Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors

Jie Wang,[1,8] Jiali Zhuang,[1,8] Sowmya Iyer,[1,2,8] XinYing Lin,[1,8] Troy W. Whitfield,[1] Melissa C. Greven,[1] Brian G. Pierce,[1] Xianjun Dong,[1] Anshul Kundaje,[3] Yong Cheng,[4] Oliver J. Rando,[1] Ewan Birney,[5] Richard M. Myers,[6] William S. Noble,[7] Michael Snyder,[4] and Zhiping Weng[1,9]

[1]*Program in Bioinformatics and Integrative Biology, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA;* [2]*Bioinformatics Program, Boston University, Boston, Massachusetts 02215, USA;* [3]*Department of Computer Science, Stanford University, Stanford, California 94305, USA;* [4]*Department of Genetics, Stanford University, Stanford, California 94305, USA;* [5]*Vertebrate Genomics Group, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, United Kingdom;* [6]*HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA;* [7]*Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA*

Chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) has become the dominant technique for mapping transcription factor (TF) binding regions genome-wide. We performed an integrative analysis centered around 457 ChIP-seq data sets on 119 human TFs generated by the ENCODE Consortium. We identified highly enriched sequence motifs in most data sets, revealing new motifs and validating known ones. The motif sites (TF binding sites) are highly conserved evolutionarily and show distinct footprints upon DNase I digestion. We frequently detected secondary motifs in addition to the canonical motifs of the TFs, indicating tethered binding and cobinding between multiple TFs. We observed significant position and orientation preferences between many cobinding TFs. Genes specifically expressed in a cell line are often associated with a greater occurrence of nearby TF binding in that cell line. We observed cell-line–specific secondary motifs that mediate the binding of the histone deacetylase HDAC2 and the enhancer-binding protein EP300. TF binding sites are located in GC-rich, nucleosome-depleted, and DNase I sensitive regions, flanked by well-positioned nucleosomes, and many of these features show cell type specificity. The GC-richness may be beneficial for regulating TF binding because, when unoccupied by a TF, these regions are occupied by nucleosomes in vivo. We present the results of our analysis in a TF-centric web repository Factorbook (http://factorbook.org) and will continually update this repository as more ENCODE data are generated.

[Supplemental material is available for this article.]

The genome encodes the information required for building an organism, including genes that encode proteins and functional RNAs, and more importantly, the instructions for when, where, under what conditions, and at what levels genes are expressed. Elaborate regulation of gene expression is a key driving force for organismal complexity (Levine and Tjian 2003). Transcription factors (TFs) are a family of proteins that can execute the instructions for transcriptional regulation by interacting with RNA polymerases to activate or repress their actions (Maston et al. 2006). The fidelity of transcriptional regulation ultimately relies on TFs, which can bind directly to genomic DNA with specific sequences via their DNA-binding domains, or indirectly through interactions with other DNA-binding TFs. The regulation of most genes requires many TFs, which may form large complexes, and a TF typically regulates many genes.

[8]These authors contributed equally to this work.
[9]Corresponding author
E-mail Zhiping.Weng@umassmed.edu
Article and supplemental material are at http://www.genome.org/cgi/doi/10.1101/gr.139105.112. Freely available online through the *Genome Research* Open Access option.

In eukaryotic cells, transcription is regulated in the context of chromatin, whereby genomic DNA is packaged into nucleosomes, and TFs must compete with nucleosomes for accessibility to genomic DNA. It was discovered early on that some loosely packaged regions of chromatin were hypersensitive to cleavage by DNase I, and these regions might harbor regulatory DNA (Weintraub and Groudine 1976). The advent of high-throughput genomic techniques allowed systematic mapping of nucleosomes, and more recent studies showed that most genomic DNA is nucleosomal and that functional TF binding sites tend to be located in nucleosome-depleted regions (Guertin and Lis 2010; John et al. 2011; Li et al. 2011). Nonetheless, some TFs are capable of remodeling nucleosomes in the absence of additional factors, and other TFs can recruit nucleosome remodelers to reposition or evict nucleosomes and expose TF binding sites (Berger 2007; Clapier and Cairns 2009). Furthermore, it was reported that TF binding sites are flanked by multiple well-positioned nucleosomes (Fu et al. 2008; Valouev et al. 2011).

Transcriptional regulation has been studied at the single-gene level for several decades. TFs recognize 8- to 21-base pair (bp) degenerate sequence motifs (Matys et al. 2003; Bryne et al. 2008), but in vivo a given TF typically only associates with a small

22:1798–1812 © 2012, Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/12; www.genome.org

subset of the genomic sites that match its binding motif. ChIP-seq is a technique for mapping TF binding regions genome-wide in living cells. The method combines chromatin immunoprecipitation (ChIP), using TF-specific antibodies, with high-throughput sequencing (seq) (Robertson et al. 2007). Dozens of ChIP-seq data sets of mammalian TFs have been reported in the literature by individual labs (Biggin 2011; MacQuarrie et al. 2011). The ENCODE Consortium has generated 457 ChIP-seq data sets on 119 TFs in 72 cell lines (Supplemental Table S1) and determined transcription levels, nucleosome occupancy, and DNase I hypersensitivity in a subset of these cell lines (The ENCODE Project Consortium 2011). We analyzed this rich collection of data to characterize the sequence features of TF binding sites and determine the local chromatin environment around them.

Results

Identification of sequence motifs and TF binding sites

As described in Supplemental Methods, we built a computational pipeline (Supplemental Fig. S1) to discover enriched sequence motifs de novo using the 500 highest ranked peaks (the training set) in each ChIP-seq data set and assessing the quality of the motifs in two ways using the remaining peaks (nonoverlapping testing sets). We used the MEME-ChIP software suite (Machanick and Bailey 2011) for motif discovery (up to five motifs per ChIP-seq data set), and 1092 motifs passed both of our quality assessment filters. For each data set, we define the most significant motif (lowest E-value computed by MEME) as the primary motif, and the remaining significant motifs as secondary motifs. We manually merged redundant motifs and consistently named motifs discovered in multiple data sets, taking into account previous literature and protein family information on the DNA-binding domains of the TFs.

In the end, we identified 79 unique motifs (Supplemental Fig. S2 for sequence logos; Supplemental Table S2 for position-specific scoring matrices in MEME format), 67 of which were in the JASPAR or TRANSFAC repositories (Matys et al. 2003; Bryne et al. 2008), while 12 were unannotated but highly significant (we named these motifs UA1–UA12). Among the 119 TFs, 87 are involved in Pol II-mediated transcription and have a DNA-binding domain; these TFs are classified as sequence-specific. The motif that reflects the sequences recognized by the DNA-binding domain of a sequence-specific TF is defined as the canonical motif of the TF. Note that the primary motif (the most enriched motif discovered by MEME) was not necessarily the canonical motif, although in most cases it was. The reason that we discovered fewer motifs than the total number of sequence-specific TFs is that some TFs belong to the same family and have indistinguishable motifs (e.g., SP1 and SP2) and other TFs are components of a functional complex (e.g., the heterodimer of USF1 and USF2 is the functional TF, called USF). We also divided the peaks for each TF into two sets, those within 2 kb of transcription start sites (TSS-proximal) and those >2 kb away from TSS (TSS-distal). We then performed motif-finding, using the top 500 TSS-proximal peaks and the top 500 TSS-distal peaks separately, and found consistent motifs between the two sets (data not shown).

For several TFs, ChIP-seq was performed by multiple labs or using multiple antibodies, and we discovered nearly identical motifs for different data sets of the same TF (Supplemental Fig. S3), indicating that the ChIP-seq data sets are of high quality. As described in Supplemental Methods, the motifs we discovered tend to be more highly enriched in ChIP-seq peaks than annotated motifs in databases (Supplemental Fig. S4) or motifs derived with an in vitro method (protein binding microarray) (Supplemental Fig. S5; Badis et al. 2009).

We identified significant sequence motifs for 86 of the 87 sequence-specific TFs (Fig. 1A). An AG tandem repeat was detected for the remaining TF (ZZZ3), but we cannot confidently assess the quality of this motif because the two ZZZ3 ChIP-seq data sets have few peaks (740 peaks in GM12878 and 193 peaks in HeLa-S3) (Supplemental Table S1). The canonical motifs of 76 sequence-specific TFs have been annotated, and we identified a significant motif that matched the annotated canonical motif for each TF. We also identified significant noncanonical motifs for 70 of the 76 sequence-specific TFs, suggesting that other TFs also bind to the peaks of these TFs or that these TFs bind to their target DNA by tethering onto other TFs. We found significant motifs for 23 of the 25 non-sequence-specific TFs (Fig. 1A), which, by definition, bind via tethering.

We computed two measures for each motif we discovered in each data set: the percentage of ChIP-seq peaks in the [–150 bp, +150 bp] window around the peak summit that contained a significant site for the motif (FIMO P-value $< 1 \times 10^{-4}$) and the distribution of the absolute distances between the nearest edge of motif sites and the peak summit. We plotted these two measures with respect to the ranks of the peaks (ranked according to the ChIP-seq signal), using SPI1 in GM12891 cells as an example (Fig. 1B). Over 80% of the top 20,000 peaks contained sites for the motif of SPI1 (PU.1), and this percentage decreased to around 75% for the bottom of the ~40,000-member peak list. For comparison, we scanned the two 300-bp regions flanking the 300-bp peak window, and 14.9% of the flanking regions contained PU.1 sites. The median distance from the motif site to the peak summit was 7 bp for the top-ranked peaks. This distance increased to 13 bp for the bottom-ranked peaks, still much smaller than the 75-bp distance expected for motifs uniformly distributed in a [–150 bp, +150 bp] window. Thus, the majority of these bottom-ranked peaks are likely bound by SPI1 in living cells. Supplemental Figure S6 is a gallery of these figures for each of the 408 data sets for which a significant motif was identified, indicating strong enrichment of one or more de novo discovered motifs in the vast majority of the data sets, especially those data sets that correspond to the 87 sequence-specific TFs (first section of Supplemental Fig. S6).

