of the 22 homodimers, comprehensive DNA binding specificity is reported for the first
214 time for human ATF2, ATF3, ATF6, ATF6B, CEBPA, CREB1, FOSL1, JUN, MAFB, and
215 NFE2L1. Hierarchical clustering of the 102 bZIP-DNA interactomes readily identified six
216 previously known classes of bZIP binding sites (TRE, CAAT, PAR, MARE, CRE, CRE-
217 L) (Figure 1B-C). Notably, several bZIP homodimers (ATF6, ATF6B, CREB3L1 and
218 JUN) enriched more than one motif (Supplementary file 2). Examining the cognate sites
219 bound by heterodimers highlighted the ability of some heterodimers to bind homodimer
220 motifs as well as a range of other heterodimer-specific motifs. Such binding to multiple
221 motifs is reminiscent of previous studies that reported bZIP dimers binding to different
222 sites with different affinities (Badis et al., 2009; Kim and Struhl, 1995; Konig and
223 Richmond, 1993). Interestingly, several heterodimers that bind classic bZIP homodimer
224 motifs such as TRE, CRE-L or CRE displayed clear differences in their preference for a

subset of sequences categorized under a single consensus motif (for example, compare motifs **9** and **10** in Figure 1 with the CRE-L site). This was also true for different homodimers (e.g. compare CRE-L binding profiles for ATF6, CREB3, CREB3L1, and XBP1 in Supplementary file 2). Thus, the binding data reported here reveals a sequence sub-structure to classic consensus motifs. Moreover, the sub-structure highlights differences in DNA-binding specificity between closely related dimers.

Three classes of bZIP heterodimer motifs were identified and are illustrated in Figure 1C: "*Conjoined*" sites for which half-sites preferred by each contributing monomer are juxtaposed (such as the CRE-CAAT site represented by motif **1**, or the MARE-CRE site of motif **7**), "*Variably-spaced*" sites for which half-sites overlap (as is the case in motifs **2** and **4**), and "*Emergent*" sites for which binding preferences could not have been readily predicted based on the half-site preferences of each partner (motifs **3**, **5**, **8**, **9** and **10**). In other words, an *emergent* site arises as a consequence of heterodimer formation and is not simply comprised of the conjoined or variably-spaced half-sites preferred by each monomer. An elegant study of Hox-Exd heterodimers identified "latent" sites that were preferred by different Hox factors when they bound DNA in conjunction with Exd (Slattery et al., 2011). Preferences for different sequences at the interface of half-sites or sequences flanking the half sites were observed for different classes of Hox-Exd heterodimers. In our studies, we observed a change in half-site preference of certain bZIPs when they bound DNA as heterodimers. In some instances, homodimers bound with low affinity to sites that emerged as high affinity sites in the context of a heterodimer whereas in other cases entirely new site preferences emerged. We

classified such newly acquired binding preferences as *emergent* sites because they are not readily inferred from the binding preferences of homodimers.

While a large fraction of heterodimers bind conjoined sites, it was surprising to find that closely related heterodimers such as FOS•CEBPG and FOS•CEBPE preferred different arrangements of half-sites, with the former heterodimer preferring the 8 bp conjoined CRE-CAAT site (motif **1** $^{5'}$TGACGCAA$^{3'}$) and the latter preferring the 7 bp variably-spaced TRE-CAAT site (motif **4** $^{5'}$TGAGCAA$^{3'}$). Figure 1 – figure supplement 5 highlights the unexpectedly poor correlation between the binding preferences of these two heterodimers and between the binding preferences of FOSL1•CEBPG and FOSL1•CEBPE. Similarly, other heterodimers bound both conjoined and variably spaced motifs (see JUNB•ATF3, and MAFB•ATF5 in Supplementary file 2), however the preference for one arrangement over the other was not amenable to predictions based on the binding preferences of each contributing partner of the heterodimer.

Emergent sites pose a particular challenge for current models of DNA binding site predictions that are based on protein homology (Weirauch et al., 2014). Emergent cognate sites for heterodimers can be subdivided into two categories: (i) “gain-of-specificity” motifs that display a change in half-site preferences for a bZIP or (ii) motifs that display a “loss-of-specificity” for one half-site. An example of the first category includes a switch in the half-site preferences of BATF family members, from a CRE half-site ($^{5'}$ TGAC$^{3'}$) that is preferred in homodimers to a CRE-L ($^{5'}$ CCAC$^{3'}$) half-site that is preferred by many BATF-containing heterodimers (compare motifs **3** and **8-10,** and see

examples in Supplementary file 2 such as BATF2•ATF3, BATF2•JUN, BATF3•ATF3, BATF3•ATF4). An example of the second category is DDIT3•CEBPG binding to $^{5'}$ ATTGCA$^{3'}$ (motif **5**) (Ubeda et al., 1996), with heterodimers displaying no apparent requirement for one half-site. Overall, for the 80 bZIP heterodimers with binding motifs reported here, 72% of the motifs can be classified as conjoined, 16% as emergent, and 12% as variably-spaced. Nine out of the 80 heterodimers (11%) enriched two motifs (Supplementary file 2). For example, BATF•CEBPG enriched both CRE-CAAT and CRE-L-CAAT motifs.

Specificity and Energy Landscapes (SELs) reveal the entire spectrum of cognate sites bound by heterodimers

To examine the full specificity spectrum of individual bZIP dimers, we displayed DNA binding data as Specificity and Energy Landscapes (SELs) (Carlson et al., 2010; Tietjen et al., 2011). In an SEL, all possible sequences of a given length are arranged within concentric circles based on their homology to a seed motif. The seed motif is often derived from position weight matrices (PWMs) of the most enriched sequences (Figure 2A). The innermost circle contains all sequences that have an exact sequence match to the seed motif (0-mismatch). As each enriched sequence placed in this ring is an exact match to the seed motif, the source of varying CSI intensity (z-score) is the contribution of the sequences flanking the seed motif. The 1-mismatch ring contains all sequences that differ from the seed motif at any one position, or a Hamming distance of one. The subsequent rings, going outwards, display sequences with increasing number of mismatches to the seed motif. The height and color of each point represents the CSI intensity for the corresponding sequence. As noted above, CSI intensity correlates with binding affinity where measured (Figure 1 – figure supplement 2) (Carlson et al., 2010; Hauschild et al., 2009; Puckett et al., 2007; Tietjen et al., 2011; Warren et al., 2006). Although there are far more low-affinity sequences than enriched sequences (as depicted by the illustrative histogram in Figure 2A), the moderate-to-low affinity sites (low CSI intensity) are often overlooked by motif searching algorithms. Such sequences readily emerge in SEL display of the entire binding data (Carlson et al., 2010; Tietjen et al., 2011). In Figure 2A, we illustrate how SELs are built and we note that an SEL can

be constructed using any sequence as a seed motif. The choice of a different seed motif simply alters the placement of the sequences on the landscape without changing the underlying binding preferences of a protein for a given sequence.

SEL plots of 102 bZIP homo- and heterodimers reveal that the impact of flanking sequence context and the range of different cognate sites bound by most bZIPs is far richer than might be inferred from motifs represented as PWMs (Supplementary file 2). In Figure 2B, the SELs of JUN•ATF3 and ATF4•CEBPG illustrate the broader insights that emerged from examining specificity profiles of these two heterodimers. JUN•ATF3 binds a CRE site composed of conjoined half-sites for JUN and ATF3. Visualizing the entire JUN•ATF3-DNA interactome via an SEL shows that the binding of JUN•ATF3 heterodimer to CRE is significantly influenced by the sequence context that flanks the motif (see affinity variations in the 0-mismatch ring). Additionally, the 3-mismatch ring of the SEL identifies several high-intensity peaks corresponding to additional cognate sites. As indicated, one of these is a variably spaced site, and another is an emergent site $^{5'}$TGACGCAT$^{3'}$. Thus, the SEL highlights that this single heterodimer binds multiple classes of cognate sites. On the other hand, the SEL for ATF4•CEBPG shows that the seed motif $^{5'}$ATGCGCAAT$^{3'}$ bound by this heterodimer is relatively insensitive to context effects (0-mismatch ring). The 1-mismatch ring indicates that both half-sites are not equally tolerant of mismatches, with mismatches in the $^{5'}$TGA$^{3'}$ core of the CRE site dramatically reducing binding whereas the $^{5'}$CAA$^{3'}$ site is tolerant of deviations at the first position of the half-site but sensitive to deviations in the $^{5'}$AA$^{3'}$ positions. Similar

insights can be obtained from the SELs for each of the102 bZIP dimers that are reported in Supplementary file 2.

Our compendium of SEL plots greatly extend previous reports that bZIP dimers bind a range of sequences with different degrees of affinity (Badis et al., 2009; Kim and Struhl, 1995; Konig and Richmond, 1993). To examine whether the set of sequences that are pointed out in SELs of JUN•ATF3 and ATF4•CEBPG from Figure 2B are bound by homodimers or any bZIP in our compendium, we displayed the relative preferences of each dimer for this set of binding sites in a heatmap (Figure 2C). Each column displays the relative preference of each of the 102 bZIP dimers for different sequences, including half-sites of all six classical homodimer motifs. An examination of row 3, which displays preferences of all 102 dimers for the emergent site $^{5'}$TGACGCAT$^{3'}$, indicates that this site is highly preferred by JUN•ATF3 and to some extent by JUN•CEBPG but not by homodimers formed by ATF3, CEBPG or JUN (denoted by asterisks). While not as exclusive as the JUN•ATF3 emergent site, the conjoined CRE-CAAT site is primarily targeted by heterodimers formed by CEBP family of bZIPs. The heatmap does indicate that this heterodimer-preferred site permits low affinity binding by the CEBPG homodimer. Interestingly, data in row 8 reveals that substituting $^{5'}$CAAT$^{3'}$ with $^{5'}$**T**AAT$^{3'}$ in the CRE-CAAT site perturbs binding by CEBP heterodimers in a non-uniform manner, unmasking hidden differential sequence preferences of related heterodimers that are opaque to current models that use protein homology to predict cognate site preferences. The C-to-T substitution also expands the repertoire of bZIPs that bind this mutated site. DBP, HLF and NFIL3 as homo- and hetero- dimers display an

unmistakable affinity for this modified CRE-CAAT site that recreates the PAR half-site that is a target of this set of bZIPs. On the other hand, a different substitution at the same position ($^{5'}$**G**AAT$^{3'}$) dramatically reduces the binding of all bZIPs to this version of the CRE-CAAT site (row 9). Furthermore, the importance of the sequences flanking a binding site (rows 4 and 5) or the contribution of each half-site to the binding of bZIPs (rows 8-10) is also made evident by the heatmap. In essence, SEL plots alongside comparative heatmaps of affinities of proteins for a range of cognate sites bring a new appreciation for diversity of DNA sequences that can be targeted by a given factor.

EMSA-FRET analysis to validate heterodimer binding to different cognate sites

To validate the ability of heterodimers to bind cognate sites identified by CSI analysis, we used an electrophoretic mobility shift assay (EMSA) in which a FRET signal distinguished homo- vs. heterodimers in protein-DNA complexes (Figure 3A; Materials and methods) (Reinke et al., 2013). We first used EMSA-FRET to assay bZIP dimers formed by mixing fluorescein and TAMRA labeled versions of 16 proteins drawn from different bZIP families. For 15 homodimers for which we could detect binding to DNA, the mixed-dye homodimer could be easily distinguished from both of the single-dye homodimers, as shown for CEBPG and ATF3 homodimers binding to CAAT and CRE sites, respectively (Figure 3A). This assay was then used to demonstrate that the ATF3•CEBPG heterodimer bound the conjoined CRE-CAAT site better than either parental homodimer (Figure 3A). Furthermore, swapping fluorophores did not alter the binding properties of the resulting heterodimer (last panel of Figure 3A). DNA fluorescence coincides with the protein FRET signals, confirming that protein-DNA complexes were being observed in the EMSA gels (Figure 3 – figure supplement 1A).

We used this EMSA-FRET assay to quantify the DNA binding of 83 bZIP homodimers and heterodimers comprised of 16 proteins. Each heterodimer was systematically examined with DNA sites that were constructed by conjoining the preferred half-site(s) for each bZIP. Figure 3B shows EMSA-FRET data for six heterodimers and corresponding homodimers binding to heterodimer-specific sites (three conjoined sites and three emergent sites). For these sites the CSI intensity for the heterodimers is

higher than the scores for either of the two contributing homodimers. The EMSA-FRET data demonstrate clearly that neither the JUN nor the ATF3 homodimers associate with the emergent site identified for JUN•ATF3 (Figure 3B and Figure 3 – figure supplement 2). Similarly, emergent sites identified for ATF4•CEBPA and ATF4•JUN, and several conjoined sites such as TRE-CAAT for ATF3•CEBPA, CRE-L-CAAT for BATF3•CEBPA, and CRE-CRE-L for BATF3•JUN, were validated by EMSA-FRET as *bona fide* heterodimer-specific cognate sites that show weaker binding, or no binding, by the contributing homodimers (Figure 3B). EMSA-FRET data also validate the ability of BATF3 to bind emergent CRE-L half-sites as a heterodimer (in addition to the CRE site preferred by the homodimer). The complete EMSA-FRET data are presented in a more compact format in Figure 3 – figure supplement 1.

A striking result of our CSI analysis is that conjoined half-sites form a substantive fraction (~70%) of the cognate sites bound by heterodimers. To determine how frequently DNA half-sites derived from homodimer binding data, when presented as conjoined sites, would bind the corresponding heterodimers, we tested DNA binding by EMSA-FRET for stably interacting bZIP heterodimers (PPI: K_d < 1 μM at 21 °C), and the corresponding homodimers (Figure 3C). Consistent with CSI analysis, 52 out of 56 bZIP pairs that form stable heterodimers bound the DNA site made by conjoining the half-site preferred by each monomer. Specific binding to conjoined sites was also detected for 6 out of 27 (22%) pairs that do not stably associate in the absence of DNA (PPI: K_d > 1 μM at 21 °C). This fraction is similar to the 20% of bZIP pairs (29 out of 144) that

showed sequence-specific DNA binding in SELEX-seq experiments despite their apparent inability to dimerize in the absence of DNA.

ATF3: a case study of the influence of interacting partners on heterodimer cognate site preferences

Activating Transcription Factor 3 (ATF3) is a member of the CREB/ATF family. Initially identified as a suppressor of inflammation and the adaptive immune response in resting cells, ATF3 is now associated with numerous diseases including a variety of aggressive and widely occurring cancers (Tanaka et al., 2011; Thompson et al., 2009; Yin et al., 2008). ATF3 is able to interact with a large variety of TFs to function as a regulatory hub of cellular adaptive response (Gilchrist et al., 2006; Hai et al., 1999; Hai et al., 2010). As a homodimer, ATF binds to CRE sites and represses a wide array of genes (Hai et al., 1999; Hai et al., 2010). However, as a heterodimer with JUN or JUND, ATF3 activates transcription of targeted genes (Chu et al., 1994; Filén et al., 2010; Hsu et al., 1992).

To test the hypothesis that heterodimerization with other bZIPs might alter DNA binding specificity and possibly genomic targets, we analyzed SELEX-seq for 20 different ATF3 heterodimers spanning the full range of PPI affinities. DNA-binding specificities could be assigned with high confidence to 9 heterodimers that displayed a range of DNA sequence preferences, including affinity for the CRE site preferred by the homodimer (Figure 4A and B). Importantly, distinct DNA binding preferences among ATF3•CEBP and ATF3•BATF heterodimers and their corresponding homodimers were detected. The motifs enriched by the ATF3 homo- and heterodimers can be described in five broad categories: CRE, TRE, CRE-CAAT, CRE-L, and the emergent $^{5'}$TGACGCAT$^{3'}$ site (Figure 4B). Scatter plots illustrate instances where the CSI intensities of ATF3

heterodimers differ markedly from those of the parent homodimers (Figure 4C and 4D). For example, as evident from high CSI intensities, CRE-CAAT sites (red) are preferably bound by ATF3•CEBPG as compared to ATF3 or CEBPG (Figure 4C, top panel). Similarly, scores for TRE (green) and $^{5'}$TGACGCA$^{3'}$ (black) are higher for JUN•ATF3 than for JUN or ATF3 (Figure 4C, middle panel). BATF3•ATF3 (Figure 4C, bottom panel) and BATF2•ATF3 (Figure 4D, top panel) enrich CRE-L sites (blue), further supporting that CRE-L is an emergent site for BATF family heterodimers (also with JUN in Figure 3C). Figure 4D further highlights the differences between CRE and TRE binding by ATF3 in its homo versus heterodimer state. An important and recurring observation is that several ATF3 heterodimers (BATF3•ATF3, JUN•ATF3, and JUNB•ATF3) can associate with the CRE site that is bound by the ATF3 homodimer (Figure 4).

Heterodimer sites enriched in SELEX-seq map to occupied genomic loci in vivo

To determine the extent to which cognate sites identified by SELEX-seq can explain genome-wide occupancy in cells, we examined ChIP-seq data for ATF3 in four different human cell lines. H1 human embryonic stem cells, HEPG2 liver-derived hepatocellular carcinoma cells, and K562 erythroblastoma cells have been examined comprehensively (ENCODE, 2011). The fourth cell line, GBM1 from Glioblastoma multiforme, is an aggressive brain cancer wherein ATF3 is a tumor suppressor and its loss of function is indicative of high-grade cancer and poor prognosis (Gargiulo et al., 2013). As a first step we identified ATF3 ChIP-seq peaks and examined the overlap between the genomic sites occupied by ATF3 in all four lines. Only a small number of sites (119) were common between the four cell lines, although the number increased to 1602 genomic loci if only the ENCODE cells lines (H1, K562, and HEPG2) were examined (Figure 5A). This is a minor fraction of the over 10,000 peaks identified in K562 and about a third of the 4808 ATF3-bound sites in H1 cells.

ATF3 ChIP-seq peaks likely include both homodimer and heterodimer bound regions. To assess how well the in vitro discovered cognate sites explain bound sites in a cellular context, we used area under the curve-receiver operating characteristic (AUC-ROC) values, plotting the true-positive rate (TPR) versus false-positive rate (FPR) for peak detection (Materials and methods). ATF3 homodimer sites spanning the entire spectrum of CSI intensities (z-scores) yielded 0.67 – 0.77 AUC values (Supplementary file 1E and Figure 5 – figure supplement 1). Using the AUC-ROC approach, we

examined the ability of CSI profiles of nine different ATF3 heterodimers as well as the ATF3 homodimer to identify ATF3 ChIP-seq peaks that might represent heterodimer-bound regions. We used published RNA-seq datasets to verify the expression of the bZIP genes used for the ATF3 heterodimer analysis (Supplementary file 1F) (ENCODE, 2011; Gargiulo et al., 2013). Each of the 10 CSI datasets captures a large but varying fraction of the ATF3 peaks and, intriguingly, these data reveal that different ATF3 heterodimers perform better in different cell lines (Supplementary file 1E). For example, JUN•ATF3 gives 0.85 AUC in the Glioblastoma line, whereas BATF3•ATF3 better explains the ChIP-seq peaks in H1 and HepG2 cells with AUC of 0.69 and 0.70, respectively. While AUC-ROC curves are not robust to subtle changes, the differences we observe may reflect underlying cell line specific differences in the abundance and regulatory roles of different ATF3 heterodimers. The underlying epigenetic landscapes would further exacerbate these differences. Nevertheless, when considered together, the ATF3 homodimer combined with 9 different heterodimers can account for a much larger fraction of ChIP-seq peaks than can the homodimer alone. For example, at an FPR-cutoff of 0.10, in the Glioblastoma cell line, the ATF3 homodimer classified just 39% of the ATF3 ChIP-seq peaks as positive, whereas 85% of the peaks are classified positive by at least one of the ATF3-containing dimers at FPR 0.10. Similar analysis for other cell lines and at different FPR cutoffs is reported in Supplementary file 1G.

