Transcription factors recognize DNA shape without nucleotide recognition

Md. Abul Hassan Samee[1], Benoit G. Bruneau[1,2], Katherine S. Pollard [1,3]

1: Gladstone Institutes, 1650 Owens St., San Francisco, CA 94158

2: Department of Pediatrics and Cardiovascular Research Institute, University of California, San Francisco, CA 94158

3: Division of Bioinformatics, Institute for Human Genetics, and Institute for Computational Health Sciences, University of California, San Francisco, CA 94158

Abstract

We hypothesized that transcription factors (TFs) recognize DNA shape without nucleotide sequence recognition. Motivating an independent role for shape, many TF binding sites lack a sequence-motif, DNA shape adds specificity to sequence-motifs, and different sequences can encode similar shapes. We therefore asked if binding sites of a TF are enriched for specific patterns of DNA shape-features, *e.g.*, helical twist. We developed ShapeMF, which discovers these shape-motifs *de novo* without taking sequence information into account. We find that most TFs assayed in ENCODE have shape-motifs and bind regulatory regions recognizing shape-motifs in the absence of sequence-motifs. When shape- and sequence-recognition co-occur, the two types of motifs can be overlapping, flanking, or separated by consistent spacing. Shape-motifs are prevalent in regions co-bound by multiple TFs. Finally, TFs with identical sequence motifs have different shape-motifs, explaining their binding at distinct locations. These results establish shape-motifs as drivers of TF-DNA recognition complementary to sequence-motifs.

Introduction

Diverse cellular processes, including gene regulation, chromatin organization, provirus activity, and DNA replication, depend upon proteins binding to specific genome sites, either alone or in complexes with other molecules. Protein-DNA recognition is thus fundamentally important and critically informs studies of development, disease, and evolution. Protein bound regions can be measured in living cells via imaging and genomic techniques, such as chromatin immunoprecipitation followed by sequencing (ChIP-Seq). To pinpoint binding sites within bound regions, predict binding in the absence of experimental measurements, and shed light on binding specificity, a variety of *in vitro* binding affinity assays and sequence analyses have been deployed.

The specificity of protein-DNA recognition is commonly approached as a problem of discriminating bound nucleotide sequences from other sequences (Bailey 2011, Arvey, Agius et al. 2012, Ghandi, Lee et al. 2014, Setty and Leslie 2015). While most methods focus on enriched patterns of nucleotides, the resulting sequence-motifs become more predictive of *in vitro* and *in vivo* binding when supplemented with structural data based on shape features of the bound DNA (Zhou, Shen et al. 2015, Mathelier, Xin et al. 2016). However, the role of DNA shape in binding specificity has not been investigated outside the context of sequence-motifs. Supporting the hypothesis that DNA-binding proteins can recognize DNA structure without nucleotide recognition, transcription factors (TFs) can bind sequences that do not match sequence-motifs identified using sequence-based searches (von Hippel, Revzin et al. 1974, von Hippel and Berg 1986, Berg and von Hippel 1987, von Hippel 2007, Wang, Zhuang et al. 2012, Yip, Cheng et al. 2012, Afek, Schipper et al. 2014, Slattery, Zhou et al. 2014). Furthermore, the best sequence-based discriminative methods fail to identify a subset of regions bound by TFs, including validated and predicted regulatory elements and regions harboring polymorphisms that correlate with gene expression. Finally, different nucleotides can encode similar DNA structure, so shape features have the potential to be complementary to nucleotide features (Garvie and Wolberger 2001). We therefore sought to investigate the independent role of DNA shape in protein-DNA binding.

To explore how frequently proteins recognize DNA structure and test the idea that structural specificity can occur in the absence of nucleotide recognition, we pose a novel question: are the binding sites of a protein enriched for characteristic patterns of DNA shape features, such as roll, helical twist, propeller twist, and minor groove width? If so these "shape-motifs" might explain many aspects of DNA recognition that are not accounted for by sequence-motifs alone or by DNA shape profiles within sequence-motifs. These open problems include binding to regions that lack sequence-motifs, differential binding of proteins that have very similar sequence-motifs, and low sequence information content positions in and flanking sequence-motifs. To investigate these questions, we developed a novel Gibbs sampling algorithm to discover shape-motifs *de novo* without

conditioning on the presence of a sequence-motif. Applying this method to more than 100 TFs and several different cell types, we find that most TFs recognize DNA shape independent of nucleotide recognition.

Results

Many strongly bound regions lack a sequence motif

To motivate the need for shape-motifs, we first quantified the prevalence of TF binding without nucleotide recognition. We called sequence motif hits for 110 TFs from regions they bind in human ENCODE data from K562 (chronic myelogenous leukemia) and Gm12878 (lymphoblastoid) cell lines. To broadly define sequence motifs for each TF, we used a collection of position weight matrices (PWMs) from JASPAR (Sandelin, Alkema et al. 2004), TRANSFAC (Matys, Fricke et al. 2003), *in vitro* studies (Berger, Philippakis et al. 2006, Berger, Badis et al. 2008, Badis, Berger et al. 2009, Jolma, Yan et al. 2013), and *de novo* motif-discovery methods, as were curated in (Kheradpour and Kellis 2014), plus up to five *de novo* motifs per TF that we learned from its top 2000 peaks using gkmSVM (Ghandi, Lee et al. 2014). Combining hits to any of these PWMs, we found that a large fraction of the top 2000 peaks for each TF lack a sequence motif for that TF (Table S1). For a typical TF, 29.6% of peaks have no sequence motif. This fraction varies significantly across TFs (range = 0.4–98%), but is fairly consistent between cell lines. These findings show that factors other than direct nucleotide recognition are likely influencing TF DNA recognition at many sites in the human genome.

An algorithm to discover variable-length shape-motifs *de novo* from unaligned genomic regions

We hypothesized that one reason TFs can bind to regions with no evident sequence motifs is their ability to recognize DNA shape independent of the underlying nucleotide sequence. Since different sequences can encode the same values of a DNA structural feature, shape recognition might occur in the absence of sequence motifs. To explore this idea, we first defined the concept of a TF shape-motif, which is a significantly over-represented *pattern* in the profiles of DNA shape features at the TF's binding sites as compared to non-bound regions (Fig. 1, see Methods). For instance, we would say that a TF has a minor groove width shape-motif if its binding sites are enriched for windows with a particular sequence of minor groove values (e.g., low, high, low) compared to flanking non-peaks. This definition is based directly on the shape feature and is not conditional on the presence of particular nucleotides. It therefore enables us to study shape preferences independent of sequence preferences, which has not been done in previous analyses of DNA shape in the context of TF binding.

Our two-step algorithm to find shape-motifs is called *ShapeMF* (Fig. 1). It extends a typical approach to sequence motif discovery to instead identify profiles of quantitative DNA shape feature values. The resulting shape-motifs retain all the constraints that sequence-motifs satisfy (see Methods). Before searching for shape-motifs, we translate DNA sequences of bound regions and matched unbound regions into a vector of shape features at each nucleotide. The shape features we utilize are helical twist (HelT), minor groove width (MGW), propeller twist (ProT), and roll (Roll), which are estimated from molecular dynamics simulations and available from the GBshape database (Chiu, Yang et al. 2015). These could easily be extended to include additional structural features. Our algorithm operates directly on the quantitative shape values for each site and does not use DNA sequence in any other way. Different nucleotide sequences can encode similar shape, so an enriched shape pattern in bound regions need not correspond to an enriched sequence pattern.

In the first step of the algorithm, we apply Gibbs sampling to compute local alignments of windows from shape-profiles of the bound regions (one per region) without using the unbound regions. The second step uses this alignment to compute a subset of windows whose shape defines a pattern that significantly discriminates bound from unbound regions. Different sets of unbound regions can be used to identify shape-motifs that are discriminative in different contexts. By repeating the search with different window sizes, the algorithm can identify variable-length shape-motifs. The final output includes only the non-redundant motifs.

ShapeMF is implemented in Python and freely available at: https://github.com/h-samee/shape-motif. The software takes as input sets of bound and unbound regions and outputs enriched motif profiles with p-values.

Most TFs have at least one shape-motif

We applied ShapeMF to discover shape-motifs in the K562 and Gm12878 ChIP-Seq peaks of 110 TFs (201 ChIP-Seq datasets). For each TF in each cell line, we took the strongest 2000 peaks and extracted 100 base-pair (bp) windows centered at the peak-summit (i.e., the location of maximum ChIP intensity) as bound regions and flanking 100-bp sequences sampled 200-bp away as unbound regions (see Methods). We found that >80% of the TFs in each cell-line have a shape-motif (71/84 TFs in K562, 50/63 TFs in Gm12878; Bonferroni-adjusted p-value $< 10^{-5}$; median false positive rate: 0.16–0.19) (Fig. 2A). Few TFs have shape-motifs for all features. However, most TFs have a ProT-motif, while Roll-motifs are much less common (Fig. 2B). We found shape-motifs up to 29 bp long, with an average length of 15 bp (Fig. 2C), which is much longer than a typical sequence motif (6–10 bp) (Stewart, Hannenhalli et al. 2012). In the following analyses, we focus only on the TFs for which we found shape-motifs.

Shape-motifs are prevalent and distinct from sequence-motifs

We identified all significant sequence matches to each TF's shape-motifs and sequence-motifs within its genome-wide ChIP-seq peaks in each cell line (see Methods). While we used all five gkmSVM sequence motifs for the TF, we allowed at most one shape-motif per feature, thus using at most four shape-motifs per TF in each cell line. As such, our estimates of shape-motif prevalence are likely conservative. We refer to shape-motif occurrences as *shape-sites*, sequence-motif occurrences as *sequence-sites*, and overlapping occurrences of both types of motifs as *overlapping-sites*.

This analysis identified thousands of shape-sites across the human genome, with a typical TF having 1.56 shape-sites per peak (range: 1.03-5.18 per peak), compared to 2.60 sequence-sites (range: 1.08-4.51), and 3.12 overlapping-sites (range: 1.02-6.61) (Table S2). The higher rate of sequence-sites and overlapping-sites could be driven in part by our being more conservative in calling shape-motifs than sequence-motifs. The amount of TF binding associated with each type of motif is explored below. Both sequence-sites and shape-sites are more prevalent in the top peaks compared to less significant peaks or all called peaks (Fig. S1). Shape-sites (Fig. 2D) generally occur within $\pm$30 bps of the ChIP-Seq peak-summit, as do sequence-sites and overlapping-sites. However, most shape-sites do not overlap a sequence-site (mean: 60.16%, range: 16.36-99.16%), though overlapping-sites are common for certain TFs, such as USF1, USF2, BHLHE40, CTCFL, EGR1, FOS, and YY1. The GC-content of shape-sites is very similar to that of sequence-sites (66.3% vs. 66.8% in K562, 63.8% vs. 63.4% in Gm12878) and is consistent with a previous analysis of sequence-motif hits in these cell-lines (Wang, Zhuang et al. 2012). Interestingly, overlapping-sites have particularly high GC-content (72% in K562, 69.8% in Gm12878).

Next, we built sequence logos from the shape-sites of each TF and compared these to its sequence-motif logo. This revealed that most shape-motifs are not sequence-specific (Table S3). Their average information content is less than half that of sequence-motifs (6.98 bits versus 15.58 bits for sequence-motifs in TRANSFAC (Matys, Fricke et al. 2003)). Similarly, average shape-motif information content per position is only 0.63 bits, compared to 1.27 bits for TRANSFAC sequence-motifs. The different shape features varied remarkably in their sequence specificity, especially when we consider motifs of different features based on information content per position (ICP): the maximum ICP of an MGW-motif and the minimum ICP of a Roll-motif occur at around the same value of ~0.5 bits, whereas the ICP of HelT-motifs is centered around ~0.5 bits and of ProT-motifs is more uniformly spread within 0.1-1.2 bits (Fig. S2). Across shape features, TFs differed in the information content of their shape-motifs by 3.15-fold. The TFs whose shape-motifs had high sequence information content were largely the same as those with the highest number of overlapping-sites (Fig. S3). Thus, sequence-motifs can encode DNA shape, but shape-motifs frequently occur without a consistent DNA sequence pattern.

TFs recognize DNA shape of regulatory elements in the absence of sequence motifs

We next examined ChIP-Seq peaks to determine if TF binding at each genomic location is determined by shape-sites, sequence-sites, or a combination. Peaks fall into four categories: (i) *sequence-only*, (ii) *co-occurrence*, (iii) *shape-only*, and (iv) *no motif occurrence* (Fig. 3A). While TFs vary in the proportion of peaks with shape-motifs, shape-recognition is very prevalent. Excluding TFs with no motifs of either type, the majority of peaks for most TFs have a shape-site. For a typical TF, co-occurrence peaks are the most common category, followed by shape-only peaks, with sequence-only peaks being the least common. On average, ~20% of genome-wide peaks in a ChIP-Seq dataset are shape-only. For ~60% and ~50% TFs in the K562 and Gm12878 cell-lines respectively, more than 10% of genome-wide peaks are shape-only (Fig. 3B). The TFs with the most shape-only peaks include EP300, BCL3, MYC, and PAX5. These results demonstrate that most TFs can bind DNA through shape recognition without underlying nucleotide specificity.

To more rigorously test whether shape-specificity is present in the absence of sequence-specificity, we repeated shape-motif discovery on the set of peaks where we could not find any occurrence of a gkmSVM motif. For 58/84 TFs in the K562 cell-line and 37/63 TFs in the Gm12878 cell-line, we found a shape-motif (Fig. 3C). Importantly, 65% of the shape-motifs discovered using peaks without sequence-motifs were also found using the top peaks. Of these, ~80% are similar to the motifs discovered from the top peaks (Fig. 3D). Altogether this analysis supports the conclusion that shape-specificity commonly occurs independent from nucleotide recognition and is a potential explanation to TF binding in regions that lack a sequence-motif.

To explore the functional role of TF shape recognition, we checked how likely a shape-only peak is to occur in putative regulatory regions as compared to a peak that contains a sequence motif (*sequence-based,* i.e., sequence-only or co-occurrence peak). We first considered the active regulatory categories in ENCODE segmentation predictions (ChromHMM and Segway combined) (Ernst and Kellis 2010, Ernst and Kellis 2012, Hoffman, Buske et al. 2012). Shape-only peaks are more likely than sequence-based peaks to occur in enhancers, weak enhancers, and TSS (Fig. 3E). We discovered the same trend upon analyzing enhancers from EnhancerAtlas (Gao, He et al. 2016), ENCODE FAIRE-Seq regions (Consortium 2012), regions in Hi-C contacts (Rao, Huntley et al. 2014), and promoter-proximal and -distal regions (see Methods). These results strongly suggest a role for shape-recognition in functional regulatory elements.

Shape-specificity is complementary to sequence-specificity

Because co-occurrence peaks are prevalent for most TFs, we sought to understand the relationship between shape-recognition and sequence-recognition within these peaks. We

therefore analyzed the spacing of shape-sites and sequence-sites for each TF to find conserved patterns (Fig. 4A-F, Fig. S4; see Methods). Many pairs of shape- and sequence-motifs lack any preference for occurring within, surrounding, or neighboring each other. However, 39% of shape-sites completely contain a sequence-site for the same TF with shape information encoded by and on both sides of the sequence-motif instance (Fig. 4A). Some TFs representative of this trend are EFOS, ATF3, EGR1, ETS1, and ELF1. The converse also occurs, but less frequently: 12% of shape-sites are completely contained within a sequence-site (flanked on both sides by the edges of the sequence-motif) (Fig. 4B,C). Examples include motifs for ATF1, GATA2, MAX, NR2F2, and SPI1. Thus, shape-motifs explain binding to low sequence information content positions within and flanking corresponding sequence-motifs.

We also observed many cases of side-by-side sequence-sites and shape-sites. The flanking shape-sites occur both upstream (48% of motif pairs) (Fig. 4D) and downstream (45%) (Fig. 4E) of sequence-sites, defined as up to 21 bp away with overlap up to 9 bp. In 2% of motif-pairs, the sequence- and shape-motifs do not overlap (Fig. 4F). These often occur with a conserved inter-motif spacing, which can be up to 16 bp. Examples of TFs with conserved spacing include CTCF, CTCFL, NRSF, MAX, MAFF, and ATF1. Some of these cases correspond to hetero- or homodimer binding motifs. For example, a MAFF MGW-motif encodes a half-site corresponding to the MAFF recognition motif TGCTGA (Yoshida, Ohkumo et al. 2005) (Fig. 4G), and a HelT-motif encodes the TGAGTCA motif of JUN, a well-known co-factor of MAFF (Fig. 4H). Similarly, a HelT-motif for NFYB, commonly located 15 bp upstream of NFYB's sequence-motif, encodes the same sequence-motif (Fig. 4I). On the other hand, a ProT-motif of NRSF (REST), commonly located 3 bp upstream of the sequence motif of NRSF, does not encode the sequence motif of NRSF, and to our knowledge NRSF does not have any dimerization partner (Fig. 4J). In the following subsection, we analyze this phenomenon of shape-motifs of a TF specifying binding sites for its co-factors in more detail.

Overall, these analyses suggest that shape- specificity is largely complementary to sequence- specificity, and shape- and sequence-motifs collectively define a broader genomic context defining the TF's putative binding locations. Hence, the shape-specificity of a TF cannot be fully captured by simply taking the shape-profiles underlying sequence-sites. Our results also corroborate and extend the previous findings that implicate flanking nucleotides of sequence-motifs in dictating TF-DNA binding (Dror, Golan et al. 2015) by showing that flanking regions often harbor shape-motifs despite lacking preferred nucleotide patterns.

Some TFs extensively use shape-specific binding to co-bind with other TFs

Co-binding of TFs is a common phenomenon and is critical for precise spatio-temporal regulation of gene transcription (Gerstein, Kundaje et al. 2012). However, we found that about half (mean 53%) of a TF's peaks that overlap peaks for other TFs (*co-bound* peaks) have no sequence-sites and ~27% are shape-only (Fig. S5, see Methods). We therefore sought to better understand the extent of shape-only binding as a mechanism for TFs to co-bind with other TFs. We found that shape-only binding is more common than sequence-based binding for ~28% of co-bound TF pairs (Fig. 5A, Fig. S6). TFs that mostly use shape-only binding belong to multiple different protein families and include TFs that are known to bind in the context of many other proteins (e.g., MYC, MAX, JUND, STAT5A, GATA2, CCNT2, MEF2A, PAX5, POUF2 (Gerstein, Kundaje et al. 2012)) or to function broadly in genome-wide transcriptional regulation (e.g., EP300).

By examining the prevalence of shape-recognition for each TF in peaks co-bound by every other TF, we learned that TFs use shape- and sequence-recognition differently when binding alone and with each TF partner. First, many TFs enriched for sequence-only peaks genome-wide (e.g., CHD2, REST, CEBPB, MEF2A, STAT5A, GATA2, JUND, MAX, MXI1, MEF2C, MEF2A, POUF2, RUNX3) use mostly shape-sites when co-binding with other TFs. There are also TF pairs where the modes of binding (mostly sequence-based vs. mostly shape-only) of both TFs are different in co-bound peaks compared to the their genome-wide preference (Fig. 5B, Fig. S6). However, the opposite scenario is more common: the same mode of binding is found genome-wide and in co-bound regions (Fig. S7). Also, whether a TF in a given pair will mainly have shape-only binding in their co-binding peaks typically depends on the TF itself. Finally, some TFs that mainly utilize sequence-recognition switch to shape-recognition being more dominant in regions co-bound with a TF that is mainly shape-based and vice versa (Fig. 5C, Fig. S6). We conclude that while the shape preferences of a TF are fairly consistent across the genome, there are many TF pairs where co-binding is characterized by unique shape-motifs.

Co-bound TFs may interact physically and form a complex. In such cases, it is likely that motifs of the partnering TFs occur with some bias in their inter-site spacing. Our observation that co-bound TFs commonly use shape-specific binding led us to hypothesize that some TFs of a DNA-binding TF-complex may bind DNA shape-specifically. This scenario is in contrast with the general notion of "tethering" whereby it is assumed that one or more TFs of a complex recognize the DNA by sequence-specificity and the other TFs do not recognize DNA (Fig. 5D). To assess whether and to what extent TFs that lack sequence-sites in co-bound regions use shape-sites versus tethering, we first evaluated motif spacing and found that 2633 out of 3710 (71%) TF pairs have sites that occur with a bias for short (~3 bps) inter-site spacing (Fig. S8). This high proportion raises the possibility that these TFs might form TF-complexes. Interestingly, for 1245 of these pairs,

motifs of one or both TFs are shape-motifs. These shape-sequence or shape-shape motif pairs confirm that about a third of adjacent co-binding occurs with at least one TF using shape recognition without nucleotide recognition. We also found that 57 of these pairs have previously been reported to have physical interaction (Stark, Breitkreutz et al. 2006, Ravasi, Suzuki et al. 2010) or predicted for tethered binding (Wang, Zhuang et al. 2012). Some shape-shape motif pairs (e.g., JUN-TAF7, CTCF-YY1, ETS1-TAL1) were not reported in the above studies but were validated elsewhere to have physical interaction or to co-bind (Munz, Psichari et al. 2003, Palii, Perez-Iratxeta et al. 2011, Schwalie, Ward et al. 2013). Therefore, we find an interesting line of evidence that some TFs in a DNA-binding complex may actually bind DNA in a shape-specific manner, and it is unlikely that tethered binding is the only explanation for TF complex members that lack sequence-motifs.

In a ChIP-exo based study of TBX5 and NKX2-5 occupancy in cardiac differentiation, we determined that the binding of these two TFs can be interdependent (Luna-Zurita, Stirnimann et al. 2016). Sequence motifs of TBX5 and NKX2-5 co-occur in only 17% of the regions where their ChIP-exo peaks overlap. Motivated by the above analysis, we hypothesized that TBX5 and NKX2-5 may bind shape-specifically in their co-binding regions. We applied ShapeMF on TBX5 and NKX2-5 ChIP-exo peaks and found that both TFs have shape-motifs for all four features. Importantly, we found strong relationships between the sequence and shape motifs of these TFs, which gave us a strong premise for the above hypothesis. In particular, we found that the sequences underlying the HelT motif of NKX2-5 contain the TF's sequence motif, CACTT (Fig. 5E). Likewise, the sequences underlying the ProT motif of TBX5 contain a partial sequence motif of TBX5, TGTCA (Fig. 5F). Interestingly, we also found that the sequences underlying the ProT motif of NKX2-5 contain TGTCA sequences (Fig. 5G) – implicating that in many NKX2-5 peaks, TBX5 may co-bind by recognizing the ProT pattern, without full sequence recognition.

In support of our hypothesis that TBX5 and NKX2-5 may bind shape-specifically in their co-binding regions, we found that 79% of co-binding regions have a shape-site for one TF and a sequence motif for the other TF, and 73% contain shape-motifs of both TFs. Furthermore, co-occurrences of the shape- and sequence-motif pairs of TBX5 and NKX2-5 are enriched for 0-4 bps inter-site spacing, which is in the same range as the preferential distances between TBX5 and NKX2-5 sequence motifs that we identified previously, supported by crystal structure (Fig. 5H) (Luna-Zurita, Stirnimann et al. 2016). Overall, this analysis shows that the TBX5-NKX2-5 co-occupancy occurs to a large extent by recognizing DNA shape. The use of shape-recognition in TBX5-NKX2-5 co-bound regions exemplifies the use of shape-motifs as an alternative to tethering, and is highly relevant to coregulation of a cardiac transcriptional program, as well as a potential mechanism for their haploinsufficiency in congenital heart disease.