We computed two additional measures for the motif sites within ChIP-seq peaks: the DNase I footprint and evolutionary conservation. TF binding sites tend to be located in accessible chromatin indicated by high DNase I cleavage, yet the binding by the TF protects the site from DNase I cleavage compared with its flanking positions (Neph et al. 2012). Thus, TF binding sites exhibit "valley in a peak" DNase I footprints. Such footprints become increasingly visible with greater sequencing depth, and DNase-seq data sets with a few hundred million reads are typically required to see the footprints clearly. For the motifs discovered in the K562 data sets, Supplemental Figure S7 illustrates both the DNase I footprints and conservation profile (computed with phyloP) (Pollard et al. 2010). For motifs discovered in other cell lines, only conservation profiles are shown because the DNase I data for these cell lines do not have sufficient sequencing depth for footprinting. Supplemental Figure S7 clearly shows that most motif sites in ChIP-seq peaks show distinct DNase I footprints and strong sequence conservation (solid lines), compared with motif sites outside ChIP-seq peaks (dashed lines).

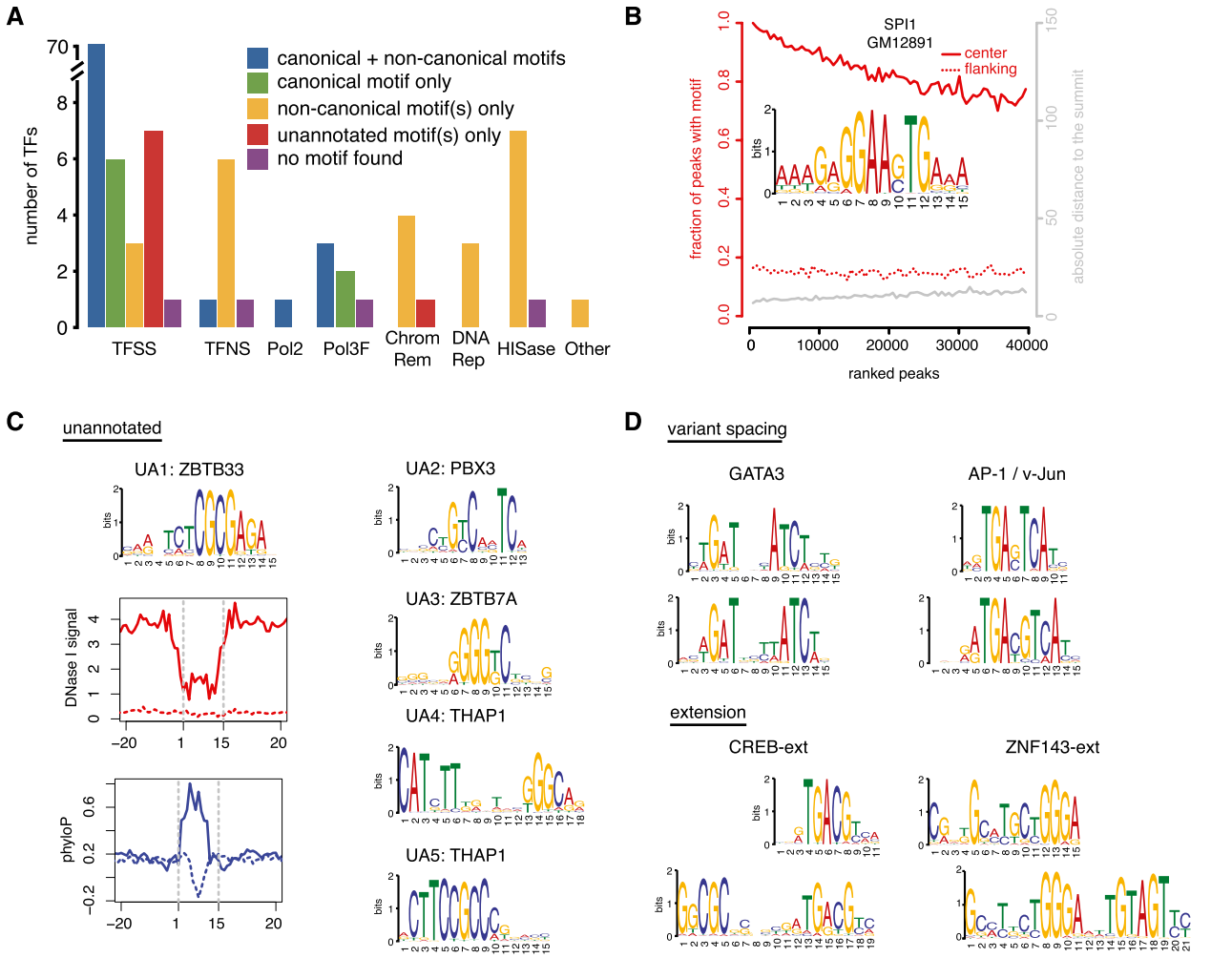

Figure 1. De novo discovery of sequence motifs. (*A*) Statistics of motif discovery among 119 TFs, classified into 87 Pol II-associated sequence-specific TFs (TFSS), eight general Pol II-associated, non-sequence-specific TFs (TFNS), Pol II (Pol2), six Pol III components and Pol III-associated TFs (Pol3F), five ATP-dependent chromatin complexes (ChromRem), three TFs involved in DNA repair (DNARep), eight histone modification complexes (HISase), and one cyclin kinase associated with transcription (Other). The TATA box binding protein (TBP) is included in the TFNS category and its canonical motif is TATA, corresponding to the blue bar. (*B*) Example result for SPI1 in GM12891 cells illustrating the percentage of peaks with the motif (*left*, *y*-axis in red) and distribution of absolute distances of the closer edge of motif sites relative to the peak summit (*right*, *y*-axis in gray), plotted against ranks of peaks (ranked by ChIP-seq signal). (*C*) Five previously unannotated motifs that are likely to be canonical motifs of four sequence-specific TFs. Also shown are DNase I footprint and sequence conservation profiles around the sites of UA1 (likely the canonical motif of ZBTB33). Motif sites in ChIP-seq peaks (solid lines) were compared with motif sites outside peaks (dashed lines). DNase I and ChIP-seq data were both from K562 cells. Sequence conservation was computed using phyloP (Pollard et al. 2010). (*D*) Motifs with variant spacing and extensions.

Previously unannotated motifs

We identified 11 high-confidence motifs that did not match any annotated motifs in the JASPAR or TRANSFAC repositories (Fig. 1C; Supplemental Figs. S2, S6). Among these motifs, UA1–UA5 are likely the canonical motifs for four TFs, and UA9 is likely the canonical motif for a factor that functions in H1-hESC cells. Supplemental Figure S7 shows that the sites of the previously unannotated motifs tend to have high evolutionary conservation and show distinct DNase I footprints.

UA1 was detected as the primary motif of three TFs (ZBTB33, BRCA1, and CHD2), as well as a secondary motif for ETS1. Because ZBTB33 is a zinc finger protein that binds methylated CpG dinucleotides (Yoon et al. 2003) and the center of UA1 contains CGCG, UA1 most likely is the canonical motif of ZBTB33. BRCA1 and CHD2 do not have a DNA-binding protein domain, suggesting that they bind ZBTB33 to perform their functions in DNA repair and genome maintenance. Indeed, the 936 ZBTB33 peaks that contain UA1 sites and the 321 BRCA1 peaks that contain UA1 sites have 312 peaks in common. Similarly, the 936 ZBTB33 peaks that contain UA1 sites and the 1022 CHD2 peaks that contain UA1 sites have 719 peaks in common.

UA2 was the primary motif for the PBX3 data set in GM12878, with 44.3% of the 7431 peaks containing at least one UA2 site. We did not identify any previously published description of the sequence motif of PBX3.

UA3 was the primary motif for the ZBTB7A data set in K562, occupying 80.1% of the 19,942 peaks. CTCF was identified as a secondary motif but with a lower enrichment than UA3 (Supplemental Fig. S6). Genecard indicates that the consensus sequence of ZBTB7A binding sites is 5′-[GA][CA]GACCCCCCCCC-3′, which is similar to the reverse complement of the UA3 consensus

5′-CNGAGACCCCNNCCC-3′. Furthermore, the motif derived from the in vitro protein binding microarray method for ZBTB7B (a paralog of ZBTB7A) is very similar to UA3 (Badis et al. 2009).

UA4 and UA5 were discovered in the THAP1 data set in K562. UA4 is a gapped motif, and it is an extended version of the motif previously reported for the THAP family of TFs (Sabogal et al. 2010). UA5 shares the "GGGC" half of UA4 but further extends it. Thus both UA4 and UA5 are likely the canonical motifs for THAP1.

UA9 was discovered as the primary motif for NANOG (in H1-hESC cells) and BCL11A (in H1-hESC cells but not in GM12878 cells). It does not resemble the previously identified NANOG motif (Chen et al. 2008). We also discovered UA9 as a secondary motif for five other TFs in H1-hESC cells (see Supplemental Fig. S6). We, therefore, suspect that UA9 is the canonical motif of a yet uncharacterized TF that functions in H1-hESC cells.