Given the cell-type specific differences in genomic sites occupied by ATF3, we scored the ATF3-bound loci for each cell line using the CSI data for 10 ATF3 dimers. Peaks were then clustered based on the FPR-cutoffs for each bound region (Figure 5B-C and

Materials and methods). All four cell lines show clear clusters of sites where one or more heterodimer detects a peak at a lower FPR compared to the ATF3 homodimer. Several such clusters are apparent for heterodimers with CEBP or BATF family members. A striking result that emerged from the analysis of the GBM1 cell line is that multiple ATF3-bound genomic loci were better described by ATF3-heterodimers than the homodimer. For GBM1, we further examined two clusters of ChIP peaks for which heterodimers scored better than the ATF3 homodimer (Figure 5C, blue and green clusters in the dendrogram). In the blue cluster, *de novo* motif discovery revealed enrichment of a CRE-CAAT motif, which is the motif with maximal CSI intensities for CEBP•ATF3 dimers. *De novo* motif search of ChIP-seq peaks in the green cluster identified the TRE motif, which is the top ranked motif for ATF3 heterodimers formed with JUNB, JUN, FOS, and FOSL1, all of which are expressed in GBM1 cells (Supplementary file 1F). This is in contrast to the CRE motif preferred by the ATF3 homodimer. Gene ontology Functional annotations of genes linked to the CRE-CAAT (blue) and TRE (green) clusters also differ substantially (Figure 5C and Supplementary file 1H). CRE-CAAT sites preferred by ATF3•CEBP heterodimers (blue cluster) enriched for gene ontology (GO) terms related to immune response and JAK-STAT signaling, whereas TRE sites (green cluster) enriched for GO terms associated to nutrient sensing, PDGF signaling, and cell junction regulation. This observation lends support to the notion that heterodimers drive cell-type and signal-specific gene networks.

503 ***Co-occupied genomic loci bear emergent and conjoined sites***

504

505 Sharpening our focus to a subset of genomic loci that are co-occupied by ATF3 and
506 another bZIP permitted us to examine whether heterodimer cognate sites were evident
507 at co-occupied genomic loci. In Tier 1 ENCODE cell lines such as H1 and K562,
508 occupancy of multiple TFs has been charted across the genome (Dunham et al., 2012).
509 We first examined loci co-occupied by ATF3 and CEBPB or JUN. In H1 embryonic stem
510 cells, we identified a region on chromosome I that shows overlapping ChIP peaks for
511 ATF3 and CEBPB (Figure 6A, top panel). This locus is also resistant to DNAse I,
512 suggesting that ATF3 and CEBPB are binding to a seemingly inaccessible part of the
513 genome. Plotting CSI intensities for a given TF across the genome generates CSI-
514 Genomescape*s* (Figure 6A-B, bottom panels; Materials and methods). CSI-
515 Genomescapes in the co-occupied region identified a high intensity site for the
516 ATF3•CEBPA heterodimer, whereas no high intensity sequences were found for ATF3
517 or CEBPA homodimers (Figure 6A). CEBPA is the closest homolog to CEBPB for which
518 CSI data were obtained. Similar CSI-Genomescape analysis of a locus with overlapping
519 ATF3 and JUN peaks readily identified the JUN•ATF3 emergent site ($^{5'}$TGACGCAT$^{3'}$).
520 This site is within DNase I accessible euchromatin, and CSI-Genomescapes provide
521 scant support for either JUN or ATF3 homodimer binding to this site (Figure 6B).

522

523 Next, we identified genomic loci that are co-occupied by ATF3 and CEBPB in H1 (1018
524 overlapping ChIP peaks) or ATF3 and JUN in K562 cells (6539 overlapping ChIP peaks;
525 Figure 6 C-D). We then used CSI data of different homo- and heterodimers to assign

CSI scores within these co-occupied regions. Violin plots clearly demonstrate that regions co-occupied by ATF3 and CEBPB have higher CSI intensities when scored with ATF3•CEBP heterodimers than when scored with ATF3 and CEBP homodimers (Figures 6E and Figure 6 – figure supplement 1). In contrast, for loci co-occupied by JUN and ATF3, violin plots indicate that cognate sites for JUN•ATF3 heterodimer perform only marginally better at explaining the genomic binding data than sites preferred by JUN or ATF3 homodimers (Figure 6F, left panel). This observation is consistent with the ability of the JUN•ATF3 heterodimer to bind consensus CRE sites that are also bound by each contributing homodimer. The perceptibly higher CSI intensity when using JUN•ATF3 cognate sites might arise from heterodimer-preferred TRE sites or heterodimer-specific emergent sites. To examine this possibility, we utilized CSI-Genomescapes to score all co-occupied regions that include emergent heterodimer-specific $^{5'}$TGACGCAT$^{3'}$ sites (39 sites). When this subset of genomic regions was examined with homodimer CSI data, the violin plots reveal the inability of ATF3 homodimer cognate sites to account for the ChIP signals whereas JUN homodimers account for some of the JUN occupancy at those regions (Figure 6F, right panel). In contrast, the ATF3•JUN heterodimer cognate sites showed the highest scores for the emergent site.

Heterodimer-specific cognate sites map to SNPs associated with diseases

Armed with 102 CSI profiles of bZIP dimers, we scrutinized 5076 non-coding single nucleotide polymorphisms (SNPs) that are associated with diseases and quantitative traits (Maurano et al., 2012). We reasoned that non-coding SNPs that are not assigned to known TF cognate sites might be explained with our compendium of new bZIP-DNA interactomes. As a first step, we calibrated our CSI data by examining SNPs that are known to alter binding by CREB1 and CEBPA (Figure 7A top panel and Figure 7 – figure supplement 1A). The minor allele of rs10993994 in the promoter of the MSMB gene has been associated with prostate cancer and it creates a cognate site that is bound by CREB1 (Lou et al., 2009). Similarly, the minor allele of rs12740374 has been associated with myocardial infarction, aberrant plasma levels of low-density lipoprotein cholesterol (LDL-C), and enhanced expression of SORT1 gene in the liver (Musunuru et al., 2010). Biochemical studies have demonstrated that the G-to-T change generates an optimal CAAT site that is bound by CEBPA. We applied CSI-Genomescape analysis to both SNPs. In both cases, the minor allele has a higher CSI intensity than the corresponding major allele, suggesting that the minor alleles of these SNPs create CEBPA and CRE binding sites (Figure 7A and Figure 7 – figure supplement 1). The CSI-Genomescape for rs7631605 site is particularly interesting because it predicts disruption of the emergent site $^{5'}$TGACGCAT$^{3'}$ (Figure 7A middle panel). This allele is associated to Alzheimer's disease and mild cognitive impairment (MCI) and elevated levels of phosphorylated Tau-181P (Han et al., 2010). Additionally, CSI-Genomescape

predicts that rs1869901, a variant associated with schizophrenia, impacts binding of FOS•CEBPE by altering a TRE-CAAT site (Figure 7A bottom panel).

A scatterplot of CSI intensity scores for FOS•JUN (AP-1) for reference (hg19) or alternate alleles reveals SNPs that create or disrupt binding sites (Figure 7B). The plot shows that nearly all of the 5076 SNPs are near the origin and do not lead to large differences in CSI scores for the FOS•JUN heterodimer. However, a striking example of predicted increase in binding is rs3758354, a SNP associated with schizophrenia, depression and bipolar disorder (Huang et al., 2010). In contrast, a decrease in FOS•JUN heterodimer binding is predicted for rs17293632, a variant linked to Crohn's autoimmune disorder (Franke et al., 2010). ChIP-seq studies in several cell lines examined by the ENCODE consortium have shown binding by FOS and JUN to both loci, providing support that these sites are accessed by bZIP proteins in a cellular context (Figure 7 – figure supplement 1B).

Extending beyond AP-1, we used 102 bZIP CSI profiles to score both alleles of the 5076 SNPs and calculated a predicted fold-change in CSI intensity, which correlates with binding affinity (Figure 1 – figure supplement 2) (Carlson et al., 2010; Puckett et al., 2007). Similar correlations also hold true for other high throughput platforms (Berger et al., 2006; Fordyce et al., 2010; Slattery et al., 2011). We added a noise factor to our scoring function to make the fold-change predictions less sensitive to low CSI intensity (Figure 7C; Materials and methods). A total of 156 SNPs yielded a greater than 2-fold difference in CSI intensity between the reference and alternate alleles (Figure 7C-D).

589 Displaying the predicted increase (blue) or decrease (red) in binding by 102 bZIP
590 dimers at 156 SNPs reveals minor alleles that are targeted by unique heterodimers as
591 well as mutations that have wide-ranging impacts on multiple bZIP dimers. For example,
592 rs10994336 is predicted to increase CSI intensity by at least 2-fold for 44 out of 102
593 bZIP pairs reported here. We also report that 80% of the identified changes impact bZIP
594 heterodimers. In the richly annotated RegulomeDB database that ties SNP impact to
595 occurrence of TF binding sites, only 20 of 156 SNPs are currently annotated with a bZIP
596 motif (Boyle et al., 2012). It is particularly important to note that many of the SNPs in the
597 database are annotated with PWMs derived from bZIP homodimers, whereas our CSI
598 intensity fold-change predictions for 22 homo- and 80 bZIP heterodimers make use of
599 the entire bZIP-DNA interactomes (all 10-mers). The clusters in Figure 7D also point to
600 potential roles of bZIP proteins in less understood diseases and provide new
601 hypotheses for the etiology of such diseases and traits.
602

Discussion

Transcription factors rarely function alone, different TFs are activated by different cellular stimuli, and specific combinations of TFs converge at specific genomic loci to regulate expression of genes (Ptashne and Gann, 2002). Such combinatorial control provides the means to integrate multiple signals and tune the expression of specific genes or sculpt genome-wide transcriptomes in a nuanced manner. The ability of different TFs to form hetero-oligomers via protein-protein and protein-DNA interactions is an essential feature of this process. While most eukaryotic TFs can bind DNA as monomers, bZIP class of TFs only binds DNA as homo- or heterodimers. The ability of bZIPs to form heterodimers appears to increase with increasing evolutionary complexity, with human bZIPs displaying more intricate heterodimerization networks than *C. elegans* and *D. melanogaster,* which in turn, exhibit more complex dimerization networks than *S. cerevisiae* (Reinke et al., 2013). Comprehensive protein-protein interaction analysis has shown that 36 human bZIP monomers can form nearly 217 heterodimers, greatly expanding the repertoire of factors that can potentially bind DNA (Reinke et al., 2013). We demonstrate that this diversity of dimers expands the DNA sequence space that can be targeted by bZIPs. Our study further reveals that nearly 20% of the non-interacting bZIP pairs examined can be induced to dimerize at cognate DNA sites, providing yet greater diversity from a modest number of contributing monomers.

Given the large repertoire of human bZIP heterodimers, this family of TFs is particularly amenable to effect combinatorial control. Indeed, this potential was recognized long ago (Bohmann et al., 1987; Franza et al., 1988; Lamb and McKnight, 1991). An ever-increasing body of evidence now implicates bZIPs in numerous aspects of cellular and organismal function. Given their importance, a systematic study of DNA binding by bZIP heterodimers is clearly essential to understanding their functions. However, despite large surveys charting the TF-DNA interactomes (Badis et al., 2009; Badis et al., 2008; Berger et al., 2008; Carlson et al., 2010; Fordyce et al., 2010; Franco-Zorrilla et al., 2014; Grove et al., 2009; Jolma et al., 2010; Jolma et al., 2013; Kamesh et al., 2015; Nitta et al., 2015; Noyes et al., 2008; Siggers et al., 2012; Wei et al., 2010; Weirauch et al., 2014), bZIP dimers were under-scrutinized with only a handful of heterodimers reported thus far (Cohen et al., 2015; Jolma et al., 2015; Mann et al., 2013). Thus, it was quite unclear prior to this work how dimerization between different bZIP partners would impact DNA recognition. The DNA-binding profiles for 80 heterodimers, which we report alongside equivalent data for 22 homodimers, is the largest bZIP heterodimer DNA binding data reported to date and provides unprecedented insight into the impact of heterodimer formation.

Guided by protein-protein interaction data, we examined the DNA binding specificities of 126 stable dimers and 144 bZIP dimers that display no dimerization even at 1 μM. These 270 bZIP pairs represent a wide survey of the 666 potential pairs that can be formed by 36 monomers. The bZIP-DNA interactomes and specificity landscapes that emerged revealed three classes of cognate sites and several heterodimers displayed an

ability to interact with more than one class of binding site. Of the three classes, conjoined half-sites were the most abundant, with nearly 72% of all heterodimers displaying some affinity for such sites. The second class contained variably-spaced half-sites, often overlapping by a single nucleotide. The final class, comprising 16% of the sites, was the least expected “emergent” class of binding sites, where new non-obvious preferences for half-sites were revealed. Emergent sites targeted by heterodimers fall into “loss of specificity” or “gain of specificity” categories, as defined above. EMSA-FRET analyses not only quantified the relative affinities of hetero- and homodimers for these sites but also revealed the widespread ability of heterodimers to associate with cognate sites that have typically assumed to be bound by homodimers. More closely examining the emergent site targeted by ATF3•JUN, we find its occurrence at multiple locations across the human genome and, more importantly, several of these sites are co-occupied by ATF3 and JUN in vivo. Further emphasizing physiological role for these non-obvious binding sites, a SNP that disrupts this site is linked to neurological diseases (Han et al., 2010).

The high granularity of our CSI data also revealed that sequences flanking well-studied homodimer motifs, such as CRE, can impart sub-structure to the motif that is recognized and preferentially bound by different bZIP pairs. Access to such nuanced specificity preferences allows better annotation of genome-wide binding data for bZIPs for which specificity profiles and high quality ChIP data exist. This is particularly relevant because it is not uncommon for ChIP or genomic DNase I footprinting experiments to identify TF-bound regions that lack matches to the consensus motifs for a given TF. Our

671 results suggest that a fraction of such in vivo occupied regions likely contain
672 heterodimer binding sites. Another important insight from our comparative analysis of
673 genome-wide binding profiles across four cell types is that a given heterodimer
674 associates with distinct set of genomic loci in each cell type. The results suggest that
675 underlying chromatin and epigenetic landscapes in different cell types may contribute
676 significantly to the sites that are accessed by bZIP dimers. In this context, the ability of
677 ATF3•CEBPB to bind a cognate site within seemingly closed chromatin is consistent
678 with the ability of bZIPs such as CEBPB and FOS•JUN to function as “pioneer” factors
679 that first associate with closed chromatin and enable binding of additional TFs to yield
680 transcriptionally active euchromatin (Biddie et al., 2011; Garber et al., 2012). Whether
681 the ability to bind just one half site is important, or whether DNA-templated dimerization
682 of bZIPs confers any added ability to bind an otherwise inaccessible enhancer in closed
683 chromatin, remains to be determined.
684
685 Finally, the specificity and binding energy profiles of 102 bZIP dimers enables a more
686 nuanced examination of SNPs that have been linked by genome-wide association
687 studies to various diseases and quantitative traits. The vast majority of SNPs associated
688 with diseases occur in non-coding regions of the genome and most are not readily
689 annotated by the available TF-DNA interactomes perhaps in part because the focus has
690 been on obtaining consensus motifs of monomeric or homodimeric TFs. Rather than
691 consensus motifs, the use of the full spectrum of binding specificity may enable more
692 accurate mapping of TF-binding sites onto SNPs that are linked to diseases and
693 phenotypic traits. Our compendium of CSI profiles accurately predicted creation of

694 known bZIP cognate sites by previously validated SNPs. Of the 156 SNPs predicted by
695 our CSI profiles to impact bZIP binding, nearly 77% were mapped to bZIP heterodimers,
696 highlighting the importance of determining protein-DNA interactomes for heterodimer
697 TFs. Nearly 64% created bZIP binding sites and were “gain of function” changes relative
698 to the human reference genome. These results are consistent with the 10-fold greater
699 abundance of bZIP heterodimers over homodimers and the observation that aberrant
700 stimulation of gene networks is arguably a greater contributor to disease etiology (Bell
701 et al., 2015; Lee and Young, 2013; Mansour et al., 2014). SNPs that disrupt binding
702 also contribute to disease, an example of this form of regulatory perturbation being the
703 loss of the emergent JUN•ATF3 binding site that is associated with Alzheimer’s and
704 other neurological, cognitive and behavioral disorders. Our bZIP-DNA interactomes
705 identify 156 SNPs that potentially impact 646 bZIP binding events, any one of these
706 could potentially contribute to the associated ailments. Not only do our data help better
707 annotate the genome, they also serve as an invaluable resource to generate
708 hypotheses on how genetic variants may contribute to the etiology of a range of
709 diseases.
710
711 The recent emergence of powerful high throughput platforms for mapping protein-DNA
712 interactomes has brought the goal of comprehensively mapping the binding specificities
713 of all individual human TFs within reach (Carlson et al., 2010; Jolma et al., 2013;
714 Stormo and Zhao, 2010; Weirauch et al., 2014). However, it is clear from our work as
715 well as the recent work of others that the binding of TFs to each other, and/or to
716 adjacent DNA sites, can influence binding specificity profiles in important ways (Ansari

717 and Peterson-Kaufman, 2011; Garvie et al., 2001; Grove et al., 2009; Jolma et al.,
718 2015; Mann et al., 2013; Siggers et al., 2012; Slattery et al., 2011). Our study heralds
719 the important next wave of specificity mapping, in which the field will tackle the effects of
720 higher-order interactions and begin to relate these to the transcriptional control of key
721 biological processes.