Shape-motifs explain genomic occupancy of TF-complexes where sequence-based models are inadequate

The above cases of shape-specific binding in regions of TF-TF co-occupancy motivated us to examine whether DNA shape may explain the genomic occupancy of TF complexes for which sequence-based models have not been able to explain complex patterns of co-binding. We focused on two well-known TF-complexes, namely the MAX homodimer and the MYC-MAX heterodimer, several aspects of whose *in vivo* occupancies have remained unresolved in sequence-based analyses (Guo, Li et al. 2014, He, Johnston et al. 2015).

The bHLHZip domain protein MAX recognizes the canonical E-box motif CACGTG. Current models suggest that MAX can bind DNA either as a homodimer or upon forming heterodimers with other bHLHZip proteins, such as MYC and MAD. Max footprints in ChIP-nexus data (He, Johnston et al. 2015) were found to span ~8 bases flanking either side of the E-box motif. Although the footprints were enriched for the E-box motif, the footprints covered the E-box motif only partially and there was no specificity in the flanking sequences. Our analysis of MAX ChIP-Seq peaks identified HelT-, MGW-, and ProT-motifs for Max, and suggest an interesting model involving DNA shape and sequence determining the specificity of MAX homodimers (Fig. 6A). In particular, sequences underlying the HelT-motif directly match the E-box sequence motif (Fig. 6B, left panel). On the other hand, the MGW- and ProT-motifs show very low sequence-specificity and in the co-occurrence peaks (i.e., where these shape-motifs co-occur with MAX's sequence motif), they occur 6–10 bases up- and 4-5 bases downstream of the E-box motif, respectively (Fig. 6B, middle and right panels). This result provides a shape-based explanation of MAX binding where specificities for HelT, MGW, and ProT explain Max binding both at the E-box motif and the inclusion of flanking sequences in Max-bound footprints.

MYC is another bHLHZip domain protein known to bind very weakly at E-box motifs as a monomer, but binds the same sequences with high affinity upon dimerization with MAX (Guo, Li et al. 2014). Our analysis of MYC K562 ChIP-Seq data supports this model, since ~75% MYC peaks overlap MAX peaks, and the intensity of MYC ChIP signal has a strong dependence on the extent of overlap (Fig. S9, see Methods). However, MAX binds in almost twice as many locations as MYC, and it is not clear how the specificity of the MYC-MAX dimer is different from that of MAX in the other MAX-bound locations. There is a hypothesis that bases flanking E-box motifs play a role in MYC-MAX co-bound regions (Nair and Burley 2003).

We hypothesized that MYC-MAX binding could be distinguishable from the binding of MAX homodimers in terms of shape-specificity. We therefore performed differential shape-motif discovery from MYC-MAX peaks utilizing MYC-unbound MAX regions as the negative control. This analysis indeed suggests a model whereby MYC-MAX binding is distinct from

the binding of MAX homodimers in terms of DNA shape (Fig. 6C). We identified HelT-, MGW-, and ProT-motifs for the MYC-MAX dimer and found that the HelT-motif for MYC-MAX encodes the E-box sequence motif (Fig. 6D, left panel), similar to the case for the MAX homodimer. Interestingly, HelT values at certain positions of the two motifs are significantly different (Fig. 6D, left panel, positions shown with asterisks). Moreover, the MGW- and the ProT-motifs for MYC-MAX are different from those of MAX in terms of length, pattern, the sequences underlying those motifs, and their spacing with the E-box motif (Fig. 6D, middle and right panels). In the co-occurrence peaks, the MGW-motif is often located 12-14 bp upstream of the E-box motif and the ProT-motif occurs 1-3 bp downstream. It is known that crystallized structures of the MYC-MAX dimer and the MAX homodimer are different despite their apparent resemblances (Nair and Burley 2003). Combining this with our shape-motif analysis, we speculate that the structural differences between the MYC-MAX dimer and the MAX homodimer cause subtle differences in their DNA-binding specificities that are largely accommodated by changes in MGW and ProT profiles. Overall, the above examples suggest that TF complexes, like individual TFs, utilize shape recognition as a mechanism to discern their binding locations.

Non-targeted TF motifs (aka "zingers") can be shape-specific

Hunt and Wasserman recently reported *zinger* motifs: sequence-motifs of a small group of TFs enriched across the binding locations of multiple other TFs (Worsley Hunt and Wasserman 2014). Along the same lines, we next asked if there are *shape-zingers*, i.e., shape-motifs enriched across ChIP-Seq peaks for many TFs. We tested for enrichment of shape-sites for each TF within the top 2000 peaks of every other TF. Results were consistent when we repeated the analyses using all peaks without a sequence-site for the other TF. We found that ~25 ENCODE TFs are shape-zingers, including previously reported sequence-zingers, such as JUN, FOSL1, and the ETS-family TFs EBF1 and ELK1. However, most shape-zingers are not sequence-zingers (Fig. 7, Fig. S10). Importantly, some of these novel shape-based zingers (e.g., GATA, ARID3A, P300, PAX5) are known to act as global regulators or regulators of large gene networks (Goodman and Smolik 2000, Medvedovic, Ebert et al. 2011, Rhee, Lee et al. 2014, Lentjes, Niessen et al. 2016). This finding suggests that shape-specificity enables these regulators to recognize a larger set of locations (and thus regulate more genes) than would be possible based on sequence-specificity alone. Consistent with the "loading station" model suggested by Hunt and Wasserman, all our shape-zingers (except P300 in the Gm12878 cell-line) show enrichment within peaks of CTCF, RAD21, and SMC3.

TFs within the same class recognize distinct shape-motifs

TFs within the same class of DNA-binding domain often recognize statistically indistinguishable sequence motifs, although such TFs still bind distinct locations in the

genome. It has been shown that sequence-based models of TF-occupancy achieve improved performance for TFs within the same class if the models use shape-features underlying sequence-motif hits (Mathelier, Xin et al. 2016, Yang, Orenstein et al. 2017). However, it is not clear whether TFs within a class show preferences for distinct shape-patterns and/or for distinct combinations of shape-features. To test this possibility, we applied ShapeMF on ENCODE ChIP-Seq datasets of the bHLH and bZIP class TFs (see Methods). For each TF within a class, we considered only those ChIP-peaks that do not overlap with peaks of any other TF within that class so that we could identify the shape-specificity intrinsic to each TF. Our analysis not only found that most TFs within these two classes recognize distinct motifs of different shape features, but also that some shape-motifs do not encode the common sequence-motif of that class (Fig. 8, Fig. S11). Overall, we conclude that preference for distinct shape-patterns, and sometimes for distinct combinations of shape-features, is a mechanistic explanation of how TFs within the same class of DNA-binding domains bind at distinct locations genome-wide.

TFs bind shape motifs *in vitro*

To complement and validate our analyses of shape-motifs in ChIP-Seq peaks, we analyzed their occurrence in oligonucleotides that have been assayed for *in vitro* binding. Nineteen ENCODE TFs were investigated via HT-SELEX (Jolma, Yan et al. 2013, Yang, Orenstein et al. 2017) and have shape-motifs that are 15-bp or less, which is short enough to be present in these libraries. We found that shape-motifs are enriched in the final round (i.e., preferentially bound) oligonucleotides for all 19 TFs. Furthermore, shape-motif prevalence is in good agreement between HT-SELEX and ChIP-seq data (correlation 0.54; Fig. S12). We additionally found enrichment of shape-motifs for MAX and MYC amongst bound versus unbound oligonucleotides in genomic-context protein binding microarray (gcPBM) data (Gordan, Shen et al. 2013, Afek, Schipper et al. 2014). Shape-motifs of MAX and MYC frequently co-occur with their sequence-motifs, as expected since these gcPBMs were designed to contain sequence-motifs (Zhou, Shen et al. 2015). Thus, despite several design features that bias currently available *in vitro* data against shape-specific binding (see Discussion), we observe a strong and consistent signal that TFs bind shape-motifs when they do occur in assayed oligonucleotides.

Discussion

Analyzing *in vivo* binding data of hundreds of human TFs with a novel algorithm that treats DNA as a structure rather than a string of nucleotides, we showed that TFs frequently bind DNA by recognizing specific patterns of DNA shape features. These shape-motifs can occur independently from the TF's nucleotide sequence-motifs, and ChIP-Seq peaks that contain a shape-motif but no sequence-motifs are as abundant as peaks with only sequence-motifs or with both. Shape motifs also shed light on TFs that bind low information-content sites and

weak matches to sequence-motifs. Thus, in addition to confirming the importance of DNA shape in the context of sequence-motifs as shown in several recent studies (Abe, Dror et al. 2015, Zhou, Shen et al. 2015, Mathelier, Xin et al. 2016), our results establish shape-recognition as an independent and sometimes exclusive mechanism for TF-DNA binding within regulatory regions.

Our analyses also reveal important functional and mechanistic consequences of shape-based TF-DNA binding. Binding of TFs within the same family to distinct instances of a shared sequence-motif can be explained by TF-specific shape-motifs that overlap or flank the sequence-motif. Similarly, TFs (e.g., MAX) that bind different instances of a sequence motif as monomers, homodimers, or heterodimers appear to recognize different DNA shapes in each of these contexts. These examples suggest that DNA shape may be a general mechanism to increase the information content of binding sites beyond that encoded by sequence-motifs, which is insufficient for eukaryotic TFs to uniquely recognize specific sites in genomic DNA (Wunderlich and Mirny 2009). We also find that co-binding TF pairs frequently utilize shape-based binding, providing a mechanism beyond tethering to explain co-binding in regions that lack one or both sequence-motifs. Finally, TF's in crystal structures generally contact nucleotide bases at sequence-specific binding sites and the DNA backbone at non-sequence-specific sites (Aishima and Wolberger 2003, von Hippel 2004, Romanuka, Folkers et al. 2009). Our results suggest that shape-specific sites contain preferential "pockets" – defined by shape-motifs – where a TF can stabilize and interact with the DNA backbone. It is also plausible that such stabilization is facilitated by enhanced electrostatic potential at the location of shape-motif occurrences (Rohs, Jin et al. 2010). Importantly, these new insights would have been missed if we had only searched for shape patterns within the context of sequence-motifs.

An important methodological contribution of our manuscript is ShapeMF, a *de novo* shape-motif discovery algorithm. ShapeMF enabled us to pursue the hypothesis that some TFs have intrinsic preferences for shape-motifs and such preferences can be discovered without taking sequence information into account. It is challenging to design such an algorithm since it requires discovering variable-length shape-patterns *de novo* from unaligned sequences with the criterion that the discovered shape-motifs are comparable to sequence motifs in terms of discriminating bound from unbound regions. Our solution was to implement Gibbs sampling with a notion of similarity between the shapes of two DNA sequences that is appropriate for quantitative features, rather than the discrete four-letter nucleotide alphabet. We considered several alternative solutions that have been used on related problems. For example, we might have discretized shape features and then directly applied *de novo* sequence motif algorithms, as in (Greenbaum, Parker et al. 2007). However, it was not clear how to appropriately bin and/or smooth shape features, or how to characterize the background distribution for these features. Time series "shapelet"

discovery algorithms are relevant to our problem (Ye and Keogh 2009, Grabocka, Schilling et al. 2014, Hou, Kwok et al. 2016), but their efficacy has been shown for datasets where discriminative shapelets appear very frequently (so that sampling a very small subset of the data would suffice to yield shape-motifs) or where the constituent time series are aligned. These scenarios do not hold for TF occupancy data, so shapelet algorithms would likely require a computationally intensive brute-force scheme. In contrast, ShapeMF does not (a) discretize data, (b) use any empirical background distribution of shape features, nor (c) assume that the given peaks/windows are aligned.

The focus of our study was to systematically test if DNA shape provides signals of "intrinsic" TF-DNA binding specificity (Stormo and Zhao 2010) and if these are independent of nucleotide sequence. We therefore did not adopt the approach of developing an optimal, holistic classifier of bound versus unbound regions. State of the art approaches to this classification problem perform well and can score a given sequence for its affinity. But a method such as ShapeMF is needed to discover the specificity signal that characterizes a TF (Setty and Leslie 2015). Adding shape-motifs to discriminative classifiers of TF bound regions does have potential to improve accuracy, since we found that many peaks with shape-motifs lack sequence-motifs and unbound regions with sequence-motifs can lack shape-motifs.

The intrinsic binding specificity of shape-motifs should be comprehensively studied *in vitro*. We conducted an initial evaluation using HT-SELEX and gcPBM data and found consistent evidence for *in vitro* binding. However, current *in vitro* data have some limitations for evaluating shape-motifs and disentangling the contributions of shape versus nucleotide recognition. Critically, existing oligonucleotide libraries do not contain sufficient coverage of shape-motifs. For example, 86% of the HT-SELEX oligonucleotides we analyzed are 20 bps or shorter (Jolma, Yan et al. 2013, Yang, Orenstein et al. 2017), and universal PBMs (uPBMs) that are designed to compactly cover all k-mers typically have values of k=8 or 10 bp (Berger and Bulyk 2009). Relatedly, HT-SELEX and uPBM oligonucleotides do not fully capture genomic context that may be important to shape-specific binding. These limitations could be overcome with improvements in the throughput of these technologies to accommodate more, longer sequences or by designing gcPBM libraries to contain oligonucleotides with better representation of shape-motif containing genomic regions, including more regions that contain shape-motifs but no sequence-motifs. It will be important to quantify binding affinities for many TFs and shape-motifs with these and other emerging technologies (e.g., microfluidics).

Several other future directions are suggested from our results. One goal is to quantify the amount of information that each TF utilizes from the shape domain and how specific this utilization of DNA shape is to different contexts, including chromatin domains, co-binding, and subsets of target genes. Another important direction will be to re-analyze *in vitro* DNA-

binding data to assess whether high-affinity sequences are enriched for shape-based binding. Additionally, the mechanisms of shape-based binding suggested by our results should be evaluated in light of all available TF-DNA structures. It will also be interesting to determine the extent to which shape-based binding is an alternative explanation to tethering. In terms of methodology, we could likely improve the performance of ShapeMF by adding more shape features and/or analyzing these jointly rather than individually. Our current analysis suggests that multivariate analysis of shape-features would require careful algorithm design since a TF may not have a motif for every feature and the motifs may differ in size and location. Finally, the notion that shape-based TF binding affinity could be conserved without sequence conservation opens the door to a whole new view of regulatory evolution and the opportunity to develop shape aware measures of DNA change and its functional impact on human disease that has at its basis abnormal TF function.

Data and Methods

The ShapeMF tool for *de novo* shape motif discovery from shape-data

Definitions and notation. Let $\boldsymbol{S} = \{s_i\}_{i=1}^{N}$ denote a set of peaks of a TF T. Let $\boldsymbol{\psi} = \{\psi_i\}_{i=1}^{N}$ denote the corresponding shape-data for a feature F (roll, helical twist, propeller twist, or minor groove width in the current GBshape-based implementation). Thus each ψ_i is a sequence of real numbers $\psi_{i,j}$ denoting the value of feature F at position j in peak s_i. A shape *pattern* of length l is a l-length sequence $\boldsymbol{P} = ((m_k, d_k))_{k=1}^{l}$ of 2-tuples of real numbers. We say that $\boldsymbol{P}$ *occurs* in peak s_i if there is a l-length sequence window $\boldsymbol{W_i} = (\psi_{i,j})_{j=t_i}^{t_i+l-1}$ starting at position t_i in the shape-data ψ_i such that $m_{j-t_i+1} - d_{j-t_i+1} \leq \psi_{i,j-t_i+1} \leq m_{j-t_i+1} + d_{j-t_i+1}$ for $t_i \leq j \leq t_i + l - 1$.

Algorithm. ShapeMF uses a two-step approach to first compute a shape pattern $\boldsymbol{P}$ from the shape-data of positive peaks in a training dataset of matched positive and negative control peaks and then modify the pattern to one that maximizes F-score between the positive and control peaks in the training data. Finally, the modified pattern is called a motif if its Bonferroni-corrected hypergeometric p-value, computed on a separate validation dataset of matched positive and control peaks, is significant.

Step 1. From the shape-data $\boldsymbol{\psi}$ of positive peaks in the training data, we first compute a set of windows $\boldsymbol{W}$ (one window $\boldsymbol{W_i}$ in each ψ_i) such that the sum of pairwise Euclidean distances between the windows, *i.e.*, $D = \sum_{x \neq y} \text{Euclidean}(\boldsymbol{W_x}, \boldsymbol{W_y})$, is minimized. We use Gibbs sampling to compute such a set of windows. In particular, we start by selecting $\boldsymbol{W_i}$'s randomly. To select a new window from ψ_i to replace $\boldsymbol{W_i}$, let $\boldsymbol{V_{i,k}} = (\psi_{i,j})_{j=k}^{k+l-1}$ denote the l-length window in ψ_i that starts at position k, and let $D_{i,k} = \sum_{x \neq y \neq i} \text{Euclidean}(\boldsymbol{W_x}, \boldsymbol{W_y}) + \sum_{j \neq i} \text{Euclidean}(\boldsymbol{W_j}, \boldsymbol{V_{i,k}})$ denote the new value of D if

$\boldsymbol{V_{i,k}}$ replaces $\boldsymbol{W_i}$ in the current set of windows. We then sample a window $\boldsymbol{V_{i,k}}$ from ψ_i with probability $\exp(-D_{i,k}) \Big/ \sum_k \exp(-D_{i,k})$ and update $\boldsymbol{W_i}$ with the sampled $\boldsymbol{V_{i,k}}$. We iterate through the ψ_i's in the order they appear in $\boldsymbol{\psi}$, and for each ψ_i, we update the window $\boldsymbol{W_i}$ following the above steps. We continue to repeat this iterative updating until convergence in the value of D. In our implementation, we decide that the value of D has converged if it does not change in two successive iterations, or if the improvement in the value of D is negligible for ten consecutive iterations. We assume that the improvement is negligible when the current and the previous values of D satisfies: $D_{prev}(1-\epsilon) < D_{current} < D_{prev}$, where $\epsilon = 10^{-5}$.

Step 2. We next compute a pattern $\boldsymbol{P} = ((m_k, d_k))_{k=1}^{l}$ from the set $\boldsymbol{W}$ of windows computed above. Let the window $\boldsymbol{W_i}$ start at the position t_i in ψ_i. We then take $m_k = \text{mean}(\{\psi_{i,t_i+k-1}\}_{i=1}^{N})$ and $d_k = \alpha \times \text{standard\_deviation}(\{\psi_{i,t_i+k-1}\}_{i=1}^{N})$, where α is a constant whose optimum value is computed as follows. We try each value of α in the range $[0.1, 2]$, in increments of 0.1, and compute the corresponding pattern $\boldsymbol{P_\alpha}$ from $\boldsymbol{W}$. We quantify the goodness of each $\boldsymbol{P_\alpha}$ as a discriminator between the positive and control peaks of the training data by its $F_{1/3}$-score. Note that the $F_{1/3}$-score here is a more conservative objective than used typically in classification settings, yet we wanted to weigh precision much higher than recall so that the number of false positives remains low. We then choose $\boldsymbol{P}$ to be the pattern $\text{argmax}_\alpha(F_{1/3}(\boldsymbol{P_\alpha}))$.

Motif identification. Using an independent validation set of positive and control peaks, patterns are tested for enrichment in positive peaks, as is done in other discriminative motif finding tools (Bailey 2011). Let $n_{\boldsymbol{P}}^{+}$ and $n_{\boldsymbol{P}}^{-}$ denote the number of validation positive and control peaks, respectively, where pattern $\boldsymbol{P}$ has an occurrence. We say that $\boldsymbol{P}$ is a *motif* of feature F (or a 'F-motif') for TF T if a hypergeometric test parameterized by $2N, N, n_{\boldsymbol{P}}^{+} + n_{\boldsymbol{P}}^{-}$, and $n_{\boldsymbol{P}}^{+}$ yields a significant p-value after Bonferroni correction. We use a Bonferroni corrected p-value threshold of 10^{-5}. Shape-motifs $\boldsymbol{P}$ that meet this criterion are retained for further analysis, and others are discarded.

Finding variable-length motifs. The above steps 1 and 2 compute a motif for a given length l. In our analysis we have considered all values of l between 5 and 30. For computational efficiency, we first compute motifs for values of l that are multiples of 5. For all other values of l, we take the starting positions t_i's of the motifs computed for length $\left\lfloor l/5 \right\rfloor$ as our initial guess for starting positions and search for motifs within the positions $t_i - l$ and $t_i + l$.

Calling redundant motifs. In the last step, we eliminate redundant motifs. We first partition all motifs according to their lengths: two motifs of lengths l_1 and l_2 are put in the

same partition if $\lfloor l_1/5 \rfloor = \lfloor l_2/5 \rfloor$. For two motifs $\boldsymbol{P_1}$ and $\boldsymbol{P_2}$, if $\boldsymbol{P_1}$ has lower false positive rate than $\boldsymbol{P_2}$ on validation data and the occurrences of $\boldsymbol{P_1}$ "cover" at least 75% of the occurrences of $\boldsymbol{P_2}$, then we assume that $\boldsymbol{P_2}$ is redundant and discard $\boldsymbol{P_2}$. An occurrence of $\boldsymbol{P_1}$ *covers* an occurrence of $\boldsymbol{P_2}$ if the occurrence of $\boldsymbol{P_1}$ overlaps with at least 75% length of the occurrence of $\boldsymbol{P_2}$. This strategy to remove redundant motifs is akin to the one utilized in (Beer and Tavazoie 2004).

Data

We downloaded uniformly processed ChIP-Seq datasets (narrowPeak format) from the *ENCODE Downloads* section at UCSC (http://genome.ucsc.edu/ENCODE/downloads.html) and genome-wide shape-data (bigwig format) from the FTP interface of the GBshape database (http://rohsdb.cmb.usc.edu/GBshape/). We used only the ChIP-Seq datasets with `quality=good` and `treatment=None`. We also discarded the histone deacetylase (HDAC) datasets from consideration, since our focus here was on TFs.

We followed the strategy of Setty and Leslie (Setty and Leslie 2015) to create training and validation data comprising positive and control peaks from each ChIP-Seq dataset. For positive peaks, we took 100 base pair (bp) windows centered at the peak summits. For negative control peaks, we took 100-bp non-peak windows located 100 bp upstream of the positive peaks. Negative control peaks that intersect with positive peaks were discarded along with the corresponding downstream positive peak. For learning motifs, we used the top 2000 positive peaks (ranked by the `signalValue` field) and their associated control peaks, or all positive-control peak pairs if less than 2000 remain. We then randomly shuffled the positive peaks and split them into equal halves (and likewise for control peaks) to obtain our training and validation data.

We used `bwtool` (Pohl and Beato 2014) to extract shape profiles of the positive and control peaks from the bigwig files.

Identifying promoter-proximal and distal regions

We followed the strategy of Setty and Leslie (Setty and Leslie 2015) to identify promoter-proximal and –distal regions. From UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables), we collected the coordinates (hg19) of RefSeq genes (`group=genes and gene predictions, track=refseq genes, table=refgene, region=genome`). We then select promoter-proximal regions as the 2-kilobase regions flanking each gene, and the distal regions as the windows spanning 10-kb to 1-Mb regions flanking each gene.

Sequence motif analysis

We applied the tool `gkm-SVM` (Ghandi, Lee et al. 2014) to compute the sequence motifs that could discriminate between the positive and the control peaks in each training data set. We note that `gkm-SVM` outputs the scores of all 10-mers as predicted by a support vector machine (svm) trained to discriminate between the positive and the control peaks. Thus, `gkm-SVM` does not directly provide a description of a TF's specificity. To overcome this issue, the `gkm-SVM` package includes an algorithm that iteratively learns a specified number of sequence motifs from the svm scores. We used this algorithm to learn five motifs for each ChIP-seq dataset. Note that, in the original work featuring `gkm-SVM` (Ghandi, Lee et al. 2014), the authors used three motifs to describe specificity of TFs. By utilizing five motifs, we in fact allowed redundancy and presumably weak (low information content) motifs to be included in our analysis in order to ensure that we have a broad sequence-based definition of TF binding sites. We then used the tool `fimo` (Grant, Bailey et al. 2011) to identify all occurrences of these motifs in the positive peaks.