We also identified two motifs that allow alternative spacing: The two GATA3 half sites, AGAT and ATCT, can be either 3 or 4 bp apart, and the two half sites of the AP-1 motif can be either 1 or 2 bp apart. The variant spacing of AP-1 was previously detected by the in vitro protein binding microarray method (Badis et al. 2009; Supplemental Fig. S5), reflecting intrinsic flexibility of the two leucine zippers of the heterodimeric AP-1 TF. The variant spacing of GATA3 has not been reported previously. We identified extensions of four annotated motifs—CREB, ZNF143, GATA1, and CTCF (Fig. 1D; Supplemental Fig. S6). ZNF143-ext (Myslinski et al. 2006) and CTCF-ext (Ohlsson et al. 2001; Rhee and Pugh 2011) have been documented before. GATA1-ext is the motif for the TAL–GATA1 complex (Xu et al. 2003). The extension for CREB has not been reported.

Comparison of bound vs. unbound motif sites

Although the ChIP-seq peaks are highly enriched in motifs, there are still many motif sites outside peaks (unbound motif sites). For example, there are, on average, 430 times more unbound motif sites (sites outside peaks; FIMO P-value $< 1 \times 10^{-4}$) than bound motif sites (sites within peaks) for the TFs with ChIP-seq data in K562 cells. We asked whether there were any sequence or chromatin features that could distinguish bound sites from unbound sites (see Supplemental Methods for details). Indeed, we found that the regions surrounding bound sites were more DNase I hypersensitive and enriched in TF motifs, compared with the regions surrounding unbound sites, as shown in Supplemental Figure S8 for the five cell lines with the most ChIP-seq data sets, one heat map per cell line. The histogram of $\log_2$ (enrichment) has a heavier right-side tail in all cell lines, indicating an overall enrichment among all pairwise comparisons (Supplemental Fig. S8). As expected, regions around bound A-box sites are enriched in B-box sites and vice versa, consistent with these sites being the TFIIIC motifs in tRNA genes (Oettel et al. 1998). The bound regions of most motifs are enriched in sites of the same motif. Several motifs such as NRF1 are enriched in the bound sites of the majority of motifs across the cell lines.

Cobinding and tethered binding between different TFs

Many eukaryotic genes are coregulated by multiple TFs in a cell-type–specific manner (Maston et al. 2006). For 70 of the 87 sequence-specific TFs, we discovered the canonical motifs as well as significant secondary motifs that were distinct from the canonical motifs of the TFs in question and that correspond to the canonical motifs of other TFs. Two scenarios may result in secondary motifs: Two TFs bind to neighboring sites (cobinding), or one TF protein binds to another that, in turn, binds to DNA (tethered binding). To distinguish between these scenarios, we computed the percentages of peaks in a ChIP-seq data set that contain sites for the canonical TF only, a noncanonical TF only, or both, and then we sorted the data sets by the percentages of peaks with only noncanonical motif sites (Fig. 2A; see Supplemental Table S3 for the underlying data). We reasoned that if sites of a noncanonical motif were frequently found to be in the same ChIP-seq peaks as canonical motif sites (hence, adjacent to them), the two TFs are likely to interact at the protein level and influence each other in binding to their DNA sites. Conversely, if the majority of the peaks contain only sites for noncanonical motifs, then tethered binding is a more plausible model. In this fashion, we identified 151 potential tethered binding and 104 cobinding sequence-specific TF pairs (255 in total). We then compared the pairs we discovered with experimentally detected pairs reported in a mammalian two-hybrid study (Ravasi et al. 2010) and in the BIOGRID database (Stark et al. 2006) and found evidence for physical interaction for 27 (10.6%) of the pairs. Eighteen of the 151 tethered binding predictions were validated in the mammalian two-hybrid data. We randomly picked 151 TF pairs for 5000 trials, and on average, 4.19 pairs were validated in the mammalian two-hybrid experiments (maximum 13 pairs), indicating that our predicted TF pairs were highly significant (P-value $< 2 \times 10^{-4}$). Thus, our results both recapitulated previously reported observations and revealed novel potential interactions that can be tested by experimentation (see Supplemental Table S3 for summary of all pairs).

SP1 (or SP2) and NF-Y (heterodimer of NF-YA and NF-YB) constitute an example of cobinding. (SP1 and SP2 mostly bind to common sites in the genome, and their motifs are indistinguishable.) The SP2 ChIP-seq data set in K562 cells contains 3025 peaks, of which 2496 peaks contain SP1/2 sites and 1711 peaks contain NF-Y sites, sharing 1512 peaks. Furthermore, 1562 of the 1711 NF-Y site-containing peaks overlap with the peaks in the NF-Y ChIP-seq data set in K562 cells, confirming that these NF-Y sites are, indeed, bound by the NF-Y protein. Thus, SP2 and NF-Y prefer to cobind neighboring sites in the genome. Similarly, the SP1 ChIP-seq data sets in K562 and GM12878 cells indicate cobinding between SP1 and NF-Y. In the next section, we show that NF-Y and SP1/2 binding sites tend to be within 30 bp of one another. Indeed, Roder et al. showed that NF-Y and SP1 proteins bind to each other and that they cobind to the same promoter (Roder et al. 1999). In another example, YY1 was shown to interact with MYC, and cooperatively, the two TFs regulate the expression of the *ITGA3* gene of the α3β1-integrin complex in human osteosarcoma cells (de Nigris et al. 2007). We found that the cobinding of YY1 and MYC was not limited to the *ITGA3* promoter. In K562 cells, 484 MYC peaks (11.9%) contain both YY1 and MYC motifs, and these peaks are bound by YY1 as well. Novel predictions of cobinding TF pairs include ESRRA and HNF4 in HepG2 cells and NFKB and SPI1 in GM12878 cells (Supplemental Table S3).

Tethered binding facilitates combinatorial regulation by additional sequence-specific TFs and is necessary for recruiting chromatin remodelers and other regulatory proteins that do not bind DNA directly. We describe examples of tethered binding between sequence-specific TFs here, and discuss the tethering of non-sequence-specific TFs onto sequence-specific TFs in a subsequent section. ATF3 contains a basic leucine zipper DNA-binding domain whose canonical motif is the CREB motif. However, 51.2%, 48.6%,

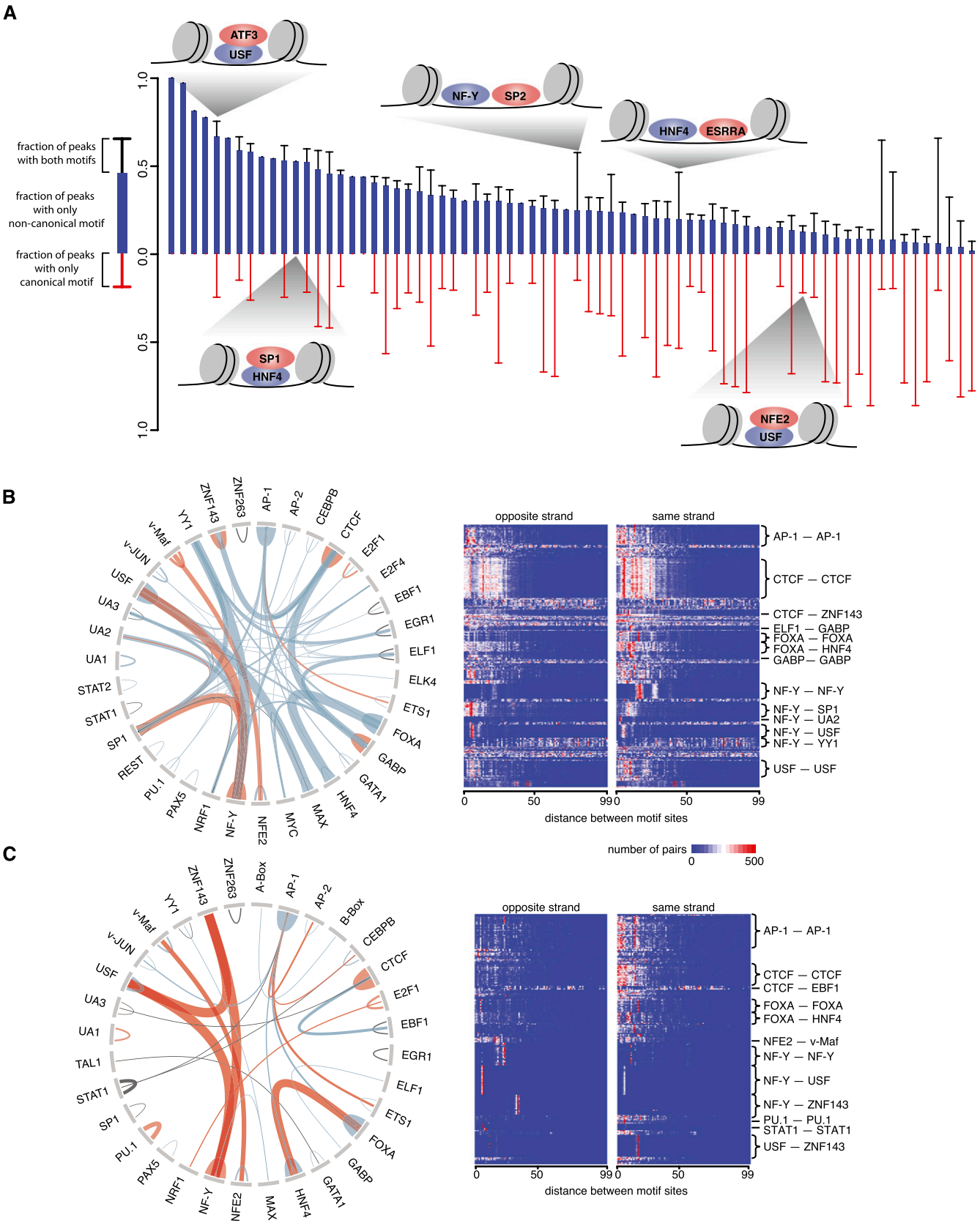

Figure 2. Interactions between TFs. (*A*) Different modes of interaction between TFs are shown. Each bar indicates the canonical TF and one non-canonical TF whose motifs were identified in the same ChIP-seq data set, and the red, blue, and black segments of the bar indicate percentage of peaks in the ChIP-seq data set that contain only canonical motif sites, only noncanonical motif sites, or both. Cartoons depict examples of different models for TF-TF interactions. (*B*) Circos plot (Krzywinski et al. 2009) on the *left* depicts pairs of motifs (connected by an arch) with significant distance preferences between their sites. The thickness of a connection is proportional to the normalized frequency of the pair. A connection is depicted as blue, black, or red when the motif pair is discovered in different data sets, the same data set, or both, respectively. The heat map on the *right* shows the distributions of distances between motif pairs. Each row is a motif pair in a particular ChIP-seq data set, and each column represents an edge-to-edge distance (from 0 bp to 99 bp). (*C*) Similar to *B* except showing motif pairs discovered in repetitive regions.