Materials and methods

bZIP Cloning, Expression, and Labeling

Human bZIP proteins containing the basic-region and coiled-coiled domains with an N-terminal 6x His tag and a C-terminal intein-chitin binding domain were expressed as described previously (Reinke et al. 2013). Sequences are in Supplementary file 1A. Briefly, *E. coli* RP3098 cells transformed with bZIP clones were grown in 0.5 L LB cultures at 37 °C to OD600 = 0.4-0.8. Expression was induced with the addition of 0.5 mM IPTG (Isopropyl β-D-thiogalactopyranoside) and cultures incubated for 3-4 hours at which point cells were pelleted. Cells pellets were resuspended in 20 mM HEPES pH 8.0, 500 mM NaCl, 2 mM EDTA (ethylenediaminetetraacetic acid), 1 M guanidine-HCl, 0.2 mM PMSF (phenylmethylsulfonyl fluoride), and 0.1% Trition X-100). Cells were then sonicated and the lysate poured over a column of 1 ml chitin beads to bind the protein (NEB, Ipswich, MA). The column was then washed and equilibrated in EPL buffer (50 mM HEPES pH 8.0, 500 mM NaCl, 200 mM MESNA (2-mercaptoethanesulfonic acid), 1 M guanidine-HCl). The bZIP domain was then cleaved from the intein and labeled with biotin on the C-terminus by incubation for at least 16 hours in 1 ml EPL buffer containing 1 mg/ml cysteine-lysine-biotin peptide (CELTEK Peptides, Nashville, TN). The cleaved and biotin-labeled proteins were then eluted from the column using EPL buffer without MESNA and then diluted 5-fold into denaturing buffer (6 M guanidine-HCl, 5 mM imidazole, 0.5 M NaCl, 20 mM TRIS, 1 mM (DTT) Dithiothreitol, pH 7.9) and bound to a column containing 1 ML Ni-NTA beads (QIAGEN, Hilden, Germany). Columns were washed and proteins eluted with 60% ACN (Acetonitrile) /0.1% TFA (Trifluoroacetic

745 acid). The labeled proteins were then lyophilized, resuspended, and desalted using
746 spin-columns (Bio-Rad, Hercules, CA). Proteins were stored in 10 mM potassium
747 phosphate pH 4.5 at -80 °C. Peptide concentrations were determined by measuring
748 absorbance at 280 nM in 6 M guanidine-HCl/100 mM sodium phosphate pH 7.4. The
749 fluorescein and TAMRA labeled proteins used in gel-shift assays were generated as
750 described previously (Reinke et al., 2013).
751
752 **Cognate Site Identification (CSI) by HT-SELEX**
753 Cognate binding sites for bZIP homo- and heterodimers were determined by SELEX-
754 seq (Jolma et al., 2010; Tietjen et al., 2011; Zhao et al., 2009; Zykovich et al., 2009). A
755 DNA library (Integrated DNA Technologies, Inc.) with a central randomized 20 bp region
756 (10^{12} possible sequences), flanked by constant sequences used for amplification was
757 used (Supplementary file 1B). In vitro selections were performed as follows. For bZIP
758 homodimers, purified, C-terminal biotinylated-bZIP proteins (50 nM) were added to 100
759 nM of DNA library (Binding buffer: 1x PBS (10 mM PO_4^{3-}, 137 mM NaCl, 2.7 mM KCl),
760 pH 7.6, 2.5 mM DTT, 50 ng/ul poly dI-dC, 0.1% BSA) and incubated at room
761 temperature for 1 hour. The DNA library concentration and volume (20 μL) were such
762 that there was a high probability of sampling at least 1 copy of every 20-mer sequence
763 (10^{12} permutations). bZIP-DNA complexes were enriched with streptavidin coated
764 magnetic beads (Dynabeads, Invitrogen) following the manufacturer's protocol. After
765 pull-down, three quick washes with 100 μL ice-cold binding buffer were performed to
766 remove unbound DNA. Beads were resuspended in a PCR master mix (EconoTaq®
767 PLUS 2X Master Mix, Lucigen) and the DNA was amplified for 15 cycles. Amplified DNA

768 was column purified (QIAGEN), quantified by absorbance at 260 nm, and used for
769 subsequent binding rounds. Three rounds of selection were performed. For bZIP
770 heterodimers, one bZIP partner had a C-terminal biotin tag. bZIP-DNA complexes were
771 pulled down with streptavidin coated magnetic beads. Several steps were followed to
772 decrease DNA binding by competing homodimers: (1) a 1:3 molar ratio with an excess
773 of the non-biotinylated bZIP was used to shift the thermodynamic equilibrium from the
774 biotin-labeled homodimer; (2) the biotin-bZIP used for pull-down was chosen as the
775 more weakly interacting homodimer of the two interaction partners. As a convention,
776 when naming bZIP heterodimers, the bZIP-biotin is listed first, unless otherwise stated.
777 After 3 rounds of selection, an additional PCR was done to incorporate Illumina
778 sequencing adapters and a unique 6 bp barcode for multiplexing. The starting library
779 (Round 0) was also barcoded. Up to 180 samples were combined and sequenced in a
780 single Illumina GAIIx or HiSeq2000 lane.
781
782 **Sequencing Data Analysis**
783 Illumina sequencing yielded ~180 million reads per lane. Reads were de-multiplexed by
784 requiring an exact match to the 6 bp barcode and truncated to include only the 20 bp
785 derived from the random portion of the library. On average, we obtained 850,000 reads
786 per barcode. The occurrence of every k-mer (lengths 8 through 14 bp) was counted
787 using a sliding window of size k. To correct for biases in our starting DNA library, we
788 took the ratio of the counts of every k-mer to the expected number of counts in the
789 starting library. The starting library was modeled using a 5^{th}-order Markov Model derived
790 from the sequencing reads corresponding to the starting library (Round 0) (Slattery et

al., 2011). We then calculated a CSI intensity (z-score = $(x - \mu)/\sigma$) for each k-mer, using the distribution of k-mer enrichment values for that dimer. The most enriched 10, 12, and 14 bp subsequences were used to derive position weight matrix (PWM) motifs using MEME. Samples that failed to enrich specific sequences relative to the starting library (Round 0) or that only enriched low-complexity sequences were not included in further analysis. Data files for 20 bp reads and normalized 10 bp sequences are available at https://ansarilab.biochem.wisc.edu/computation.html.

Previously reported bZIP-DNA interaction data were downloaded from study PRJEB3289 in the European Nucleotide Archive (http://www.ebi.ac.uk/ena/data/view/PRJEB3289) (Jolma et al., 2013). 20 bp reads for bZIP proteins and their corresponding 20 bp DNA library (round 0) were analyzed as described previously.

Homodimer and heterodimer clustering

Binding profiles were defined for each bZIP pair using the CSI intensities (z-scores) of 1222 unique 10-mer sequences. This set of 10-mers is composed of the 50 highest-scoring sequences for each dimer. Unsupervised hierarchical clustering of pair-wise binding profile similarities, assessed by Pearson's correlation coefficient r, was done using R. Dendrograms and heatmaps were generated using the *heatmap.2* function in the gplots R-package. Heterodimers were labeled as such if the bZIP-DNA complex was pulled-down by a biotinylated bZIP that does not binds DNA as a homodimer in our experimental conditions. If the bZIP used for pull-down of the bZIP heterodimer also bound DNA as a homodimer, the observed DNA specificity was assigned to the

heterodimer only if the heterodimer specificity landscape was different (t-test p <0.05) from the homodimer specificity, assessed by correlation scores (Figure 1 – figure supplement 4).

Sequence Logos

PWMs were derived from the 1000 most enriched 12-mer sequences (ranked by z-score) for each bZIP pair, using the MEME (Bailey and Elkan, 1994). The most enriched 14-mer sequences were used for MAF dimers. MEME was run with following parameters: **-dna -mod anr -nmotifs 10 -minw 8 -maxw 18 -time 7200 -maxsize 60000 –revcomp.**

Specificity and Energy Landscapes (SELs)

Specificity and Energy Landscapes (SELs) display high-throughput protein-DNA (or protein-RNA) binding data for both array and sequencing methods (Campbell et al., 2012; Carlson et al., 2010; Tietjen et al., 2011). The organization of data in SEL is detailed in Figure 2A. The SELs shown in this work were generated from 10-, 12-, or 14-mer intensity files. Seed motifs were derived from PWM-derived DNA logos or from the highest intensity k-mer, and are shown on top of each SEL. The length of the seed motifs has to be smaller than the k-mer length of the CSI intensity file. The software to generate SELs is provided as Source Code Files (SEL_10MER and SEL12MER_14MER).

Electrophoretic Mobility Shift Assay (EMSA) – FRET

An electrophoretic mobility shift assay (EMSA) with fluorescence resonance energy transfer (FRET) readout was used to validate bZIP heterodimer binding to DNA. The assay relies on the ability to observe FRET between two fluorophores, TAMRA and fluorescein, as well as to detect each fluorophore in the absence of FRET (Figure 3A). The assay also measures DNA fluorescence to ensure that protein-DNA complexes are being examined. Two versions of each bZIP were made, one conjugated to TAMRA and the other to fluorescein. We observed that the fluorophores reproducibly retard (TAMRA) or increase (fluorescein) the mobility of the bZIP protein that they are attached to and thus assist in resolving each heterodimer with respect to the two homodimers formed by contributing partners. The sequences of all the DNA sites used are listed in Supplementary file 1D. Each site was flanked by 6 constant nucleotides on each side (GAGTCC-site-CCGTAG). Oligos modified on the 5′ end with the dye TYE 665 (IDT, Coralville, IA) were annealed with an unlabeled reverse-complement oligo. Binding reactions contained 50 nM of each fluorescein- and TAMRA-labeled proteins, 10 nM annealed dye-labeled DNA in 20 µl of binding buffer (50 mM potassium phosphate pH 7.4, 150 mM KCl, 0.1% BSA, 0.1% Tween-20, 5 ng/µl poly (dI-dC), 0.5 mM TCEP). Samples were mixed, incubated at 37 °C for 30 minutes, and then at 21 °C for 30 minutes. NOVEX 6% DNA retardation gels were loaded with 16 µl of each sample (Life Technologies, Carlsbad, CA) and run at 300V for 20-22 minutes at 22-25 °C. Gels were then imaged using a Typhoon 9500 scanner (GE Healthcare Bio-Sciences Corp., Piscataway, NJ) with separate channels for fluorescein, TAMRA, TYE 665, and FRET. Bleed through between channels was corrected using the spectral-unmixing plugin in ImageJ (http://rsb.info.nih.gov/ij/). The amount of DNA bound for the homodimers was

calculated by quantifying the DNA signal that corresponded to all three bound species (fluorescein homodimer, TAMRA homodimer, and mixed-dye homodimer). For the heterodimers, the amount of DNA bound was calculated by quantifying the DNA signal that corresponded to the mixed dye heterodimer. The amount of bound DNA was divided by the amount of unbound DNA run without protein added. For each heterodimer, the interaction was measured twice, with the fluorescein and TAMRA dye on different proteins, and the average of the two measurements is reported.

ChIP-seq Data

ChIP-seq peaks from the ENCODE project used in this work were downloaded from ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/ (Dunham I et al 2012). Overlapping genomic regions of ChIP-seq peaks were determined and extracted using bedops (Neph et al., 2012). For ATF3 ChIP-seq in GBM1 cells, aligned reads (.bam file) were downloaded from GEO (GSE33912). ATF3 peaks were called using the MACS tool (Zhang et al., 2008) in the Galaxy (Goecks et al., 2010) platform using default parameters. Overlapping ATF3-bound regions between different cell lines (Figure 5) were determined using the ChIPpeakAnno R-package (Zhu et al., 2010).

CSI Genomescapes: scoring in vivo bound sites with in vitro data

A CSI Genomescape is a plot generated by assigning in vitro CSI intensities (z-scores) to genomic regions. To generate the CSI Genomescapes in Figures 6 and 7, a 10 bp sliding window was used to score reported ChIP-seq peaks using quantile-normalized

CSI intensities for different bZIP dimers as follows: Given a bZIP pair and a ChIP-seq peak, the peak was assigned the maximum CSI intensity for any 10-mer within the reported peak.

Receiver Operating Characteristic (ROC)

CSI Genomescapes of ChIP-seq data sets were then used to generate Receiver Operating Characteristic (ROC) curves to reflect how well the in vitro binding data for different bZIPs explains the ChIP-seq data. In this analysis, ChIP-seq peaks were used as a true positive set, whereas two regions of equal length ±5 kbp from the center of each peak (that did not overlap another ChIP-seq peak) were chosen to make the true negative set. The fraction of regions in the positive vs. negative sets with scores above a varying CSI intensity cutoff were plotted to generate ROC curves (True Positive Rate vs. False Positive Rate). ATF3-bound regions (ChIP-seq peaks) were scored with the CSI intensities for ATF3 homodimer or for ATF3-containing heterodimers to generate the areas under the curves. Heatmaps and clustergrams in Figure 5 were made by hierarchical clustering of the lowest FPR-cutoff values at which peaks were detected as positives using the CSI intensities of the ATF3 containing dimers. ROC curves and heatmaps were generated in MATLAB.

De novo motifs and functional annotation of ChIP-seq peaks

Motif finding within ChIP-seq peaks was done with MEME-ChIP with default settings (Machanick and Bailey, 2011). Enrichment of functional annotations of genomic regions was done with Genomic Regions Enrichment of Annotations Tool (GREAT) with default settings (McLean et al., 2010). Gene Ontology annotations that are significantly

enriched (FDR < 0.05) by both binomial and hypergeometric test are shown. The False Discovery Rate (q-value) is corrected for multiple hypothesis tests.

Single Nucleotide Polymorphism (SNP) scoring

SNPs linked to diseases or quantitative traits by GWAS were obtained from the Supplemental Table S2 from Maurano et al., which reports human SNPs associated to diseases and quantitative traits (Maurano et al., 2012). For each SNP, we considered 21 bp region centered on the SNP (10 bp on each side) and assigned a score using the CSI intensity data all 10-mers. We scored both alleles using a 10 bp sliding window and assigning the highest CSI intensity (z-score) in the 21 bp fragment; each 21 bp region was scored with twelve 10 bp windows. We calculated a predicted fold-difference in CSI intensity between a given SNP and its reference allele (hg19) using the following formula:

$$\frac{(\text{CSI Intensity for alternate allele} - \text{Minimum CSI Intensity} + A)}{(\text{CSI Intensity for reference allele (hg19)} - \text{Minimum CSI Intensity} + A)}$$

where the A = (Maximum CSI Intensity – Minimum CSI Intensity) * F, Minimum CSI Intensity = minimum CSI Intensity (z-score) among the scored SNPs, Maximum CSI Intensity = maximum CSI Intensity (z-score) among the scored SNPs. And *F is a noise factor which* was varied from 1% to 90%, from lower to higher stringency in estimating the predicted difference in CSI intensity. We added a noise factor (F) to the formula to make the fold-change prediction less sensitive to low CSI scores and decrease the number of false positives predictions.

Acknowledgements

We thank Professor Parmesh Ramanathan and members of the Ansari and Keating laboratories for helpful discussions, Christos Kougentakis for technical assistance with EMSA assays, Laura Vanderploeg for help with the artwork, and Marie Adams from the University of Wisconsin Biotechnology Center DNA Sequencing Facility. This study was supported by NIH award R01 GM096466 to A.E.K, and NIH grants R01 CA133508 and U01 HL099773, and the W. M. Keck Medical Research Award to A.Z.A. J.A.R.M. was supported by the National Human Genome Research Institute (NHGRI) training grant of the Genome Sciences Training program (T32 HG002760).

939 **References**

940 Ansari, A.Z., and Peterson-Kaufman, K.J. (2011). A Partner Evokes Latent Differences
941 between Hox Proteins. Cell *147*, 1220-1221.

942 Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A.,
943 Chan, E.T., Metzler, G., Vedenko, A., Chen, X.Y.*, et al.* (2009). Diversity and
944 Complexity in DNA Recognition by Transcription Factors. Science *324*, 1720-1723.

945 Badis, G., Chan, E.T., van Bakel, H., Pena-Castillo, L., Tillo, D., Tsui, K., Carlson, C.D.,
946 Gossett, A.J., Hasinoff, M.J., Warren, C.L.*, et al.* (2008). A Library of Yeast
947 Transcription Factor Motifs Reveals a Widespread Function for Rsc3 in Targeting
948 Nucleosome Exclusion at Promoters. Molecular Cell *32*, 878-887.

949 Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization
950 to discover motifs in biopolymers. Proceedings / International Conference on Intelligent
951 Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems
952 for Molecular Biology *2*, 28-36.

953 Bell, R.J.A., Rube, H.T., Kreig, A., Mancini, A., Fouse, S.D., Nagarajan, R.P., Choi, S.,
954 Hong, C., He, D., Pekmezci, M.*, et al.* (2015). The transcription factor GABP selectively
955 binds and activates the mutant TERT promoter in cancer. Science *348*, 1036-1039.

956 Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena-Castillo, L.,
957 Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T.*, et al.* (2008). Variation in
958 homeodomain DNA binding revealed by high-resolution analysis of sequence
959 preferences. Cell *133*, 1266-1276.

960 Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.X.S., Estep, P.W., and Bulyk, M.L.
961 (2006). Compact, universal DNA microarrays to comprehensively determine
962 transcription-factor binding site specificities. Nat Biotechnol *24*, 1429-1435.
963 Biddie, S.C., John, S., Sabo, P.J., Thurman, R.E., Johnson, T.A., Schiltz, R.L., Miranda,
964 T.B., Sung, M.H., Trump, S., Lightman, S.L.*, et al.* (2011). Transcription factor AP1
965 potentiates chromatin accessibility and glucocorticoid receptor binding. Mol Cell *43*,
966 145-155.
967 Bohmann, D., Bos, T.J., Admon, A., Nishimura, T., Vogt, P.K., and Tjian, R. (1987).
968 Human protooncogene c-jun encodes a DNA-binding protein with structural and
969 functional-properties of transcription factor AP-1. Science *238*, 1386-1392.
970 Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M.,
971 Karczewski, K.J., Park, J., Hitz, B.C., Weng, S.*, et al.* (2012). Annotation of functional
972 variation in personal genomes using RegulomeDB. Genome Res *22*, 1790-1797.
973 Campbell, Z.T., Bhimsaria, D., Valley, C.T., Rodriguez-Martinez, J.A., Menichelli, E.,
974 Williamson, J.R., Ansari, A.Z., and Wickens, M. (2012). Cooperativity in RNA-Protein
975 Interactions: Global Analysis of RNA Binding Specificity. Cell Reports *1*, 570-581.
976 Carlson, C.D., Warren, C.L., Hauschild, K.E., Ozers, M.S., Qadir, N., Bhimsaria, D.,
977 Lee, Y., Cerrina, F., and Ansari, A.Z. (2010). Specificity landscapes of DNA binding
978 molecules elucidate biological function. Proceedings of the National Academy of
979 Sciences of the United States of America *107*, 4544-4549.
980 Chu, H.M., Tan, Y., Kobierski, L.A., Balsam, L.B., and Comb, M.J. (1994). Activating
981 transcription factor-3 stimulates 3',5'-cyclic adenosine monophosphate-dependent gene
982 expression. Molecular Endocrinology *8*, 59-68.

983 Ciofani, M., Madar, A., Galan, C., Sellars, M., Mace, K., Pauli, F., Agarwal, A., Huang,
984 W.D., Parkurst, C.N., Muratet, M.*, et al.* (2012). A Validated Regulatory Network for
985 Th17 Cell Specification. Cell *151*, 289-303.
986 Cohen, D.M., Won, K.-J., Nguyen, N., Lazar, M.A., Chen, C.S., and Steger, D.J. (2015).
987 ATF4 licenses C/EBPβ activity in human mesenchymal stem cells primed for
988 adipogenesis. eLife *4*.
989 Collins, C., Wang, J., Miao, H., Bronstein, J., Nawer, H., Xu, T., Figueroa, M., Muntean,
990 A.G., and Hess, J.L. (2014). C/EBPα is an essential collaborator in Hoxa9/Meis1-
991 mediated leukemogenesis. Proceedings of the National Academy of Sciences *111*,
992 9899-9904.
993 Costa, R.H., Kalinichenko, V.V., Holterman, A.X.L., and Wang, X.H. (2003).
994 Transcription factors in liver development, differentiation, and regeneration. Hepatology
995 *38*, 1331-1347.
996 Deng, T.L., and Karin, M. (1993). junb differs from c-jun in its DNA-binding and
997 dimerization domains, and represses c-jun by formation of inactive heterodimers. Genes
998 Dev *7*, 479-490.
999 Deppmann, C.D., Alvania, R.S., and Taparowsky, E.J. (2006). Cross-species annotation
1000 of basic leucine zipper factor interactions: Insight into the evolution of closed interaction
1001 networks. Mol Biol Evol *23*, 1480-1492.
1002 Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C., Doyle, F., Epstein, C.B.,
1003 Frietze, S., Harrow, J., Kaul, R.*, et al.* (2012). An integrated encyclopedia of DNA
1004 elements in the human genome. Nature *489*, 57-74.

ENCODE (2011). A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol *9*, e1001046.

Filén, S., Ylikoski, E., Tripathi, S., West, A., Björkman, M., Nyström, J., Ahlfors, H., Coffey, E., Rao, K.V.S., Rasool, O.*, et al.* (2010). Activating Transcription Factor 3 Is a Positive Regulator of Human IFNG Gene Expression. The Journal of Immunology *184*, 4990-4999.

Fordyce, P.M., Gerber, D., Tran, D., Zheng, J.S., Li, H., DeRisi, J.L., and Quake, S.R. (2010). De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. Nat Biotechnol *28*, 970-976.

Franco-Zorrilla, J.M., Lopez-Vidriero, I., Carrasco, J.L., Godoy, M., Vera, P., and Solano, R. (2014). DNA-binding specificities of plant transcription factors and their potential to define target genes. Proceedings of the National Academy of Sciences of the United States of America *111*, 2367-2372.

Franke, A., McGovern, D.P., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R.*, et al.* (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nature genetics *42*, 1118-1125.

Franza, B.R., Rauscher, F.J., Josephs, S.F., and Curran, T. (1988). The Fos complex and Fos-related antigens recognize sequence elements that contain AP-1 binding-sites. Science *239*, 1150-1153.

Garber, M., Yosef, N., Goren, A., Raychowdhury, R., Thielke, A., Guttman, M., Robinson, J., Minie, B., Chevrier, N., Itzhaki, Z.*, et al.* (2012). A high-throughput

chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. Mol Cell *47*, 810-822.

