Supplemental Materials

Sequence Features and Chromatin Structure around the Genomic Regions Bound by 119 Human Transcription Factors

Running title: Analysis of Sites of 119 TFs in the Human Genome

Jie Wang1*, Jiali Zhuang1*, Sowmya Iyer1,2*, XinYing Lin1*, Troy W. Whitfield1, Melissa Greven1, Brian G. Pierce1, Xianjun Dong1, Anshul Kundaje3, Yong Cheng4, Oliver J. Rando1, Ewan Birney5, Richard M. Myers6, William S. Noble7, Michael Snyder4, Zhiping Weng1

* These authors contributed equally to this work
Supplemental Methods:

**ENCODER ChIP-seq, MNase-seq, and DNase-seq datasets**

All of the ChIP-seq datasets were processed with the ENCODE uniform ChIP-seq processing pipeline as described in the ENCODE companion manuscript (Gerstein et al. 2012). Briefly the SPP algorithm (Kharchenko et al. 2008) was applied to each ChIP-seq and matching input DNA control datasets to identify a list of genomic regions, called peaks, which showed statistically significant binding by the TF, indicated by significant enrichment of ChIP signal over control. For each peak, SPP reports the signal (enrichment of the ChIP signal over control) and summit (the position in the peak with the strongest signal). ENCODE generated at least two biological replicates for each ChIP-seq dataset, and the uniform processing pipeline assessed the consistency of the peaks called for the two replicates and determines an optimal number of reproducible peaks using the irreproducible discovery rate (IDR) approach (Li et al. 2011). We used this list of reproducible peaks as the input of our analysis. The signal and summit were computed by pooling the ChIP-seq data on the replicates. The peaks are 242±56 bp in size (**Table S1**), and in this paper we identified the precise locations of TF binding sites (8–21 bp long).

Among the 119 TFs, 87 have a DNA-binding domain and hence are expected to bind to DNA specifically. To facilitate the determination of the canonical sequence motifs and comparison of the motifs of TFs whose DNA-binding domains belong to the same protein family, we determined the three-dimensional fold classification for the DNA-binding domains of these TFs, using a combination of sequence alignment, literature searching, and threading (**Table S1**). The canonical motifs for 76 of these 87 sequence-specific TFs have been annotated in the JASPAR (Bryne et al. 2007) or TRANSFAC (Matys et al. 2003) databases. Some of these TFs share have indistinguishable canonical motifs, e.g., SP1 and SP2. Excluding Pol II, Pol III, and Pol III-associated factors, we classify the remaining 25 TFs as non-sequence-specific because they do not contain a sequence-specific DNA-binding domain. The TF ChIP-seq data were collected in 72 cell lines, with most datasets from the three Tier 1 cell lines (K562, GM12878, and H1-hESC) and three Tier 2 cell lines (HeLa-S3, HepG2 and HUVEC). Many TFs were assayed in several cell lines, e.g., CTCF in 49 cell lines and Pol II in 27 cell lines, and some under multiple treatment conditions (e.g., stimulation with interferons or retinoic acid). Some TFs were assayed in the same cell line by multiple labs.

The ENCODE consortium has generated several other types of data which we integrated with the TF ChIP-seq data in this study (ENCODE Project Consortium et al. 2011). First, the nucleosome occupancy profiles in two Tier 1 cell lines (GM12878 and K562) were determined by digesting chromatin with micrococcal nuclease (MNase), isolating mono-nucleosomes, and deeply sequencing the nucleosomal DNA (Kundaje et al. 2012). Second, RNA expression levels were determined by RNA-seq for a large number of cell lines (Djebali et al. 2012). Third, the DNA accessibility for a larger number of cell lines were determined by digesting chromatin with the DNase I enzyme followed by high-throughput sequencing (Neph et al. 2012).

**De novo sequence motif discovery in ChIP-seq peaks**

We constructed a de novo motif discovery pipeline as illustrated in **Fig. S1**. We separated the peaks in each ChIP-seq dataset into independent training and testing sets to ensure the quality of
the motif discovery. The peaks in each ChIP-seq dataset were ranked by their ChIP signal computed by the SPP algorithm (Kharchenko et al. 2008), which indicates the enrichment of ChIP signal over input DNA. The top 500 peaks were reserved for motif discovery (the training set), and the remaining peaks (ranked 501 and onward) were used as for assessing the quality of discovered motifs (the testing set).