As a second source of sequence-motif hits, we took the genome-wide annotations for occurrences of sequence-motifs curated by Kheradpour and Kellis (Kheradpour and Kellis 2014). In their curated collection, Kheradpour and Kellis used all motifs from available motif libraries and also motifs from several *de novo* motif finders.

Co-binding regions for a given pair of TFs

For two TFs f_1 and f_2, we first identify the peaks $\boldsymbol{S_1}$ of f_1 that intersects with a peak of f_2, and likewise the peaks $\boldsymbol{S_2}$ of f_2 that intersects with a peak of f_1. We then merge the genomic regions denoted by $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ to obtain the regions where f_1 and f_2 co-bind.

Calling shape-zingers

A shape-motif $\boldsymbol{P_1}$ of a TF f_1 is also a motif for a TF f_2 if $\boldsymbol{P_1}$ can discriminate the positive peaks (top 2000 or all if there are less than 2000 positive peaks) of f_2 from control peaks with a hypergeometric p-value $< 10^{-10}$. A TF f_1 is a shape-zinger if one or more of its motifs are motifs for at least 30% TFs with a shape-motif in the same cell-line. We chose the fraction 30% following Hunt and Wasserman (Worsley Hunt and Wasserman 2014) who reported sequence-zingers to be enriched in 30-60% datasets.

Analysis of spacing bias between motif pairs

We followed the strategy of Ng et al. (Ng, Schutte et al. 2014) for analysis of biased spacing between a given pair of motifs. For the given motif pairs, we first compute the distances between their non-overlapping neighboring (adjacent) occurrences. We arbitrarily decide one motif as primary and the other as secondary. A distance between the primary-

secondary motif pair is positive if the primary-motif occurs upstream of the secondary-motif, zero if they occur at the same location, and it is negative otherwise. We test each distance between -50 to +50 bps, and call a distance to be significant if the binomial p-value (after Bonferroni correction) is significant ($< 10^{-5}$).

Shape-motif analysis for TF families

We took the classification of TFs into families according to their DNA-binding domains from `TFClass` (Wingender, Schoeps et al. 2013). For each TF in a family, we first selected the sets of peaks that do not overlap with the peaks of any other TF in the same family. We then used ShapeMF on the remaining sets of peaks to identify the shape motifs of each TF.

Acknowledgement

NIH/NHLBI Bench to Bassinet Program UM1HL098179, to K.S.P. and B.G.B, and by William H. Younger, Jr. (B.G.B.)

Figure Legends

Figure 1. Overview of ShapeMF. (A) Shape-motif discovery involves comparing TF-bound regions (positives; solid lines) to flanking non-bound regions (negatives; dashed lines). For each region (different colors) and each shape feature (MGW, HelT, Roll, ProT), ShapeMF extracts the profile of quantitative feature values across the region. The Gibbs sampler then identifies a set of short windows (5 to <30 bp) from the profiles of positive sequences that have similar patterns of the shape feature. In the second step, this initial set of positive region windows is refined so that the resulting windows share a shape pattern that has the maximum accuracy to discriminate between positives and negatives (Methods). The range of feature values from this refined set of windows defines the shape-motif. For visualization purposes, we also generate sequence logos from the sequences underlying the occurrences of a shape-motif and the range of feature values in the 50-bp regions flanking up- and downstream its occurrences (both shown below the shape-motif). (B) Difference in the approach of identifying a shape-motif occurrence vs. a sequence-motif occurrence. A shape-motif occurs in a sequence if it contains a window whose feature values at every position fit within the ranges defined by the shape-motif. A sequence-motif occurs in a sequence if it contains a window that is significantly similar to the multinomial model defined by the sequence-motif.

Figure 2. Most TFs have a shape-motif in their ChIP-seq peaks. (A) Heatmap of negative-$\log_{10}$ transformed Bonferroni corrected p-values for the four types of shape-motifs of each TF. White cells indicate no significant motif. 'K' or 'G' after a TF's name denotes K562 or Gm12878 cells, respectively. (B) Numbers of TFs with each of the four types of shape-motifs. (C) Length distribution of shape-motifs. (D) Boxplots of relative distances of sequence-, shape-, and overlapping-sites from ChIP-seq peak-summits.

Figure 3. Shape-motifs are common and occur independently of sequence-motifs. (A) The proportion of ChIP-seq peaks that are shape-only, common, and sequence-only for each TF and cell line. (B) Shape-only peaks comprise a considerable fraction of genome-wide peaks for many TFs. (C) For most TFs, shape-motifs can be discovered from the peaks that lack a sequence motif. (D) Most (~65%) of these shape-motifs can discriminate between top peaks and their flanking non-peaks, and many (~50%) are indeed similar to the shape-motifs discovered from top peaks. (E) Shape-only peaks are generally as abundant as sequence-based (sequence-only and common) peaks in different types of regulatory regions.

Figure 4. (A-F) Different scenarios of shape- and sequence-motif co-occurrence found enriched in datasets from the K562 cell-line (see Fig. S4 for the same information from the Gm12878 cell-line). Each scenario is shown with a schematic and an example from our analysis. A schematic uses a cartoon DNA double helix (from http://veleta.rosety.com), a

sequence ACTGACA, and a hypothetical shape pattern to show that a shape-motif can completely contain a sequence-motif (A), a sequence-motif can completely contain a shape-motif (B,C), a shape-motif can overlap with a sequence-motif and flank up- or downstream (D,E), and shape- and sequence-motifs can co-occur with some inter-site spacing (F). Each schematic is accompanied by a real example that includes the sequence-motif (the first of the five sequence-motifs computed by gkmSVM; our analysis uses all five gkmSVM sequence-motifs), the shape-motif, and the inter-site distance that we found enriched in our analysis. We also show the sequence-logo created from sequences underlying the shape-motif, and the range of shape-feature values in the flanking 50-bp regions (up- and downstream) of the shape-site. (G-J) Examples where a shape-motif encodes the binding location of a TF's dimerization partner.

Figure 5. (A) Co-binding TF pairs often utilize shape-specific binding. For each TF pair (f_1, f_2), the cell at row f_1 and column f_2 of the heatmap shows whether f_1 binds more shape-specifically or sequence-specifically with f_2 in the regions where they co-bind in K562 (Gm12878 in Fig. S6). Whether f_1's binding is more shape-specific or sequence-specific ("binding mode") in the f_1-f_2 co-binding regions is defined as the number of shape-only peaks being more than that of sequence-based (sequence-only and common) peaks in their co-binding regions. To show whether a TF's genome-wide binding is more shape-specific or sequence-specific, we use colored bars (following the same scale as in the heatmap) adjacent to the row and the column corresponding to that TF. The colored bars for a TF show whether its genome-wide peaks are more shape-only or sequence-based. The binding mode of a TF in a co-binding region may alter depending on its partner: both (B) or one (C) TF may alter binding mode. (D) Schematic comparing the tethering model with a model of TF co-binding where one or more TFs bind by recognizing DNA shape. (E-G) Shape-motifs of NKX2-5 and TBX5. For each shape-motif, we show the logo created from its underlying sequences and the range of shape-feature values in the flanking 50-bp regions of shape-sites (as in Fig. 4). (H) Occurrences of TBX5 and NKX2-5 shape-motifs in a 22 bp DNA sequence (from mouse *Nppa* promoter) where the two TFs are known to bind. Crystal structure of the ternary complex comprising TBX5, NKX2-5, and DNA is from our previous study.

Figure 6. Shape-motifs suggest models for genomic occupancy of (A) MAX homodimer and (B) MYC-MAX heterodimer. The HelT-, MGW-, and ProT-motifs enriched under (C) MAX ChIP-Seq peaks vs. negative controls and (D) MYC vs. MYC-unbound MAX-peaks. For each shape motif, we show the logo created from its underlying sequences and flanking 50-bp regions (as in Fig 4). HelT values in the MYC-MAX motif (left panel in D) differ significantly from the MAX HelT motif (left panel in C) in positions 1 (Kolgomorov-Smirnov test p-value < 0.05), 2 (p-value<0.005), 4 (p-value < 0.005), and 5 (p-value < 0.01).

Figure 7. Shape-zinger TFs (nodes) and the fraction (color-coded) of other ChIP-Seq datasets where their shape-motifs are enriched in. Shape-zingers are the TFs with shape-motifs enriched in the largest numbers (at least 30%, see Methods) of other TF's ChIP-seq datasets. Enrichment was calculated in the top 2000 peaks vs. non-peaks (Methods). TFs corresponding to nodes with a filled circle were previously reported as sequence-zingers. For brevity, edges connecting shape-zinger pairs are omitted.

Figure 8. bZIP TFs have similar sequence-motifs but unique shape-motifs. Shape-motifs for five bZIP proteins NRF1, JUND, JUNB, CEBPB, and FOSL1 are shown (H, M, and P denote motifs for HelT, MGW, and ProT, respectively). A feature is not mentioned for a TF if the TF does not have a shape-motif for that feature. For each TF, the first of its five gkmSVM motifs is shown to aid comparison with the logos created from sequences underlying the TF's shape-motifs.

Supplementary Figure Legends

Supplementary Figure 1. Fractions of top 500, 1000, 2000, 5000, and all peaks of a TF's ChIP-Seq dataset that contain an occurrence of the TF's sequence-motif (orange circle) and shape-motif (green-circle). Data plotted for TFs assayed in (A) K562 and (B) Gm12878 cell-lines.

Supplementary Figure 2. Histograms showing number of motifs of the four different features having different values of Information Content per Position (ICP). The ICP statistic for a shape-motif is derived from the sequences underlying the occurrences of the shape-motif (Methods).

Supplementary Figure 3. Scatterplot showing a monotonically increasing relationship between the fraction of overlapping sites of a TF and the mean information content derived from sequence-logos underlying its shape-motifs. Each circle denotes a TF and the mean information content is the mean of the information contents of its different shape-motifs. Data shown for TFs in (A) K562 and (B) Gm12878 cell-lines.

Supplementary Figure 4. Different scenarios of shape- and sequence-motif co-occurrence found enriched in datasets from the Gm12878 cell-line.

Supplementary Figure 5. Histograms showing number of co-binding TF-pairs (f_1, f_2) for different fractions of peaks of f_1 lacking a sequence-motif of f_1 (left panel) or containing a shape-motif of f_1 (right panel) in (A) K562 and (B) Gm12878 cell-lines.

Supplementary Figure 6. Heatmaps similar to Figure 5 showing results of our co-binding analysis on TF ChIP-Seq data in Gm12878 cell-line. (A) Co-binding TF pairs often utilizing shape-specific binding. The binding mode of a TF in a co-binding region may alter depending on its partner: both (B) or one (C) TF may alter binding mode.

Supplementary Figure 7. Heatmaps similar to Figure 5 showing that a TF generally maintains its genomewide binding mode when co-binding with other TFs in (A) K562 and (B) Gm12878 cell-lines.

Supplementary Figure 8. Significantly enriched inter-site distances between (A) any types of motif-pairs and (B) shape-sequence or shape-shape motif-pairs. Instances from K562 and Gm12878 cell-lines were combined in each panel.

Supplementary Figure 9. Moving average plot (window size = 250) showing monotonically decreasing relationship between rank of MYC peaks and their % of overlap (bps) with MAX peaks.

Supplementary Figure 10. Histograms of the numbers of TFs whose shape-motifs are enriched in different fractions of ChIP-Seq datasets in (A) K562 and (B) Gm12878 cell-lines. Shape-zingers are defined as the TFs with enrichment in at least 30% datasets (Methods).

Supplementary Figure 11. Shape-motifs of bZIP and bHLH proteins (for which ChIP-Seq assays were performed by ENCODE in the K562 cell-line and we could find a shape-motif) showing that TFs within the same family extensively utilize different shape-motifs (H, M, and P denote motifs for HelT, MGW, and ProT, respectively), and/or combinations of different shape-features to recognize their target binding sites. A feature is not mentioned for a TF if the TF does not have a shape-motif for that feature. Seven of 13 bZIP TFs and 4 of 7 bHLH TFs were found to have a shape-motif in this analysis.

Supplementary Figure 12. Scatterplot showing fraction of ChIP-peaks of a TF with its shape-motifs vs. fraction of the TF's HT-SELEX oligonucleotides with its shape-motifs. Green/red data points indicate TFs with mean-length of shape-motifs smaller/longer than 15 bps.

References

Abe, N., I. Dror, L. Yang, M. Slattery, T. Zhou, H. J. Bussemaker, R. Rohs and R. S. Mann (2015). "Deconvolving the recognition of DNA shape from sequence." Cell **161**(2): 307-318.

Afek, A., J. L. Schipper, J. Horton, R. Gordan and D. B. Lukatsky (2014). "Protein-DNA binding in the absence of specific base-pair recognition." Proc Natl Acad Sci U S A **111**(48): 17140-17145.

Aishima, J. and C. Wolberger (2003). "Insights into nonspecific binding of homeodomains from a structure of MATalpha2 bound to DNA." Proteins **51**(4): 544-551.

Arvey, A., P. Agius, W. S. Noble and C. Leslie (2012). "Sequence and chromatin determinants of cell-type-specific transcription factor binding." Genome Res **22**(9): 1723-1734.

Badis, G., M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C. F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes and M. L. Bulyk (2009). "Diversity and complexity in DNA recognition by transcription factors." Science **324**(5935): 1720-1723.

Bailey, T. L. (2011). "DREME: motif discovery in transcription factor ChIP-seq data." Bioinformatics **27**(12): 1653-1659.

Beer, M. A. and S. Tavazoie (2004). "Predicting gene expression from sequence." Cell **117**(2): 185-198.

Berg, O. G. and P. H. von Hippel (1987). "Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters." J Mol Biol **193**(4): 723-750.

Berger, M. F., G. Badis, A. R. Gehrke, S. Talukder, A. A. Philippakis, L. Pena-Castillo, T. M. Alleyne, S. Mnaimneh, O. B. Botvinnik, E. T. Chan, F. Khalid, W. Zhang, D. Newburger, S. A. Jaeger, Q. D. Morris, M. L. Bulyk and T. R. Hughes (2008). "Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences." Cell **133**(7): 1266-1276.

Berger, M. F. and M. L. Bulyk (2009). "Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors." Nat Protoc **4**(3): 393-411.

Berger, M. F., A. A. Philippakis, A. M. Qureshi, F. S. He, P. W. Estep, 3rd and M. L. Bulyk (2006). "Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities." Nat Biotechnol **24**(11): 1429-1435.

Chiu, T. P., L. Yang, T. Zhou, B. J. Main, S. C. Parker, S. V. Nuzhdin, T. D. Tullius and R. Rohs (2015). "GBshape: a genome browser database for DNA shape annotations." Nucleic Acids Res **43**(Database issue): D103-109.

Consortium, E. P. (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.

Dror, I., T. Golan, C. Levy, R. Rohs and Y. Mandel-Gutfreund (2015). "A widespread role of the motif environment in transcription factor binding across diverse protein families." Genome Res **25**(9): 1268-1280.

Ernst, J. and M. Kellis (2010). "Discovery and characterization of chromatin states for systematic annotation of the human genome." Nat Biotechnol **28**(8): 817-825.

Ernst, J. and M. Kellis (2012). "ChromHMM: automating chromatin-state discovery and characterization." Nat Methods **9**(3): 215-216.

Gao, T., B. He, S. Liu, H. Zhu, K. Tan and J. Qian (2016). "EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types." Bioinformatics **32**(23): 3543-3551.

Garvie, C. W. and C. Wolberger (2001). "Recognition of specific DNA sequences." Mol Cell **8**(5): 937-946.

Gerstein, M. B., A. Kundaje, M. Hariharan, S. G. Landt, K. K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander, R. Min, P. Alves, A. Abyzov, N. Addleman, N. Bhardwaj, A. P. Boyle, P. Cayting, A. Charos, D. Z. Chen, Y. Cheng, D. Clarke, C. Eastman, G. Euskirchen, S. Frietze, Y. Fu, J. Gertz, F. Grubert, A. Harmanci, P. Jain, M. Kasowski, P. Lacroute, J. Leng, J. Lian, H. Monahan, H. O'Geen, Z. Ouyang, E. C. Partridge, D. Patacsil, F. Pauli, D. Raha, L. Ramirez, T. E. Reddy, B. Reed, M. Shi, T. Slifer, J. Wang, L. Wu, X. Yang, K. Y. Yip, G. Zilberman-Schapira, S. Batzoglou, A. Sidow, P. J. Farnham, R. M. Myers, S. M. Weissman and M. Snyder (2012). "Architecture of the human regulatory network derived from ENCODE data." Nature **489**(7414): 91-100.

Ghandi, M., D. Lee, M. Mohammad-Noori and M. A. Beer (2014). "Enhanced regulatory sequence prediction using gapped k-mer features." PLoS Comput Biol **10**(7): e1003711.

Goodman, R. H. and S. Smolik (2000). "CBP/p300 in cell growth, transformation, and development." Genes Dev **14**(13): 1553-1577.

Gordan, R., N. Shen, I. Dror, T. Zhou, J. Horton, R. Rohs and M. L. Bulyk (2013). "Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape." Cell Rep **3**(4): 1093-1104.

Grabocka, J., N. Schilling, M. Wistuba and L. Schmidt-Thieme (2014). Learning time-series shapelets. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, New York, USA, ACM**:** 392-401.

Grant, C. E., T. L. Bailey and W. S. Noble (2011). "FIMO: scanning for occurrences of a given motif." Bioinformatics **27**(7): 1017-1018.

Greenbaum, J. A., S. C. Parker and T. D. Tullius (2007). "Detection of DNA structural motifs in functional genomic elements." Genome Res **17**(6): 940-946.

Guo, J., T. Li, J. Schipper, K. A. Nilson, F. K. Fordjour, J. J. Cooper, R. Gordan and D. H. Price (2014). "Sequence specificity incompletely defines the genome-wide occupancy of Myc." Genome Biol **15**(10): 482.

He, Q., J. Johnston and J. Zeitlinger (2015). "ChIP-nexus enables improved detection of in vivo transcription factor binding footprints." Nat Biotechnol **33**(4): 395-401.

Hoffman, M. M., O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes and W. S. Noble (2012). "Unsupervised pattern discovery in human chromatin structure through genomic segmentation." Nat Methods **9**(5): 473-476.

Hou, L., J. T. Kwok and J. M. Zurada (2016). Efficient learning of timeseries shapelets. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, Arizona, AAAI Press**:** 1209-1215.

Jolma, A., J. Yan, T. Whitington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J. M. Vaquerizas, R. Vincentelli, N. M. Luscombe, T. R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja and J. Taipale (2013). "DNA-binding specificities of human transcription factors." Cell **152**(1-2): 327-339.

Kheradpour, P. and M. Kellis (2014). "Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments." Nucleic Acids Res **42**(5): 2976-2987.

Lentjes, M. H., H. E. Niessen, Y. Akiyama, A. P. de Bruine, V. Melotte and M. van Engeland (2016). "The emerging role of GATA transcription factors in development and disease." Expert Rev Mol Med **18**: e3.

Luna-Zurita, L., C. U. Stirnimann, S. Glatt, B. L. Kaynak, S. Thomas, F. Baudin, M. A. Samee, D. He, E. M. Small, M. Mileikovsky, A. Nagy, A. K. Holloway, K. S. Pollard, C. W. Muller and B. G. Bruneau (2016). "Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and Coordinates Cardiogenesis." Cell **164**(5): 999-1014.

Mathelier, A., B. Xin, T. P. Chiu, L. Yang, R. Rohs and W. W. Wasserman (2016). "DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo." Cell Syst **3**(3): 278-286 e274.

Matys, V., E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele and E. Wingender (2003). "TRANSFAC: transcriptional regulation, from patterns to profiles." Nucleic Acids Res **31**(1): 374-378.

Medvedovic, J., A. Ebert, H. Tagoh and M. Busslinger (2011). "Pax5: a master regulator of B cell development and leukemogenesis." Adv Immunol **111**: 179-206.

Munz, C., E. Psichari, D. Mandilis, A. C. Lavigne, M. Spiliotaki, T. Oehler, I. Davidson, L. Tora, P. Angel and A. Pintzas (2003). "TAF7 (TAFII55) plays a role in the transcription activation by c-Jun." J Biol Chem **278**(24): 21510-21516.

Nair, S. K. and S. K. Burley (2003). "X-ray structures of Myc-Max and Mad-Max recognizing DNA. Molecular bases of regulation by proto-oncogenic transcription factors." Cell **112**(2): 193-205.

Ng, F. S., J. Schutte, D. Ruau, E. Diamanti, R. Hannah, S. J. Kinston and B. Gottgens (2014). "Constrained transcription factor spacing is prevalent and important for transcriptional control of mouse blood cells." Nucleic Acids Res **42**(22): 13513-13524.

Palii, C. G., C. Perez-Iratxeta, Z. Yao, Y. Cao, F. Dai, J. Davison, H. Atkins, D. Allan, F. J. Dilworth, R. Gentleman, S. J. Tapscott and M. Brand (2011). "Differential genomic targeting of the transcription factor TAL1 in alternate haematopoietic lineages." EMBO J **30**(3): 494-509.

Pohl, A. and M. Beato (2014). "bwtool: a tool for bigWig files." Bioinformatics **30**(11): 1618-1619.

Rao, S. S., M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander and E. L. Aiden (2014). "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." Cell **159**(7): 1665-1680.

Ravasi, T., H. Suzuki, C. V. Cannistraci, S. Katayama, V. B. Bajic, K. Tan, A. Akalin, S. Schmeier, M. Kanamori-Katayama, N. Bertin, P. Carninci, C. O. Daub, A. R. Forrest, J. Gough, S. Grimmond, J. H. Han, T. Hashimoto, W. Hide, O. Hofmann, A. Kamburov, M. Kaur, H. Kawaji, A. Kubosaki, T. Lassmann, E. van Nimwegen, C. R. MacPherson, C. Ogawa, A. Radovanovic, A. Schwartz, R. D. Teasdale, J. Tegner, B. Lenhard, S. A. Teichmann, T. Arakawa, N. Ninomiya, K. Murakami, M. Tagami, S. Fukuda, K. Imamura, C. Kai, R. Ishihara, Y. Kitazume, J. Kawai, D. A. Hume, T. Ideker and Y. Hayashizaki (2010). "An atlas of combinatorial transcriptional regulation in mouse and man." Cell **140**(5): 744-752.

Rhee, C., B. K. Lee, S. Beck, A. Anjum, K. R. Cook, M. Popowski, H. O. Tucker and J. Kim (2014). "Arid3a is essential to execution of the first cell fate decision via direct embryonic and extraembryonic transcriptional regulation." Genes Dev **28**(20): 2219-2232.

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

Romanuka, J., G. E. Folkers, N. Biris, E. Tishchenko, H. Wienk, A. M. Bonvin, R. Kaptein and R. Boelens (2009). "Specificity and affinity of Lac repressor for the auxiliary operators O2 and O3 are explained by the structures of their protein-DNA complexes." J Mol Biol **390**(3): 478-489.

Sandelin, A., W. Alkema, P. Engstrom, W. W. Wasserman and B. Lenhard (2004). "JASPAR: an open-access database for eukaryotic transcription factor binding profiles." Nucleic Acids Res **32**(Database issue): D91-94.

Schwalie, P. C., M. C. Ward, C. E. Cain, A. J. Faure, Y. Gilad, D. T. Odom and P. Flicek (2013). "Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes." Genome Biol **14**(12): R148.

Setty, M. and C. S. Leslie (2015). "SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps." PLoS Comput Biol **11**(5): e1004271.