Gargiulo, G., Cesaroni, M., Serresi, M., de Vries, N., Hulsman, D., Bruggeman, S.W., Lancini, C., and van Lohuizen, M. (2013). In Vivo RNAi Screen for BMI1 Targets Identifies TGF-beta/BMP-ER Stress Pathways as Key Regulators of Neural- and Malignant Glioma-Stem Cell Homeostasis. Cancer Cell *23*, 660-676.

Garvie, C.W., Hagman, J., and Wolberger, C. (2001). Structural studies of Ets-1/Pax5 complex formation on DNA. Molecular Cell *8*, 1267-1276.

Gilchrist, M., Thorsson, V., Li, B., Rust, A.G., Korb, M., Kennedy, K., Hai, T., Bolouri, H., and Aderem, A. (2006). Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. Nature *441*, 173-178.

Goecks, J., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome biology *11*, R86.

Grove, C.A., De Masi, F., Barrasa, M.I., Newburger, D.E., Alkema, M.J., Bulyk, M.L., and Walhout, A.J.M. (2009). A Multiparameter Network Reveals Extensive Divergence between C. elegans bHLH Transcription Factors. Cell *138*, 314-327.

Hai, T., and Curran, T. (1991). Cross-family dimerization of transcription factors FOS, JUN, and ATF/CREB alters DNA-binding specificity. Proceedings of the National Academy of Sciences of the United States of America *88*, 3720-3724.

Hai, T., Wolfgang, C.D., Marsee, D.K., Allen, A.E., and Sivaprasad, U. (1999). ATF3 and stress responses. Gene Expression *7*, 321-335.

1049 Hai, T., Wolford, C.C., and Chang, Y.-S. (2010). ATF3, a Hub of the Cellular Adaptive-
1050 Response Network, in the Pathogenesis of Diseases: Is Modulation of Inflammation a
1051 Unifying Component? Gene Expression *15*, 1-11.
1052 Han, M.R., Schellenberg, G.D., Wang, L.S., and Alzheimer's Dis Neuroimaging, I.
1053 (2010). Genome-wide association reveals genetic effects on human A beta(42) and tau
1054 protein levels in cerebrospinal fluids: a case control study. BMC Neurol *10*, 14.
1055 Hauschild, K.E., Stover, J.S., Boger, D.L., and Ansari, A.Z. (2009). CSI-FID: High
1056 throughput label-free detection of DNA binding molecules. Bioorg Med Chem Lett *19*,
1057 3779-3782.
1058 Herdegen, T., and Waetzig, V. (2001). AP-1 proteins in the adult brain: facts and fiction
1059 about effectors of neuroprotection and neurodegeneration. Oncogene *20*, 2424-2437.
1060 Hsu, J.C., Bravo, R., and Taub, R. (1992). Interactions among LRF-1, JunB, c-Jun, and
1061 c-Fos define a regulatory program in the G1 phase of liver regeneration. Molecular and
1062 Cellular Biology *12*, 4654-4665.
1063 Huang, J., Perlis, R.H., Lee, P.H., Rush, A.J., Fava, M., Sachs, G.S., Lieberman, J.,
1064 Hamilton, S.P., Sullivan, P., Sklar, P.*, et al.* (2010). Cross-disorder genomewide
1065 analysis of schizophrenia, bipolar disorder, and depression. The American journal of
1066 psychiatry *167*, 1254-1263.
1067 Jain, J., McCaffrey, P.G., Valgeardher, V.E., and Rao, A. (1992). Nuclear factor of
1068 activated T-cells contains Fos and Jun. Nature *356*, 801-804.
1069 Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M.,
1070 Vaquerizas, J.M., Yan, J., Sillanpaa, M.J.*, et al.* (2010). Multiplexed massively parallel

SELEX for characterization of human transcription factor binding specificities. Genome Res *20*, 861-873.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G.*, et al.* (2013). DNA-Binding Specificities of Human Transcription Factors. Cell *152*, 327-339.

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

Jung, K.A., and Kwak, M.K. (2010). The Nrf2 System as a Potential Target for the Development of Indirect Antioxidants. Molecules *15*, 7266-7291.

Kamesh, N., Lambert, S.A., Yang, A.W.H., Riddell, J., Mnaimneh, S., Zheng, H., Albu, M., Najafabadi, H.S., Reece-Hoyes, J.S., Bass, J.I.F.*, et al.* (2015). Mapping and analysis of Caenorhabditis elegans transcription factor sequence specificities. eLife, e06967-e06967.

Kim, J., and Struhl, K. (1995). Determinants of half-site spacing preferences that distinguish AP-1 and ATF/CREB bZIP domains. Nucleic Acids Res *23*, 2531-2537.

Kittler, R., Zhou, J., Hua, S.J., Ma, L.J., Liu, Y.W., Pendleton, E., Cheng, C., Gerstein, M., and White, K.P. (2013). A Comprehensive Nuclear Receptor Network for Breast Cancer Cells. Cell Reports *3*, 538-551.

Konig, P., and Richmond, T.J. (1993). The X-ray structure of the GCN4-bZIP bound to ATF/CREB site DNA shows the complex depends on DNA flexibility. Journal of molecular biology *233*, 139-154.

1093 Lamb, P., and McKnight, S.L. (1991). Diversity and specificity in transcriptional
1094 regulation: the benefits of heterotypic dimerization. Trends Biochem Sci *16*, 417-422.
1095 Lee, T.I., and Young, R.A. (2013). Transcriptional Regulation and Its Misregulation in
1096 Disease. Cell *152*, 1237-1251.
1097 Lopez-Bergami, P., Lau, E., and Ronai, Z.e. (2010). Emerging roles of ATF2 and the
1098 dynamic AP1 network in cancer. Nat Rev Cancer *10*, 65-76.
1099 Lou, H., Yeager, M., Li, H.C., Bosquet, J.G., Hayes, R.B., Orr, N., Yu, K., Hutchinson,
1100 A., Jacobs, K.B., Kraft, P.*, et al.* (2009). Fine mapping and functional analysis of a
1101 common variant in MSMB on chromosome 10q11.2 associated with prostate cancer
1102 susceptibility. Proceedings of the National Academy of Sciences of the United States of
1103 America *106*, 7933-7938.
1104 Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA
1105 datasets. Bioinformatics *27*, 1696-1697.
1106 Male, V., Nisoli, I., Gascoyne, D.M., and Brady, H.J.M. (2012). E4BP4: an unexpected
1107 player in the immune response. Trends in Immunology *33*, 98-102.
1108 Mann, I.K., Chatterjee, R., Zhao, J.F., He, X.M., Weirauch, M.T., Hughes, T.R., and
1109 Vinson, C. (2013). CG methylated microarrays identify a novel methylated sequence
1110 bound by the CEBPB vertical bar ATF4 heterodimer that is active in vivo. Genome Res
1111 *23*, 988-997.
1112 Mansour, M.R., Abraham, B.J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin,
1113 A.D., Etchin, J., Lawton, L., Sallan, S.E., Silverman, L.B.*, et al.* (2014). An oncogenic
1114 super-enhancer formed through somatic mutation of a noncoding intergenic element.
1115 Science *346*, 1373-1377.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H.Z., Brody, J., *et al.* (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science *337*, 1190-1195.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol *28*, 495-U155.

Miller, M. (2009). The Importance of Being Flexible: The Case of Basic Region Leucine Zipper Transcriptional Regulators. Current Protein & Peptide Science *10*, 244-269.

Murphy, T.L., Tussiwand, R., and Murphy, K.M. (2013). Specificity through cooperation: BATF-IRF interactions control immune-regulatory networks. Nat Rev Immunol *13*, 499-509.

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V., Li, X.Y., Li, H., Kuperwasser, N., Ruda, V.M., *et al.* (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature *466*, 714-U712.

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., *et al.* (2012). BEDOPS: high-performance genomic feature operations. Bioinformatics *28*, 1919-1920.

Nitta, K.R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., Hens, K., Toivonen, J., Deplancke, B., Furlong, E.E.M., *et al.* (2015). Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. eLife *4*.

1137 Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., Brodsky, M.H., and
1138 Wolfe, S.A. (2008). Analysis of homeodomain specificities allows the family-wide
1139 prediction of preferred recognition sites. Cell *133*, 1277-1289.
1140 Ptashne, M., and Gann, A. (2002). Genes and Signals (Cold Spring Harbor Laboratory
1141 Press).
1142 Puckett, J.W., Muzikar, K.A., Tietjen, J., Warren, C.L., Ansari, A.Z., and Dervan, P.B.
1143 (2007). Quantitative microarray profiling of DNA-binding molecules. Journal of the
1144 American Chemical Society *129*, 12310-12319.
1145 Reinke, A.W., Baek, J., Ashenberg, O., and Keating, A.E. (2013). Networks of bZIP
1146 Protein-Protein Interactions Diversified Over a Billion Years of Evolution. Science *340*,
1147 730-734.
1148 Siggers, T., Chang, A.B., Teixeira, A., Wong, D., Williams, K.J., Ahmed, B., Ragoussis,
1149 J., Udalova, I.A., Smale, S.T., and Bulyk, M.L. (2012). Principles of dimer-specific gene
1150 regulation revealed by a comprehensive characterization of NF-kappa B family DNA
1151 binding. Nat Immunol *13*, 95-U123.
1152 Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T.Y., Rohs, R.,
1153 Honig, B., Bussemaker, H.J.*, et al.* (2011). Cofactor Binding Evokes Latent Differences
1154 in DNA Binding Specificity between Hox Proteins. Cell *147*, 1270-1282.
1155 Stormo, G.D., and Zhao, Y. (2010). Determining the specificity of protein-DNA
1156 interactions. Nature Reviews Genetics *11*, 751-760.
1157 Tanaka, Y., Nakamura, A., Morioka, M.S., Inoue, S., Tamamori-Adachi, M., Yamada, K.,
1158 Taketani, K., Kawauchi, J., Tanaka-Okamoto, M., Miyoshi, J.*, et al.* (2011). Systems

Analysis of ATF3 in Stress Response and Cancer Reveals Opposing Effects on Pro-Apoptotic Genes in p53 Pathway. PLoS One *6*, 12.

Thanos, D., and Maniatis, T. (1995). Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. Cell *83*, 1091-1100.

Thompson, M.R., Xu, D., and Williams, B.R. (2009). ATF3 transcription factor and its emerging roles in immunity and cancer. Journal of molecular medicine (Berlin, Germany) *87*, 1053-1060.

Tietjen, J.R., Donato, L.J., Bhimisaria, D., and Ansari, A.Z. (2011). Sequence-specificity and energy landscapes of DNA-binding molecules. In Methods in Enzymology, Vol 497: Synthetic Biology, Methods for Part/Device Characterization and Chassis Engineering, Pt A, C. Voigt, ed., pp. 3-30.

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.

Ubeda, M., Wang, X.Z., Zinszner, H., Wu, I., Habener, J.F., and Ron, D. (1996). Stress-induced binding of the transcriptional factor CHOP to a novel DNA control element. Mol Cell Biol *16*, 1479-1489.

Vinson, C.R., Hai, T.W., and Boyd, S.M. (1993). Dimerization specificity of the leucine zipper-containing bzip motif on DNA-binding - prediction and rational design. Genes Dev *7*, 1047-1058.

Warren, C.L., Kratochvil, N.C.S., Hauschild, K.E., Foister, S., Brezinski, M.L., Dervan, P.B., Phillips, G.N., and Ansari, A.Z. (2006). Defining the sequence-recognition profile of

DNA-binding molecules. Proceedings of the National Academy of Sciences of the United States of America *103*, 867-872.

Wei, G.-H., Badis, G., Berger, M.F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A.R.*, et al.* (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. Embo Journal *29*, 2147-2160.

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.

Xie, D., Boyle, A.P., Wu, L.F., Zhai, J., Kawli, T., and Snyder, M. (2013). Dynamic trans-Acting Factor Colocalization in Human Cells. Cell *155*, 713-724.

Yamamoto, T., Kyo, M., Kamiya, T., Tanaka, T., Engel, J.D., Motohashi, H., and Yamamoto, M. (2006). Predictive base substitution rules that determine the binding and transcriptional specificity of Maf recognition elements. Genes to Cells *11*, 575-591.

Yin, X., DeWille, J.W., and Hai, T. (2008). A potential dichotomous role of ATF3, an adaptive-response gene, in cancer development. Oncogene *27*, 2118-2127.

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

Zhao, Y., Granas, D., and Stormo, G.D. (2009). Inferring Binding Energies from Selected Binding Sites. Plos Computational Biology *5*, e1000590.

Zhu, L.J., Gazin, C., Lawson, N.D., Pages, H., Lin, S.M., Lapointe, D.S., and Green, M.R. (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics *11*, 10.

Zykovich, A., Korf, I., and Segal, D.J. (2009). Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. Nucleic Acids Res *37*, 7.

Figure Legends

Figure 1. Overview of human bZIP homodimer and heterodimer DNA-binding specificities. (A) Summary of SELEX-seq results categorized by protein-protein interaction (PPI) affinity (Reinke et al., 2013). Specificity profiles were classified as resulting in a motif arising from DNA binding by either a homodimer (brown) or heterodimer (dark brown), or not resulting in a motif (white). Some profiles could not be unambiguously assigned to a homo vs. heterodimer (light brown). (B) Pairwise comparisons of the DNA binding preferences of 102 bZIP dimers (22 homodimers and 80 heterodimers) using the z-scores for 1222 unique 10 bp sequences corresponding to the 50 top ranked sequences for each dimer. Throughout the paper the biotinylated bZIP is listed first when describing a heterodimer. (C) Representative motifs bound by bZIP homodimers and heterodimers reported in this study. Heterodimer motifs were grouped as Conjoined, Variable spacer, and Emergent. The color code defined here for half sites (colored arrows above motifs) is used throughout the figures.

Figure 1 – figure supplement 1.

Cognate site identification by SELEX-sequencing. (A) In CSI by SELEX-seq a DNA library with a randomized 20 bp region is incubated with a bZIP pair in which one bZIP partner (light grey) was biotinylated and the other partner (light grey) was labeled with fluorescein (blue star). bZIP partners were mixed in 3:1 molar ratios with the biotinylated partner at the lower concentration. Affinity purification using the less abundant biotinylated partner enriched for heterotypic dimers. (B) Reproducibility of CSI by

SELEX-Seq. Scatter plots of CSI intensities (z-scores) for all 10-mers for replicate samples and (C) reciprocal samples.

Figure 1 –figure supplement 2.

ATF3 CSI Intensity (z-score) correlates with equilibrium association constant. (A) DNA sequence of oligonucleotides used for determining binding constants. (B) Correlation between CSI intensity (z-score) and association constant (K_a) for ATF3. Binding constants were measured by EMSA. Error bars are ± S.D. of at least duplicate measurements. (C) Representative autoradiographs of EMSA experiments from which binding constants were calculated using non-linear regression.

Figure 1 – figure supplement 3.

Pairwise comparison of bZIP homodimers reported in this study and bZIP dimers reported by Jolma et al. (Jolma et al., 2013). (A) Hierarchical clustering was performed using the CSI intensities (z-scores) of 871 unique 10 bp sequences corresponding to the 50 top ranked sequences identified from each dimer. Corresponding bZIP pairs are highlighted in matching color. (B) Scatter plots comparing CSI intensity (z-score) for all 10-mers of bZIP dimers from this study with bZIP dimers previously reported by Jolma et al (Jolma et al., 2013). (top left) BATF3 vs. BATF3; (top right) CEBPG vs. CEBPG; (bottom left) ATF4 vs. ATF4; (bottom right) ATF4 vs. ATF4•CEBPG.

Figure 1 –figure supplement 4.

bZIP Heterodimer specificity. Pearson's correlations (r) of all 10-mers between replicate experiments of bZIP dimers (top), and correlations between a bZIP heterodimer and the bZIP homodimer that was used to pull-down the heterodimer. The average (± standard deviation) Pearson's correlation (r) for 8 replicate samples was 0.8 ± 0.1.

Figure 1 – figure supplement 5.

DNA sequence preferences for FOS•CEBPE, FOS•CEBPG, FOSL1•CEBPE and FOSL1•CEBPG. Left, PPI affinity for the corresponding heterodimer is shown. Middle, MEME motifs are represented as DNA logos. Right, 2-dimensional scatter plots comparing the CSI intensities for all 10-mers. CRE/CAAT (TGACGTAA) sites are colored red and TRE/CAAT (TGAGCAA) sites are colored orange.

Figure 2. Specificity and Energy Landscapes (SELs) and motifs for bZIP heterodimers. (A) SEL displays CSI intensities for all sequence permutations of a given binding site size (k-mers). Sequences are organized with respect to any selected seed motif, however a k-mer representing PWM-derived motif is typically used. CSI intensities correlate with equilibrium binding affinities. As an example, the arrangement of 6-mer sequences for a simplified 4-mer seed motif is shown. The innermost circle displays the intensities for all sequences that have an exact match to the seed motif (0-mismatch ring). In this ring, sequences are arranged in a clockwise manner with sequences that include residues 5′ of the seed motif at the start, sequences with residues that flank both 5′ and 3′ ends in the middle, and 3′ flanking sequences at the

end (context). The subsequent 1-mismatch ring contains the sequences that differ at one position from the seed. The sequences are organized clockwise starting with mismatches at the first position and ending with mismatches at the last position of the motif. Within each sector, the mismatches at a given position (indicated by x) are organized in alphabetical order (A, C, G, and T). The 2-mismatch ring contains all permutations with two positional differences with the seed, similarly ordered. (B) Left, SEL for JUN•ATF3 heterodimer using CRE ($^{5'}$TGACGTCA$^{3'}$) as the seed motif. By displaying the 10 bp sequence space, preferred sequences become apparent. Peaks corresponding to emergent and variably-spaced sites are identified by arrows. Right, SEL displaying 12 bp sequences for ATF4•CEBPG heterodimer using CRE-CAAT ($^{5'}$ATGACGCAAT$^{3'}$) as the seed motif. (C) Heatmap of the relative CSI intensities of 102 bZIP dimers (columns) for the 10 sites highlighted in Figure 2B as well as constituent half-sites of the 6 classic bZIP motifs (rows). Displayed is the maximum CSI intensity of all the 10mers matching the site. bZIP dimers are listed in the same order as in Figure 1B. ATF3, ATF4, CEBPG and JUN homodimers are marked by asterisks. While bZIPs do not bind as monomers to half-sites, the occurrence of bZIP half-sites within motifs is displayed in the second set of rows to enable comparison between the half-site preferences versus the CSI intensity for motifs that display these half-sites in different combinations or in different contexts.

Figure 3. Influence of bZIP protein dimerization on DNA binding. (A) EMSA-FRET assay used to quantify bZIP heterodimers and homodimers binding to DNA. Fluorescein and TAMRA are depicted as blue and green stars, respectively. In the EMSA gel,

homodimers give rise to pseudo-colored blue (fluorescein) or green (TAMRA) signals, whereas heterodimers give a FRET signal that is pseudo-colored red. (B) EMSA-FRET results for bZIP dimers binding to selected heterodimer-specific emergent sites (brown) and conjoined half-sites (blue). Bar graphs show the percent of the indicated DNA oligomer bound by each dimer. The PPI strength of each dimer is indicated with gray-scale circles sized according to the K_d for a given protein-protein interaction. Homodimers are marked with an asterisk (*). (C) EMSA-FRET results for bZIP dimers tested for binding to DNA sites composed of conjoined half-sites. Left, dimers tested against two different sites composed of conjoined half-sites. Right, dimers tested against a single site. Data are displayed as in B.

Figure 3 –figure supplement 1.