46.9%, and 49.5% of the peaks in the four ATF3 ChIP-seq data sets (in GM12878, H1-hESC, HepG2, and K562 cells, respectively) contain USF sites but not CREB sites, and 98.35%–98.95% of these peaks overlap with peaks of USF1 or USF2 in the respective cell lines, indicating that they are bound by USF (USF is the heterodimer of USF1 and USF2). USF belongs to the basic helix-loop-helix leucine zipper family, and its motif does not resemble the CREB motif (Supplemental Fig. S2). Thus, the above analysis suggests that ATF3 tethers to USF, which binds DNA directly. Although ATF3 and USF have been reported to coregulate the same promoters (Runkel et al. 1991), it has yet to be shown that these TFs physically interact with each other independent of target DNA. As another example, in K562 cells, NFE2 binds 75.98% of its ChIP-seq peaks directly and tethers through USF for 15.59% of the peaks. Interestingly, the peaks that involve tethered binding show the strongest ChIP signal for this data set (Supplemental Fig. S6), perhaps because the NFE2 proteins that bind to DNA directly adopt a different conformation from those involved in tethered binding, and because the antibody used to perform the ChIP experiments has a lower affinity for NFE2 in the direct DNA-binding conformation. This is an exception because, for most TFs, the peaks that contain sites for the canonical motif show the strongest ChIP-seq signal (Supplemental Fig. S6). Other examples of tethered binding include SP1 tethering to HNF4 in HepG2 cells and STAT3 tethering to CEBPB in HeLa cells, both with literature support (Kardassis et al. 2002; Zhou et al. 2010). Novel predictions of tethering include TCF12 to FOXA and HNF4 in HepG2 cells, IRF1 to NF-Y in K562 cells, SREBF1 to RFX5 in HepG2 cells, and SIX5 to ZNF143 in GM12878 cells (Supplemental Table S3).

As a special case of cobinding, TFs that belong to the same protein family usually share identical or similar motifs and may compete for sites that match both motifs. MYC/MAX and USF (the USF1/USF2 heterodimer) both contain basic helix-loop-helix leucine zipper DNA-binding domains, but they do not cross-dimerize (Sawadogo et al. 1999). Their motifs share the CACGTG core, but the USF motif contains two additional nucleotides (gtCACGTG) that have a moderate sequence preference (Supplemental Fig. S2). The USF motif was discovered as a secondary motif in all five MAX ChIP-seq data sets, and 77.37%–92.75% of USF sites identified in the MAX data sets overlap with peaks in the USF1 or USF2 ChIP-seq data sets in the same cell line. These results suggest that USF and MYC/MAX compete for these sites. It was reported that both USF and MYC/MAX can bind an E-box motif in the promoter of the hamster *cad* gene, but only the binding of MYC/MAX is required for the transcription of *cad* (Boyd and Farnham 1997).

Distance and orientation preferences between the sites of cobinding TFs

Cobinding TFs bind to neighboring sites in the genome. For some TFs, multiple molecules of the same TF also can occupy neighboring sites. We asked whether these neighboring sites prefer to be on the same strand or opposite strands and whether they prefer to be in a specific range of distances. In addition to the analysis presented in the previous section, which compared the canonical motif with each noncanonical motif discovered in the same data set, we also compared motifs discovered in different data sets collected using the same cell line. In Figure 2B,C, we summarize the heterotypic and homotypic TF pairs that show statistically significant orientation or distance preferences separately in nonrepetitive and repetitive regions of the genome (the underlying data are in Supplemental Table S4). Out of the 78 motifs discovered from ChIP-seq data sets, 36 motifs (92 pairs; 62 heterotypic pairs and 30 homotypic pairs) are included in Figure 2B, suggesting that preferred arrangements of nearby TF binding sites are a common phenomenon. The neighboring sites for many heterotypic TF pairs (e.g., CTCF–NF-Y, ELF1–GABP, and FOXA–HNF4) as well as the neighboring homotypic sites of many TFs (e.g., AP-1, CTCF, and USF) show a strong preference for an edge-to-edge distance of <30 bp and varying degrees of preference for one orientation over the other. For example, neighboring NF-Y sites prefer to be in the same orientation. NF-Y also prefers one orientation to the other when cobinding with SP1, PBX3 (its motif is UA2), and USF. We hypothesized that these 92 TF pairs are more likely to represent protein–protein interactions than the TF pairs we identified in the previous section without testing for position or orientation preferences. Indeed, 14 heterotypic pairs and 17 homotypic pairs (33.7%) were detected in the aforementioned mammalian two-hybrid study (Ravasi et al. 2010) or in the BIOGRID database (Stark et al. 2006).

TFs tend to bind gene-rich regions of the genome due to their role in regulating target gene expression (Carroll et al. 2006). Nonetheless, repetitive elements are known to harbor functional TF binding sites, especially when such elements occur near genes. We systematically compared our compilation of TF binding sites with all repeats annotated in the human genome, and the results are summarized in Figure 3A. We confirmed the previously reported enrichment of STAT1, NF-Y, and CTCF binding sites in various repetitive elements (Bourque et al. 2008; Schmid and Bucher 2010), and we uncovered many more TFs whose binding sites are enriched in certain repetitive elements, e.g., UA1 sites in THE1B and THE1D retrotransposons. It was shown that a long terminal repeat (LTR) region of the THE1D retrotransposon was recruited as an alternative promoter for the human *IL2RB* gene and that the activity of this alternative promoter is regulated by DNA methylation (Cohen et al. 2011). The UA1 motif we identified in ZBTB33 peaks contains a prominent CGCG center (Fig. 1C) and ZBTB33 is known to bind methylated CpG dinucleotides (Yoon et al. 2003), raising the interesting possibility that the THE1B/D retrotransposons spread ZBTB33 binding sites across the genome and that the regulation of the newly recruited target genes can be modulated by the DNA methylation mechanism. Figures 2C and 3B summarize all motif pairs that show statistically significant distance or orientation preference in repetitive regions of the genome. The NF-Y–USF site pairs that typically have an end-to-end distance of 5–6 bp are nearly all located in the MLT1 family of retrotransposons. Similarly, the NF-Y–NF-Y site pairs at a 9-bp distance are found most often in LTR12 retrotransposons. There are 181 copies of the MLT1J transposon in the genome that contain sites for the NF-Y, USF, and ZNF143 motifs simultaneously, bound directly by NF-Y, USF, and ZNF143 TFs, respectively. The relative distance among the sites are nearly invariant (Fig. 2C), indicating recent duplications of MLT1J. Our results suggest a mechanism whereby retrotransposons amplify functional TF site pairs across the genome through transposition, potentially bringing new genes under the regulation of those TFs.

Cell-type–specific binding of sequence-specific TFs

The majority of the ENCODE ChIP-seq data was produced using five cell lines: K562, GM12878, HepG2, H1-hESC, and HeLa. Integrating ChIP-seq data with RNA-seq data for these five cell lines, we asked whether genes that are preferentially expressed in a given

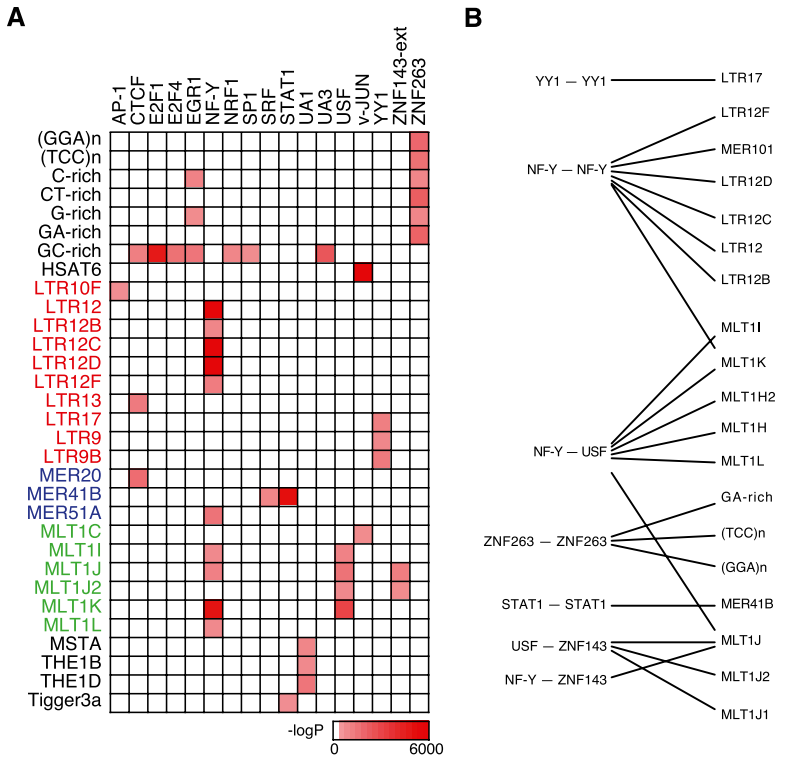

Figure 3. Binding sites of certain TFs or TF pairs are enriched in repeats. (*A*) Enrichment of TF binding sites in repetitive elements. The redness of each grid point is proportional to the negative logarithm of enrichment *P*-value. Repetitive elements are color-coded by family. (*B*) Enrichment of motif pairs that strongly prefer a narrow distance range in various repetitive elements (Fig. 2C).