Slattery, M., T. Zhou, L. Yang, A. C. Dantas Machado, R. Gordan and R. Rohs (2014). "Absence of a simple code: how transcription factors read the genome." Trends Biochem Sci **39**(9): 381-399.

Stark, C., B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz and M. Tyers (2006). "BioGRID: a general repository for interaction datasets." Nucleic Acids Res **34**(Database issue): D535-539.

Stewart, A. J., S. Hannenhalli and J. B. Plotkin (2012). "Why transcription factor binding sites are ten nucleotides long." Genetics **192**(3): 973-985.

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

von Hippel, P. H. (2004). "Biochemistry. Completing the view of transcriptional regulation." Science **305**(5682): 350-352.

von Hippel, P. H. (2007). "From "simple" DNA-protein interactions to the macromolecular machines of gene expression." Annu Rev Biophys Biomol Struct **36**: 79-105.

von Hippel, P. H. and O. G. Berg (1986). "On the specificity of DNA-protein interactions." Proc Natl Acad Sci U S A **83**(6): 1608-1612.

von Hippel, P. H., A. Revzin, C. A. Gross and A. C. Wang (1974). "Non-specific DNA binding of genome regulating proteins as a biological control mechanism: I. The lac operon: equilibrium aspects." Proc Natl Acad Sci U S A **71**(12): 4808-4812.

Wang, J., J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield, M. C. Greven, B. G. Pierce, X. Dong, A. Kundaje, Y. Cheng, O. J. Rando, E. Birney, R. M. Myers, W. S. Noble, M. Snyder and Z. Weng (2012). "Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors." Genome Res **22**(9): 1798-1812.

Wingender, E., T. Schoeps and J. Donitz (2013). "TFClass: an expandable hierarchical classification of human transcription factors." Nucleic Acids Res **41**(Database issue): D165-170.

Worsley Hunt, R. and W. W. Wasserman (2014). "Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets." Genome Biol **15**(7): 412.

Wunderlich, Z. and L. A. Mirny (2009). "Different gene regulation strategies revealed by analysis of binding motifs." Trends Genet **25**(10): 434-440.

Yang, L., Y. Orenstein, A. Jolma, Y. Yin, J. Taipale, R. Shamir and R. Rohs (2017). "Transcription factor family-specific DNA shape readout revealed by quantitative specificity models." Mol Syst Biol **13**(2): 910.

Ye, L. and E. Keogh (2009). Time series shapelets: a new primitive for data mining. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. Paris, France, ACM**:** 947-956.

Yip, K. Y., C. Cheng, N. Bhardwaj, J. B. Brown, J. Leng, A. Kundaje, J. Rozowsky, E. Birney, P. Bickel, M. Snyder and M. Gerstein (2012). "Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors." Genome Biol **13**(9): R48.

Yoshida, T., T. Ohkumo, S. Ishibashi and K. Yasuda (2005). "The 5'-AT-rich half-site of Maf recognition element: a functional target for bZIP transcription factor Maf." Nucleic Acids Res **33**(11): 3465-3478.

Zhou, T., N. Shen, L. Yang, N. Abe, J. Horton, R. S. Mann, H. J. Bussemaker, R. Gordân and R. Rohs (2015). "Quantitative modeling of transcription factor binding specificities using DNA shape." Proceedings of the National Academy of Sciences.

Figure 1

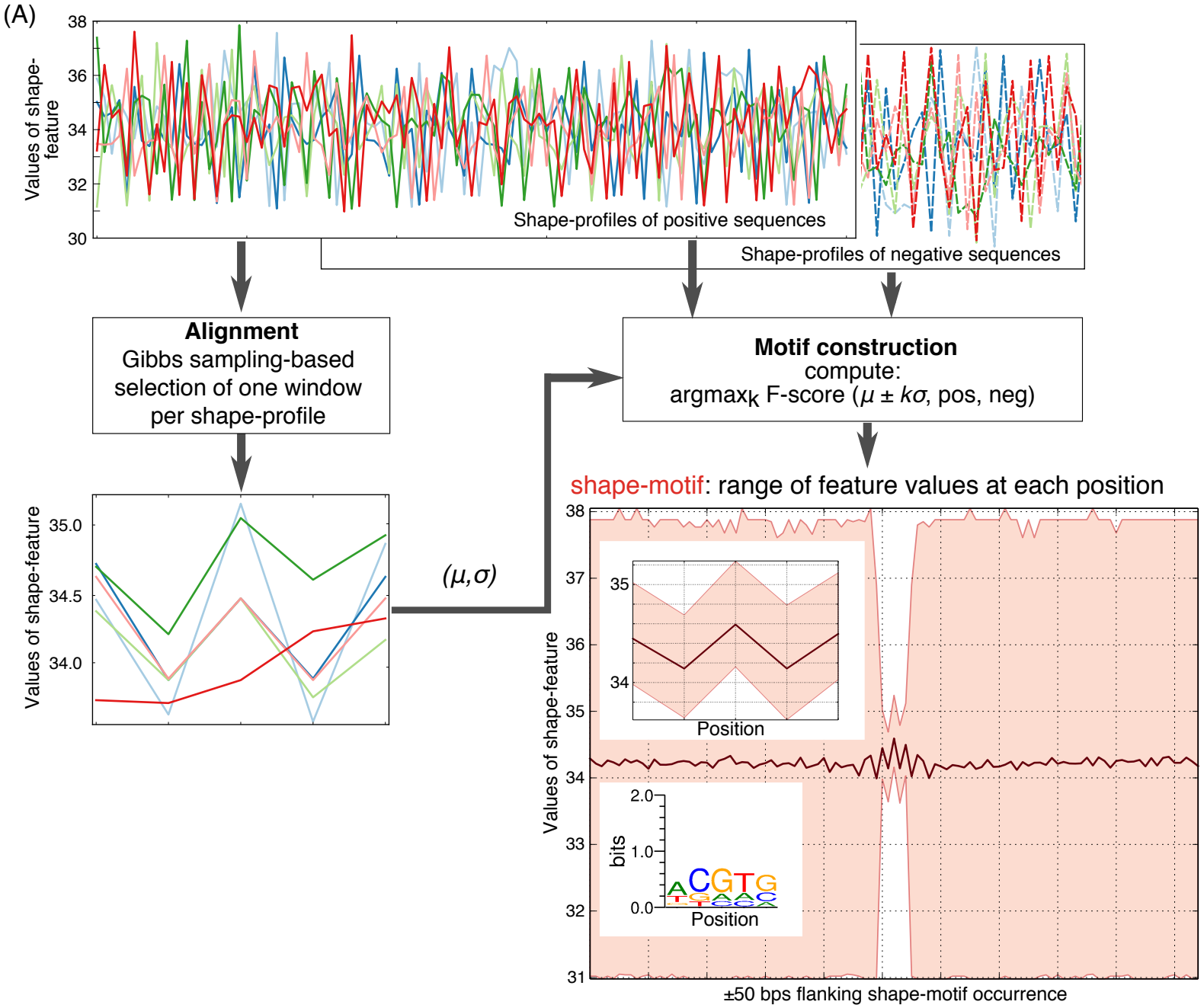


Figure 2

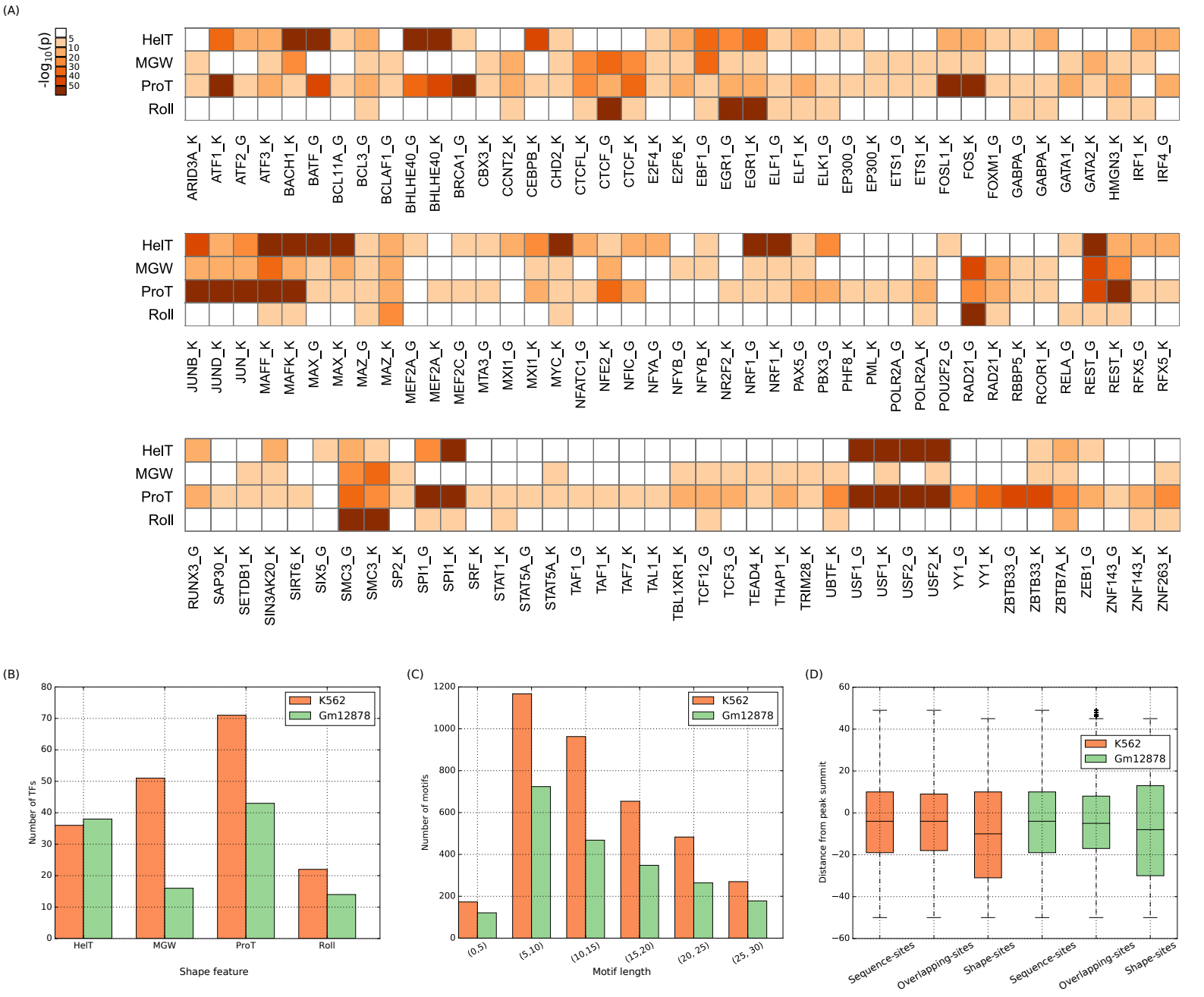
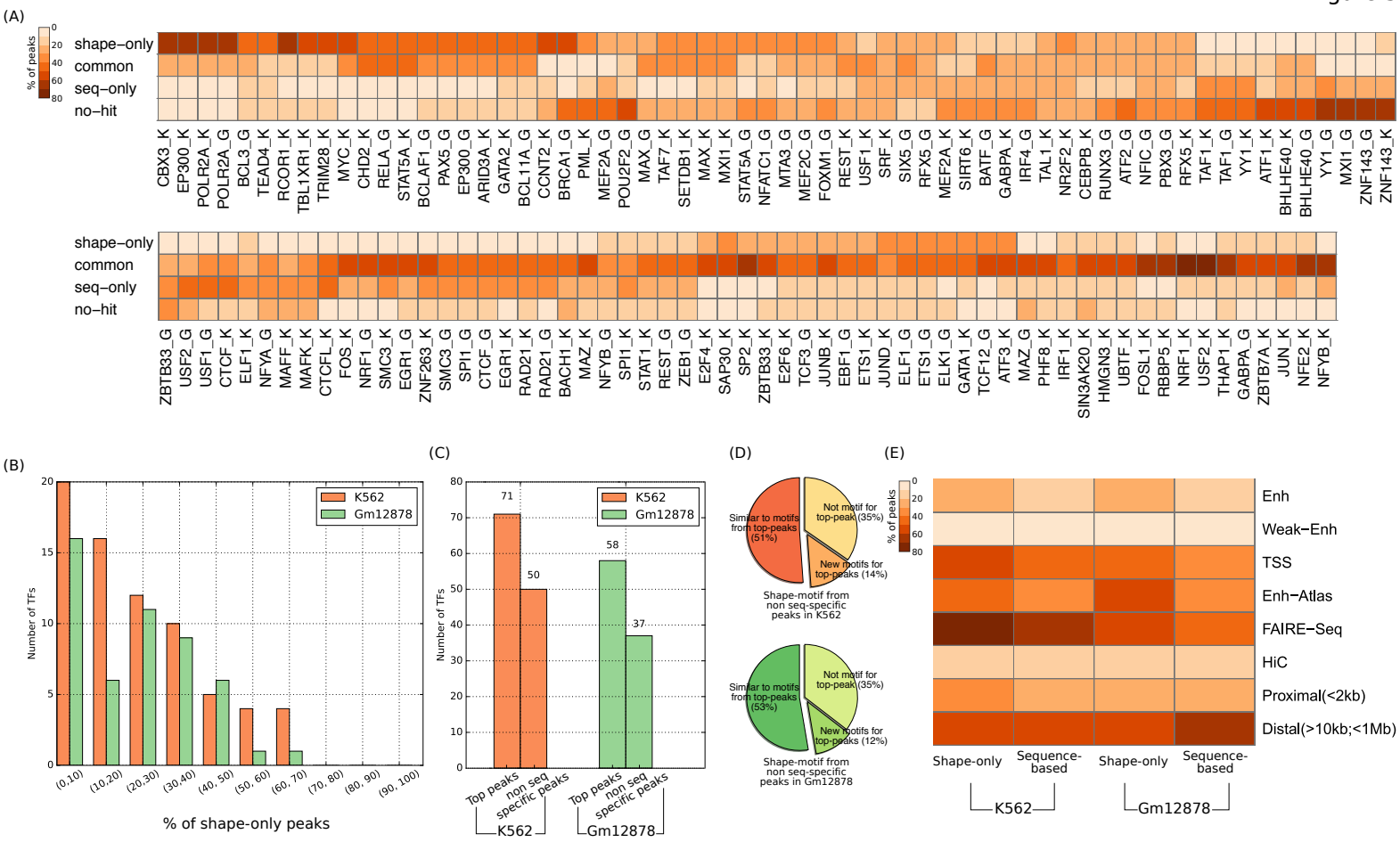
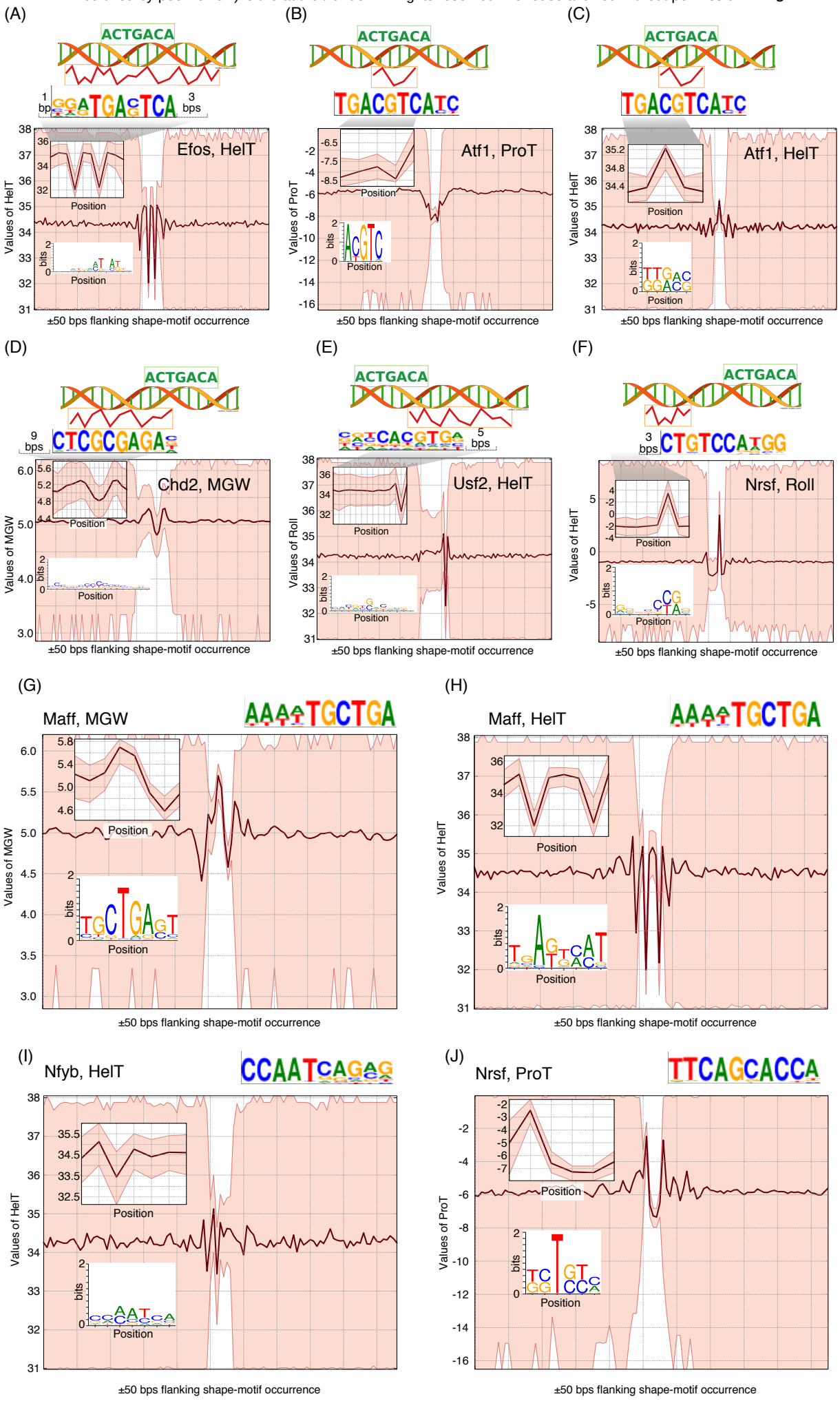
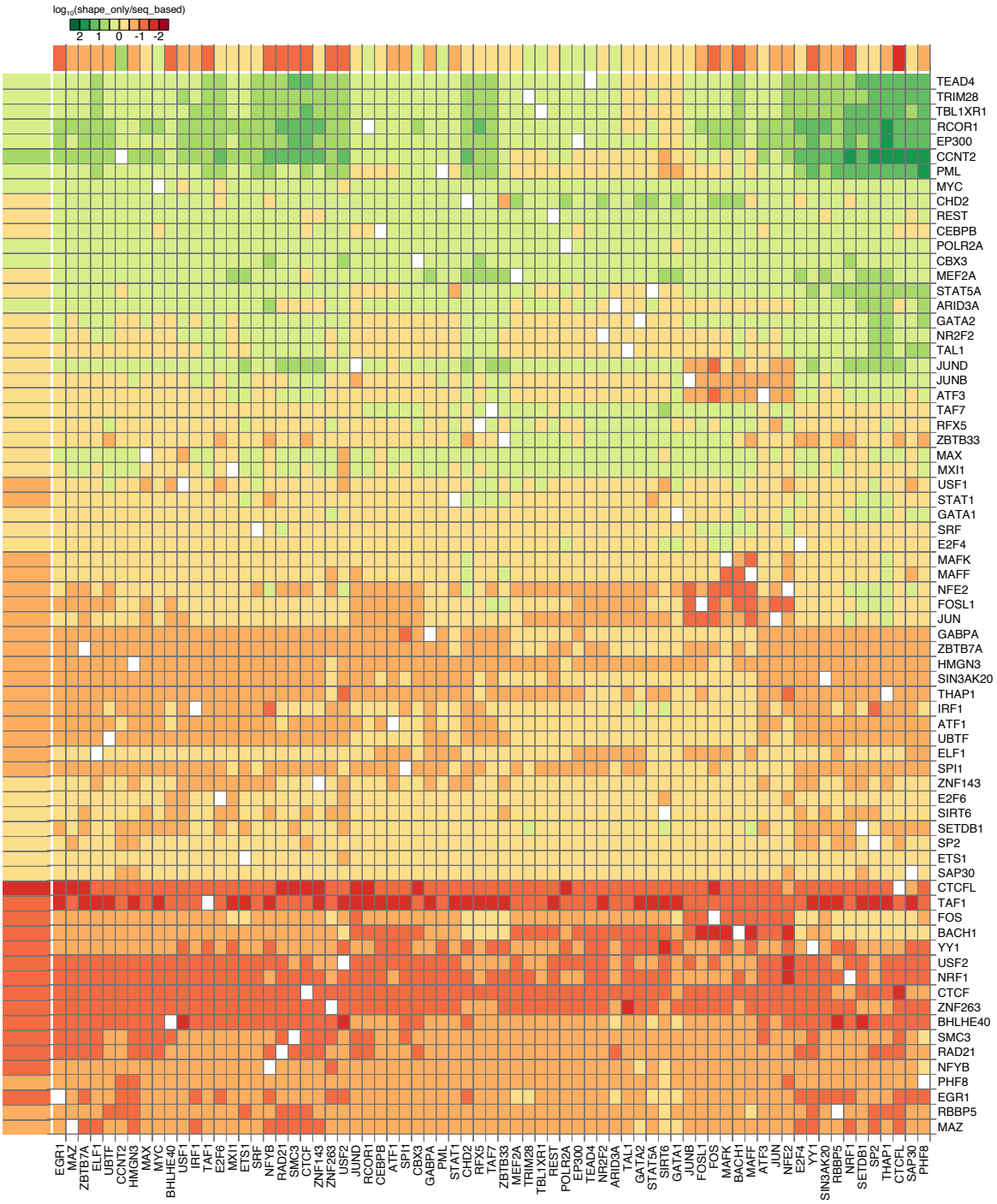


Figure 3

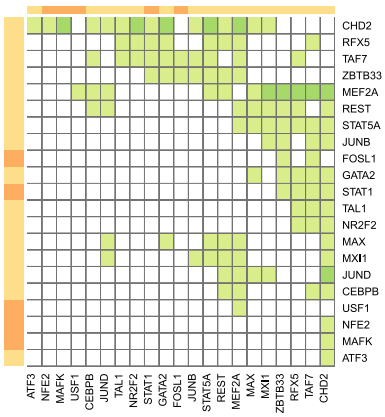




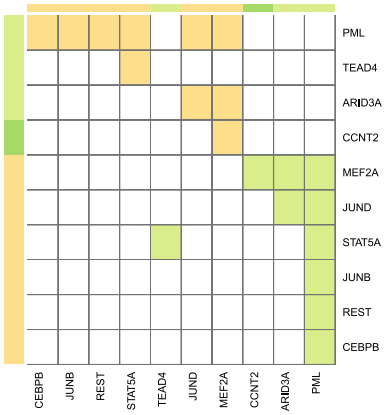
(A) Many TFs bind shape-specifically when co-binding with other TFs