Influence of bZIP protein dimerization on DNA binding. (A) Detecting heterodimer DNA complexes using an EMSA-FRET assay. Top, Fluorescein signal in blue, TAMRA signal in green, and FRET signal in red. Bottom, TYE 665 labeled DNA site. (B) Three examples to explain the notation used in part C summarize data for DNA binding by homodimers and heterodimers composed of (left) ATF3 and DBP, (middle) ATF3 and CEBPA and (right) BATF3 and JUN. Within each example, rows indicate different bZIP dimers. The top row describes the homodimer formed by the first-mentioned bZIP, the bottom row is for the other homodimer, and the middle row contains data for the heterodimer. Within each example, each column represents binding to a different DNA site composed of two half-sites. DNA binding affinity is indicated using a green-scale heatmap with key indicating % binding at far right. The color of the cell border indicates strength of the protein-protein interaction as measured previously by FRET, indicated by

yellow-scale heatmap at right. ATF3•DBP example: top row is ATF3 binding to CRE-PAR, middle row is ATF3 • DBP heterodimer binding to CRE-PAR, bottom is DBP homodimer binding to CRE-PAR. ATF3•CEBPA example: Top row is ATF3 homodimer, middle row is ATF3 • CEBPA heterodimer, and bottom row is CEBPA homodimer. Binding is to CRE-CAAT in left column and TRE-CAAT in right column. BATF3•JUN example: Top row is BATF3 homodimer, middle row is BATF3 • JUN heterodimer and bottom row is JUN homodimer. Binding is to CRE-CRE in left column and CRE-CREA in right column. (C) Complete set of EMSA-FRET data. Examples in B are included in this grid and other cells can be interpreted analogously.

Figure 3 –figure supplement 2.

Heterospecific binding of DNA Top, DNA sequences composed of optimal half sites. Bottom, comparison of an optimal DNA site to a heterodimer-specific non-optimal DNA site. DNA sequences for EMSA-FRET experiments are reported in Supplementary file 1D.

Figure 4. ATF3 heterodimers bind a range of distinct cognate sites.

(A) Hierarchical clustering of pairwise comparisons of DNA-binding specificity (10-mers) for ATF3 homodimer and 9 ATF3-containing heterodimers. (B) DNA logos showing the MEME motifs derived from the top 1000 12-mer sequences for ATF3 homodimer and ATF3-containing heterodimers. Colored squares next to dimer names indicate PPI strength using the scale from Figure 3. (C) 3-dimensional and (D) 2-dimensional scatter plots comparing the DNA binding specificities of bZIP homodimers vs. ATF3-containing

heterodimers. Scatter plots of quantile-normalized CSI intensities (z-scores) of ATF3 dimers for 80,000 10-mers are shown.

Figure 5. ATF3 binds to different genomic regions using diverse motifs.

(A) Venn diagram of the numbers of ATF3-bound regions determined by ChIP-seq in different cell lines. (B) Heatmap of the False Positive Rate (FPR)-cutoffs at which ATF3 ChIP-seq peaks (rows) are detected as positive for ATF3 or ATF3-dimer binding. Peaks were scored using CSI intensities of the ATF3 homodimer or ATF3-containing heterodimers (columns) in H1hESCs, K562, and HEPG2 cells, and clustered by FPR-cutoffs across all dimers. (C) Same as (B) for the glioblastoma multiforme (GBM1) cell line. Highlighted clusters (blue and green) contain DNA motifs preferred by different ATF3 dimers and are enriched with different Gene Ontology Biological Process terms. False Discovery Rates (q-values) for each GO term are shown. See Supplementary file 1H.

Figure 5 –figure supplement 1.

ROC curves. (A) Area Under the Receiver Operating Characteristic curve (AUC-ROC) values for the intersection of ChIP-seq peaks determined using in vitro specificity profiles of the corresponding bZIP heterodimer, as described in Materials and methods. x-axis: False-Positive Rate; y-axis: True-Positive Rate (TPR). ChIP-seq peaks from specified cell lines were downloaded from the ENCODE project. (B) ROC curves and AUC values for ChIP-seq peaks (all peaks) determined using DNA binding specificity profiles of the corresponding bZIP homodimer.

Figure 6. bZIP heterodimer DNA sites are bound in vivo.

(A) ChIP-seq traces for ATF3 (blue) and CEBPB (orange) and DNase I hypersensitivity (black) trace for in H1 human embryonic stem cells. Below, CSI-Genomescape for bound genomic regions for ATF3 and CEBPA homodimers and ATF3•CEPBA heterodimer. CEBPA and CEBPB share 76% identity over their bZIP domain. (B) ChIP-seq traces for ATF3 (blue) and JUN (green) and DNAse I hypersensitivity trace (black) in K562 cells. Below, CSI-Genomescape for bound genomic region for ATF3 and JUN homodimers, and for JUN•ATF3 heterodimer. (C) Venn diagram of bound regions (ChIP-seq peaks) for ATF3 and CEBPB in H1hESC and for (D) ATF3 and JUN in K562 cells. (E) Violin plots of CSI-seq scores for the ChIP-seq peaks derived from the intersection of ATF3 and CEBPB ChIP peaks (1018 overlapping peaks) in H1 stem cells using in vitro data for ATF3, CEBPA, CEBPB (from Jolma et al.) (Jolma et al., 2013), CEBPE, CEBPG homodimers and ATF3•CEBPA, ATF3•CEBPE, and ATF3•CEBPG heterodimers. CSI intensities were quantile normalized. (F) Violin plots of CSI-seq scores for the ChIP-seq peaks derived from the intersection of ATF3 and JUN ChIP peaks (left, 6539 overlapping peaks) in K562 cells, left. Violin plots for the subset of overlapping peaks of ATF3 and JUN containing a match for the heterodimer-specific site TGACGCAT (39 peaks), right. Peaks were scored using ATF3 and JUN homodimers, and JUN•ATF3 heterodimers.

Figure 6 – figure supplement 1.

CSI intensities for bound genomic regions. Violin plots of CSI intensities (z-scores) for (A) Negative regions were taken from ± 5 kb from the center of each ATF3 and CEBPB overlapping ChIP peaks in H1 cells. (B) Negative regions were taken from ± 5

kb from the center of each ATF3 and JUN overlapping ChIP peaks in K562 cells. (C) ATF3 ChIP-seq peaks after removing peaks that overlap with CEBPB in H1 cells. (D) CEBPB ChIP-seq peaks after removing peaks that overlap with ATF3 in H1 cells.

Figure 7. bZIP heterodimers and human diseases and traits

(A) CSI-Genomescape predicts increased binding by CREB1 to the alternate allele of rs10993994 and decreased binding to alternate alleles of rs7631605 and rs1869901 by JUN•ATF3 and FOS•CEBPE heterodimers, respectively. (B) Scatterplot of FOS•JUN predicted CSI intensities for reference and alternative alleles of 5076 autosomal SNPs linked to human diseases and quantitative traits identified in genome-wide association studies. SNPs and disease/traits classifications are from Maurano et al. (Maurano et al., 2012). (C) (left) Number of SNPs predicted to increase or decrease bZIP binding by 2-fold at different stringency levels determined by noise factor *F* (see Materials and methods). The F values at which a 2-fold difference in CSI score is predicted for rs12740374 (#) and rs10993994 (*) are indicated in red. (right) Distribution of predicted fold changes in bZIP binding for GWAS SNPs using CSI Intensities, using *F* = 25. Dashed lines mark a 2-fold change. Red lines indicate the predicted change in binding of CREB1 and CEBPA to rs10993994 (*) and rs12740374 (#). (D) Predicted fold-change in CSI score of sequences centered at SNPs linked to disease or quantitative traits. A total of 156 SNPs have a predicted increase (red) or decrease (blue) of ≥2-fold in CSI score for at least one bZIP dimer, when *F* = 25 (Materials and methods, and Supplementary file 1l). Fold-changes are relative to the reference genome hg19. Rows

(SNPs) are organized by class of disease/trait. Columns (bZIP dimers) are clustered by DNA specificity as in Figure 1.

Figure 7 – figure supplement 1.

Genomescapes, transcription factor binding and chromatin environment for selected SNPs. (A) Left, CSI Genomescape and right, UCSC genome browser screen shots of the genomic and chromatin context of SNPs rs12740374 and rs10993994. (B) UCSC genome browser screen shots of the genomic and chromatin context of SNPs rs3758354 and rs17293632. UCSC genome browser tracks for ChIP-seq peaks for selected bZIPs in ENCODE cell lines, ChIP-seq signal for histone 3 lysine 27 acetylation (H3K27Ac marks), and DNAse I hypersensitive regions.

Figure Supplements

Figure 1 – figure supplement 1. Cognate site identification by SELEX-sequencing.

Figure 1 –figure supplement 2. ATF3 CSI Intensity (z-score) correlates with equilibrium association constant.

Figure 1 – figure supplement 3. Pairwise comparison of bZIP homodimers reported in this study and bZIP dimers reported by Jolma et al. (Jolma et al., 2013).

Figure 1 –figure supplement 4. bZIP Heterodimer specificity.

Figure 1 – figure supplement 5. DNA sequence preferences for FOS•CEBPE, FOS•CEBPG, FOSL1•CEBPE and FOSL1•CEBPG.

Figure 3 –figure supplement 1. Influence of bZIP protein dimerization on DNA binding.

Figure 3 –figure supplement 2. Heterospecific binding of DNA.

Figure 5 –figure supplement 1. ROC curves.

Figure 6 – figure supplement 1. CSI intensities for bound genomic regions.

Figure 7 – figure supplement 1. Genomescapes, transcription factor binding and chromatin environment for selected SNPs.

Source data files

Figure 1-source data 1. Data for Figure 1C. Pairwise comparison (Pearson's correlation) of the DNA binding preferences of 102 bZIP dimers using the CSI intensity for 1222 10 bp sequences

Figure 2-source data 1. Data for Figure 2C. Relative CSI intensity for 102 bZIP dimers for different DNA binding sites and half-sites.

Source code files

3.6_SEL_bZIPs_2017_01_29_SEL10MER.zip

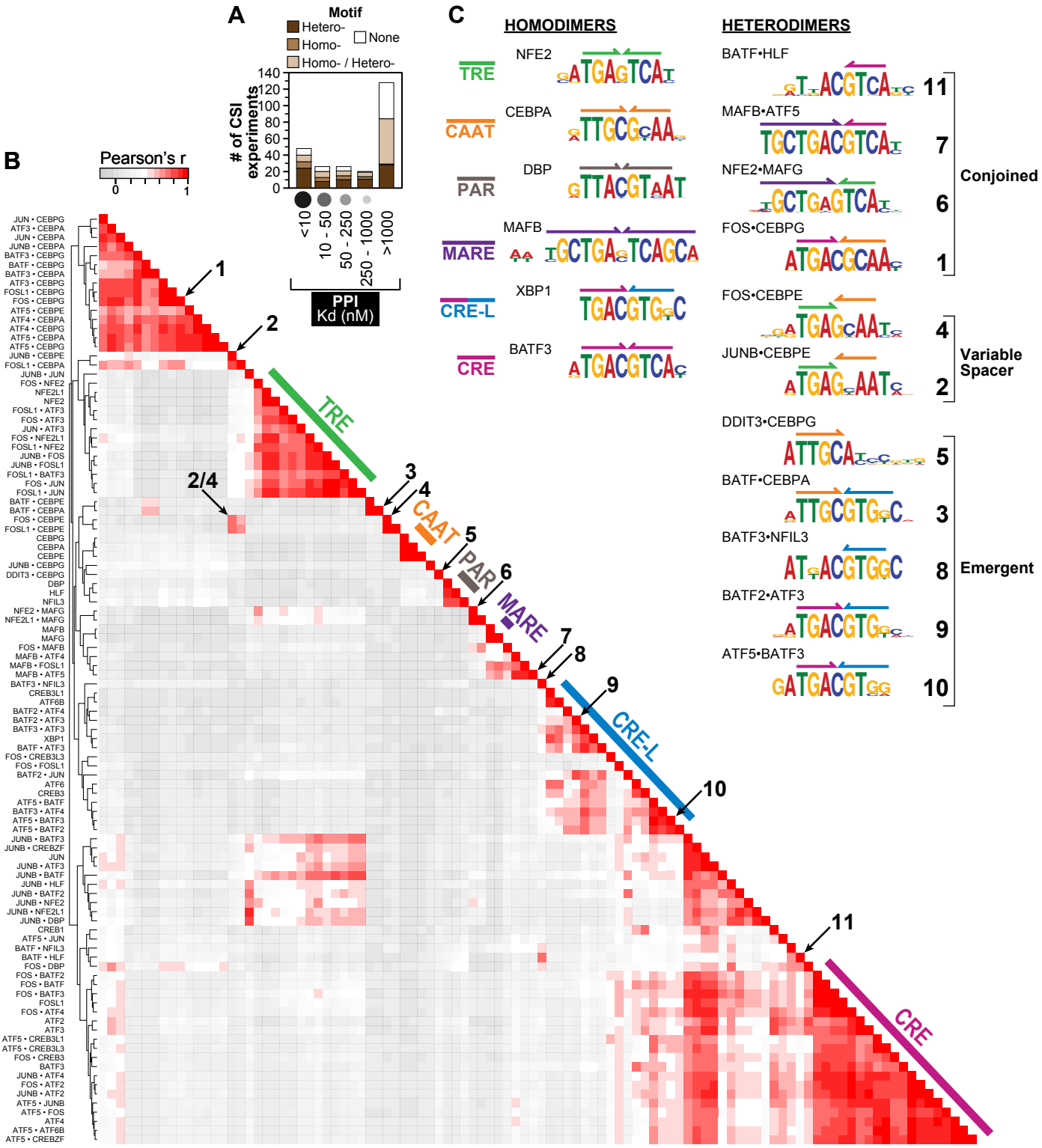
3.7_perl_bZIPs 2017_01_30_SEL12_14MER.zip

Supplementary Files

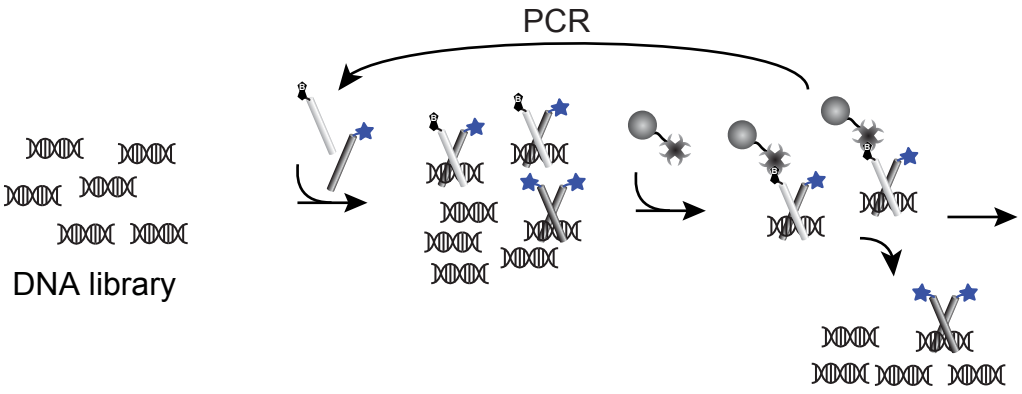
Supplementary File 1A-1I

Supplementary File 2. MEME motifs and Sequence Specificity and Energy Landscapes (SEL) for human bZIP homodimers and heterodimers.

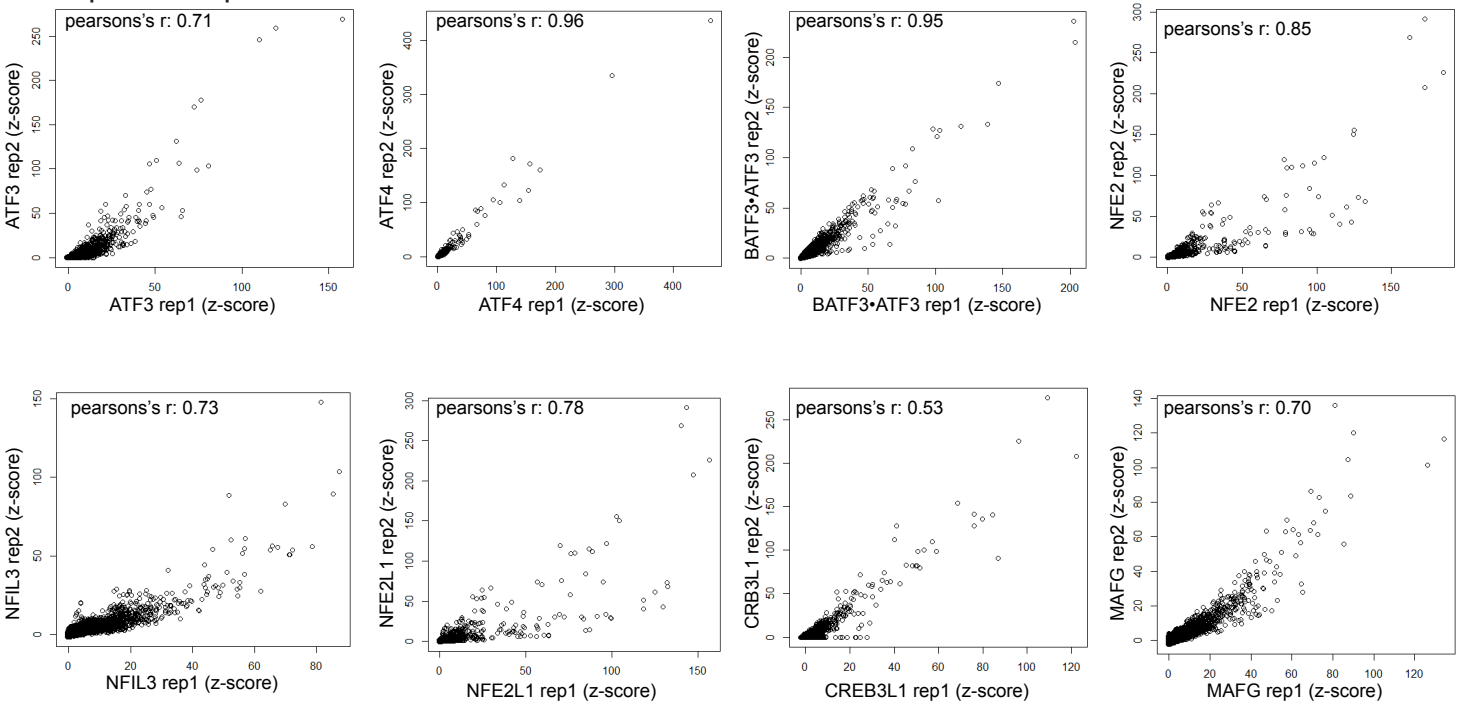
Figure 1


A
Motif
Hetero-
Homo-
Homo- / Hetero-
None
# of CSI experiments
<10
10 - 50
50 - 250
250 - 1000
>1000
PPI Kd (nM)
B
Pearson's r
C
HOMODIMERS
HETERODIMERS
TRE
NFE2
CAAT
CEBPA
PAR
DBP
MAFB
MARE
XBP1
CRE-L
BATF3
CRE
BATF•HLF
MAFB•ATF5
NFE2•MAFG
FOS•CEBPG
FOS•CEBPE
JUNB•CEBPE
DDIT3•CEBPG
BATF•CEBPA
BATF3•NFIL3
BATF2•ATF5
ATF5•BATF3
Conjoined
Variable Spacer
Emergent
1
2
2/4
3
4
5
6
7
8
9
10
11

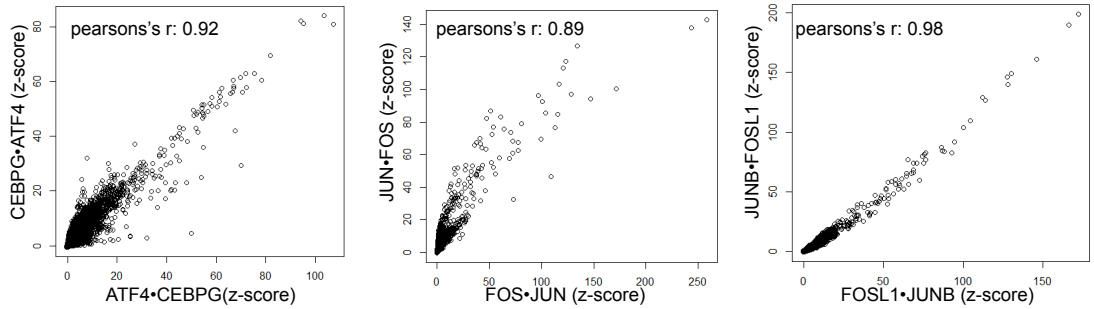
A Cognate Site Identification by SELEX-Seq