cell line (defined by the average expression level in one cell line being more than 10-fold higher than that in any of the remaining four cell lines) show enriched TF binding sites in the corresponding cell line. This is, indeed, the case for a large fraction of genes, and Figure 4A shows five examples, one per cell line. (1) *FCER2* (the low-affinity receptor for IgE) is a key gene for B-cell function. It is highly and specifically expressed in GM12878. Its promoter region and gene body are bound by nine TFs in GM12878, including SPI1. (2) The G protein-coupled receptor *GPRC5A* plays a role in epithelial cell differentiation. It is highly and specifically expressed in HeLa cells, and accordingly, its promoter region and gene body are bound by seven TFs in HeLa cells. (3) The Abd-B homeobox family member *HOXB9* is a sequence-specific transcription factor. It is highly and specifically expressed in K562 cells, and accordingly, its promoter regions and gene body are bound by seven TFs including GATA1-TAL1 in K562 cells. (4) *SERPINA1* encodes a serine protease inhibitor, and defects in this gene can cause liver diseases. It is four orders of magnitude more highly expressed in HepG2 than in the other four cell lines. FOXA, HNF4, RXRA, TCF7L2, and eight other TFs bind near this gene in HepG2 but not in other cell lines. (5) *AC104304* encodes for a putative teratocarcinoma-derived growth factor that plays an important function in embryonic development. It is highly expressed in H1-hESC and bound by eight TFs, including NANOG.

We then asked whether the noncanonical motifs we discovered also reflect cell type specificity. Figure 4B plots the noncanonical motifs (circles) detected in the ChIP-seq data sets of sequence-specific TFs for each of the five cell lines (squares) with the most ENCODE ChIP-seq data sets. Cell-line–specific, noncanonical motifs are placed close to their respective cell lines in Figure 4B. We defined cell-line–specific motifs as those that were discovered three times more often in one cell line than in any other cell line. The remaining noncanonical motifs are placed in the center of the figure, and these motifs correspond to TFs that cooperate with other sequence-specific TFs across multiple cell lines. The thickness of the solid line connecting a noncanonical motif to a cell line indicates the proportion of data sets in that cell line that revealed the motif as a noncanonical motif.

We highlight several motifs that were frequently discovered as noncanonical motifs in a particular cell line. (1) PU.1 was most frequently discovered in GM12878 cells. Its corresponding TF SPI1, a member of the ETS family, activates gene expression during myeloid and B-lymphoid cell development. The *SPI1* gene is expressed in both GM12878 and K562 cells (RPKM = 4 and 9, respectively), but not in the other three cell lines (RPKM < 0.1). On the other hand, another member of the ETS family, *SPIB*, is only expressed in GM12878 cells, and the *SPIB* gene shows extensive TF binding sites specifically in GM12878 cells (bottom inset in Fig. 4B). SPIB and SPI1 have the same canonical motif (PU.1) and are both essential for B cell development (Sokalski et al. 2011). (2) GATA1 was the most frequently discovered noncanonical motif in K562 cells. It is bound by the GATA family of TFs, which are essential for erythroid development by regulating the fetal-to-adult switch of hemoglobin production (Takahashi et al. 1997). The *GATA1* gene is highly expressed in K562 cells but not in the other four cell lines and shows extensive binding sites only in the K562 cell line (top right inset in Fig. 4B). (3) FOXA and HNF4 are the most frequently identified noncanonical motifs in HepG2 cells. Their corresponding TFs (FOXA1 and HNF4) are activators of many liver-specific genes and are essential for hepatocyte function (Lemaigre and Zaret 2004). Both the *FOXA1* and *HNF4* genes are more than 10-fold more highly expressed and show more extensive TF binding sites in the HepG2 cell line than in the other four cell lines (*FOXA1* is shown in the middle right inset in Fig. 4B). (4) The SOX2-OCT4 combined motif was the most frequently identified noncanonical motif in H1-hESC cells. OCT4 is the canonical motif of POU5F1, a POU homeodomain-containing TF required for embryonic stem cell pluripotency. Their corresponding TFs (POU5F1 and SOX2) form a protein–protein complex and are required for embryonic stem cell pluripotency (Kashyap et al. 2009). Both *POU5F1* and *SOX2* are exclusively expressed in H1-hESC cells and extensively regulated by a large number of TFs, including by themselves (*POU5F1* is shown in the lower left inset of Fig. 4B).

Tethered binding of non-sequence-specific TFs

In Figure 4B, we also included all non-sequence-specific TFs (diamonds) for which there are ChIP-seq data in these cell lines. Dashed lines connect non-sequence-specific TFs to the motifs

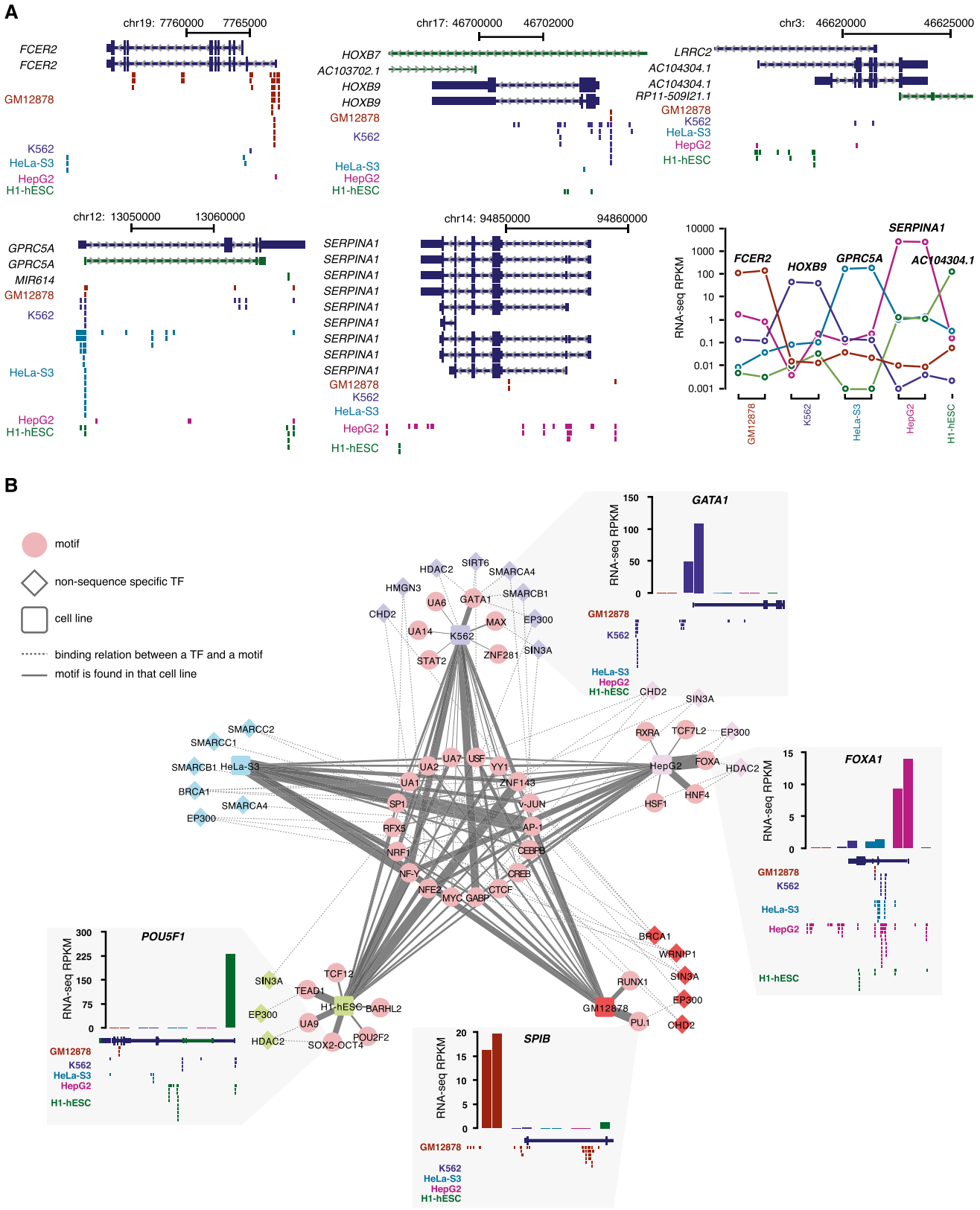

Figure 4. Cell-type–specific binding of sequence-specific and non-sequence-specific TFs. (*A*) Abundant TF binding sites are observed near cell-line–specific transcripts. Binding sites are shown as vertical bars and colored by cell line (dark blue for K562, red for HepG2, brown for GM12878, green for H1-hESC, and cyan for HeLa-S3). (*Bottom, right*) Expression levels (in RPKM) for example cell-line–specific transcripts across the five cell lines with the most ChIP-seq data. (*B*) Secondary motifs identified in the ChIP-seq data sets of sequence-specific TFs and their enrichment in the ChIP-seq peaks of non-sequence-specific TFs. The five cell lines are indicated with color-coded squares, noncanonical motifs of sequence-specific TFs are shown in pink circles and a solid line connecting each motif to the respective cell line. The thickness of the solid line is proportional to the normalized frequency in which a non-canonical motif is discovered in a particular cell line. Non-sequence-specific TFs are shown in diamonds whose colors match the color of the cell line if there is a ChIP-seq data set of the TF in that cell line. Dashed lines connect non-sequence-specific TFs and noncanonical motifs, indicating that a noncanonical motif of a sequence-specific TF is enriched in the ChIP-seq peaks of the non-sequence-specific TF. (Four *insets*) Expression profiles of sequence-specific TFs whose canonical motifs are found to be specific to a cell type and the TF binding sites around the genes that encode these TFs in the appropriate cell line. The expression levels in each cell line are assigned a similar color as the cell line. For four cell lines, two biological replicates were available for RNA-seq data; hence, there are two bars for each of these cell lines. Only one biological replicate was available for H1-hESC.