We used the MEME-ChIP software suite (Machanick and Bailey 2011) to identify enriched sequence motifs in the [−50 bp, +50 bp] window around the summits of the top 500 peaks (the training set) for each ChIP-seq dataset. MEME-ChIP is based on MEME, a de novo motif discovery algorithm using expectation maximization (Bailey et al. 2006). We asked MEME to report up to five significant motifs per dataset.

To assess the quality of each motif discovered by MEME, we performed two comparisons between the occurrence of the motif in the testing ChIP-seq peaks and control genomic DNA, for the 437 ChIP-seq datasets with 600 or more peaks. In the first comparison, we constructed a testing set that included the peaks ranked 501–1000 (or all peaks ranked beyond 500 if the dataset had fewer than 1000 peaks in total), in a [−150 bp, +150 bp] window around each peak summit (called Testing Set 1). We custom made 100 control sets, which were randomly chosen 300-bp-long genomic regions that matched the GC content of the testing set and did not overlap with any peaks in that particular ChIP-seq dataset (called Control Set 1). We asked whether the motif discovered in the training set of ChIP-seq peaks was significantly enriched in Testing Set 1 than in Control Set 1. We used the FIMO algorithm (Grant et al. 2011) to scan both the testing and control sets, and report the numbers of regions with at least one matching site at a p-value cutoff of 1e−4. The significance of a motif was evaluated with a t-test and the Bonferroni corrected p-value was required to pass the cutoff of 1e−5. The total number of tests used for the Bonferroni correction was 2185, which was the total number of motifs discovered by MEME in the 437 ChIP-seq datasets with 600 or more peaks. 1140 motifs passed this first comparison.

In the second comparison, we asked whether the motif was more enriched in the 300-bp window around the peak summits than in flanking regions. The testing set included all peaks ranked 501 and beyond, in a [−150 bp, +150 bp] window around each peak summit (called Testing Set 2). We used the two 300-bp-long regions flanking the peaks as the control set—the left flanking region (the [−450 bp, −150 bp] window with respect to the peak summit) and the right flanking region (the [+150 bp, +450 bp] window with respect to the peak summit)—and averaged the results of the two regions (called Control Set 2). Again, we used FIMO to scan each motif in the testing and control regions and reported the total numbers of regions with at least one matching site at a p-value cutoff of 1e−4. We deemed a motif as significant if at least 10% of the testing peaks had at least one site and if the percentage of testing peaks with a site was more than 25% higher than the percentage of flanking regions with a motif site. We used a 10% cutoff for motif occurrence in peaks, with the consideration that some secondary motifs may occur in a subset of the peaks. 1045 motifs passed the second comparison also and were used in the subsequent analyses.

For the 20 datasets that had fewer than 600 peaks, we performed both comparisons, using all peaks as the testing set. Additional 47 motifs were identified and also included in the subsequent analysis.
Comparison with annotated motifs in JASPAR and TRANSFAC, followed by manual curation

We compared the discovered motifs with annotated motifs in the JASPAR (Bryne et al. 2007) and TRANSFAC (Matys et al. 2003) databases using TOMTOM (Gupta et al. 2007). To determine if there were significant differences between our de novo discovered motifs and their respective annotated motifs in TRANSFAC or JASPAR, we computed the difference in percentages of regions in the above described Testing Set 2 and Control Set 2 containing at least one significant site for the discovered or annotated motif. Fig. S3 shows that the percentage is higher for 61.8% discovered motifs than for TRANSFAC or JASPAR motifs (two-tailed Wilcoxon signed rank test p-value=1.46e–8), indicating that ChIP-seq data can be used to accurately derive sequence motifs for TFs.

We also compared the motifs that do not resemble annotated motifs with TOMTOM (Gupta et al. 2007). We manually merged similar motifs and identified 78 distinct motifs, which included 11 previously unannotated motifs (UA1–UA11) that we felt most confident about because they were highly enriched in one or more ChIP-seq datasets or were supported by the literature. We labeled the remaining motifs as “No match” in Fig. S6.