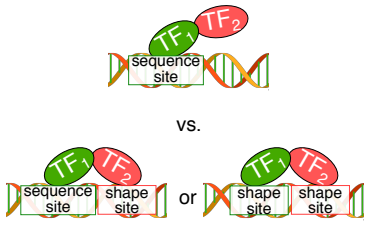
(B) Both TFs switch binding mode



(C) One TF switches binding mode

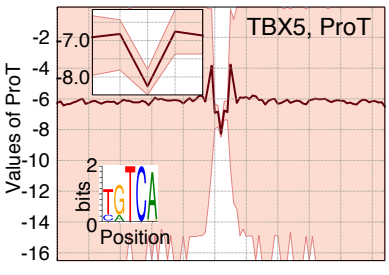


(D) Shape-specific co-binding as an alternative mechanism to tethering

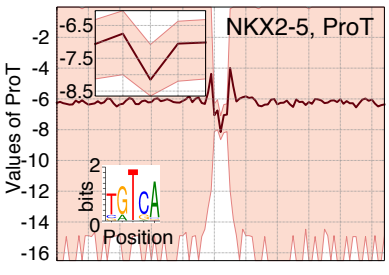


(E)

(F)



(G)



±50 bps flanking shape motif occurrence

(H)

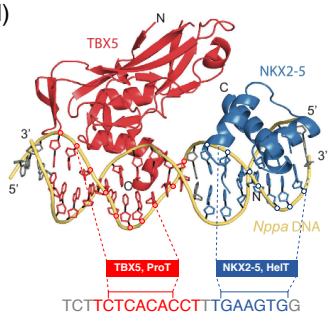


Figure 6

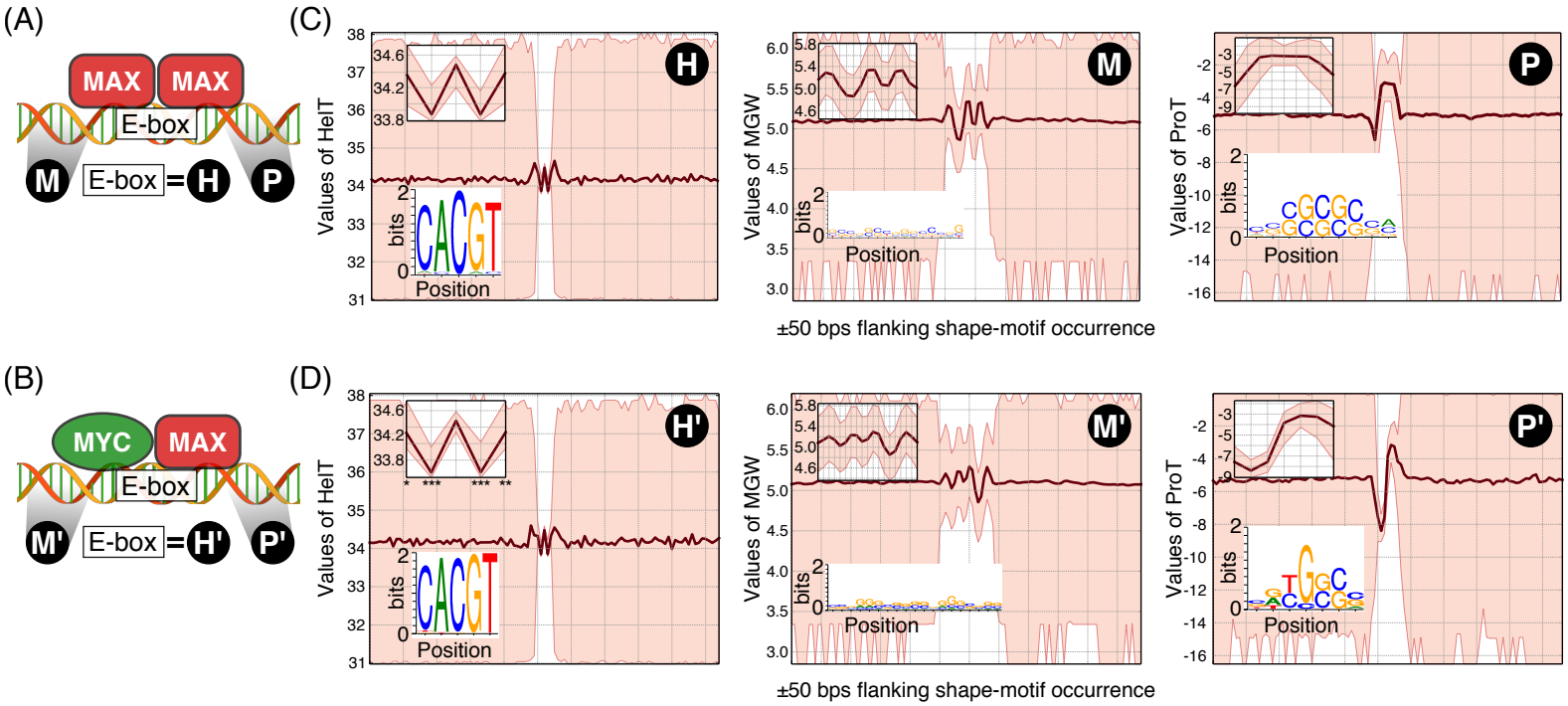


Figure 7

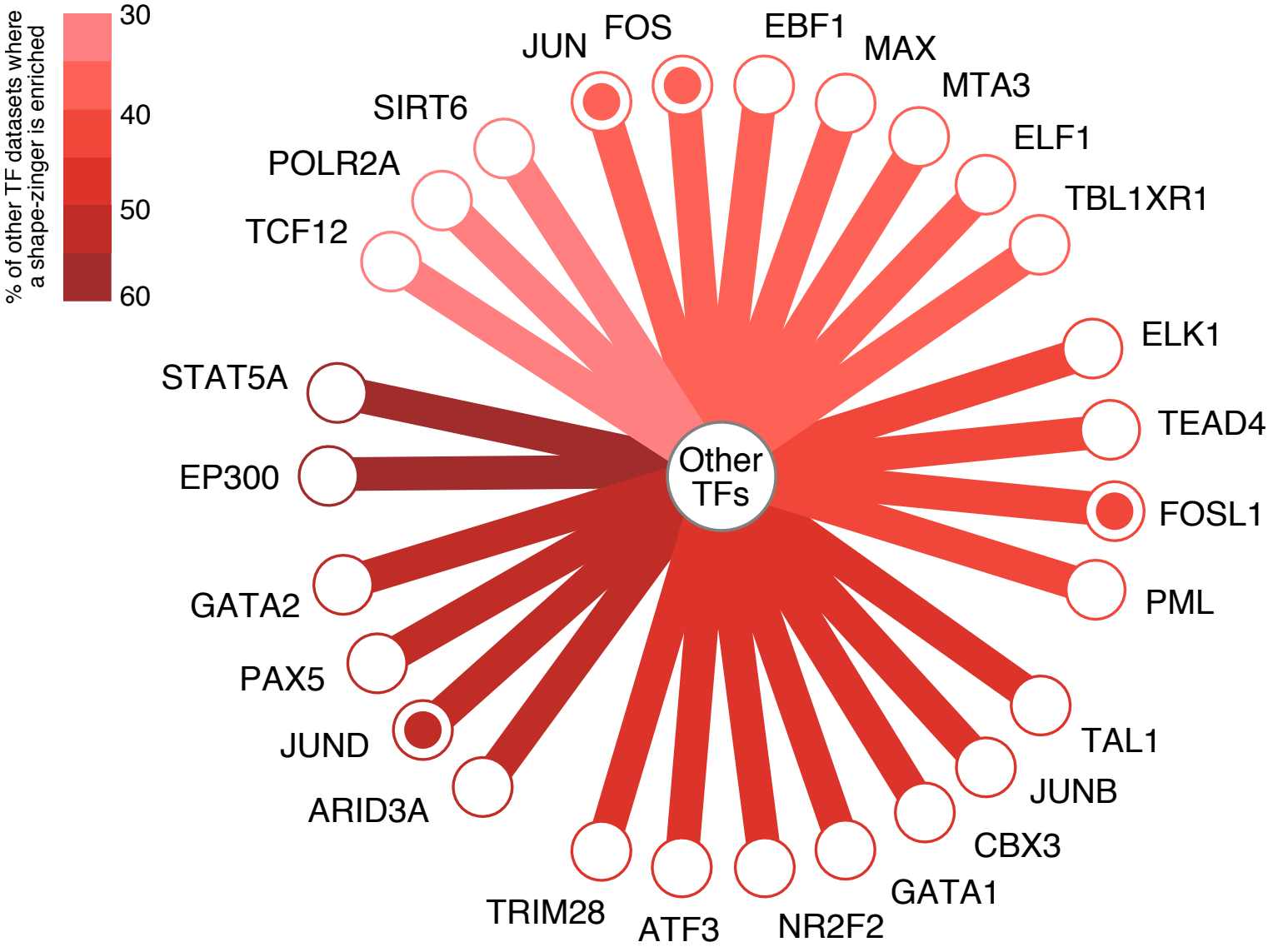
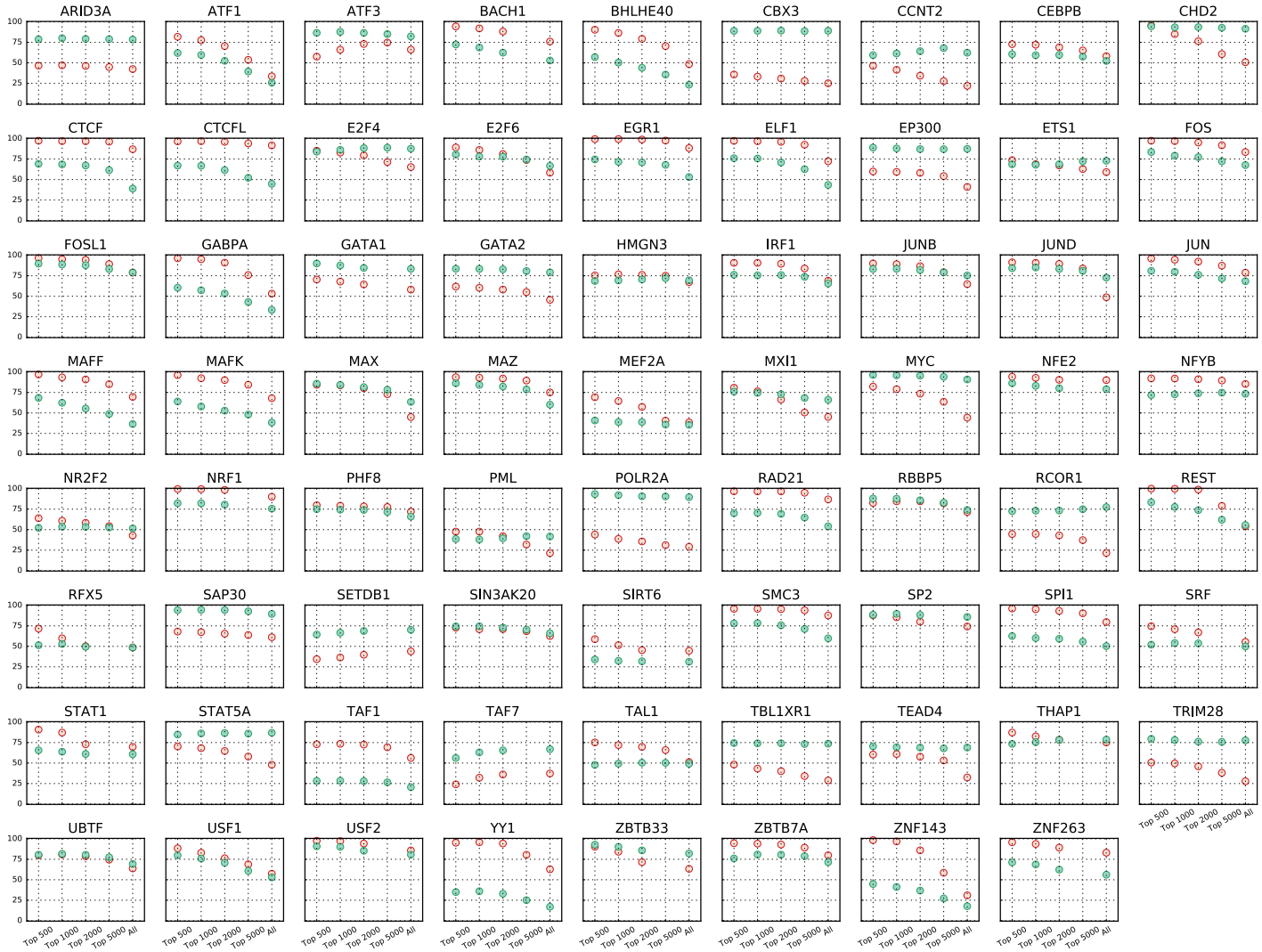