B Replicate experiments



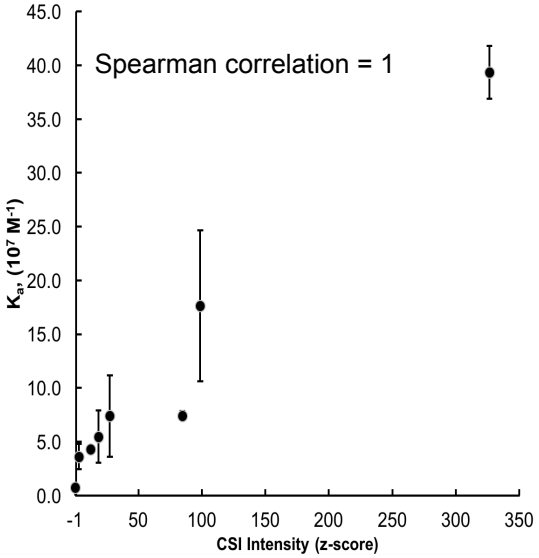
C Reciprocal experiments



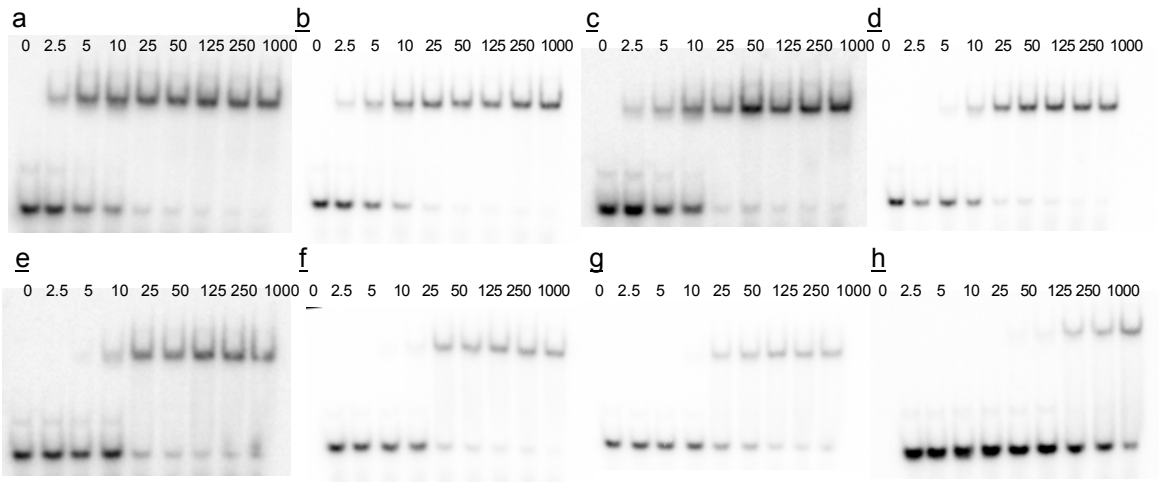
A

	Sequence	Ka, 10^7 M^{-1}
a	gagtccaGATGACGTCATCtccgtag	39.4 ± 3
b	gagtccaGATTACGTCATCtccgtag	17.6 ± 7
c	gagtcccCATGACGTCATGgccgtag	7.4 ± 0.1
d	gagtcccCATTACGTCATGgccgtag	7.4 ± 2
e	gagtccaGATGACGTCAAAaccgtag	5.5 ± 0.4
f	gagtccaGACGACGTCATCtccgtag	4.3 ± 2
g	gagtcccCACGACGTCATGgccgtag	3.6 ± 1
h	gagtcctTTTGACGTCAAAaccgtag	0.7 ± 0.1

B



C



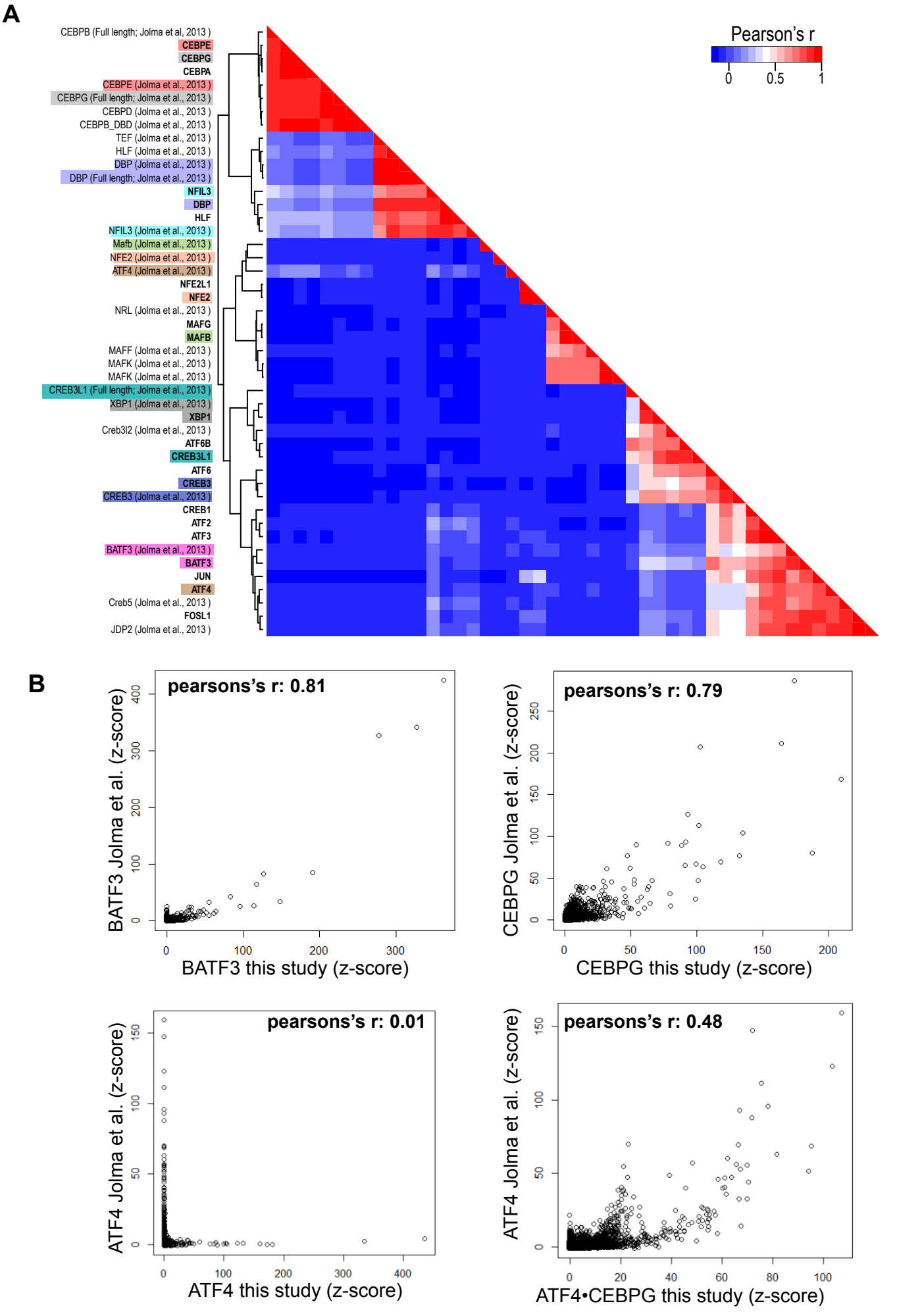
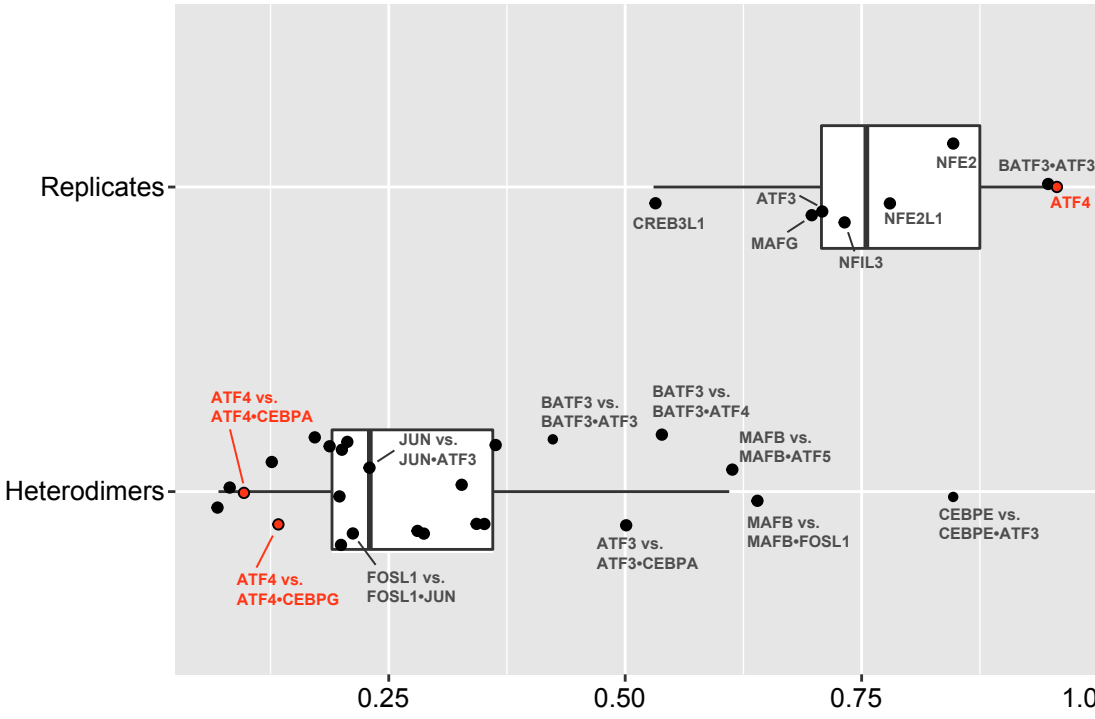

A
Pearson's r
0
0.5
1
CEBPB (Full length; Jolma et al, 2013)
CEBPE
CEBPG
CEBPA
CEBPE (Jolma et al., 2013)
CEBPG (Full length; Jolma et al., 2013)
CEBPD (Jolma et al., 2013)
CEBPB_DBD (Jolma et al., 2013)
TEF (Jolma et al., 2013)
HLF (Jolma et al., 2013)
DBP (Jolma et al., 2013)
DBP (Full length; Jolma et al., 2013)
NFIL3
DBP
HLF
NFIL3 (Jolma et al., 2013)
Mafb (Jolma et al., 2013)
NFE2 (Jolma et al., 2013)
ATF4 (Jolma et al., 2013)
NFE2L1
NFE2
NRL (Jolma et al., 2013)
MAFG
MAFB
MAFF (Jolma et al., 2013)
MAFK (Jolma et al., 2013)
MAFK (Jolma et al., 2013)
CREB3L1 (Full length; Jolma et al., 2013)
XBP1 (Jolma et al., 2013)
XBP1
Creb3l2 (Jolma et al., 2013)
ATF6B
CREB3L1
ATF6
CREB3
CREB3 (Jolma et al., 2013)
CREB1
ATF2
ATF3
BATF3 (Jolma et al., 2013)
BATF3
JUN
ATF4
Creb5 (Jolma et al., 2013)
FOSL1
JDP2 (Jolma et al., 2013)
B
pearsons's r: 0.81
BATF3 Jolma et al. (z-score)
BATF3 this study (z-score)
pearsons's r: 0.79
CEBPG Jolma et al. (z-score)
CEBPG this study (z-score)
pearsons's r: 0.01
ATF4 Jolma et al. (z-score)
ATF4 this study (z-score)
pearsons's r: 0.48
ATF4 Jolma et al. (z-score)
ATF4•CEBPG this study (z-score)

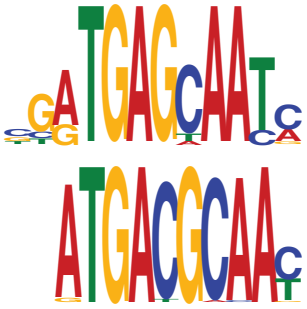
Figure 1-figure supplement 4

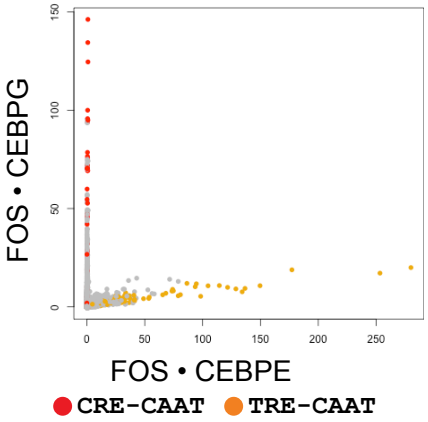


Comparison between FOS•CEBPE and FOS•CEBPG heterodimers

FOS•CEBPE
(239nM)

FOS•CEBPG
(26nM)

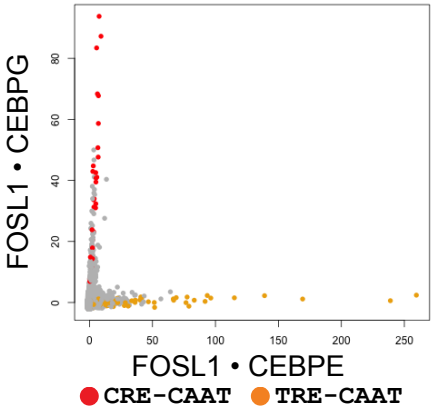


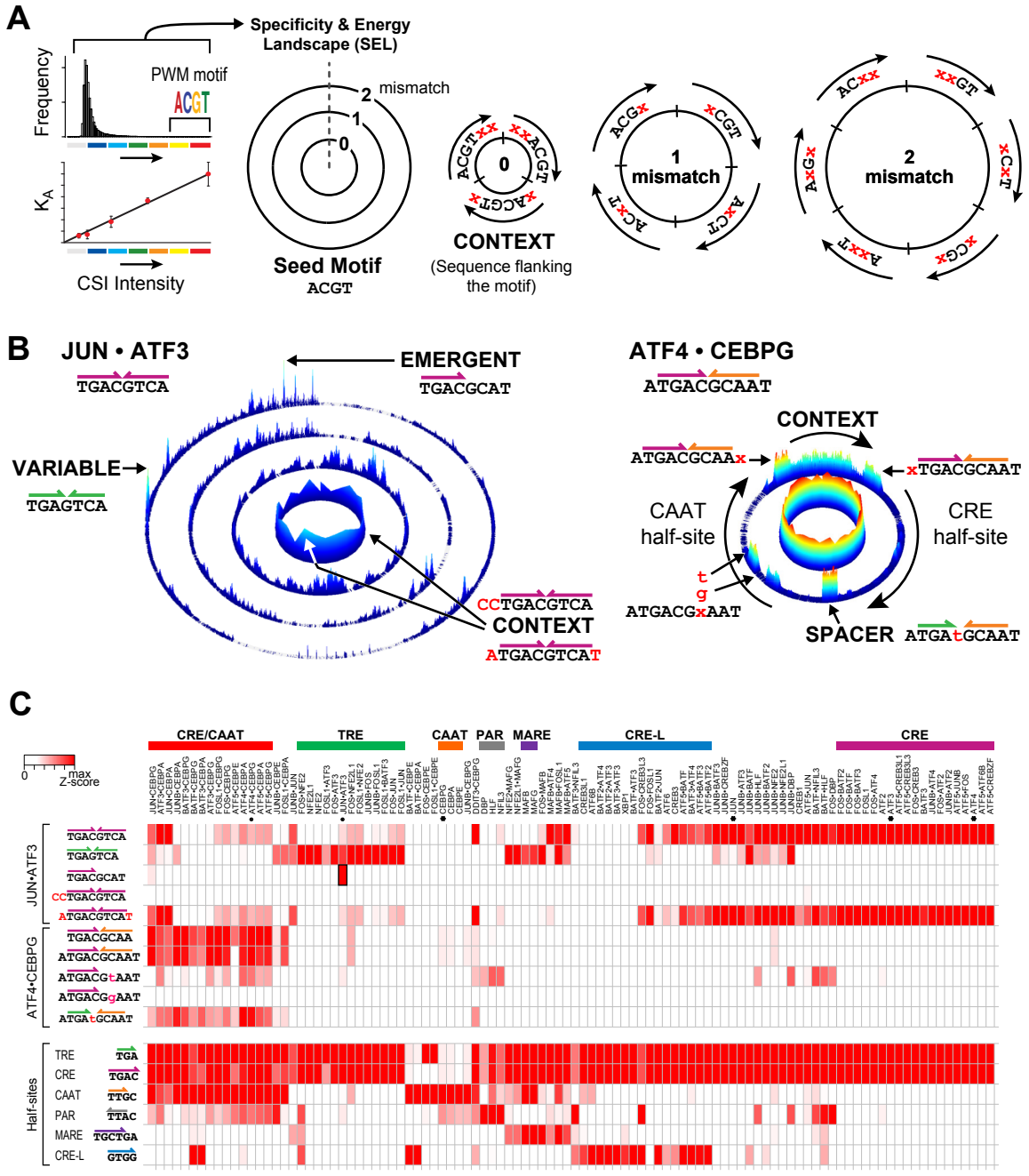


Comparison between FOSL1•CEBPE and FOSL1•CEBPG heterodimers

FOSL1•CEBPE
(266 nM)

FOSL1•CEBPG
(103 nM)




A
Specificity & Energy Landscape (SEL)
Frequency
PWM motif
ACGT
K_A
CSI Intensity
mismatch
2
1
0
Seed Motif
ACGT
CONTEXT
(Sequence flanking the motif)
1 mismatch
2 mismatch
B
JUN • ATF3
TGACGTCA
EMERGENT
TGACGCAT
VARIABLE
TGAGTCA
CCTGACGTCA
CONTEXT
ATGACGTCAT
ATF4 • CEBPG
ATGACGCAAT
CONTEXT
ATGACGCAAx
xTGACGCAAT
CAAT half-site
CRE half-site
t
g
ATGACGxAAT
SPACER
ATGAtGCAAT
C
CRE/CAAT
TRE
CAAT
PAR
MARE
CRE-L
CRE
max Z-score
JUN•ATF3
ATF4•CEBPG
Half-sites
TRE
CRE
CAAT
PAR
MARE
CRE-L
TGA
TGAC
TTGC
TTAC
TGCTGA
GTGG