Comparison with motifs derived from in vitro protein-DNA binding data

Bulyk and colleagues invented an in vitro method called protein binding microarray (PBM) which uses microarrays containing all possible 10-bp sequences to identify the binding specificities of biochemically purified DNA-binding domains of TF proteins. They assayed more than 100 TFs, and reported multiple distinct sequence motifs for roughly half of them (Pollard et al. 2010; Badis et al. 2009). Their collection and our collection have 17 TFs in common, corresponding to 46 ChIP-seq datasets in our collection. We compared the motifs obtained from the top 500 ChIP-seq peaks (the training set as described above) and the motifs derived from PBM (obtained from their UniPROBE database (Newburger and Bulyk 2009)), judged by the percentage of the remaining ChIP-seq peaks (testing set 2 as described above) containing the motif, and the median distance of the motif sites to the peak summits (Fig. S5). For 12 of the 17 common TFs, the primary motifs from the two methods are similar. For nine of these 12 TFs, the motifs derived from the top ChIP-seq peaks are more enriched in the remaining ChIP-seq peaks than the corresponding UniPROBE motifs, while the motifs derived from the two methods are equally enriched for the remaining three TFs. Furthermore, for nine of the 12 TFs, UniPROBE contains secondary motifs distinct from the primary motifs, but none of these secondary motifs were enriched in the ChIP-seq peaks. For the remaining three TFs (MAX, AP2α, and AP2γ), the UniPROBE secondary motifs resemble their respective primary motifs, but are less enriched in the ChIP-seq peaks than the primary motifs.

The remaining five of the common TFs exhibit greater diversity in the motifs discovered by the two methods. The motifs discovered in the top ChIP-seq peaks of GATA3 represent variant spacing. The primary motif resembles the annotated GATA motif in JASPAR (Bryne et al. 2007) but with a 4-bp extension. In the secondary ChIP-seq motif, this 4-bp extension is located 1-bp closer to the annotated GATA motif. UniPROBE’s primary motif resembles the annotated GATA motif, and its secondary motif resembles the secondary ChIP-seq motif. However, both UniPROBE motifs are less enriched than the respective ChIP-seq motifs. JunD presents another
example of differential spacing. UniPROBE’s primary JunD motif has a 2-bp spacing between the two half-sites, and the secondary motif has a 1-bp spacing. One variant is more enriched than the other depending upon the cell type, and the more enriched variant was discovered in the top ChIP-seq peaks. UniPROBE’s secondary motif for HNF4A resembles the primary ChIP-seq motif as well as the annotated HNF4A motif in JASPAR (Bryne et al. 2007), although it is less enriched than the ChIP-seq motif. UniPROBE’s primary motif for HNF4A is not enriched in ChIP-seq peaks. The NF-Y motif is highly enriched in all three ChIP-seq datasets for IRF3, but neither UniPROBE motif is enriched. Similarly, PU.1 and AP-1 are enriched in the IRF4 ChIP-seq dataset, but neither UniPROBE motif is enriched. The ChIP-seq results for IRF3 and IRF4 suggest tethered binding, i.e., NF-Y (PU.1) binds directly to the genomic regions corresponding to the IRF3 peaks, while IRF3 (IRF4) binds NF-Y (PU.1) via protein-protein interaction, thereby binding to the same regions indirectly.

Zhao and Stormo developed a new method for deriving sequence motifs from PBM data called Binding Energy Estimation by Maximum Likelihood for PBM (BEEML-PBM; Zhao and Stormo 2011)). They reported that a single motif (summarized by a position weight matrix) by BEEML-PBM outperformed a combination of primary and secondary motifs from Badis et al. in (Badis et al. 2009) for the PBM data of most TFs. We obtained the position weight matrices of the 17 TFs [http://ural.wustl.edu/~zhaoy/beeml/](http://ural.wustl.edu/~zhaoy/beeml/), converted them to the position specific scoring matrices in MEME format using the beeml2meme tool in the MEME package ([http://meme.sdsc.edu/meme/doc/beeml2meme.html](http://meme.sdsc.edu/meme/doc/beeml2meme.html)), and performed the above analysis (Fig. S5). Except for MAX, for which BEEM and ChIP-seq derived motifs were equally enriched, the BEEML motifs were less enriched than the ChIP-seq derived motifs in other datasets. The enrichments for the BEEML-PBM motifs were similar to those of the primary UniPROBE motifs for 11 TFs, somewhat worse than those of the primary UniPROBE motifs for three TFs (EGR1, MAFK, and SRF), and somewhat better than those of the primary UniPROBE motifs for three other TFs (MAX, TFAP2A, and TFAP2C).