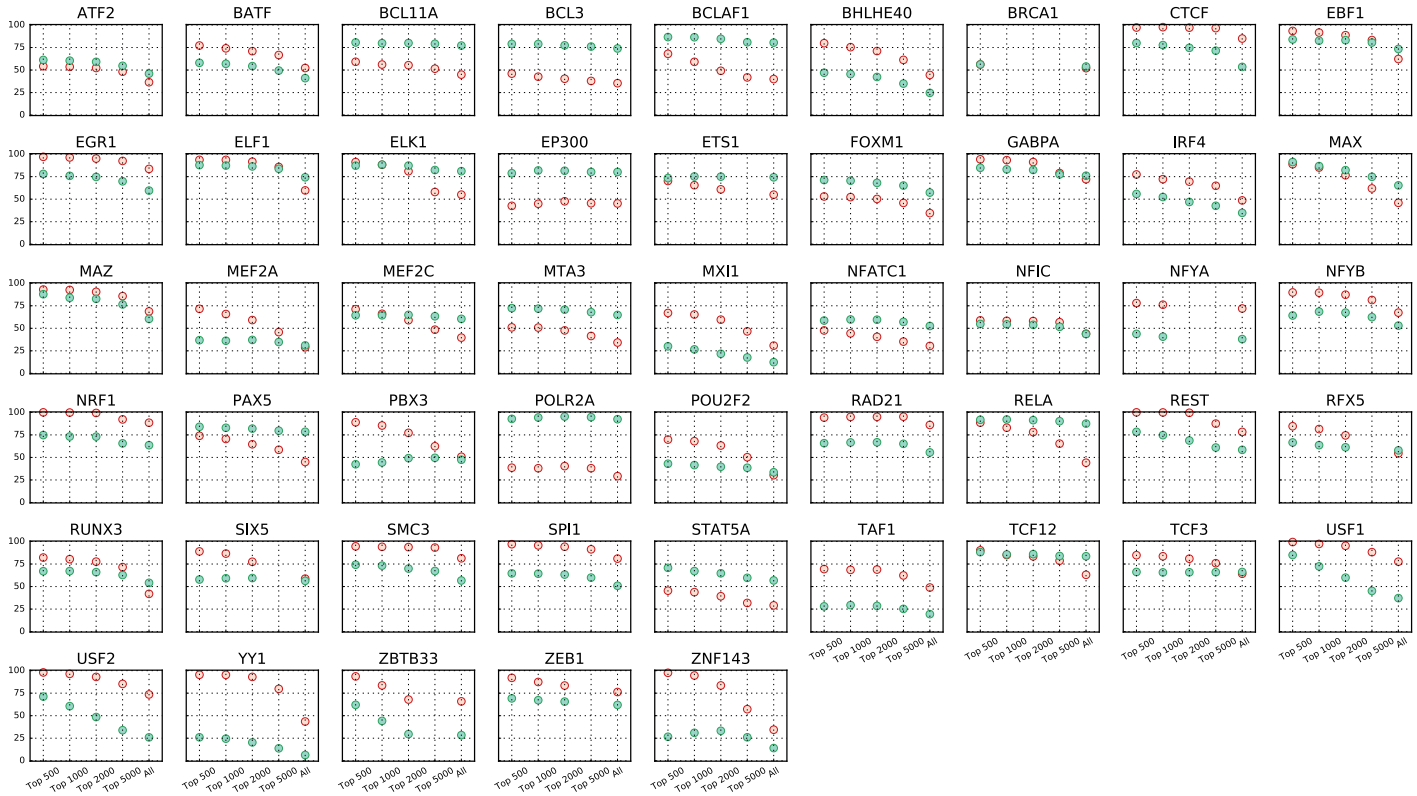
Figure 8

Supplementary Figure 1

(A) K562 cell-line

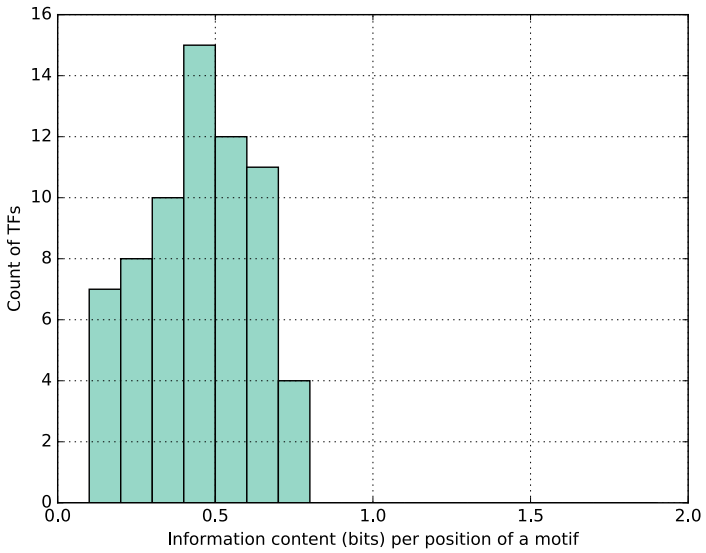


(B) Gm12878 cell-line

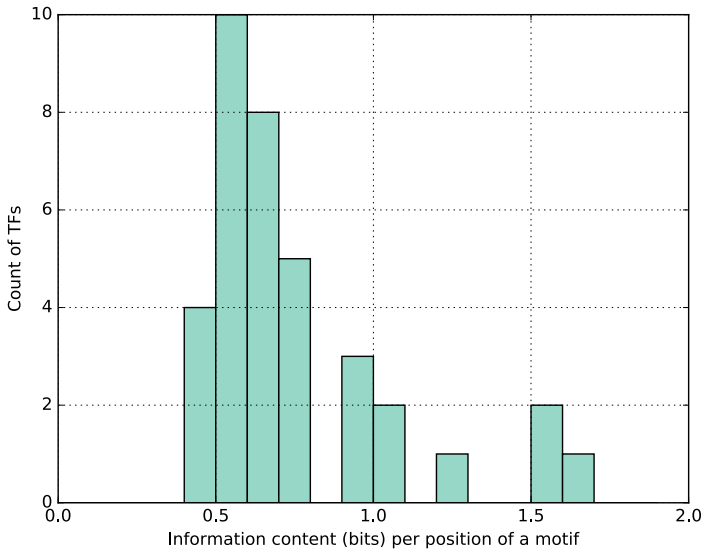


Supplementary Figure 2

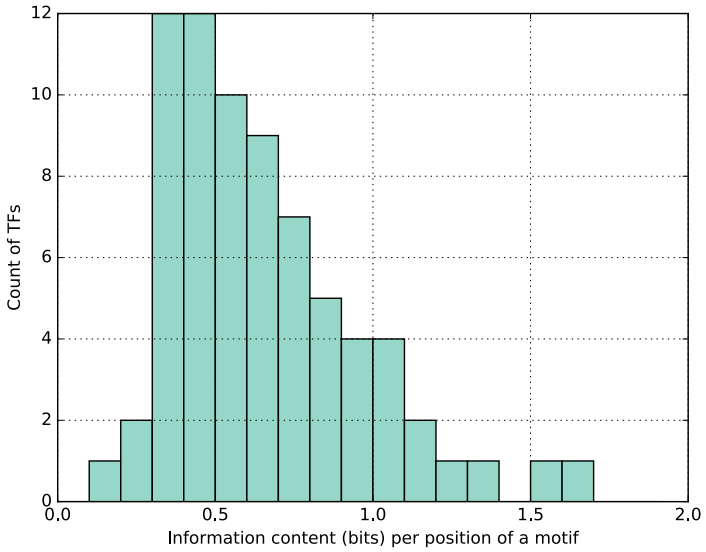
(A) MGW



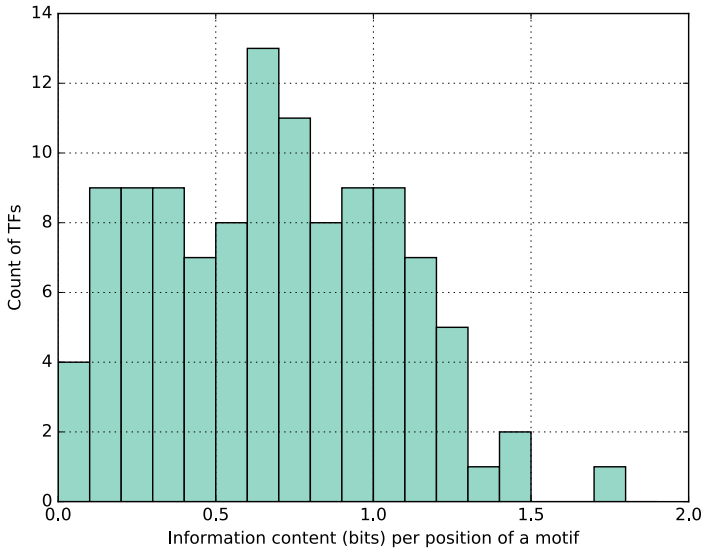
(B) Roll



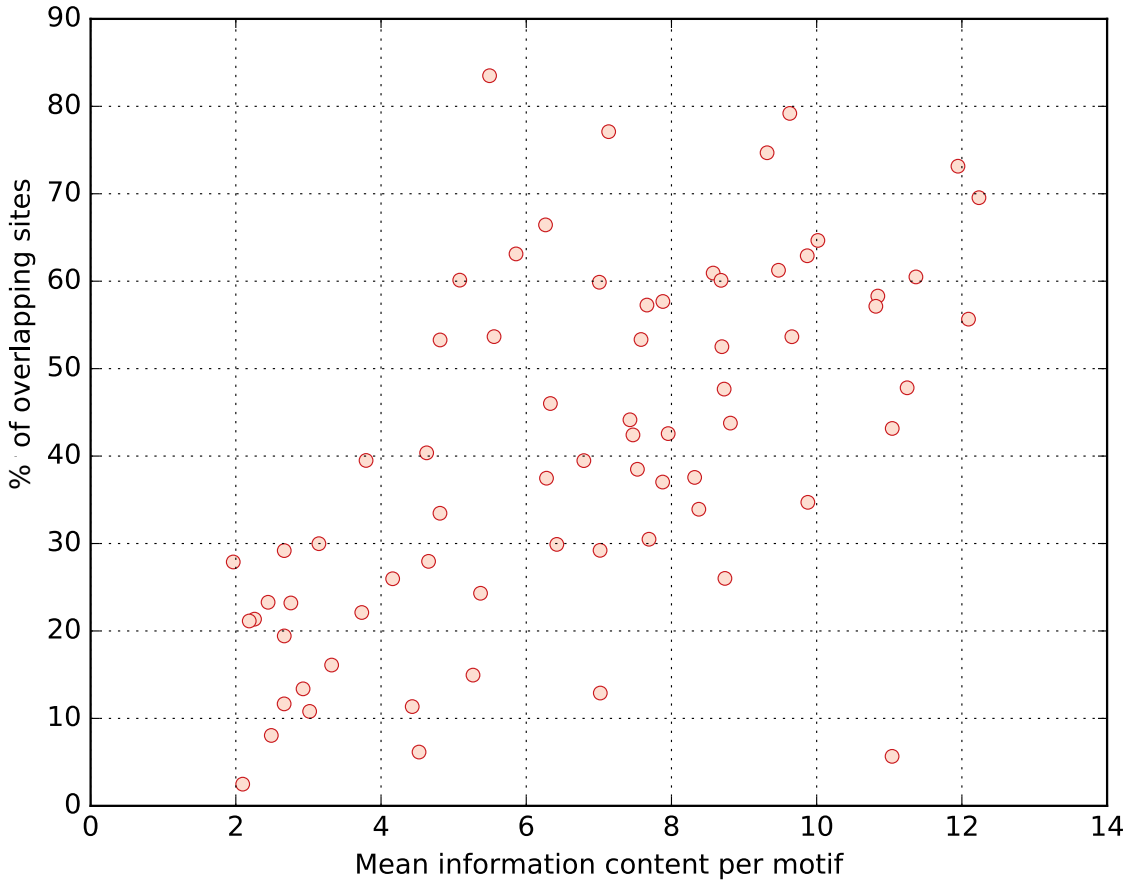
(C) HelT



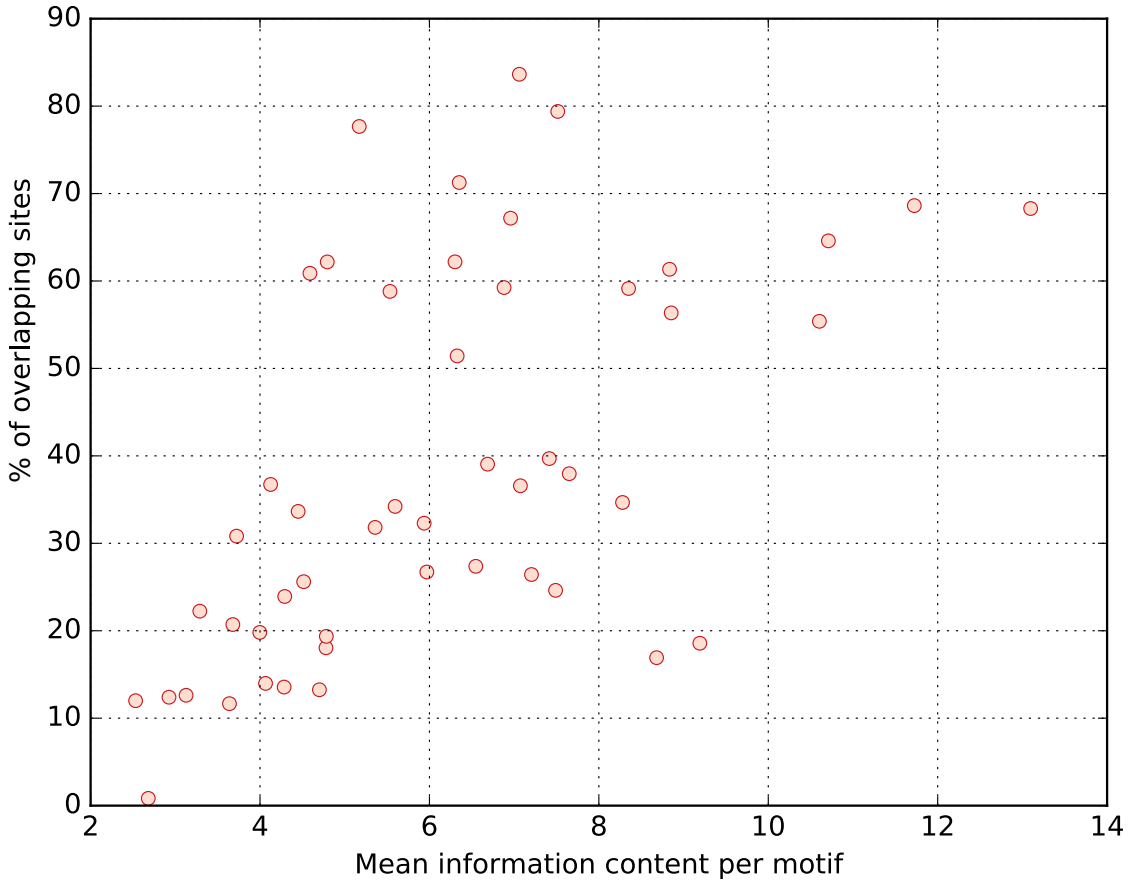
(D) ProT



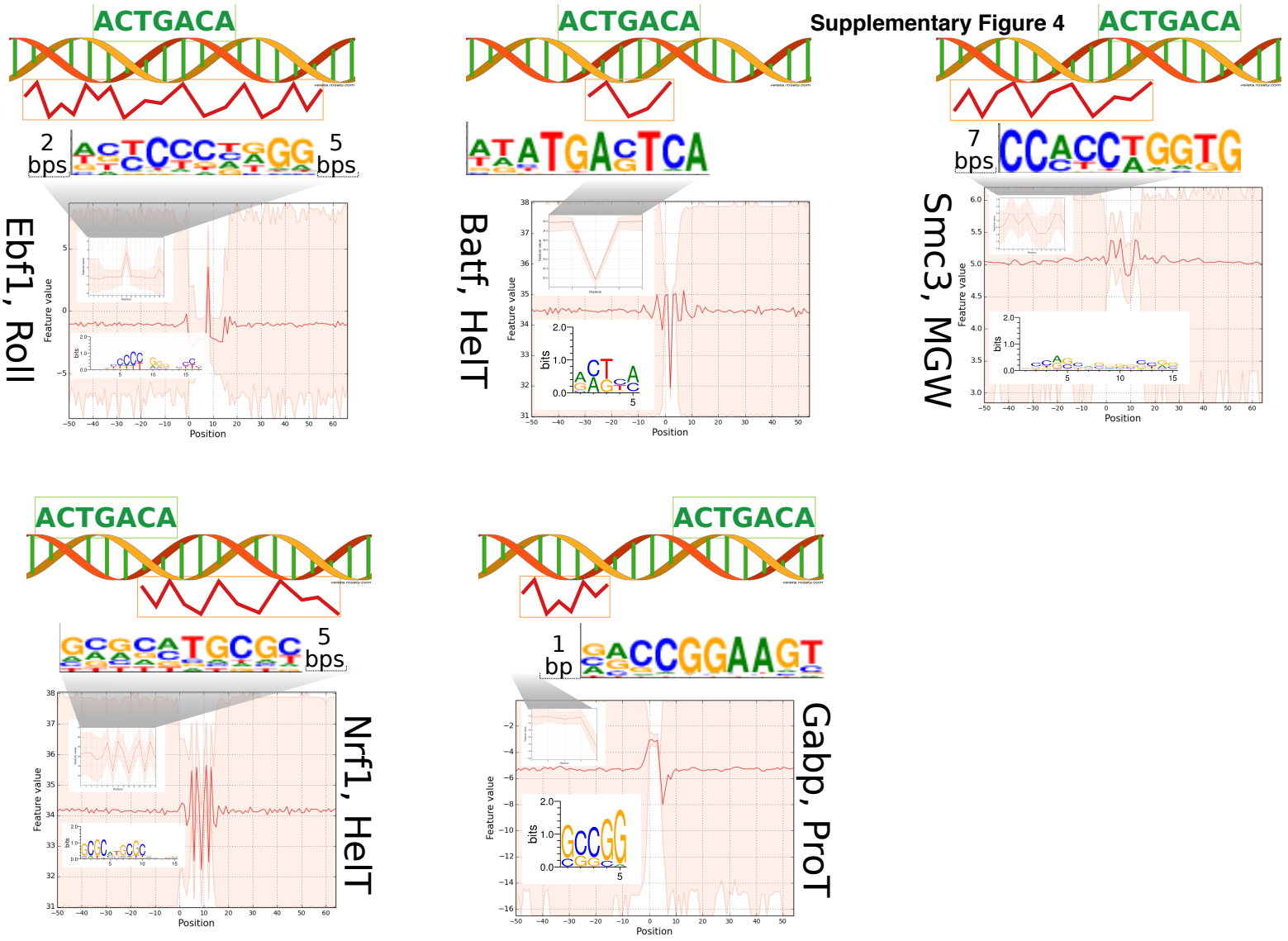
(A) Motifs detected in K562 Cell-line



(B) Motifs detected in Gm12878 Cell-line

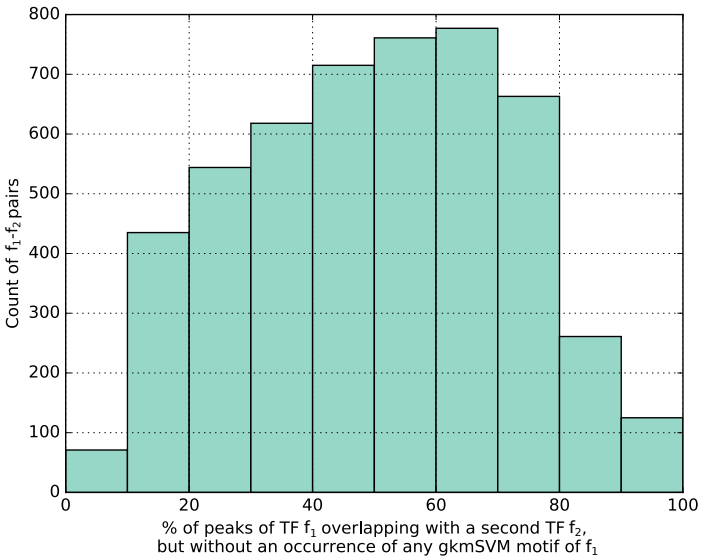


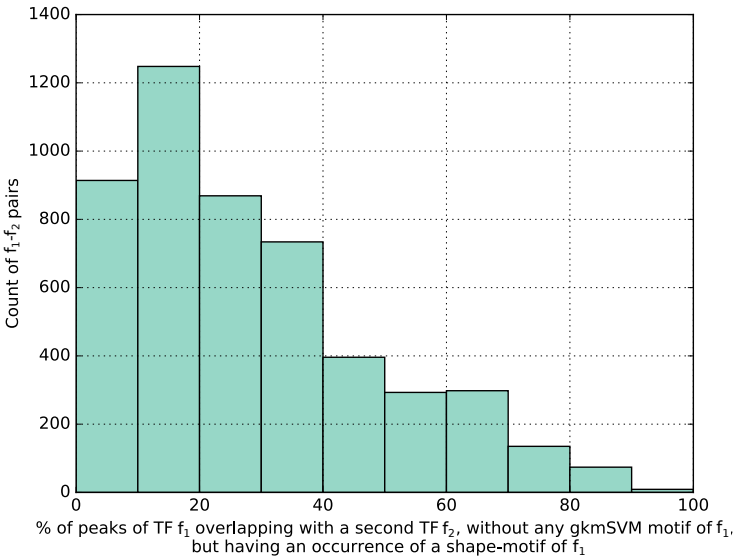
Supplementary Figure 4



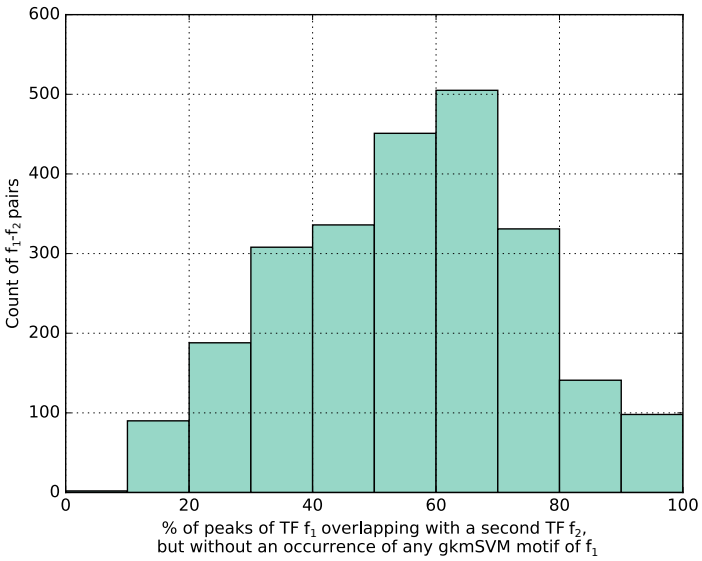
Supplementary Figure 5

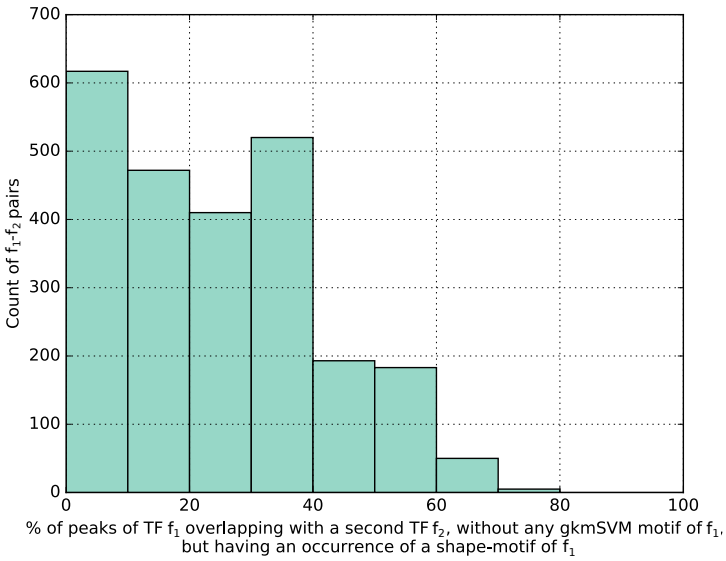
(A) K562 cell-line

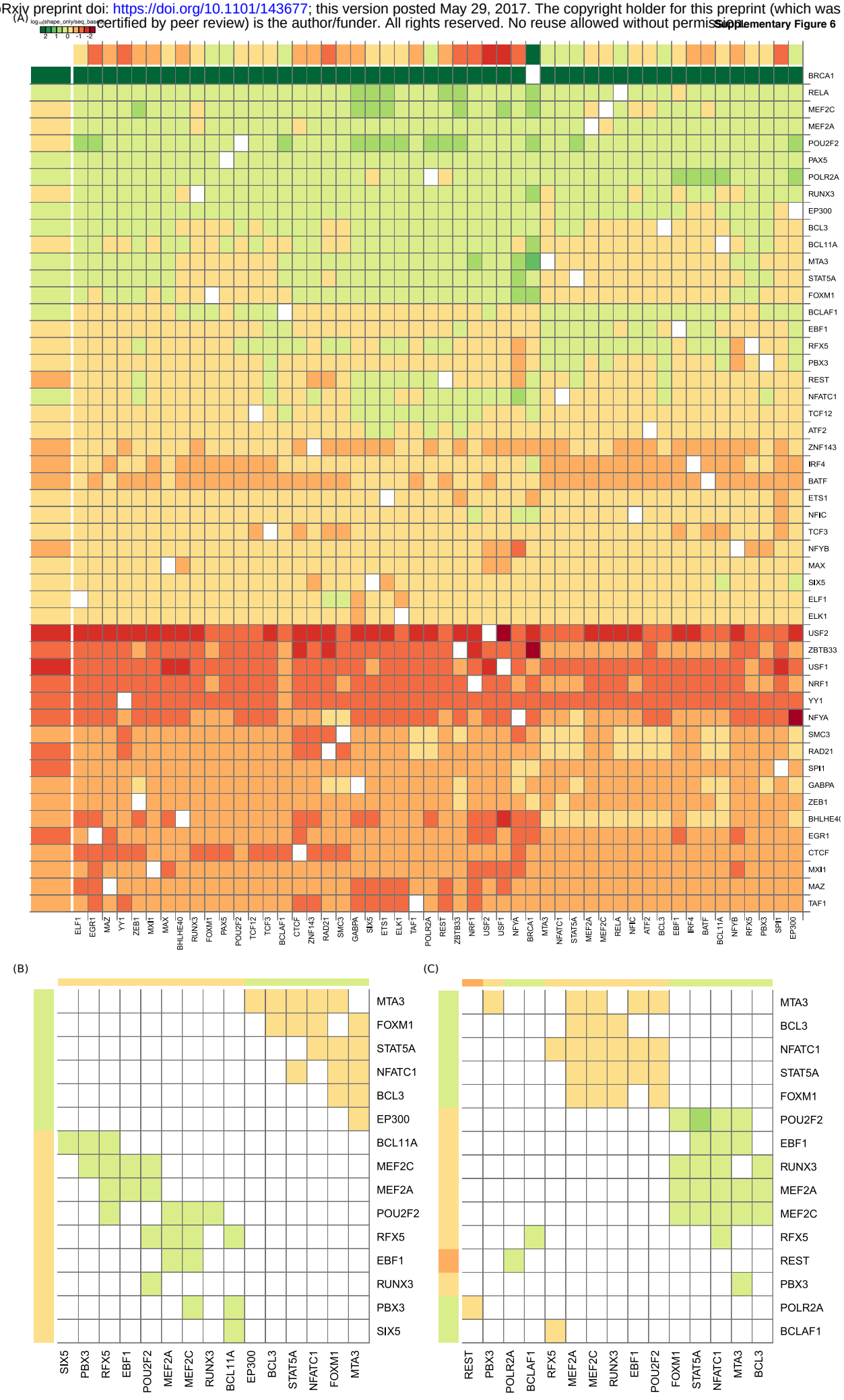