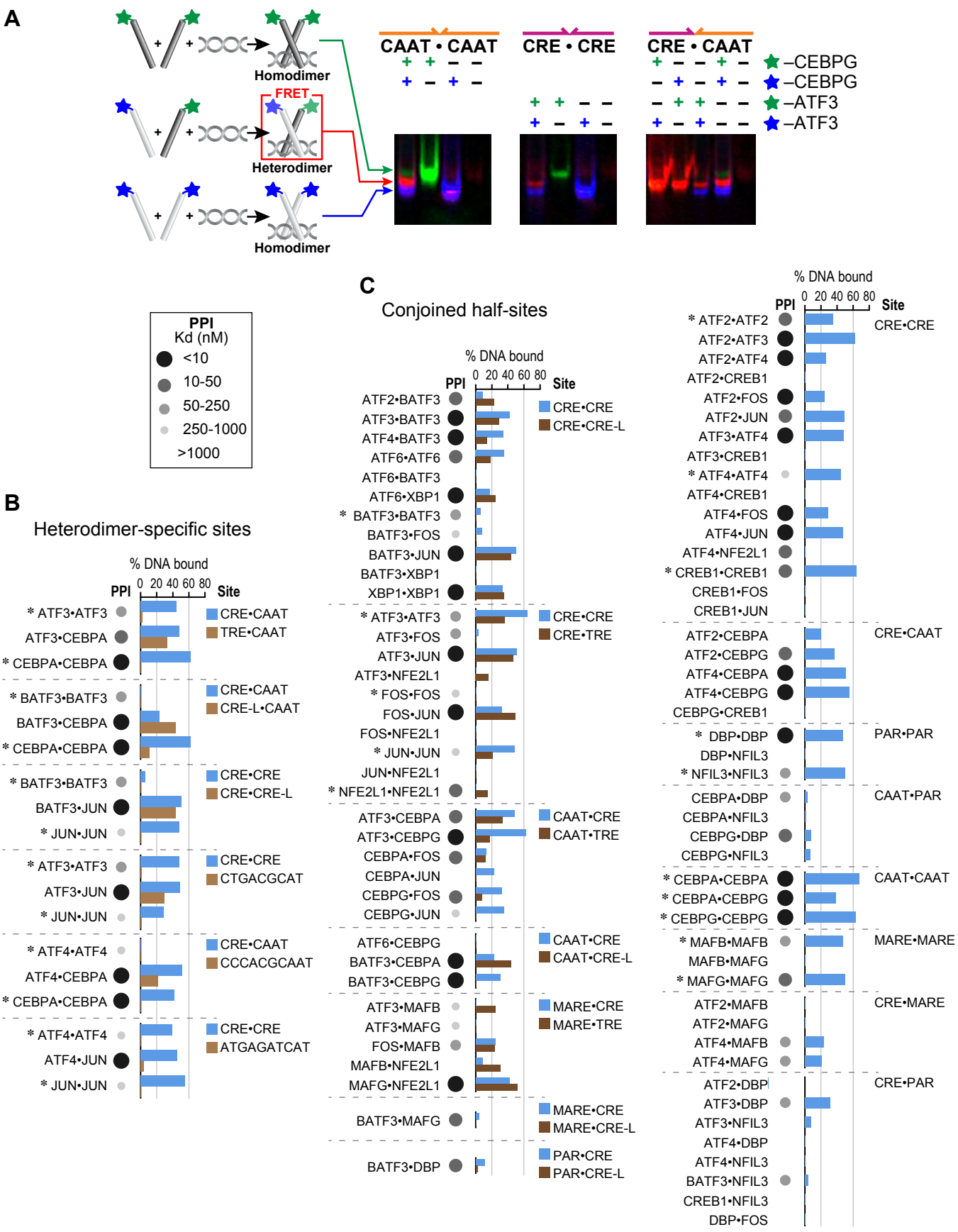

A
Homodimer
FRET
Heterodimer
Homodimer
CAAT • CAAT
CRE • CRE
CRE • CAAT
–CEBPG
–CEBPG
–ATF3
–ATF3
PPI
Kd (nM)
<10
10-50
50-250
250-1000
>1000
B
Heterodimer-specific sites
% DNA bound
0 20 40 60 80
PPI
Site
* ATF3•ATF3
ATF3•CEBPA
* CEBPA•CEBPA
CRE•CAAT
TRE•CAAT
* BATF3•BATF3
BATF3•CEBPA
* CEBPA•CEBPA
CRE•CAAT
CRE-L•CAAT
* BATF3•BATF3
BATF3•JUN
* JUN•JUN
CRE•CRE
CRE•CRE-L
* ATF3•ATF3
ATF3•JUN
* JUN•JUN
CRE•CRE
CTGACGCAT
* ATF4•ATF4
ATF4•CEBPA
* CEBPA•CEBPA
CRE•CAAT
CCCACGCAAT
* ATF4•ATF4
ATF4•JUN
* JUN•JUN
CRE•CRE
ATGAGATCAT
C
Conjoined half-sites
% DNA bound
0 20 40 60 80
PPI
Site
ATF2•BATF3
ATF3•BATF3
ATF4•BATF3
ATF6•ATF6
ATF6•BATF3
ATF6•XBP1
* BATF3•BATF3
BATF3•FOS
BATF3•JUN
BATF3•XBP1
XBP1•XBP1
CRE•CRE
CRE•CRE-L
* ATF3•ATF3
ATF3•FOS
ATF3•JUN
ATF3•NFE2L1
* FOS•FOS
FOS•JUN
FOS•NFE2L1
* JUN•JUN
JUN•NFE2L1
* NFE2L1•NFE2L1
CRE•CRE
CRE•TRE
ATF3•CEBPA
ATF3•CEBPG
CEBPA•FOS
CEBPA•JUN
CEBPG•FOS
CEBPG•JUN
CAAT•CRE
CAAT•TRE
ATF6•CEBPG
BATF3•CEBPA
BATF3•CEBPG
CAAT•CRE
CAAT•CRE-L
ATF3•MAFB
ATF3•MAFG
FOS•MAFB
MAFB•NFE2L1
MAFG•NFE2L1
MARE•CRE
MARE•TRE
BATF3•MAFG
MARE•CRE
MARE•CRE-L
BATF3•DBP
PAR•CRE
PAR•CRE-L
% DNA bound
0 20 40 60 80
PPI
Site
* ATF2•ATF2
ATF2•ATF3
ATF2•ATF4
ATF2•CREB1
ATF2•FOS
ATF2•JUN
ATF3•ATF4
ATF3•CREB1
* ATF4•ATF4
ATF4•CREB1
ATF4•FOS
ATF4•JUN
ATF4•NFE2L1
* CREB1•CREB1
CREB1•FOS
CREB1•JUN
CRE•CRE
ATF2•CEBPA
ATF2•CEBPG
ATF4•CEBPA
ATF4•CEBPG
CEBPG•CREB1
CRE•CAAT
* DBP•DBP
DBP•NFIL3
* NFIL3•NFIL3
PAR•PAR
CEBPA•DBP
CEBPA•NFIL3
CEBPG•DBP
CEBPG•NFIL3
CAAT•PAR
* CEBPA•CEBPA
* CEBPA•CEBPG
* CEBPG•CEBPG
CAAT•CAAT
* MAFB•MAFB
MAFB•MAFG
* MAFG•MAFG
MARE•MARE
ATF2•MAFB
ATF2•MAFG
ATF4•MAFB
ATF4•MAFG
CRE•MARE
ATF2•DBP
ATF3•DBP
ATF3•NFIL3
ATF4•DBP
ATF4•NFIL3
BATF3•NFIL3
CREB1•NFIL3
DBP•FOS
CRE•PAR

Figure 3-figure supplement 1

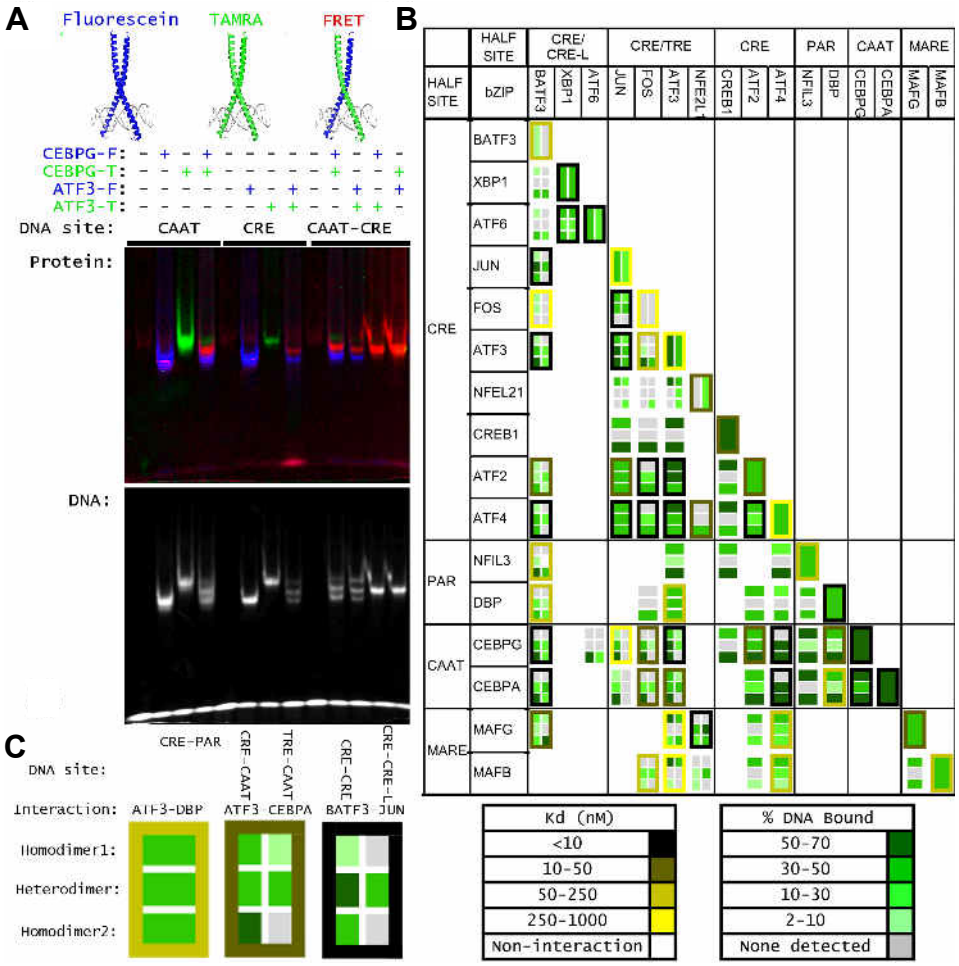
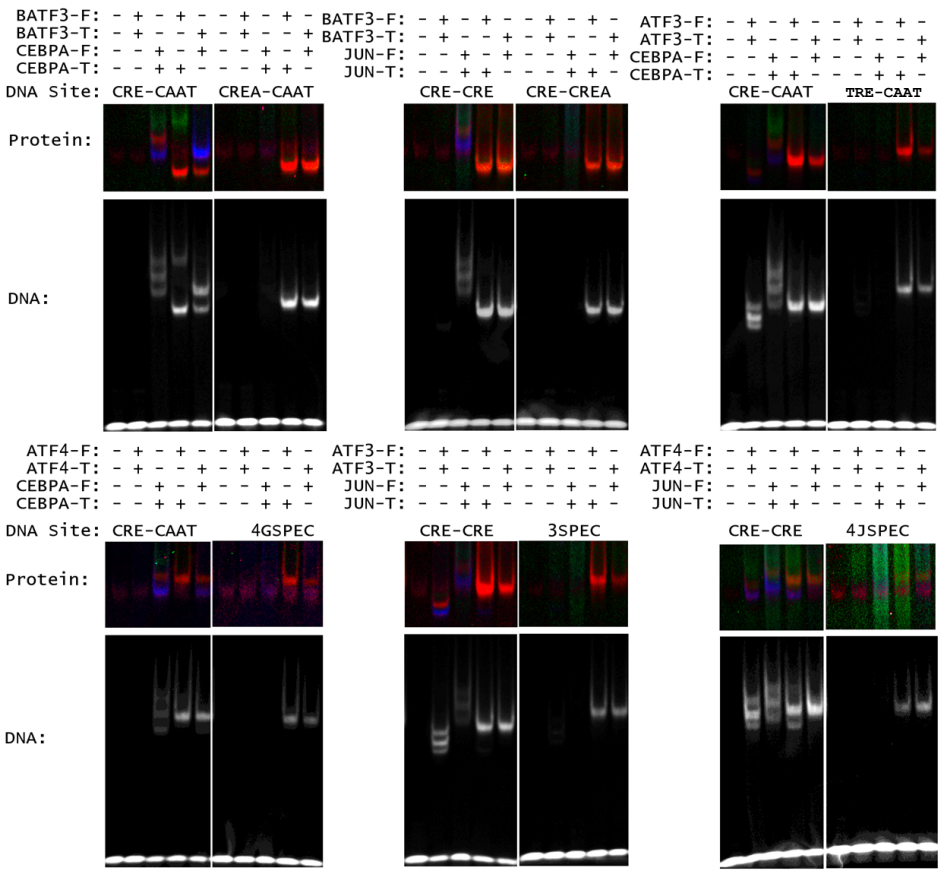
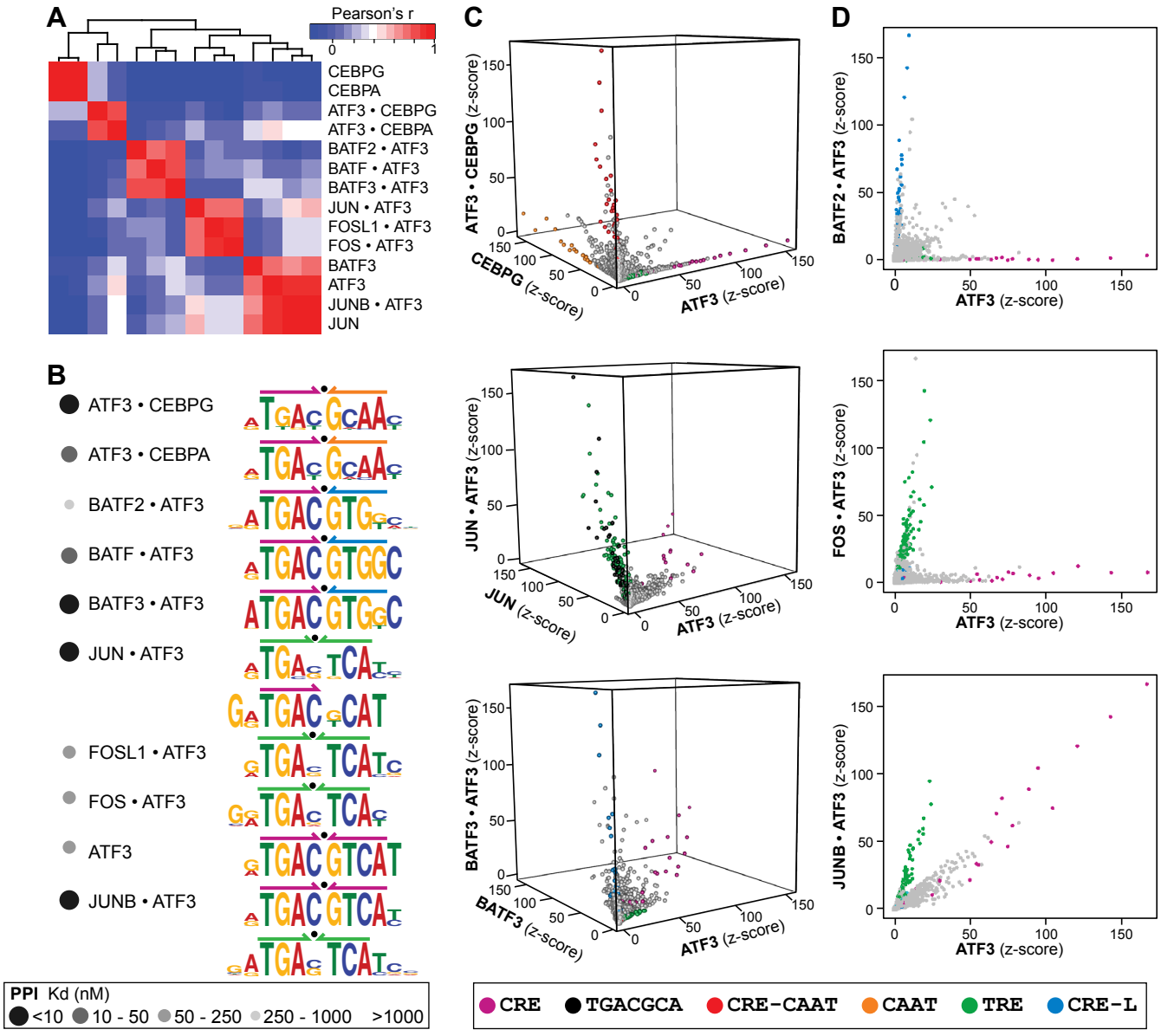
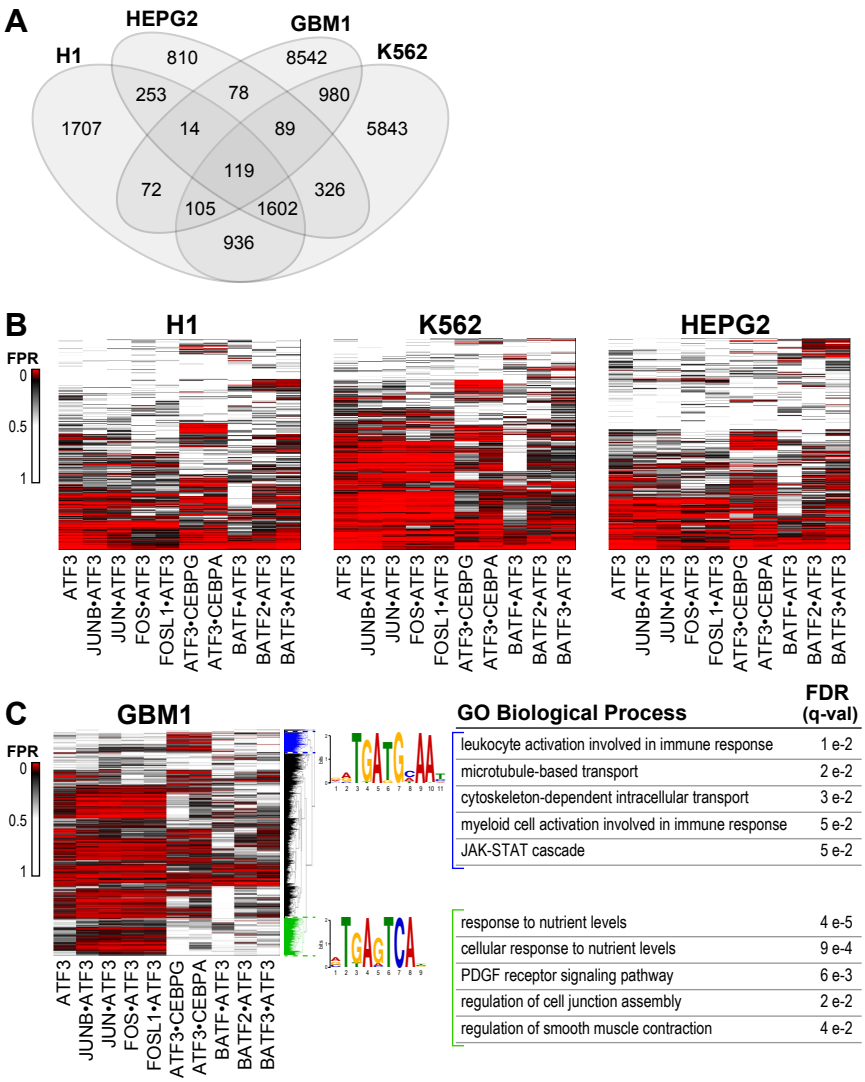
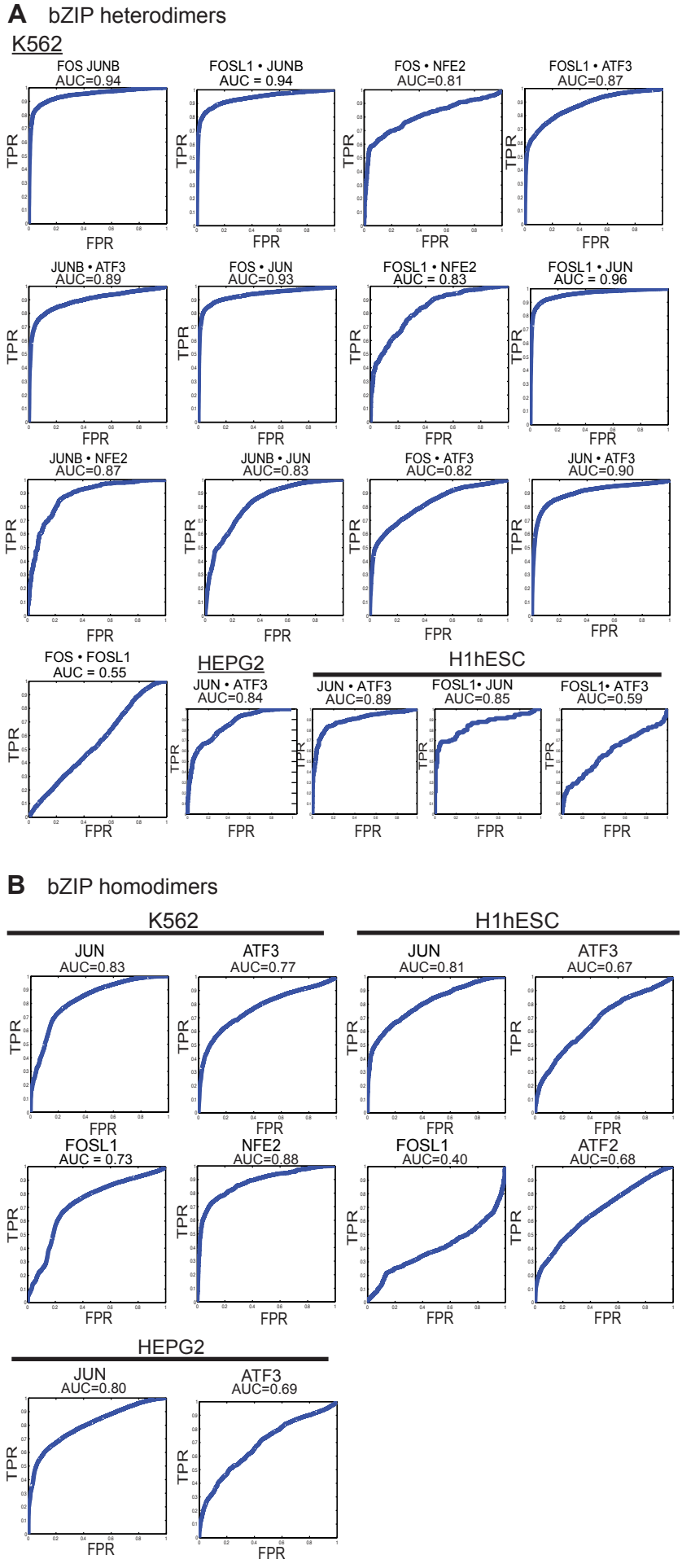