In summary, the motifs discovered in the top ChIP-seq peaks are more enriched in the remaining ChIP-seq peaks than motifs discovered by the in vitro PBM method. For five of the 17 TFs, the two methods differ qualitatively in their motifs, and the ChIP-seq motifs were more enriched than the motifs derived from in vitro binding data, with ChIP-seq motifs of two TFs reflecting tethered binding in a cell type-dependent manner.

**Comparison of bound vs. unbound motif sites**
We scanned the human genome (FIMO p-value<1e−4) with each of the motifs in Fig. 2S (for example, motif j) and defined the sites that are in the ChIP-seq peaks in a particular cell line as bound sites and other sites as unbound. Thus the definition of bound and unbound sites is specific to a cell line. Then we retrieved the 300-bp regions centered on these sites and asked: (1) whether the bound regions are enriched in the motif sites of any other motifs in our collection (for example motif i), with enrichment defined as (average number of sites for motif i, regardless whether the sites were bound by TF i or not, for each bound regions of motif j)/(average number of sites for motif i for each unbound regions of motif j). (2) whether the bound regions were more hypersensitive to DNase I cleavage than unbound regions, also quantified in enrichment, defined as (average DNase I cleavage in bound regions of motif j)/(average DNase I cleavage in unbound regions of motif j). The enrichments of all motif pairs and with DNase I
hypersensitivity in the same cell line are plotted as a heatmap (Fig. S8), with the element in the $i$-th row and $j$-th column indicating the log$_2$(enrichment) of motif located in the $j$-th column in the bound regions vs. unbound regions of the motif located in the $i$-th row. Note that the bottom right element (DNase vs. DNase) is undefined.

**Different modes of interactions between TFs**

For each ChIP-seq dataset, we computed the percentage of peaks that contain a site for the canonical motif but not a non-canonical motif, the percentage of peaks that contain a site for a non-canonical motif (FIMO p-value < 1e-4), and the percentage of peaks that contain sites for both canonical and non-canonical motifs. These are plotted in Fig. 2a and the underlying data is in Table S3. High percentages of peaks with only non-canonical motif sites suggest tethered binding. High percentages of peaks that contain sites for both canonical and non-canonical motifs suggest co-binding. Some TFs show both tethered and co-binding in the same dataset.

**Distance and/or orientation preferences between motif sites**

For all pairs of motifs discovered in the same ChIP-seq dataset, we computed the edge-to-edge distance and relative orientation (same or opposite strands) between their sites in the same peak. For all pairs of motifs discovered in different ChIP-seq datasets, we performed the same calculation for sites that were within 100 bp of each other. We then compared the observed distribution of edge-to-edge distance between motif sites with the background distribution, which was derived by assuming each motif is uniformly distributed in the peak and on either strand. A p-value was computed using the Kolmogorov-Smirnov test. The analysis was also applied to peaks in non-repetitive regions and repetitive regions of the genome separately. A total of 24,283 and 14,881 tests were performed for non-repetitive and repetitive regions respectively, and an FDR cutoff of 0.025 was applied after multiple-testing correction. For motif pairs in repetitive regions that had exactly the same distance as the mode distance of the observed distribution, we tested their enrichment in various repetitive elements using a hypergeometric test, and an FDR cutoff of 0.025 was applied after multiple-testing correction.