discovered in their ChIP-seq peaks. Two non-sequence-specific TFs show cell-line–specific enrichment in motifs: the enhancer-binding protein EP300 and the histone deacetylase HDAC2. There are seven data sets for EP300 in seven different cell lines and three data sets for HDAC2 in three different cell lines. Distinct motifs were found in different cell lines: SPI1 for EP300 in GM12878 cells; GATA1 (and GATA1-ext) for both EP300 and HDAC2 in K562 cells; FOXA and HNF4 for HDAC2, and FOXA and TCF7L2 for EP300 in HepG2 cells; SOX2-OCT4 and UA9 for HDAC2, and TEAD1 for EP300 in H1-hESC cells; and CEBPB, AP-1, and CREB for EP300 in HeLa cells. As described in the previous section, many of these motifs were most frequently and specifically observed as secondary motifs for sequence-specific TFs in the respective cell lines. Because non-sequence-specific TFs do not bind DNA directly, they tether onto sequence-specific TFs to bind target DNA. EP300 is known to interact with AP-1 and CEBPB (Mink et al. 1997; C-C Wang et al. 2007) and HDAC2 with TAL1-GATA (the motif is GATA1-ext) (Hu et al. 2009). Our results highlight that the interactions of EP300 and HDAC2 with sequence-specific TFs are highly cell type dependent.

We further analyzed the subsets of peaks of the sequence-specific TFs through which EP300 and HDAC2 might tether in a cell-type–specific fashion. For example, in GM12878, 4.9% (1745 out of 35,821) of the SPI1 peaks were associated with EP300 (i.e., overlapping with the EP300 peaks in the same cell line by at least 1 bp). We asked whether these peaks differed from the remaining 95.1% of SPI1 peaks by performing two types of analysis. First, we tested whether these EP300 or HDAC2-associated peaks were differentially enriched in any of our collection of 78 motifs (with enrichment defined in the Supplemental Methods). Second, we performed de novo motif finding on the top 500 of the subsets of peaks (ranked by ChIP-seq signal) using MEME-ChIP.

For EP300-associated peaks, we observed differential enrichment in a number of known motifs (Supplemental Fig. S8), and a subset of these motifs were also found by MEME-ChIP. For example, the 1745 SPI1 peaks that were associated with EP300 were more enriched in 18 motifs than all SPI1 peaks in GM12878 (cf. the PU.1–EP300 row and the PU.1 row in the GM12878 heat map in Supplemental Fig. S8), and MEME-ChIP identified AP-1, SP1, and RUNX1, in addition to the canonical motif of SPI1 (PU.1), which were among the 18 motifs. Similar results were observed for the GATA1 peaks that were associated with EP300 in K562 cells and the FOXA1 peaks that were associated with EP300 in HepG2 cells (see Supplemental Fig. S8 for results on enrichment; the MEME-ChIP results are not shown).

For HDAC2-associated peaks, we did not observe differential enrichment of any annotated motifs (for example, cf. the GATA1 row and the GATA1–HDAC2 row in the K562 heat map of Supplemental Fig. S8). For the HNF4 peaks associated with HDAC2 in HepG2 cells, MEME-ChIP only identified the HNF4 motif. Intriguingly, MEME-ChIP identified UA9 from the subset of POU5F1 peaks associated with HDAC2 in H1-hESC cells (in addition to POU5F1's canonical motif SOX2-OCT4). In comparison, MEME-ChIP did not identify UA9 from the overall top 500 POU5F1 peaks but identified UA9 from the overall top 500 HDAC2 peaks in H1-hESC cells (Supplemental Fig. S6), indicating that UA9 is specific to the regions bound by HDAC2. Indeed, 199 of the 473 HDAC2-associated POU5F1 peaks (42.1%) contain significant UA9 sites, compared with 23.1% in the POU5F1 peaks not associated with HDAC2 (Supplemental Fig. S9). Furthermore, MEME-ChIP identified another previously unannotated palindromic motif (Supplemental Fig. S9). This motif was significantly enriched in the HDAC2-associated GATA1 peaks in K562 cells compared to a set of randomly chosen genomic regions with matching GC% and lengths (P-value = 1.1×10^{-8}, using only the 1078 peaks that were not used as input to MEME-ChIP). This motif was distinct from all of the 78 motifs that were identified in the overall top 500 ChIP-seq peaks; thus, we named it UA12 and added it to our collection (Supplemental Fig. S2). Indeed, 36.9% of the HDAC2-associated GATA1 peaks in K562 cells contained UA12 sites (Supplemental Fig. S9), which was higher than the percentage of GATA1 peaks not associated with HDAC2 (32.1%) and the flanking regions (26.1%).

The ChIP-seq peaks of several non-sequence-specific TFs were enriched in the same motifs, regardless of cell line. The best example is the chromodomain helicase DNA-binding protein CHD2, for which the UA1 motif (likely the canonical motif of ZBTB33) was prominently discovered in GM12878, HepG2, and K562 cells, suggesting that CHD2 functions by interacting with ZBTB33. REST is highly enriched in the ChIP-seq peaks of SIN3A in H1-hESC cells. Accordingly, SIN3A is known to associate with REST and repress neuronal genes in nonneuronal cells (Huang et al. 1999).

The ChIP-seq peaks of many TFs are flanked by positioned nucleosomes

After analyzing the motif content of TF binding peaks, we set out to investigate the chromatin structure around these peaks. It is well known that transcription has a profound impact on nucleosome occupancy: Active TSSs in all eukaryotes are flanked by an upstream nucleosome-depleted region and several well-positioned downstream nucleosomes (Radman-Livaja and Rando 2010). We previously reported that the binding sites of the insulator binding protein CTCF were flanked by an array of strongly positioned nucleosomes, shown as a periodic oscillatory pattern in the average nucleosome occupancy profile centered on CTCF binding sites (Fu et al. 2008). Another study showed that NRSF (also called REST) binding sites are flanked by positioned nucleosomes in CD4+ T cells, CD8+ T cells, and granulocytes (Valouev et al. 2011).

In order to investigate where TF binding peaks were located with respect to nucleosomes, we computed an average nucleosome occupancy profile centered on the peak summits of each TF with available ChIP-seq data in GM12878 or K562 cells (Fig. 5A,B for YY1 in GM12878 cells; Supplemental Fig. S10 for all data sets). We had ChIP-seq data for 51 TFs in GM12878 cells, 73 TFs in K562 cells, and 32 TFs in both cell lines. Some TFs were tested by multiple labs in the same cell line, and we included all these data sets. To account for the impact of transcription, we computed the average nucleosome profile anchored on TSS-proximal and TSS-distal peak summits separately. Nucleosome profiles anchored on TSS-proximal peaks were oriented such that the nearest transcript is downstream from the anchor. We further stratified peaks in each data set as top, middle, and bottom thirds according to the ChIP-seq signal, reflecting the extent to which a peak is bound by the TF (averaged over a population of cells). We distinguish nucleosome occupancy and nucleosome positioning, with occupancy defined as the area under the occupancy profile and positioning defined as the regularity of the oscillatory pattern in the occupancy profile. Thus, the regions around TSS-proximal summits tend to show lower nucleosome occupancy and lower nucleosome positioning than regions around TSS-distal summits (cf. Fig. 5A,B; similarly the proximal and distal panels in Supplemental Fig. S10; note the difference in the y-scale). This difference may reflect the effects of RNA polymerase on chromatin structure (Weiner et al. 2010). Within the proximal and distal categories, the top, middle,