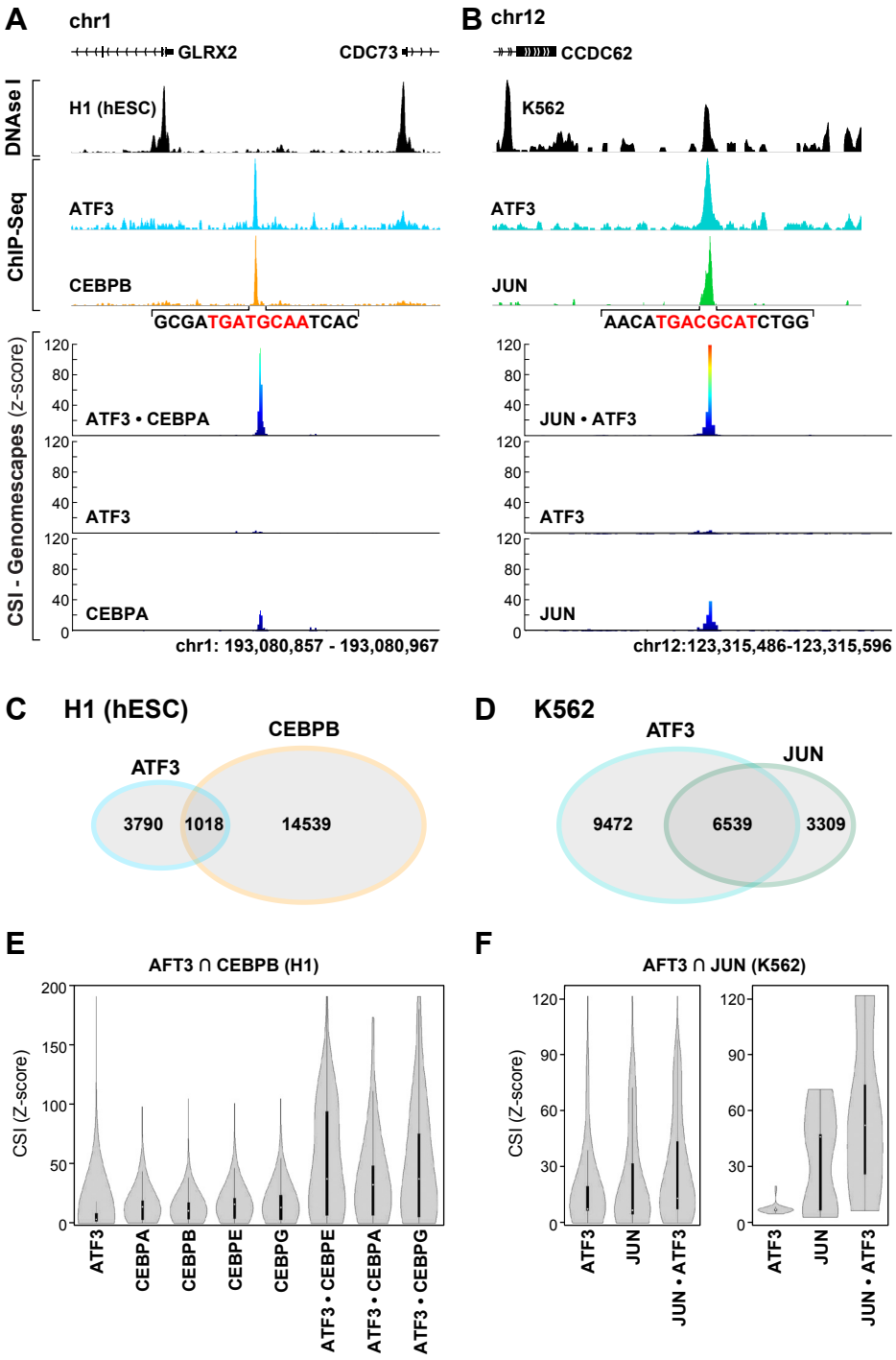
Figure 3-figure supplement 2

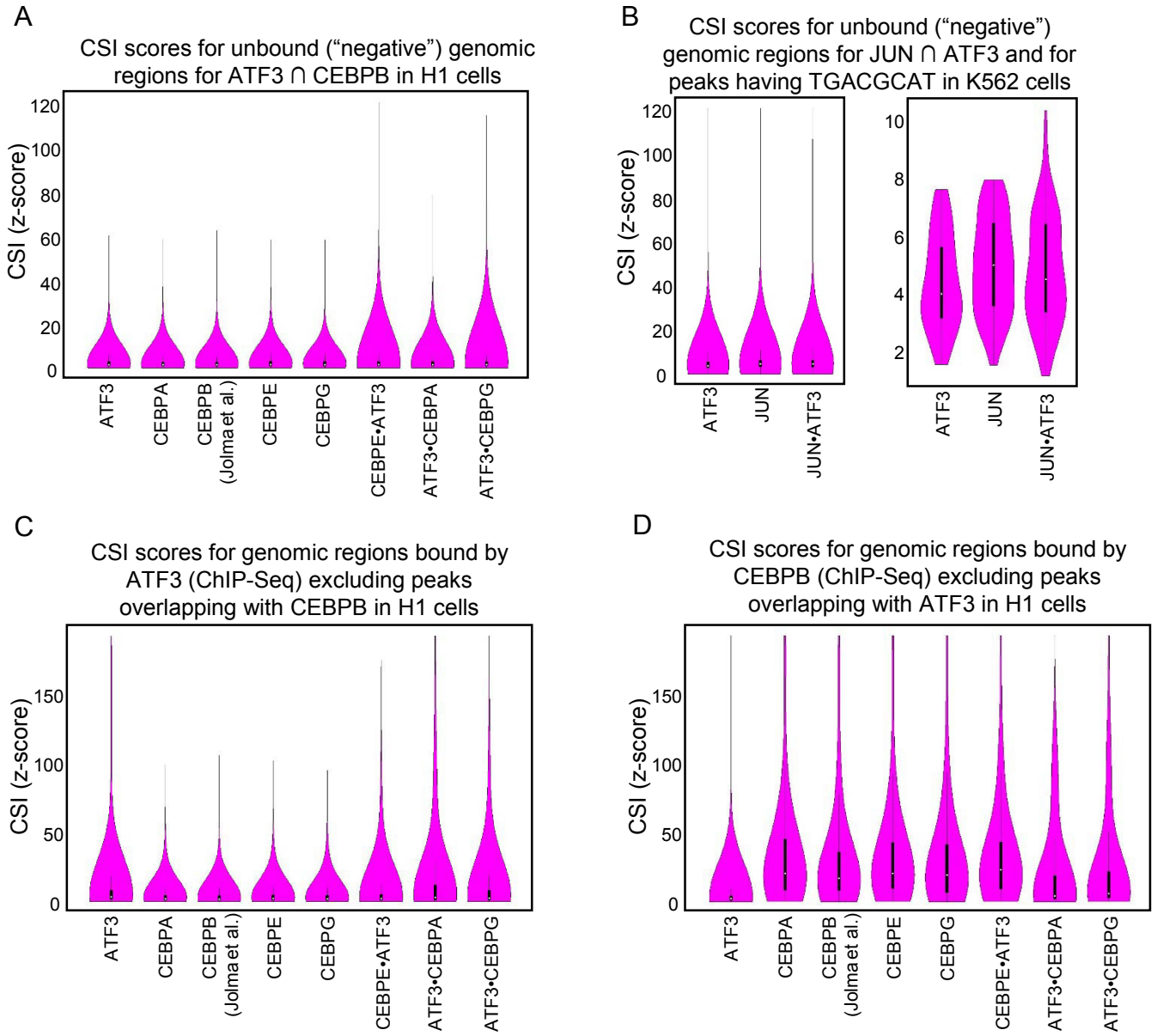



A
Pearson's r
0
1
CEBPG
CEBPA
ATF3 • CEBPG
ATF3 • CEBPA
BATF2 • ATF3
BATF • ATF3
BATF3 • ATF3
JUN • ATF3
FOSL1 • ATF3
FOS • ATF3
BATF3
ATF3
JUNB • ATF3
JUN
B
ATF3 • CEBPG
ATF3 • CEBPA
BATF2 • ATF3
BATF • ATF3
BATF3 • ATF3
JUN • ATF3
FOSL1 • ATF3
FOS • ATF3
ATF3
JUNB • ATF3
PPI Kd (nM)
<10
10 - 50
50 - 250
250 - 1000
>1000
C
ATF3 • CEBPG (z-score)
CEBPG (z-score)
ATF3 (z-score)
JUN • ATF3 (z-score)
JUN (z-score)
BATF3 • ATF3 (z-score)
BATF3 (z-score)
D
BATF2 • ATF3 (z-score)
FOS • ATF3 (z-score)
JUNB • ATF3 (z-score)
ATF3 (z-score)
CRE
TGACGCA
CRE-CAAT
CAAT
TRE
CRE-L


A
HEPG2
GBM1
H1
K562
810
8542
253
78
980
1707
14
89
5843
119
72
326
105
1602
936
B
H1
K562
HEPG2
FPR
0
0.5
1
ATF3
JUNB•ATF3
JUN•ATF3
FOS•ATF3
FOSL1•ATF3
ATF3•CEBPG
ATF3•CEBPA
BATF•ATF3
BATF2•ATF3
BATF3•ATF3
C
GBM1
FPR
0
0.5
1
GO Biological Process
FDR (q-val)
leukocyte activation involved in immune response 1 e-2
microtubule-based transport 2 e-2
cytoskeleton-dependent intracellular transport 3 e-2
myeloid cell activation involved in immune response 5 e-2
JAK-STAT cascade 5 e-2
response to nutrient levels 4 e-5
cellular response to nutrient levels 9 e-4
PDGF receptor signaling pathway 6 e-3
regulation of cell junction assembly 2 e-2
regulation of smooth muscle contraction 4 e-2


A bZIP heterodimers
K562
FOS JUNB
AUC=0.94
FOSL1 • JUNB
AUC = 0.94
FOS • NFE2
AUC=0.81
FOSL1 • ATF3
AUC=0.87
JUNB • ATF3
AUC=0.89
FOS • JUN
AUC=0.93
FOSL1 • NFE2
AUC = 0.83
FOSL1 • JUN
AUC = 0.96
JUNB • NFE2
AUC=0.87
JUNB • JUN
AUC=0.83
FOS • ATF3
AUC=0.82
JUN • ATF3
AUC=0.90
FOS • FOSL1
AUC = 0.55
HEPG2
JUN • ATF3
AUC=0.84
H1hESC
JUN • ATF3
AUC=0.89
FOSL1• JUN
AUC=0.85
FOSL1• ATF3
AUC=0.59
TPR
FPR
B bZIP homodimers
K562
JUN
AUC=0.83
ATF3
AUC=0.77
FOSL1
AUC = 0.73
NFE2
AUC=0.88
H1hESC
JUN
AUC=0.81
ATF3
AUC=0.67
FOSL1
AUC=0.40
ATF2
AUC=0.68
HEPG2
JUN
AUC=0.80
ATF3
AUC=0.69
TPR
FPR


A
chr1
GLRX2
CDC73
DNAse I
H1 (hESC)
ChIP-Seq
ATF3
CEBPB
GCGATGATGCAATCAC
CSI - Genomescapes (z-score)
ATF3 • CEBPA
ATF3
CEBPA
chr1: 193,080,857 - 193,080,967
B
chr12
CCDC62
K562
ATF3
JUN
AACATGACGCATCTGG
JUN • ATF3
ATF3
JUN
chr12:123,315,486-123,315,596
C
H1 (hESC)
CEBPB
ATF3
3790
1018
14539
D
K562
ATF3
JUN
9472
6539
3309
E
AFT3 ∩ CEBPB (H1)
CSI (Z-score)
ATF3
CEBPA
CEBPB
CEBPE
CEBPG
ATF3 • CEBPE
ATF3 • CEBPA
ATF3 • CEBPG
F
AFT3 ∩ JUN (K562)
CSI (Z-score)
ATF3
JUN
JUN • ATF3
ATF3
JUN
JUN • ATF3


A
CSI scores for unbound (“negative”) genomic regions for ATF3 ∩ CEBPB in H1 cells
CSI (z-score)
0
20
40
60
80
100
120
ATF3
CEBPA
CEBPB (Jolma et al.)
CEBPE
CEBPG
CEBPE•ATF3
ATF3•CEBPA
ATF3•CEBPG
B
CSI scores for unbound (“negative”) genomic regions for JUN ∩ ATF3 and for peaks having TGACGCAT in K562 cells
CSI (z-score)
0
20
40
60
80
100
120
ATF3
JUN
JUN•ATF3
2
4
6
8
10
ATF3
JUN
JUN•ATF3
C
CSI scores for genomic regions bound by ATF3 (ChIP-Seq) excluding peaks overlapping with CEBPB in H1 cells
CSI (z-score)
0
50
100
150
ATF3
CEBPA
CEBPB (Jolma et al.)
CEBPE
CEBPG
CEBPE•ATF3
ATF3•CEBPA
ATF3•CEBPG
D
CSI scores for genomic regions bound by CEBPB (ChIP-Seq) excluding peaks overlapping with ATF3 in H1 cells
CSI (z-score)
0
50
100
150
ATF3
CEBPA
CEBPB (Jolma et al.)
CEBPE
CEBPG
CEBPE•ATF3
ATF3•CEBPA
ATF3•CEBPG

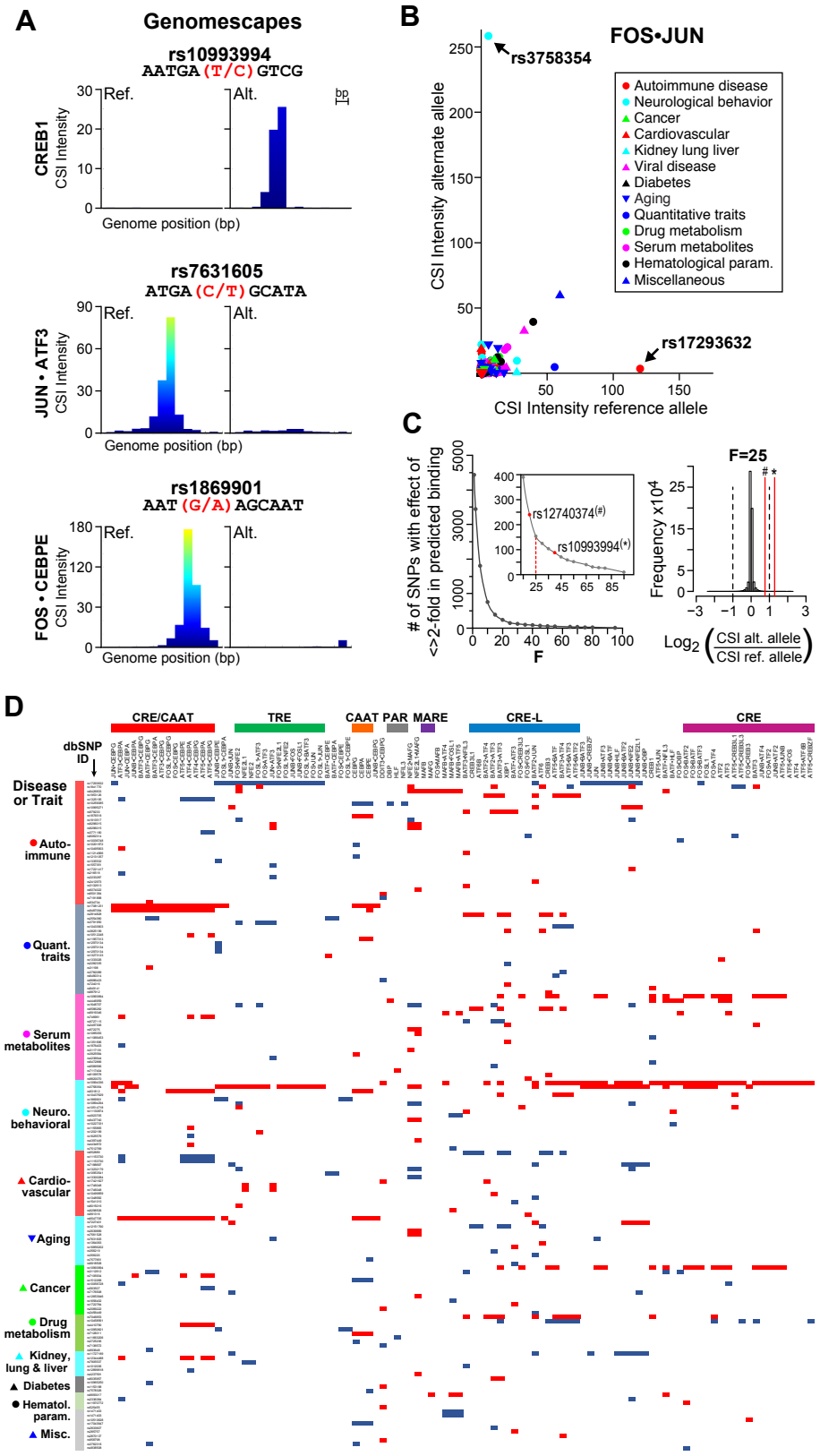

A
Genomescapes
rs10993994
AATGA (T/C) GTCG
CREB1
CSI Intensity
Ref.
Alt.
bp
Genome position (bp)
rs7631605
ATGA (C/T) GCATA
JUN • ATF3
rs1869901
AAT (G/A) AGCAAT
FOS • CEBPE
B
FOS•JUN
rs3758354
rs17293632
Autoimmune disease
Neurological behavior
Cancer
Cardiovascular
Kidney lung liver
Viral disease
Diabetes
Aging
Quantitative traits
Drug metabolism
Serum metabolites
Hematological param.
Miscellaneous
CSI Intensity alternate allele
CSI Intensity reference allele
C
# of SNPs with effect of <>2-fold in predicted binding
F
rs12740374(#)
rs10993994(*)
F=25
Frequency x10^4
Log2 (CSI alt. allele / CSI ref. allele)
D
CRE/CAAT
TRE
CAAT PAR MARE
CRE-L
CRE
dbSNP ID
Disease or Trait
Auto-immune
Quant. traits
Serum metabolites
Neuro. behavioral
Cardio-vascular
Aging
Cancer
Drug metabolism
Kidney, lung & liver
Diabetes
Hematol. param.
Misc.

Figure 7-figure supplement 1