**Cell line-specific gene expression**

RNA-seq data was processed as described in the ENCODE companion manuscript (Djebali et al. 2012). Expression levels of genes were quantified in reads per kb of transcript per million sequencing reads (RPKM). We identified genes that were at least 10-fold more highly expressed in one cell line than in any of the other four cell lines. Fig. 4a shows five examples of such cell line-specific genes and the TF binding sites around them.

**Cell type specific non-canonical motifs**

For most datasets associated with a sequence-specific TF, the most significant motif identified by MEME-ChIP was also the TF’s annotated canonical motif. We manually corrected a few exceptions and designated all the other motifs discovered by MEME-ChIP as non-canonical motifs. The enrichment of each non-canonical motif in a cell line was computed as the ratio between the total number of datasets in that cell line having the motif as a secondary motif, and the total number of datasets in that cell line. A non-canonical motif was deemed specific to a cell line if its enrichment coefficient in that particular cell line was at least three times higher than its coefficients in all other four cell lines, and is plotted around the cell line squares in Fig. 4b. Otherwise, it is plotted in the center of Fig. 4b.
**Characterization of chromatin structure around TF binding regions**

We computed and plotted average nucleosome occupancy profiles anchored on summits of ChIP-seq peaks of each TF. We also computed nucleosome occupancy profiles anchored on TF binding sites, and they look nearly identical to the nucleosome profiles anchored on peak summits (data not shown), because as in the beginning of Results, most TF binding sites are located near peak summits. To account for the impact of transcription on nucleosome positioning, we separated ChIP-seq summits into two sets: TSS-proximal (within 2 kb of a TSS) and TSS-distal (>2 kb from a TSS). We further separated the peak summits into three equal-sized sets based on their ChIP-seq signals. Thus we considered a combination of six sets of peaks in each ChIP-seq dataset for which we constructed nucleosome occupancy profiles and DNase I cleavage profiles. For datasets available on both GM12878 and K562, we constructed two sets of cell line-specific peaks: those that were in GM12878 but not in K562, and vice versa. Each set was further partitioned into two subsets, those that overlapped the ChIP-seq peaks of other TFs in a particular cell line, and those that did not overlap the ChIP-seq peaks of any other tested TFs.

We defined nucleosome depletion as the dip in nucleosome occupancy at the peak summit compared with signal between background and the peak summit. We applied a fast Fourier transform (FFT) on the nucleosome occupancy profile anchored on the top 1/3 distal peaks of each dataset (Fig. S11 contains an example FFT spectrum). Fourier transform is a well-established method that detects periodicity in signals by decomposing them into sinusoidal components of differing frequencies. FFT is an efficient algorithm that computes Fourier transforms. The output of an FFT is a power spectrum that plots the magnitude of each frequency component resulting from the decompositon of the original signal. In the context of a nucleosome profile, the magnitude of the FFT power spectrum at the frequency component that corresponds to the period of positioned nucleosomes indicates the strength of the nucleosome positioning (the higher the magnitude, more periodic the nucleosome occupancy profile). The frequency component of the power spectrum (x-axis) corresponds to all possible periods that may exist in the input signal.
Reference:


Supplemental Figures:

**Figure S1:** Flow chart of the de novo motif discovery pipeline.

**Figure S2:** Sequence logos for all de novo discovered motifs and their corresponding TFs. There are 79 motifs in total, and they are ordered alphabetically by motif name. UA1–UA12 are 12 motifs that were not annotated in TRANSFAC or JASPAR. UA12 was identified in the subset of HDAC2-associated GATA1 peaks in K562. The remaining 78 motifs were identified in the top 500 ChIP-seq peaks using our motif discovery pipeline.

**Figure S3:** Nearly identical motifs were discovered in datasets generated by different labs, with different antibodies targeting the same TF.

**Figure S4:** De novo discovered motifs were of higher quality than previously annotated motifs for majority of the ChIP-seq datasets. Motif quality was judged by the percentage of peaks in Testing Set 2 that contained a significant motif site minus the percentage of flanking regions (Control Set 2) that contained a significant motif site. Each circle is a motif in a dataset. The cluster of circles with highest percentage right under the diagonal line belongs to the CTCF motifs discovered in the CTCF, RAD21, and SMC3 datasets.