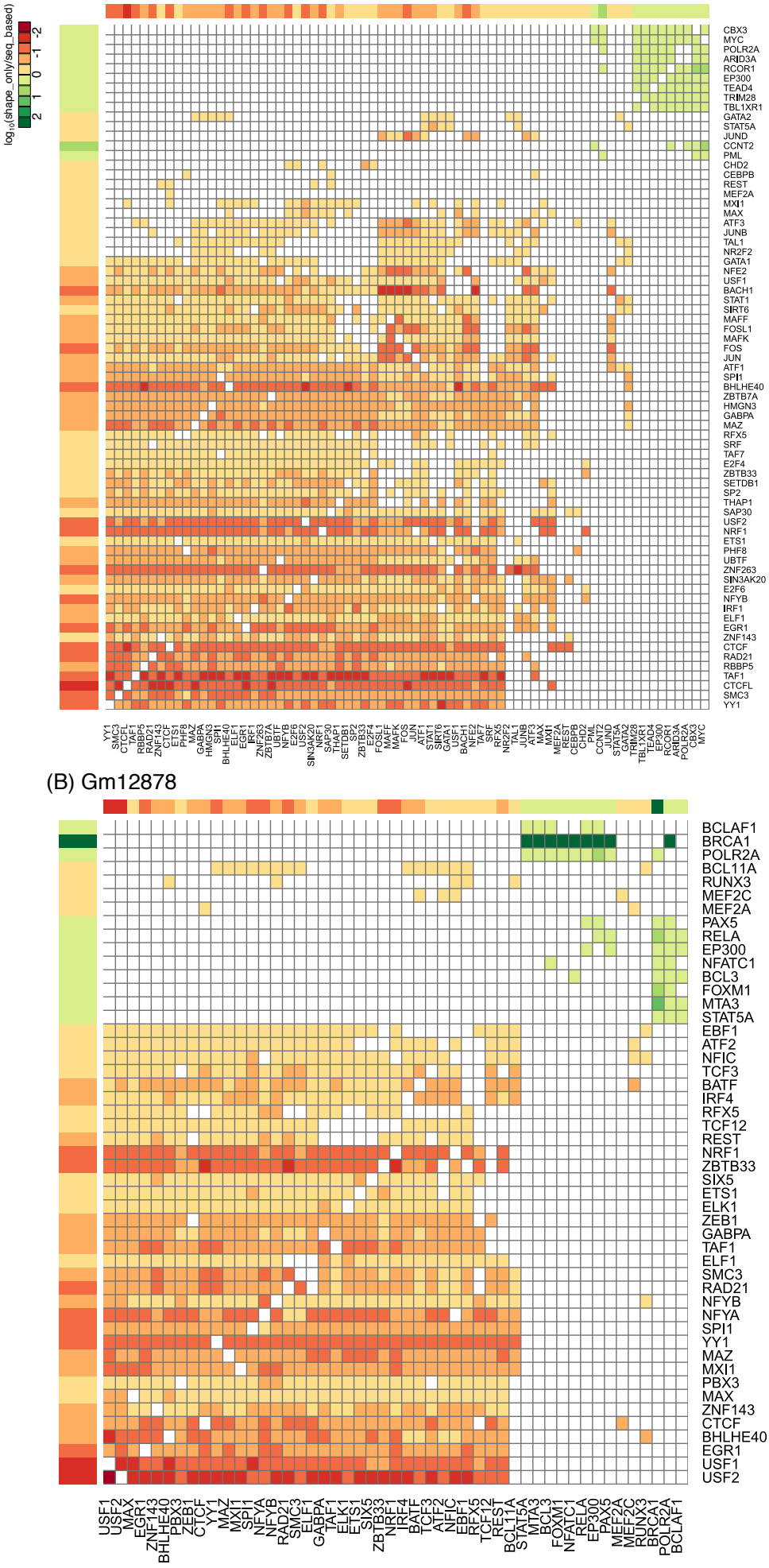
(B) Gm12878 cell-line





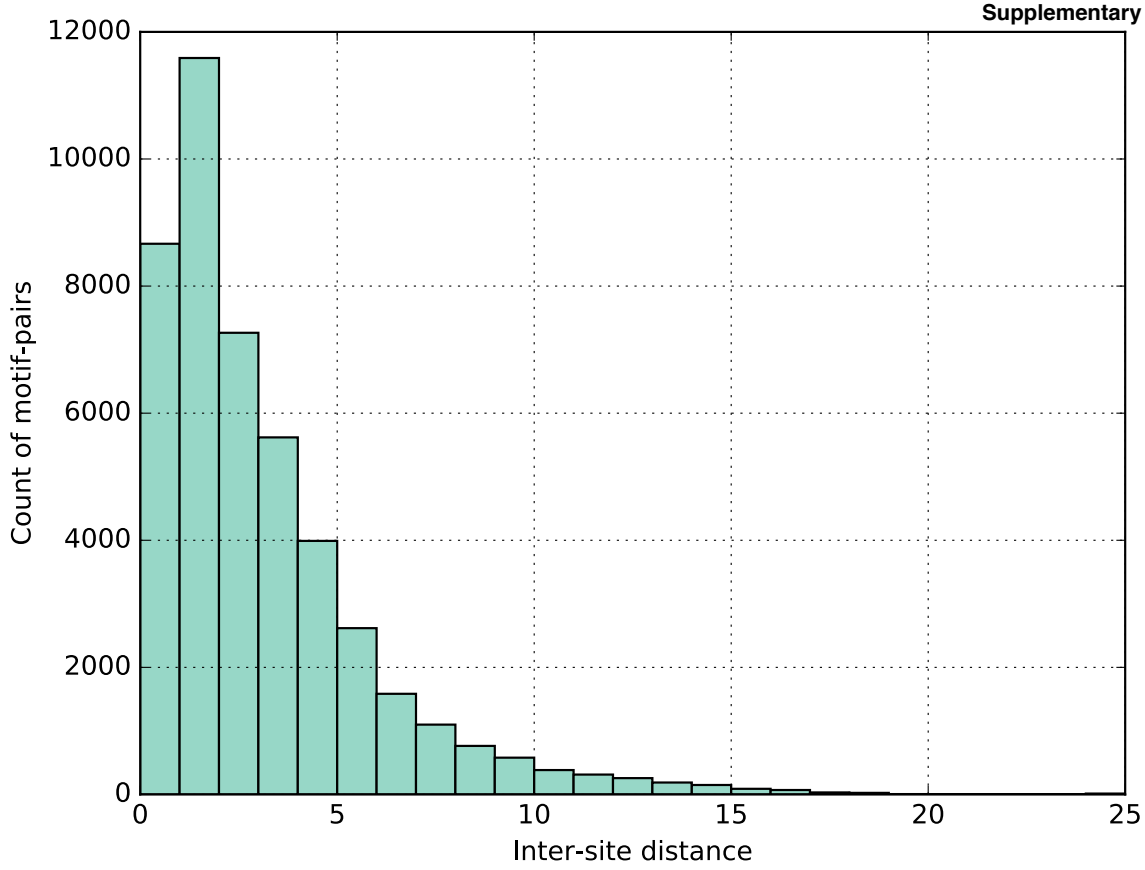


Supplementary Figure 7

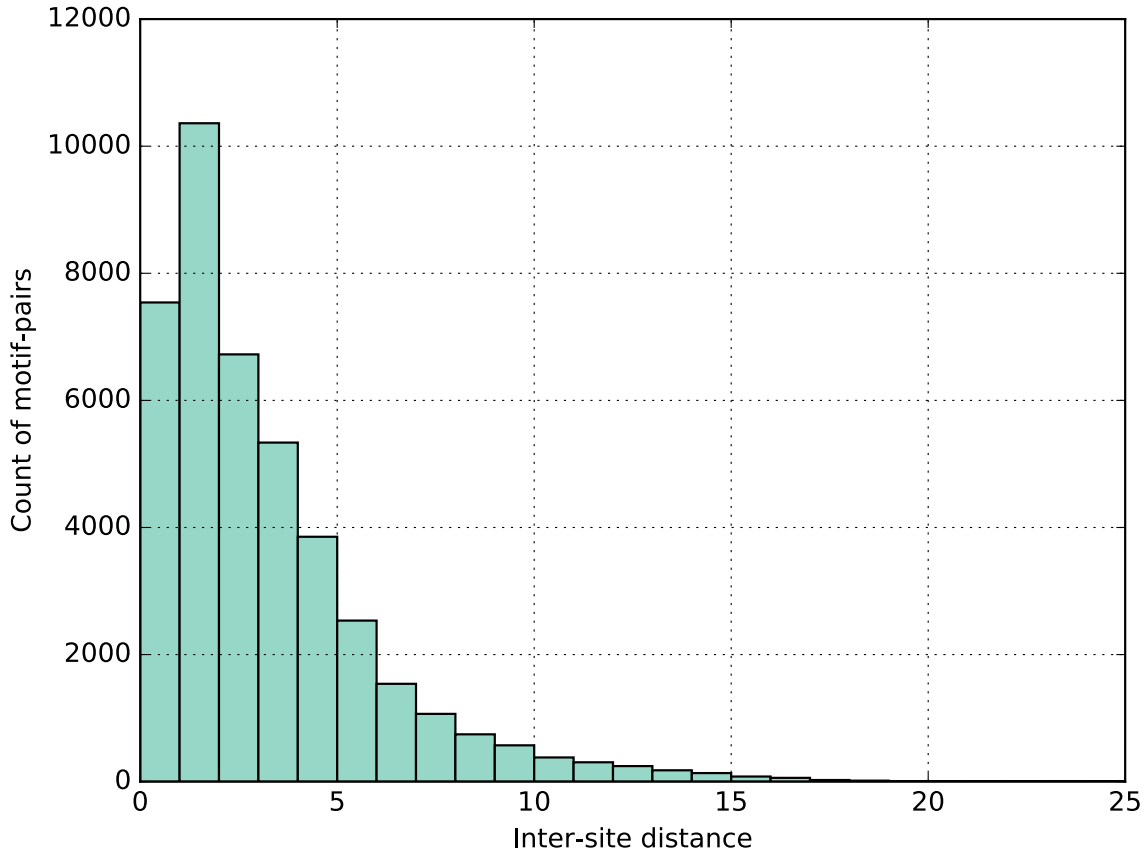


Supplementary Figure 8

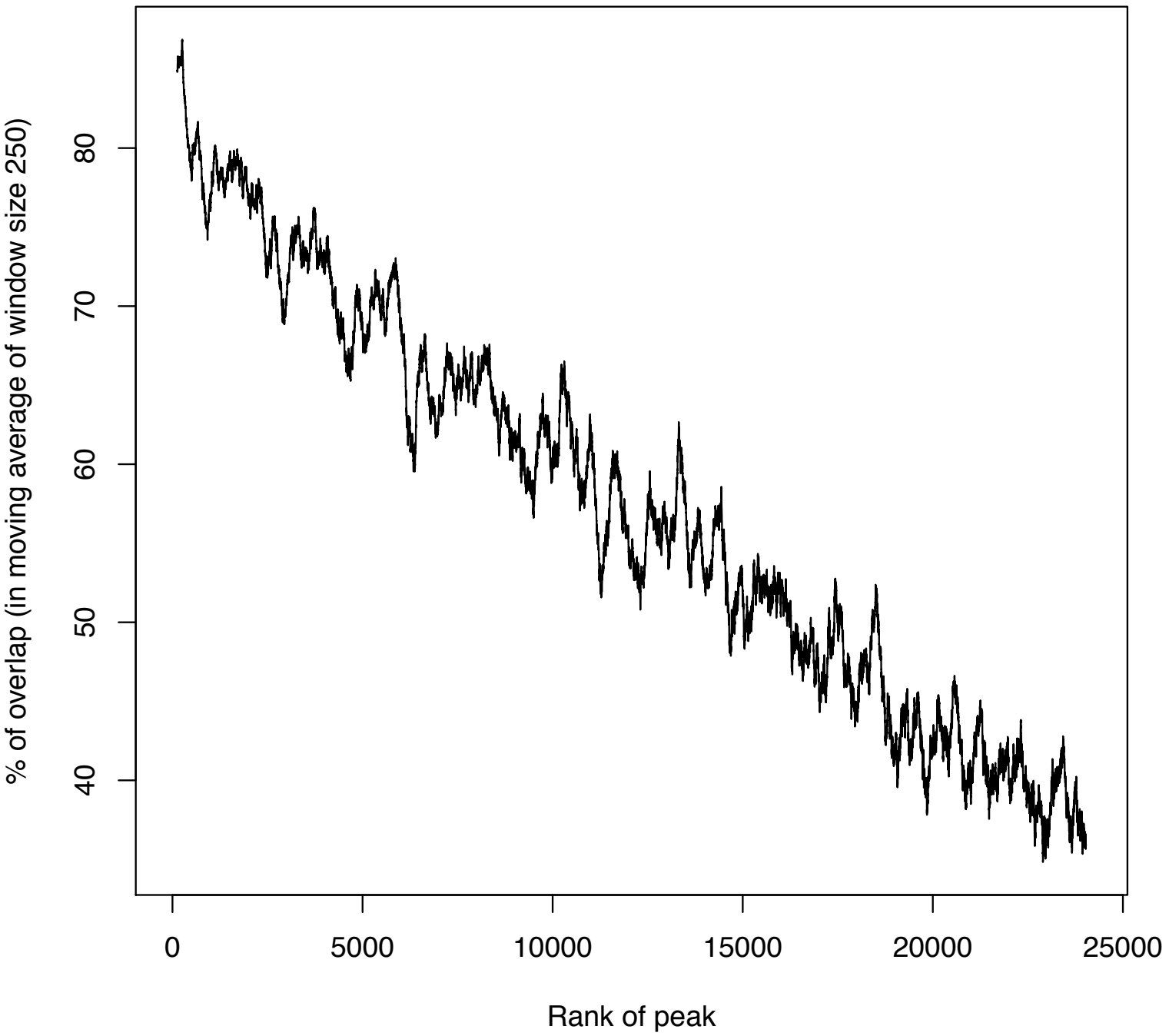
(A) Significantly enriched inter-site distances between any types of motifs of co-binding TFs



(B) Significantly enriched inter-site distances between sequence-shape or shape-shape motifs of co-binding TFs

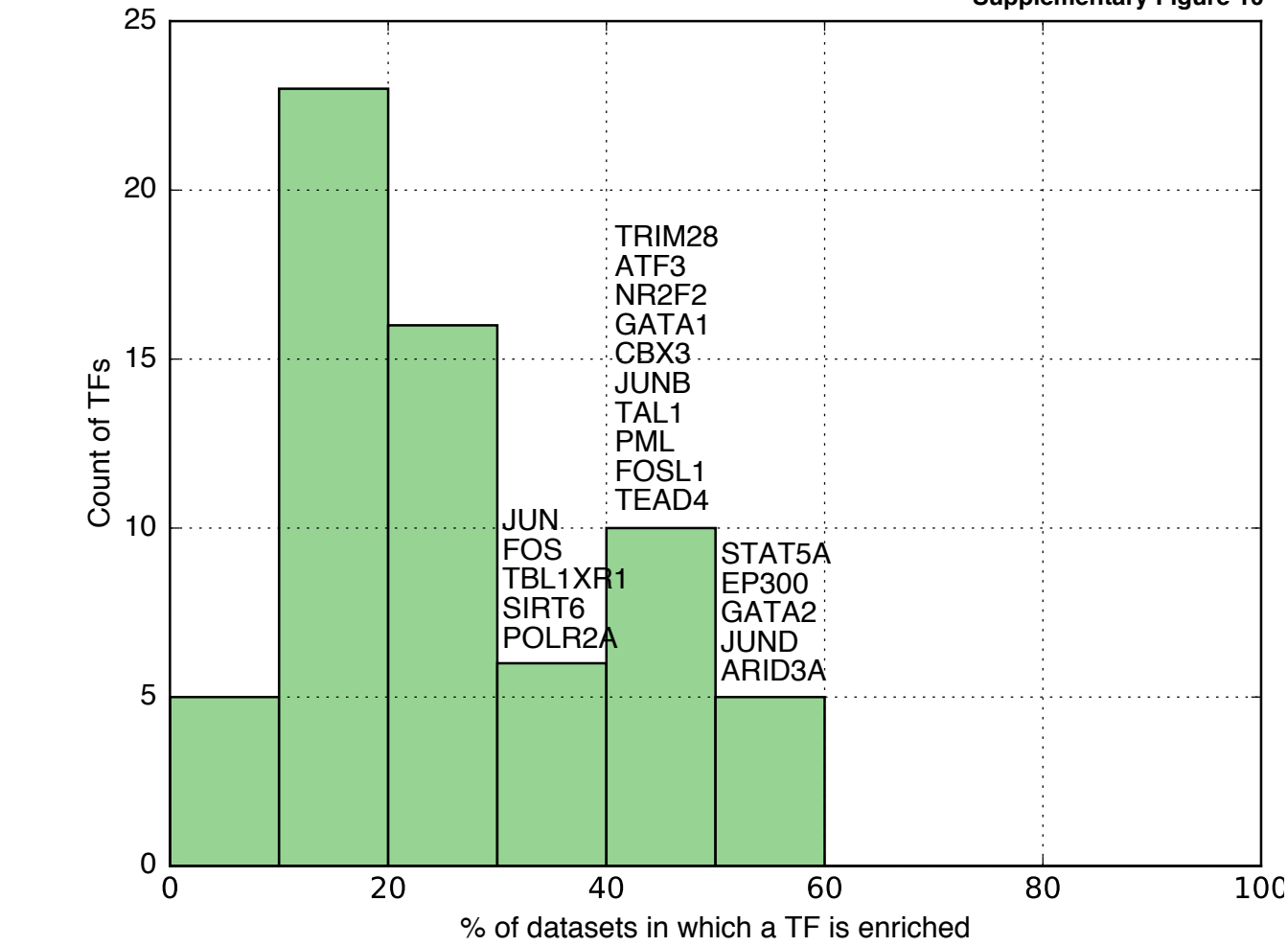


Supplementary Figure 9

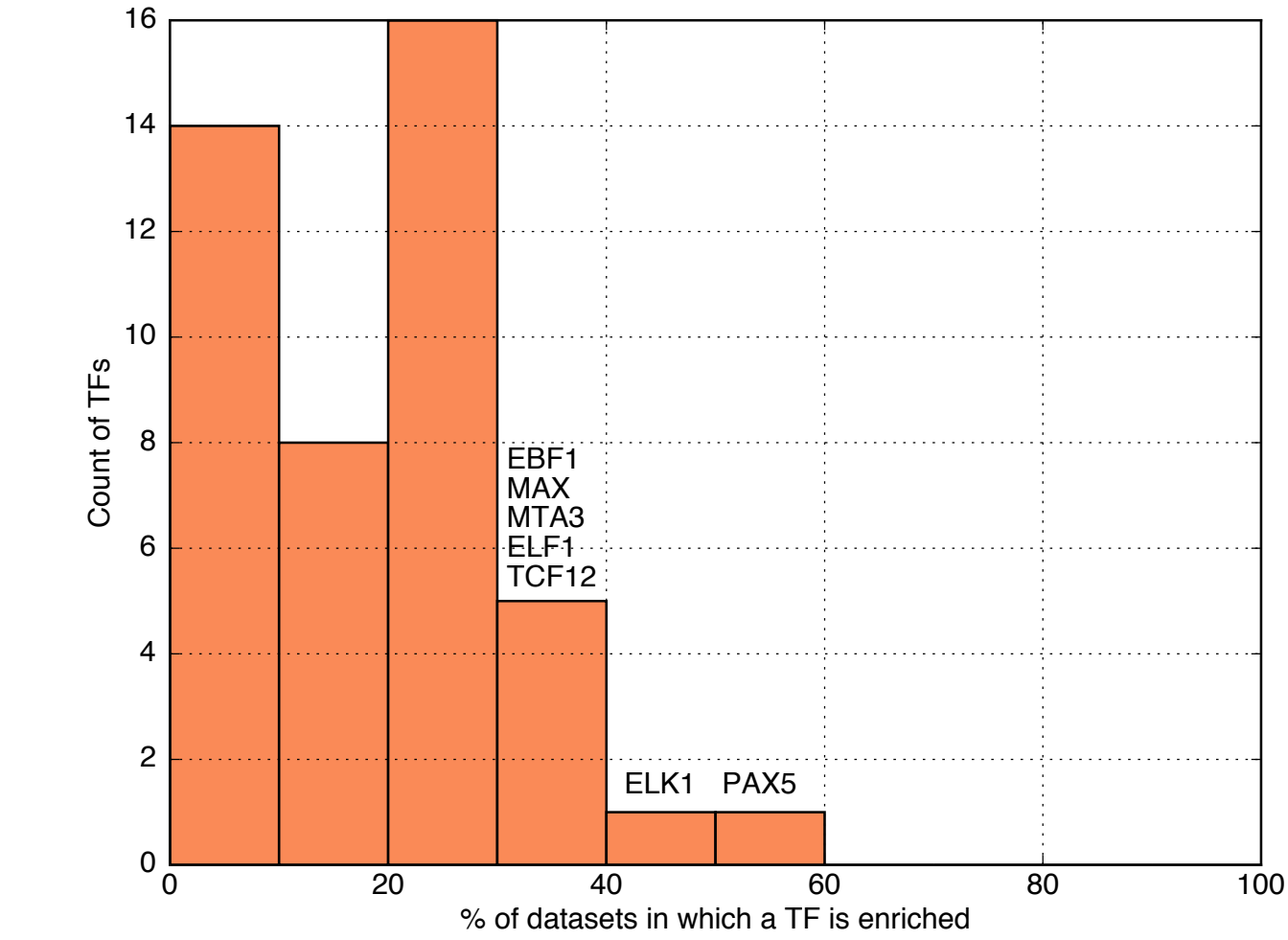


Supplementary Figure 10

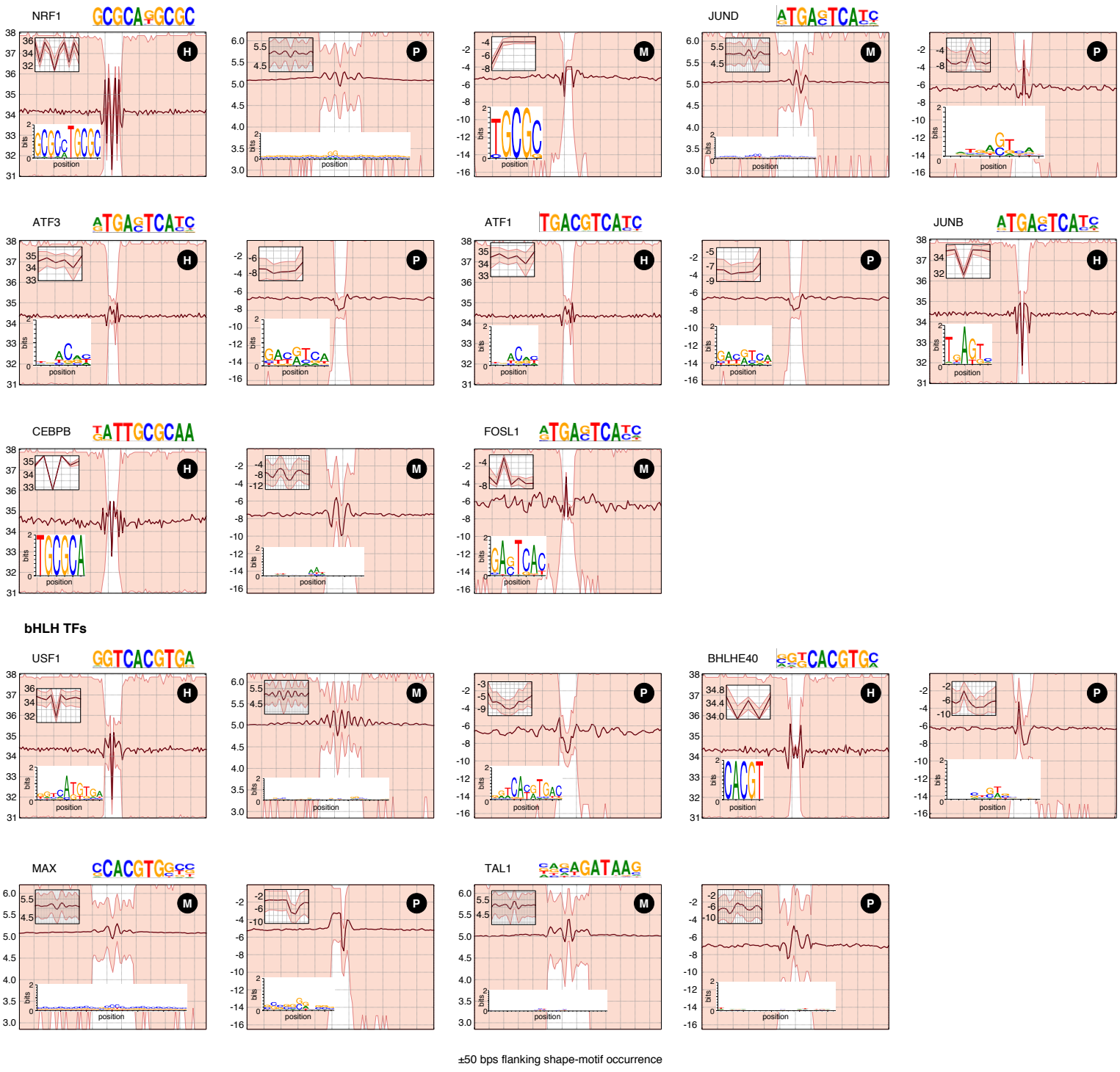
(A) K562



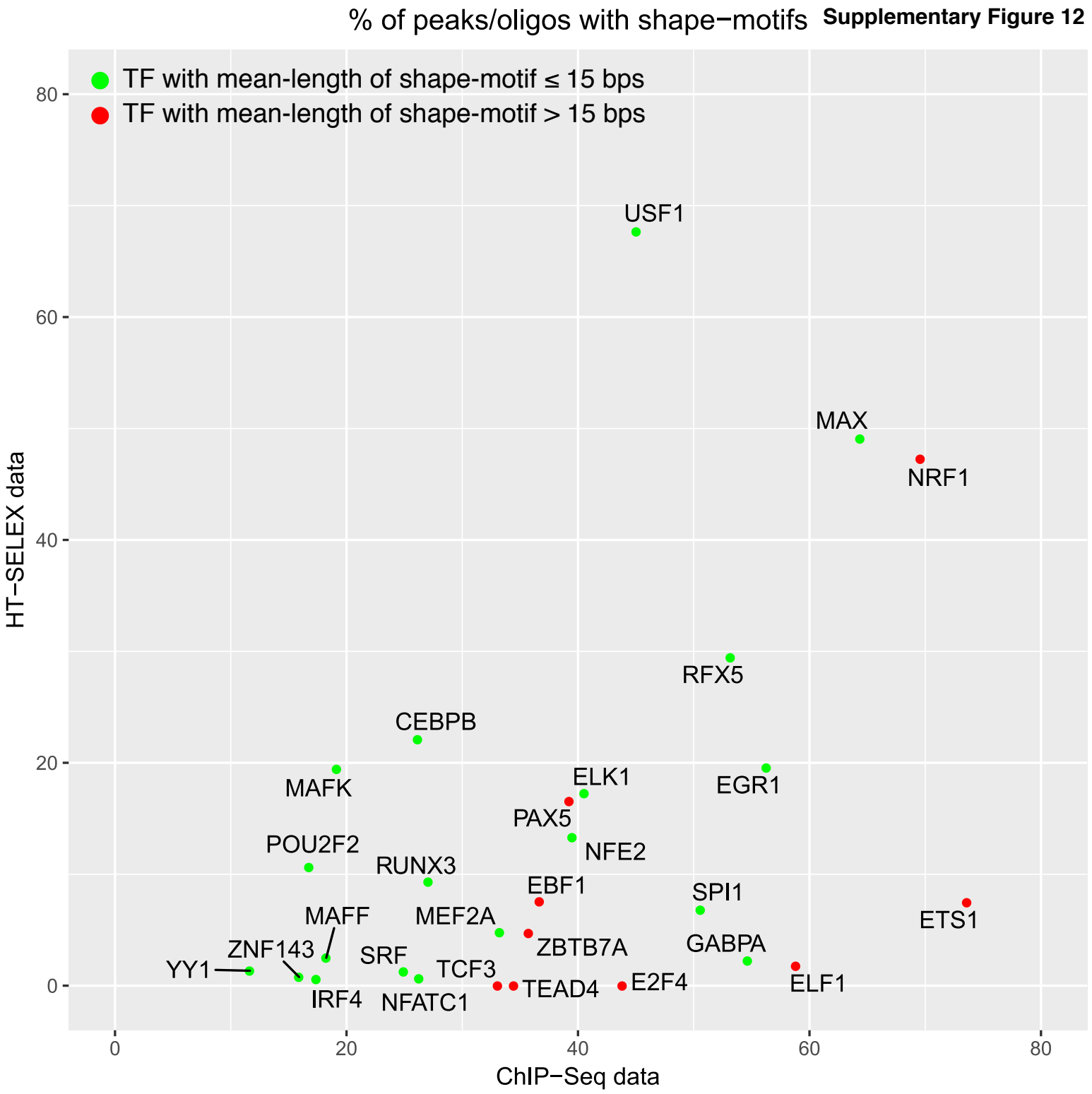
(B) Gm12878



Supplementary Figure 11



Supplementary Figure 12



Supplementary Table 1: Fraction (%) of peaks without occurrence of any sequence motif