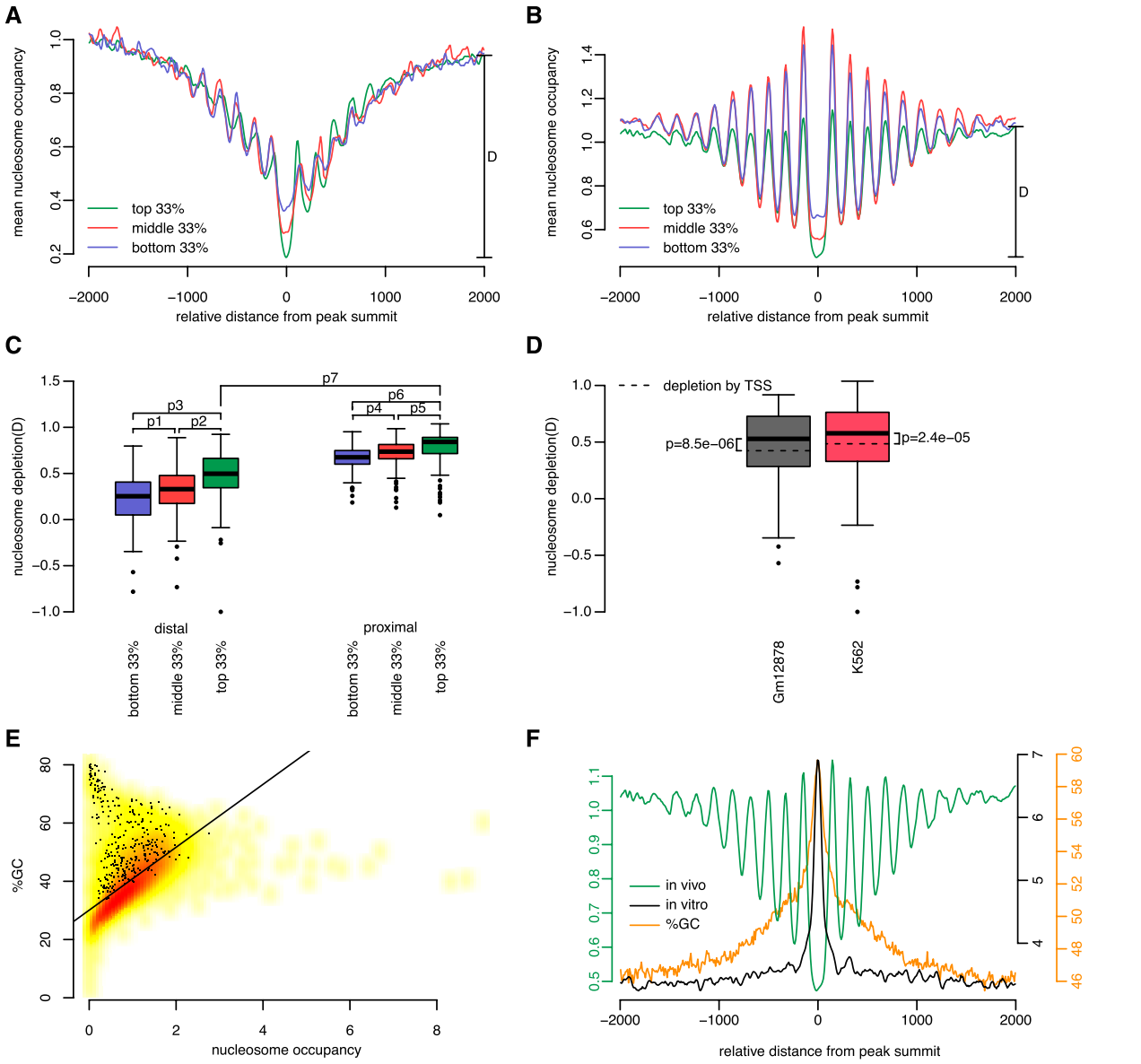

Figure 5. Chromatin structure and GC content around TF binding regions. (*A*,*B*) Nucleosome occupancy profiles anchored on the summits of TSS-proximal (*A*) and TSS-distal (*B*) peaks of YY1 grouped by ChIP-seq signal strength: top (green), middle (red), and bottom (blue) third peaks in terms of ChIP-seq signal. Nucleosome depletion for the top third peaks is shown as D in each panel. (*C*) Distribution of nucleosome depletion "D" across all tested TFs, with peaks stratified according to TSS proximity (proximal or distal) and ChIP-seq signal strength (top, middle, or bottom third). *P*-values for pairwise comparisons based on paired Wilcoxon rank-sum tests are: $P1 = 8.2 \times 10^{-17}$, $P2 = 7.6 \times 10^{-21}$, $P3 = 3.8 \times 10^{-23}$, $P4 = 8.8 \times 10^{-10}$, $P5 = 1.1 \times 10^{-9}$, $P6 = 1.1 \times 10^{-11}$, and $P7 = 6.6 \times 10^{-22}$. (*D*) TF binding is correlated with significantly more nucleosome depletion than TSS. Wilcoxon rank-sum test *P*-values are shown separately for GM12878 and K562 cells. For the box plots in *C* and *D*, only those subcategories with 200 or more peaks are included, and whiskers represent the 1.5 inter-quartile range. (*E*) Nucleosome occupancy genome-wide is correlated with GC%. The smoothed density scatter plot contains 40,000 data points; each data point is a randomly chosen 250-bp region of the human genome. (Black dots) Those regions that overlap with ChIP-seq peaks. (Black line) Least square fit. Pearson correlation coefficient = 0.62; *P*-value < 2.2×10^{-16}. (*F*) Comparison of in vivo (green) and in vitro (black) nucleosome occupancy profiles around peak summits of YY1. GC% profile around the same summits is plotted in orange. Note elevated GC% at summit coincides with high in vitro nucleosome occupancy and low in vivo nucleosome occupancy.

and bottom third peaks, which correspond to the peaks with strongest, medium, and weakest TF binding, tended to show the greatest, medium, and weakest nucleosome positioning (Fig. 5A,B; Supplemental Fig. S10). Thus regions that are more strongly bound by TFs are flanked by better-positioned nucleosomes.

The cohesin components SMC3 and RAD21 (Supplemental Fig. S10) show the most striking patterns of positioned flanking nucleosomes, similar to what we previously reported for CTCF (Fu et al. 2008), to which these factors bind (Parelho et al. 2008). Two other TFs—CTCFL (a paralog of CTCF) and ZNF143 (a zinc-finger protein with a long motif)—also show striking patterns of positioned flanking nucleosomes. The binding sites for ~70% of the tested TFs are flanked by positioned nucleosomes (Supplemental Fig. S10), indicating that this is a general phenomenon for

most TFs. To quantify the regularity of nucleosome positioning around TF binding sites, we applied Fourier transforms to the nucleosome occupancy profiles, yielding power spectra. The height of the power spectrum at the spatial frequency corresponding to the nucleosomal repeat length was used as an indicator of how periodically nucleosomes were positioned (an example power spectrum is shown in Supplemental Fig. S11B). The spectrum height correlated significantly with the extent of positioning of the –1 and +1 nucleosomes (measured as the maximum of the nucleosome occupancy profile) (Supplemental Fig. S11C). Thus, how well the –1 and +1 nucleosomes are positioned strongly predicts how periodically the flanking nucleosomes are positioned.

Most TFs bind at GC-rich, nucleosome-depleted, and DNase I-accessible regions

The nucleosome occupancy profile dips at the peak summits of most TFs (Fig. 5A,B; Supplemental Fig. S10), indicating that TFs prefer to bind nucleosome-depleted regions or that the binding of a TF excludes nucleosomes. In the vicinity of TSS-proximal summits, lower nucleosome occupancy is seen in the direction of transcription than upstream of transcription. We define nucleosome depletion as the amount that nucleosome occupancy dips at the peak summit, as compared to the nucleosome occupancy at 2 kb from the summit (considered as background). TSS-proximal summits show significantly greater nucleosome depletion than TSS-distal summits (Fig. 5C). It is well known that the binding of the transcriptional machinery to the TSS excludes nucleosomes to a considerable extent (Radman-Livaja and Rando 2010). Indeed, average nucleosome occupancy anchored on the TSS shows an overall loss of nucleosomes (Supplemental Fig. S12). Interestingly, we observed that TSS-proximal TF peak summits show a significantly greater depletion in nucleosome occupancy than do TSSs (Fig. 5D). The median nucleosome depletion at the summits of TSS-proximal peaks is 0.56 for GM12878 cells and 0.59 for K562 cells, significantly greater than the maximal nucleosome depletion around TSS (0.42 for GM12878 cells and 0.48 for K562 cells; Wilcoxon rank-sum test P-value = 7.1×10^{-28} and 1.1×10^{-22}, respectively). Within the proximal and distal categories, the top, middle, and bottom third peaks showed greatest, medium, and weakest nucleosome depletion, respectively (Fig. 5C). This result indicates that TFs and nucleosomes compete for the genomic DNA and that stronger TF binding is correlated with greater nucleosome depletion, above and beyond the effect of transcription.

The peaks of seven TFs (BRF2, HDAC8, TRIM28, SETDB1, WRNIP1, ZNF274, and ZZZ3) do not show nucleosome depletion, nor are these peaks flanked by well-positioned nucleosomes, indicating these TFs tend to bind nucleosomal DNA (Supplemental Fig. S10). Three of these TFs function with each other to repress transcription. SETDB1 is a histone methyltransferase that catalyzes H3K9me3, which signals for the silencing of euchromatic genes (Bilodeau et al. 2009). TRIM28 (commonly known as KAP1) represses transcription by recruiting SETDB1 (C Wang et al. 2007). ZNF274 is a zinc-finger containing TF that binds to the 3′ end of zinc-finger coding genes and recruits chromatin-modifying proteins such as SETDB1 and TRIM28, which leads to transcriptional repression (Frietze et al. 2010). HDAC8 is a histone deacetylase and a transcriptional repressor. We caution that the HDAC8 ChIP-seq data set had only 287 peaks. BRF2 is a component of the RNA Pol III machinery (Moqtaderi et al. 2010). WRNIP1 (commonly known as WHIP) regulates DNA synthesis. ZZZ3 is a component of the ATAC complex and a histone H3 acetyltransferase and has been shown to acetylate both free and nucleosomal H3 (Wang et al. 2008).