**Figure S5:** Motifs discovered using the top ChIP-seq peaks were more enriched in the remaining ChIP-seq peaks (the left heatmap) and their sites are located closer to the peak summit (the right heatmap) than motifs discovered using the in vitro technique protein binding microarray (obtained from the UniPROBE database). All ChIP-seq datasets for the 17 shared TFs were analyzed, and the detailed results are shown in subsequent pages, one page per dataset. On each of the subsequent pages, the top-right two panels correspond to the two most enriched motifs discovered in each ChIP-seq dataset, whenever available, the middle-right two panels correspond to the two distinct motifs obtained from UniPROBE, the bottom-right two panels correspond to the motifs by BEEML-PBM derived using two biological replicates of PBM data, and the top left panel combines all motifs to facilitate direct comparison. See Fig. S6 for descriptions of the panels.

**Figure S6:** Motif composition and distance to peak summit are plotted against peak rank (ranked by the ChIP-seq signal) for each of the 408 ChIP-seq datasets, for which a significant motif was identified. The datasets were organized by TFs, grouped as described in Materials and Methods (sequence-specific, chromatin remodeling, etc.), and the TFs within each group were ordered alphabetically. The top left panel summarizes the percentage of peaks (solid lines) and the flanking regions (dashed lines) that contain each of the motifs discovered by MEME (up to five motifs per dataset). The motifs are labeled M1-M5 in the ascending order of MEME p-values (i.e., descending order of statistical significance). The remaining panels display the percentage of motif-containing peaks (colored solid lines for peaks and colored dashed lines for flanking regions) and the absolute distance of motif sites with respect to peak summits (gray lines), shown separately for each motif. Only the motifs that passed both of the comparisons described in Materials and Methods were shown (more enriched in peaks than in randomly chosen GC-matched regions, and more enriched in peaks than in flanking regions). The plots are not
cumulative in terms of peaks. To decrease the file size of this figure, we plotted one data point for the average of every 80 peaks.

**Figure S7:** DNase I footprints and sequence conservation anchored on sites of each motif discovered in each ChIP-seq dataset. For easy viewing, the previously unannotated motifs UA1–UA11 are shown at the beginning, then all motifs are shown alphabetically by TF names, as in Fig. S6. Motif sites in ChIP-seq peaks (solid lines) were compared with motif sites outside ChIP-seq peaks (dashed lines). The DNase I footprints were average DNase I cleavage signal in the same cell line as the ChIP-seq data (the cell line is indicated on each page), and sequence conservation was computed using average phyloP scores (Pollard et al. 2010).

**Figure S8:** Enrichment of sequence motifs and DNase I hypersensitivity around bound motif sites. Color indicates log2(enrichment); see Methods for definition of enrichment. The element in the i-th row and j-th column indicates the log2(enrichment) of motif located in the j-th column in the bound regions vs. unbound regions of the motif located in the i-th row. The rows and columns are in alphabetical order by motif name. We also considered subsets of peaks for the TFs through which EP300 and HDAC2 may tether. For example, the PU.1–EP300 row in the GM12878 heatmap indicates the subset of SPI1 peaks that overlap by at least 1 bp with EP300 peaks in GM12878 cells, likewise the rows labeled with “–EP300” or “–HDAC2”.

**Figure S9:** De novo motif discovery using the subset of peaks associated with HDAC2. UA9 was identified in the HDAC2-associated POU5F1 peaks in H1-hESC cells (left panels) and UA12 was identified in the HDAC2-associated GATA1 peaks in K562 cells (right panels). The bottom four panels depict percentage of peaks with significant motif sites; see Fig. S6 for explanation on similar panels.

**Figure S10:** Nucleosome occupancy profiles anchored on the summits of TSS-proximal (left panels) and TSS-distal (right panels) peaks of each TF in GM12878 or K562 cells. Peaks are grouped according to their ChIP-seq signal strength—top (green), middle (red) and bottom (blue) thirds.

**Figure S11:** Fast Fourier transform was used to determine the extent of nucleosome positioning. a. Example nucleosome occupancy profiles in GM12878 and K562. b. The FFT spectra that correspond to the profiles in a. c. The regularity of positioning, as shown as the magnitude of the FFT spectrum at the period of positioning, is positively correlated with the occupancy of −1 or +1 nucleosomes.