TF	Cell line	Top 500 peaks	Top 1000 peaks	Top 2000 peaks	Top 5000 peaks	All peaks
ARID3A	K562	53.4	53.0	53.75	55.16	57.61
ATF1	K562	18.2	22.3	29.5	46.34	66.46
ATF2	Gm12878	45.8	46.4	47.7	51.82	63.55
ATF3	K562	42.4	34.0	26.95	25.3	33.87
BACH1	K562	5.6	8.0	11.75		24.01
BATF	Gm12878	22.8	25.9	29.05	33.32	47.74
BCL11A	Gm12878	40.8	43.9	44.6	48.52	55.09
BCL3	Gm12878	54.0	57.4	59.65	62.1	64.61
BCL3	K562	69.6	72.1			75.67
BCLAF1	Gm12878	32.0	40.9	50.85	58.18	60.03
BCLAF1	K562	50.2	57.1	64.7		70.48
BDP1	K562	97.8				98.07
BHLHE40	Gm12878	20.2	24.6	28.9	38.72	55.46
BHLHE40	K562	9.6	13.5	20.7	29.56	51.59
BRCA1	Gm12878	43.8				47.91
CBX3	K562	64.2	66.8	69.1	72.04	74.96
CCNT2	K562	53.6	58.5	65.7	72.32	78.07
CEBPB	K562	27.2	27.9	31.05	34.86	41.84
CHD1	Gm12878	46.8	53.1	56.75	61.94	62.99
CHD1	K562	34.0	33.5	33.1	36.62	41.53
CHD2	Gm12878	10.4	19.8	27.7	40.34	55.72
CHD2	K562	4.4	14.9	23.65	39.32	49.22
CTCF	Gm12878	2.2	2.3	2.25	2.6	8.65
CTCF	K562	1.2	2.6	3.25	3.52	11.14
CTCFL	K562	3.4	3.2	4.0	5.9	8.31
E2F4	Gm12878	25.6	26.8	30.5		39.8
E2F4	K562	14.8	17.0	20.5	28.66	34.75
E2F6	K562	7.0	10.9	12.45	17.32	32.65
EBF1	Gm12878	6.6	8.4	11.35	16.74	37.77
EGR1	Gm12878	3.0	3.7	4.95	7.56	16.44
EGR1	K562	0.6	0.6	1.15	2.42	11.61
ELF1	Gm12878	6.4	6.4	8.4	14.24	40.28
ELF1	K562	2.8	3.2	3.95	7.5	27.71
ELK1	Gm12878	8.8	11.8	18.95	42.04	44.99
EP300	Gm12878	57.4	55.1	52.35	54.54	54.78
EP300	K562	40.2	40.7	41.85	45.78	59.04
ETS1	Gm12878	29.8	34.5	39.1		45.0
ETS1	K562	26.4	31.2	32.8	37.16	40.91
EZH2	Gm12878	60.8	61.3	63.85		65.05
FOS	K562	17.0	16.7	15.55	13.44	15.24

TF	Cell line	Top 500 peaks	Top 1000 peaks	Top 2000 peaks	Top 5000 peaks	All peaks
FOSL1	K562	3.4	4.8	5.6	10.76	21.06
FOXM1	Gm12878	47.0	47.7	49.75	54.18	65.49
GABPA	Gm12878	5.8	6.7	8.8	21.04	27.93
GABPA	K562	3.6	4.8	9.05	24.1	46.84
GATA1	K562	29.6	32.1	35.6		41.88
GATA2	K562	35.4	38.1	39.95	46.06	51.32
GTF2B	K562	48.4	47.3	45.75		47.23
HMGN3	K562	24.6	23.3	23.95	24.94	32.8
IKZF1	Gm12878	55.2	56.1	57.95	62.2	65.61
IRF1	K562	9.2	9.3	10.4	16.1	31.03
IRF4	Gm12878	22.6	27.8	30.5	35.14	51.37
JUNB	K562	10.0	10.9	13.3	20.68	35.24
JUND	K562	5.2	5.7	8.05	16.04	47.21
JUN	K562	4.0	5.4	7.75	12.66	21.28
KAP1	K562	56.2	59.7	61.9	68.0	68.76
KDM5B	K562	49.2	47.7	45.15	46.48	52.59
MAFF	K562	3.0	6.6	9.1	14.94	30.39
MAFK	K562	3.8	7.6	9.9	15.58	31.96
MAX	Gm12878	11.2	14.9	23.5	38.08	54.08
MAX	K562	15.6	16.4	19.8	27.14	55.04
MAZ	Gm12878	7.0	7.4	9.55	14.32	31.28
MAZ	K562	6.4	7.0	8.0	10.46	24.93
MEF2A	Gm12878	28.4	34.2	40.85	54.26	71.07
MEF2A	K562	30.8	35.4	42.6	59.44	61.5
MEF2C	Gm12878	28.8	33.9	41.0	51.54	60.36
MTA3	Gm12878	49.0	49.2	52.35	58.52	65.81
MXI1	Gm12878	33.0	34.7	40.4	53.4	69.23
MXI1	K562	19.2	23.5	33.55	49.46	54.84
MYC	K562	9.8	12.3	15.9	22.74	32.71
NFATC1	Gm12878	52.4	55.6	59.45	64.76	69.57
NFE2	K562	5.8	7.1	9.55		9.9
NFIC	Gm12878	41.2	41.7	41.95	43.24	55.98
NFYA	Gm12878	22.0	23.8			28.03
NFYA	K562	15.0	15.0	15.1		21.26
NFYB	Gm12878	10.2	10.5	12.7	18.58	32.62
NFYB	K562	7.8	8.0	8.8	10.46	14.63
NR2C2	Gm12878	34.0	54.7			60.25
NR2F2	K562	36.0	39.2	41.7	45.62	57.11
NRF1	Gm12878	0.0	0.2	0.65	7.8	11.47
NRF1	K562	0.4	0.5	1.5		9.88
PAX5	Gm12878	26.0	29.5	35.4	41.5	55.05
PBX3	Gm12878	11.0	14.6	22.75	37.62	49.26

TF	Cell line	Top 500 peaks	Top 1000 peaks	Top 2000 peaks	Top 5000 peaks	All peaks
PHF8	K562	20.2	20.8	21.6	22.36	28.06
PML	Gm12878	50.2	47.3	47.75	50.14	61.3
PML	K562	52.4	52.4	57.95	68.18	78.59
POLR2A	Gm12878	49.8	46.9	45.6	47.18	51.31
POLR2A	K562	47.8	45.0	42.1	41.8	48.37
POU2F2	Gm12878	30.2	32.0	36.8	49.76	69.71
RAD21	Gm12878	5.8	5.0	4.95	4.76	13.96
RAD21	K562	3.2	3.3	3.3	4.98	13.03
RBBP5	K562	17.4	15.4	15.4	17.86	28.41
RCOR1	K562	56.0	57.4	63.0	71.9	74.12
RELA	Gm12878	11.4	17.0	21.85	34.72	55.88
REST	Gm12878	0.0	0.1	0.45	12.46	21.68
REST	K562	0.0	0.1	1.1	21.04	45.93
RFX5	Gm12878	15.4	18.5	25.55		45.43
RFX5	K562	28.4	40.2	49.9		51.34
RUNX3	Gm12878	17.8	19.6	22.35	28.42	58.11
RXRA	Gm12878	36.8	34.6			36.44
SAP30	K562	32.0	32.7	34.55	36.16	38.88
SETDB1	K562	65.6	63.6	60.2		56.13
SIN3A	Gm12878	37.2	37.1	40.9	46.48	52.61
SIN3AK20	K562	27.2	29.1	28.95	31.52	37.4
SIRT6	K562	41.2	48.5	54.6		55.4
SIX5	Gm12878	11.0	13.4	22.5		41.21
SIX5	K562	11.2	16.7	26.7		45.52
SMC3	Gm12878	5.4	5.9	6.2	6.74	18.63
SMC3	K562	4.2	4.1	4.55	5.98	12.26
SP1	Gm12878	10.0	13.0	22.1	40.08	61.79
SP1	K562	11.4	15.0	17.85	26.0	30.61
SP2	K562	12.0	14.5	19.9		25.8
SPI1	Gm12878	3.0	4.3	5.8	8.66	19.0
SPI1	K562	3.8	4.7	6.8	9.62	20.36
SRF	K562	25.4	29.1	32.95		44.54
STAT1	K562	9.2	12.8	27.05		30.28
STAT2	K562	11.6	25.9			40.46
STAT5A	Gm12878	54.6	56.2	60.75	68.2	70.94
STAT5A	K562	29.8	31.8	35.3	42.04	52.04
TAF1	Gm12878	30.6	31.5	31.15	37.68	51.08
TAF1	K562	27.0	26.5	27.45	30.74	43.74
TAF7	K562	76.0	67.9	63.95		62.77
TAL1	K562	24.8	28.1	30.25	34.22	48.8
TBL1XR1	Gm12878	47.0	48.6	51.65	56.44	63.16
TBL1XR1	K562	47.6	51.2	53.6	57.54	57.59

TF	Cell line	Top 500 peaks	Top 1000 peaks	Top 2000 peaks	Top 5000 peaks	All peaks
TBP	Gm12878	60.4	62.2	65.5	70.0	75.99
TBP	K562	46.6	46.6	45.65	49.9	58.29
TCF12	Gm12878	9.8	15.0	16.5	21.3	36.95
TCF3	Gm12878	15.2	16.4	19.1	24.08	35.78
TEAD4	K562	39.6	39.0	42.25	46.84	67.59
THAP1	K562	12.8	17.4	21.55		24.76
TRIM28	K562	49.4	50.4	54.05	61.92	72.26
UBTF	K562	9.8	9.6	9.95	16.84	19.43
USF1	Gm12878	0.6	2.7	4.75	11.82	22.38
USF1	K562	11.6	17.1	23.75	31.18	42.67
USF2	Gm12878	2.0	3.6	6.95	14.76	26.45
USF2	K562	2.4	2.8	5.8		14.27
WRNIP1	Gm12878	57.4	61.6	62.15	64.52	65.28
YY1	Gm12878	4.6	4.7	6.95	20.24	56.47
YY1	K562	3.0	2.7	3.45	12.34	36.81
ZBTB33	Gm12878	6.4	16.4	32.1		34.14
ZBTB33	K562	9.6	15.8	28.45		36.65
ZBTB7A	K562	5.4	5.9	7.0	10.78	20.11
ZEB1	Gm12878	8.0	12.6	16.55		23.72
ZNF143	Gm12878	2.4	5.4	16.45	42.8	65.56
ZNF143	K562	1.6	3.2	14.05	41.42	69.0
ZNF263	K562	4.2	6.3	10.75		16.98
ZNF274	K562	40.8	55.7			68.85

Supplementary Table 2: Average number of sequence-, shape-, and overlapping-sites per peak

TF	Cell line	Sequence-sites	Shape-sites	Overlapping-sites
ARID3A	K562	1.850983	4.011307	2.038847
ATF1	K562	2.116156	1.178377	4.960402
ATF2	Gm12878	1.661939	1.273406	2.721612
ATF3	K562	3.033527	2.009117	3.816035
BACH1	K562	3.325033	1.028633	4.468291
BATF	Gm12878	2.128514	1.196855	3.909628
BCL11A	Gm12878	2.476075	1.820723	2.466671
BCL3	Gm12878	2.417045	2.748787	2.378821
BCLAF1	Gm12878	1.597341	2.606249	1.843890
BHLHE40	Gm12878	2.546575	1.097337	5.951323
BHLHE40	K562	2.760358	1.031258	5.888942
BRCA1	Gm12878	1.077586	1.070833	1.025000
CBX3	K562	1.562367	3.590288	1.512876
CCNT2	K562	2.900970	2.273263	1.234176
CEBPB	K562	2.175771	1.330146	2.097579
CHD2	K562	2.696203	2.265794	3.522910
CTCF	Gm12878	3.436919	1.267716	5.005814
CTCF	K562	3.764529	1.393798	4.494983
CTCFL	K562	2.981050	1.579345	3.601432
E2F4	K562	3.076991	2.288074	3.251491
E2F6	K562	2.595805	1.739135	2.896812
EBF1	Gm12878	2.182432	1.459677	3.384931
EGR1	Gm12878	3.426606	1.179414	5.034096
EGR1	K562	4.093689	1.239442	4.443457
ELF1	Gm12878	2.928476	1.851005	3.061573
ELF1	K562	2.719348	1.315760	2.617989
ELK1	Gm12878	2.652687	1.448126	3.697882
EP300	Gm12878	2.098999	2.271837	2.248865
EP300	K562	2.476136	3.983849	2.577470
ETS1	Gm12878	2.725357	2.434259	2.539986
ETS1	K562	1.896573	3.165377	2.102642
FOS	K562	4.417042	1.514894	5.020009
FOSL1	K562	3.245138	1.644859	3.978272
FOXM1	Gm12878	1.508929	1.255067	2.982677
GABPA	Gm12878	3.643373	1.434904	4.034830
GABPA	K562	2.979454	1.101957	3.000306
GATA1	K562	3.135582	2.863236	3.288367
GATA2	K562	3.094256	2.775937	3.028172
HMGN3	K562	2.449555	1.555199	3.874708
IRF1	K562	3.067660	1.254019	4.311335

TF	Cell line	Sequence-sites	Shape-sites	Overlapping-sites
IRF4	Gm12878	2.258181	1.295021	3.245764
JUN	K562	3.816679	1.820437	4.696599
JUNB	K562	3.238352	1.875949	3.854695
JUND	K562	2.448237	1.894558	2.814551
MAFF	K562	2.107170	1.030609	2.281654
MAFK	K562	1.986234	1.037483	2.077616
MAX	Gm12878	2.142541	1.897633	4.317953
MAX	K562	2.536787	1.589831	3.926057
MAZ	Gm12878	2.737835	1.331350	4.165066
MAZ	K562	2.213455	1.381204	3.914225
MEF2A	Gm12878	1.961067	1.268266	1.200000
MEF2A	K562	2.155431	1.634840	1.383333
MEF2C	Gm12878	1.772852	1.264671	1.509739
MTA3	Gm12878	1.146839	1.553568	2.371668
MXI1	Gm12878	2.446319	1.051337	4.309701
MXI1	K562	2.779145	1.968308	3.921905
MYC	K562	2.130208	2.156751	3.461987
NFATC1	Gm12878	1.988894	3.680381	2.921167
NFE2	K562	3.365909	1.304368	3.288228
NFIC	Gm12878	1.459487	1.348810	2.859092
NFYA	Gm12878	4.085633	1.097403	4.026087
NFYB	Gm12878	4.442175	1.688961	4.170677
NFYB	K562	4.085709	1.425311	4.518740
NR2F2	K562	2.094563	2.484291	2.046898
NRF1	Gm12878	4.154159	1.490836	6.250297
NRF1	K562	4.096960	1.719672	6.604682
PAX5	Gm12878	1.778102	2.138147	1.996103
PBX3	Gm12878	2.409104	1.267372	2.907345
PHF8	K562	1.746489	2.622913	3.178094
PML	K562	3.196313	2.115961	3.093802
POLR2A	Gm12878	1.635739	4.181538	2.132422
POLR2A	K562	1.307531	3.081256	1.271011
POU2F2	Gm12878	4.509253	1.261058	5.444853
RAD21	Gm12878	3.433969	1.400006	4.776081
RAD21	K562	3.686156	1.279022	4.771104
RBBP5	K562	2.515934	1.686058	5.463606
RCOR1	K562	3.086114	3.380000	2.727582
RELA	Gm12878	2.256219	1.924554	2.601801
REST	Gm12878	3.500144	1.258964	5.137623
REST	K562	2.735634	1.477112	3.993125
RFX5	Gm12878	2.921343	1.518903	2.040150
RFX5	K562	2.310762	1.575659	2.848262

TF	Cell line	Sequence-sites	Shape-sites	Overlapping-sites
RUNX3	Gm12878	3.337189	1.282730	3.338013
SAP30	K562	1.469083	5.175806	2.737982
SETDB1	K562	1.761349	2.211158	3.122543
SIN3AK20	K562	1.994061	1.319902	2.951697
SIRT6	K562	2.796984	1.835714	2.962085
SIX5	Gm12878	3.567405	1.817408	2.742913
SMC3	Gm12878	3.375284	1.205541	4.510937
SMC3	K562	3.363107	1.116697	5.227886
SP2	K562	1.759151	2.542544	2.244420
SPI1	Gm12878	2.005668	1.046024	2.456784
SPI1	K562	1.922722	1.108044	2.000695
SRF	K562	3.189021	1.734824	2.465772
STAT1	K562	3.004299	1.633984	3.255127
STAT5A	Gm12878	1.230483	4.038239	3.325025
STAT5A	K562	2.591910	4.634488	2.749937
TAF1	Gm12878	2.665145	1.084440	2.681424
TAF1	K562	2.296463	1.202235	3.220945
TAF7	K562	2.407362	2.288889	2.504979
TAL1	K562	3.085925	1.855027	3.001542
TBL1XR1	K562	3.305104	4.528447	3.115244
TCF12	Gm12878	2.388021	3.442253	2.953016
TCF3	Gm12878	2.631579	2.928512	3.024879
TEAD4	K562	3.022547	2.658947	2.996171
THAP1	K562	2.364377	1.625373	3.284175
TRIM28	K562	3.013383	3.545075	2.930823
UBTF	K562	1.933504	1.668977	2.939831
USF1	Gm12878	2.264898	1.058475	3.749663
USF1	K562	1.981631	1.473624	2.437491
USF2	Gm12878	3.125816	1.039030	5.797151
USF2	K562	2.679178	1.404502	5.555013
YY1	Gm12878	2.354101	1.032558	3.325312
YY1	K562	2.681844	1.042067	3.044849
ZBTB33	Gm12878	3.754370	1.041958	6.364008
ZBTB33	K562	2.355086	1.595287	4.826633
ZBTB7A	K562	2.477973	1.730208	3.130002
ZEB1	Gm12878	3.468019	1.365584	3.651928
ZNF143	Gm12878	2.480012	1.263070	1.468606
ZNF143	K562	2.142286	1.059000	2.744661
ZNF263	K562	4.109995	1.230136	4.833759

Supplementary Table 3: Information content (bits) for each type of motif of each TF

TF	Cell line	HelT motif	MGW motif	ProT motif	Roll motif
ATF1	K562	4.354304		5.813853	
ATF2	Gm12878	3.739719		3.706825	
ATF3	K562	3.686794	5.597207	4.679672	
BACH1	K562	8.401840	8.373654	4.629956	
BATF	Gm12878	4.755129		4.831843	
BCL11A	Gm12878	2.731374		2.331119	
BCL3	Gm12878	8.307989	4.415979	8.087002	13.915730
BCLAF1	Gm12878		6.153377	5.784666	
BHLHE40	Gm12878	6.659689		6.043686	
BHLHE40	K562	4.831120		6.158475	
BRCA1	Gm12878	7.902666		13.303756	
CBX3	K562		3.974170	2.058881	
CCNT2	K562		13.223148	11.681911	8.210620
CEBPB	K562	3.033718		2.300033	
CHD2	K562	5.157121	8.678947	7.214341	
CTCF	Gm12878		9.591501	6.342796	10.569503
CTCF	K562		8.657654	8.240361	11.524574
CTCFL	K562		11.342471	14.267542	11.094728
E2F4	K562	7.629851	12.680834	12.811067	
E2F6	K562	8.058094	10.192821	11.383686	
EBF1	Gm12878	7.546488	8.776658	3.714710	8.257570
EGR1	Gm12878	8.707555	14.324367	12.780700	11.083092
EGR1	K562	8.566762	14.611101	10.949566	10.856076
ELF1	Gm12878	5.410424		8.677957	7.525383
ELF1	K562	6.040370		9.723409	9.366509
ELK1	Gm12878	5.746806		6.420904	7.473470
EP300	Gm12878	3.162812		2.687594	
EP300	K562		3.287761	2.046333	
ETS1	Gm12878		8.301907	6.676335	
ETS1	K562		10.002779	7.620823	
FOS	K562	5.182979	4.617328	7.777137	
FOSL1	K562	5.021982	5.388716	8.386084	
FOXM1	Gm12878	4.213639		3.916519	
GABPA	Gm12878	5.636910		8.110776	9.203928
GABPA	K562	5.335118		8.716376	8.357448
GATA1	K562		4.458784	3.010830	
GATA2	K562		3.702154	2.937098	
HMGN3	K562		14.511307	12.863219	8.898824
IRF1	K562	5.878582	7.743200		12.554550
IRF4	Gm12878	4.088219		4.161863	

TF	Cell line	HelT motif	MGW motif	ProT motif	Roll motif
JUN	K562	3.888448	5.250780	5.300358	
JUNB	K562	5.128315	3.345063	5.411392	
JUND	K562	4.050870	4.431581	5.952647	
MAFF	K562	6.562588	6.107544	8.285775	7.070965
MAFK	K562	6.510408	7.551979	9.655213	10.579278
MAX	Gm12878	5.598401	8.576249	8.068755	
MAX	K562	5.315556	8.767118	8.989175	
MAZ	Gm12878	7.873606	11.101825	13.699716	10.163128
MAZ	K562	7.908911	12.827615	12.629814	10.002484
MEF2A	Gm12878	2.680824			
MEF2A	K562			2.094583	
MEF2C	Gm12878	4.263320		4.305974	
MTA3	Gm12878	2.676849		4.681152	
MXI1	Gm12878	5.534359			
MXI1	K562	4.396907	10.588809	4.277373	
MYC	K562	6.051838	9.564069	7.372048	5.091212
NFATC1	Gm12878	3.754768		5.802839	
NFE2	K562	3.061157	4.814266	3.505833	
NFIC	Gm12878	3.927508		5.102922	
NFYA	Gm12878	4.589229			
NFYB	Gm12878		5.594612		
NFYB	K562	3.359020	7.754133		
NR2F2	K562			2.257454	
NRF1	Gm12878	11.949359	13.625259	13.717602	
NRF1	K562	10.288648	11.391690	12.420220	
PAX5	Gm12878	3.373766	5.922113	3.578923	
PBX3	Gm12878	4.540342		4.359496	
PHF8	K562			11.945955	
PML	K562			2.488596	
POLR2A	Gm12878			4.701124	
POLR2A	K562		6.291600	4.655549	4.853058
POU2F2	Gm12878	3.639135			
RAD21	Gm12878		5.435081	3.938752	9.532026
RAD21	K562		8.654341	6.573681	10.820522
RBBP5	K562		11.892422	8.139989	
RCOR1	K562		5.947349	3.099629	
RELA	Gm12878	2.811867			6.753642
REST	Gm12878	7.511681	7.408559	8.465631	10.018286
REST	K562	6.879823	8.100623	8.063087	8.775688
RFX5	Gm12878	2.541426		3.712873	
RFX5	K562	2.805591		3.482983	
RUNX3	Gm12878	3.491183		3.087217	

TF	Cell line	HelT motif	MGW motif	ProT motif	Roll motif
SAP30	K562			7.878489	
SETDB1	K562		8.650117	6.412054	
SIN3AK20	K562	5.404872	8.604263	8.733767	
SIRT6	K562			1.964513	
SIX5	Gm12878	5.358556			
SMC3	Gm12878	4.986087	7.022509	5.397144	10.118401
SMC3	K562	5.406092	9.008039	6.745234	10.366183
SP2	K562		7.612506	9.026121	
SPI1	Gm12878	6.526468		7.895665	6.451772
SPI1	K562	7.134858		8.053192	7.799134
SRF	K562			4.160145	
STAT1	K562			4.624593	7.933305
STAT5A	Gm12878			3.997081	
STAT5A	K562		3.488723	1.399948	
TAF1	Gm12878			8.279398	
TAF1	K562			9.316957	
TAF7	K562			5.370406	
TAL1	K562			2.184837	
TBL1XR1	K562		5.677345	3.181479	
TCF12	Gm12878		7.657051	4.972183	5.183052
TCF3	Gm12878		8.015065	5.357886	
TEAD4	K562		3.876757	1.975382	
THAP1	K562		8.715861	8.673127	
TRIM28	K562		3.201460	2.129453	
UBTF	K562		10.112320	11.750788	7.112115
USF1	Gm12878	7.276225		7.753492	
USF1	K562	5.924492	8.185294	8.175397	
USF2	Gm12878	6.325652		7.796670	
USF2	K562	5.404733	5.701908	7.890269	
YY1	Gm12878			8.853913	
YY1	K562			9.869987	
ZBTB33	Gm12878			5.170867	
ZBTB33	K562	5.630235	8.807290	5.942027	
ZBTB7A	K562	7.356040	12.470241	14.281782	9.150920
ZEB1	Gm12878	4.603663		8.050429	
ZNF143	Gm12878			9.192158	
ZNF143	K562			7.489105	9.983317
ZNF263	K562		10.278232	9.203515	9.407070