**Figure S12:** Average nucleosome occupancy profiles centered on the TSS for GM12878 and K562 cells.

**Figure S13:** In vitro nucleosome occupancy profiles (black lines), GC% (yellow lines) anchored on the summits of TSS-distal ChIP-seq peaks. For comparison, the in vivo nucleosome occupancy profiles anchored on the top one-third TSS-distal ChIP-seq peaks (strongest ChIP-seq signal; green lines; same as the green lines in Fig. S10) are also shown. Each ChIP-seq dataset is shown as a panel, alphabetically ordered by TF names.
**Figure S14**: Chromatin structure around ChIP-seq peaks occupied differentially between GM12878 and K562. Only those 32 TFs that have ChIP-seq datasets in both GM12878 and K562 are included; one row per TF. **(left panel)** Nucleosome occupancy profiles (solid lines) and DNase I cleavage profiles (dashed lines) anchored on the summits of peaks in GM12878 but not in K562. Note the average nucleosome occupancy at these peaks (x=0) is lower in GM12878 than in K562, while the average DNase I cleavage at these peaks is higher in GM12878 than in K562. **(right panel)** Same as the corresponding left panel, but around the summits of peaks in K562 but not in GM12878.

**Figure S15**: In vitro nucleosome occupancy profiles (black lines), GC% (yellow lines) anchored on the summits of TSS-distal ChIP-seq peaks, separates into three groups, shared between GM12878 and K562, in GM12878 but not in K562, and in K562 but not in GM12878. In comparison, the in vitro nucleosome occupancy profiles at flanking regions (2k bp away from the summit) are shown as dotted lines. Only those 32 TFs that have ChIP-seq datasets in both GM12878 and K562 are included.

**Figure S16**: Chromatin structure around ChIP-seq peaks unoccupied by a TF in a cell line. Only those 32 TFs that have ChIP-seq datasets in both GM12878 and K562 are included; one row per TF. **(left panel)** Nucleosome occupancy profiles in K562 anchored on the summits of the ChIP-seq peaks occupied by a TF in GM12878 but not in K562. These peaks were divided into two subsets: peaks were bound by one or more TFs in K562 (dot-dashed line) and peaks were not bound by any TF for which we had ChIP-seq data in K562 (solid line). Note high nucleosome occupancy at the summits of the unoccupied peaks (x=0) and the lack of positioned nucleosomes flanking the unbound peaks, in sharp contrast with the lack of nucleosome occupancy on the peak summits and well-positioned nucleosomes flanking the peak. **(right panel)** Same as the corresponding left panel, but around the summits of the ChIP-seq peaks occupied in K562 but not in GM12878.
Supplemental Tables:

Table S1: Summary of motif discovery in all ChIP-seq datasets.

Table S2: Position specific scoring matrices for all 79 motifs in Fig. S2, in MEME format, one motif per page. The motifs are ordered alphabetically by name, as in Fig. S2.

Table S3: Summary of interacting TF pairs, one row per TF pair in a particular cell line. Rows shaded pink are the underlying data for Fig. 2a, for which the motifs were discovered in the same ChIP-seq dataset and restricted to sequence-specific TFs, and at least 60% of the sites for the non-canonical motif are supported by the ChIP-seq data of its corresponding TF in the same cell line. Rows shaded blue are additional TF pairs, requiring that at least 25% of the sites for the non-canonical motif are supported by the ChIP-seq data of its corresponding TF in the same cell line, and include non-sequence-specific TFs. Note that whether to call motif co-occurrence depends on the motif cutoff (we currently use FIMO p-value<1e-4). It is also possible that the interaction with another TF enables a TF to bind much weaker motif sites.

Table S4: Summary of motif pairs with position and/or orientation preferences. The table has four sheets, corresponding to the underlying data for Fig. 2b left panel, Fig. 2b right panel, Fig. 2c left panel, and Fig 2c right panel. The left panels are motif-based (one motif pair per row) and the right panels are dataset-based (one motif pair in a particular cell line per row).