We next asked whether the intrinsic DNA sequence properties of ChIP-seq peaks contribute to nucleosome depletion. In an earlier study, we reported a strong correlation between GC-rich sequences and their potential to form nucleosomes (Peckham et al. 2007). In vitro data also indicate that GC-rich sequences promote nucleosome formation (Valouev et al. 2011). Indeed, there is positive correlation between nucleosome occupancy and GC content for randomly chosen 250-bp regions of the genome (r = 0.62 and P-value < 2.2×10^{-16}) (Fig. 5E). Many of those regions that overlap ChIP-seq peaks (Fig. 5E, black dots) are located above and to the left of the best-fit line, indicating that they have high GC% and low nucleosome occupancy. Compared with the average GC content of 40% in the human genome, ChIP-seq peaks are considerably more GC rich (61 ± 5% for TSS-proximal peaks and 53 ± 6% for TSS-distal peaks across the TFs). The high GC content may be due to the GC-richness of some TF motifs, but the motif sites are much smaller than peaks (8–21 bp vs. ~250 bp), and we found similar GC patterns around TF summits without a motif site (data not shown). We conclude that TFs tend to bind GC-rich regions in the genome, regardless of the distance from the TSS. These results are seemingly contradictory—GC content is highly predictive of sequences that promote nucleosome formation, yet the GC-rich sequences surrounding TF binding sites are nucleosome-depleted in vivo.

To determine whether TF binding sites are, indeed, favorable sites for nucleosome formation, we used recent data from in vitro reconstitution of human genomic DNA into nucleosomes (Valouev et al. 2011) to construct in vitro nucleosome occupancy profiles around ChIP-seq peaks, confirming that in vitro nucleosome occupancy is much higher on the peak compared with flanking regions for the vast majority of TFs (Fig. 5F for YY1; Supplemental Fig. S13 for all TFs). Thus, TFs or their cofactors (such as chromatin remodelers) prevent the formation of nucleosomes or evict nucleosomes at these GC-rich locations of the genome.

Chromatin structure around cell-line–specific TF binding regions

In order to further investigate the relationship between TF binding and chromatin structure, we examined two sets of cell-line–specific ChIP-seq peaks for each TF—the set of peaks detected in GM12878 but not in K562 and the set of peaks detected in K562 but not in GM12878. We computed nucleosome occupancy profiles and DNase I cleavage profiles anchored on the summits of these two sets of peaks separately in each cell line (Fig. 6A,B for YY1; Supplemental Fig. S14 for all TFs). Strikingly, the peaks that were occupied by a TF in GM12878 (or K562) but not occupied by the TF in K562 (or GM12878) tend to be occupied by a nucleosome in K562 (or GM12878), similar to the in vitro nucleosome profiles of these peaks (Supplemental Fig. S15). Accordingly, the increase in nucleosome occupancy is reflected in decreased DNase I cleavage in K562 (or GM12878). For many TFs, the ChIP-seq peaks that were occupied by a TF in GM12878 (or K562) but not occupied by the TF in K562 (or GM12878) were no longer flanked by positioned nucleosomes in K562 (or GM12878), yet positioned nucleosomes were observed for other TFs, albeit at a lesser extent of positioning than the nucleosomes flanking TF-occupied peaks (Supplemental Fig. S14). Thus, for the same set of genomic sequences in two cell lines, TF binding level deviates from thermodynamic preference for nucleosome formation—TF binding either was enabled by, or caused, cell-type–specific depletion of nucleosomes from intrinsically favorable genomic locations.

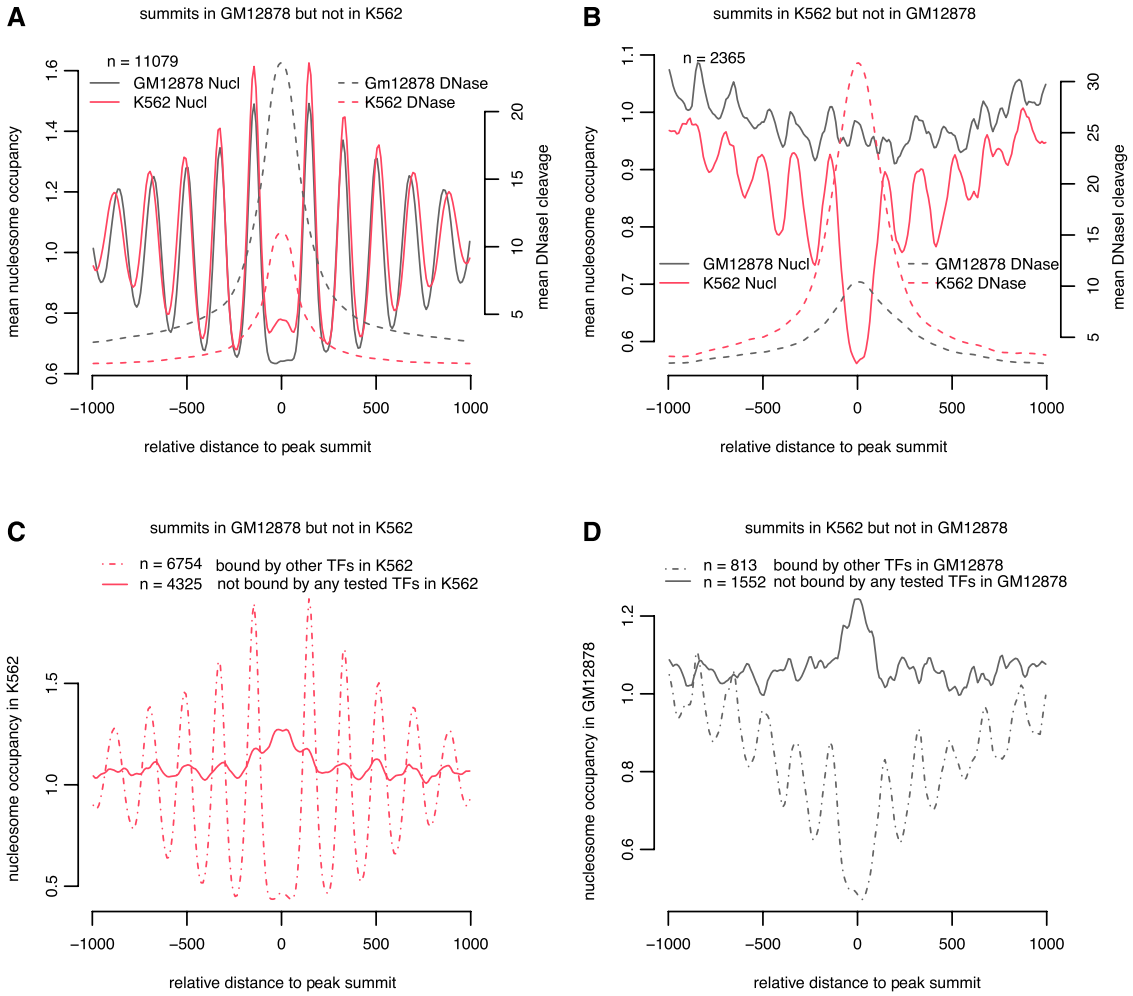

Figure 6. Chromatin structure around YY1 ChIP-seq peaks occupied differentially between GM12878 and K562. (*A*) Nucleosome occupancy profiles (solid lines) and DNase I cleavage profiles (dashed lines) anchored on the summits of YY1 peaks in GM12878 but not in K562. Note the average nucleosome occupancy 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. (*B*) Same as *A*, but around the summits of YY1 peaks in K562 but not in GM12878. (*C*) Nucleosome occupancy profiles in K562 anchored on the summits of the ChIP-seq peaks occupied by YY1 in GM12878 but not in K562. These 11,079 peaks were divided into two groups: 6754 peaks were bound by one or more TFs in K562 (dashed line), and 4325 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 to the lack of nucleosome occupancy at the peak summits and well-positioned nucleosomes flanking the peaks bound by other TFs. (*D*) Same as *C*, but around the summits of the ChIP-seq peaks occupied by YY1 in K562 but not in GM12878.

We further partitioned the peaks occupied by a TF in GM12878 (or K562) but not occupied by the TF in K562 (or GM12878) into two subsets: group 1 peaks that overlapped with one or more ChIP-seq peaks of any other TF tested in K562 (or GM12878), and group 2 peaks that did not overlap any ChIP-seq peaks in K562 (or GM12878). For the vast majority of the TFs, nucleosome occupancy profiles for the group 2 peaks show high nucleosome occupancy on the peak and no positioned nucleosomes flanking the peak. In contrast, the group 1 peaks show nucleosome depletion on the peak and well-positioned nucleosomes flanking the peak (Fig. 6C,D; Supplemental Fig. S16). The ChIP-seq data we have only cover up to 10% of TFs in a particular cell line, thus group 2 peaks can still be bound by other TFs for which we had no data, which could account for any residual pattern of nucleosome positioning. These results further strengthen the correlation between TF-binding and flanking positioned nucleosomes and indicate that such correlation can be regulated in a cell-type–specific manner.

Factorbook

The analysis presented here, in addition to forthcoming new types of analysis will be performed on an ongoing basis as more ENCODE data are produced. For example, the ENCODE Consortium has generated 743 ChIP-seq data sets as of March 2012. The results of the continual analysis are presented via a web-accessible wiki-based TF-centric repository called Factorbook (http://factorbook.org). At present, Factorbook also includes the profiles of all histone modifications assayed by the ENCODE Consortium anchored on TF ChIP-seq peaks in the corresponding cell lines.

Discussion

The ENCODE Consortium has produced the largest collection of ChIP-seq data sets of human TFs to date: 457 data sets on 119 TFs in the January 2011 freeze. The majority of these data are of high

quality, and we were able to detect sequence motifs with high confidence. Of the 87 Pol II-associated sequence-specific TFs, the canonical motifs of 76 TFs were annotated in the JASPAR or TRANSFAC databases (Matys et al. 2003; Bryne et al. 2008), and we discovered the canonical motifs de novo from ChIP-seq peaks for all of these TFs. Furthermore, we discovered 12 high-confidence unannotated motifs, and five of them are likely the canonical motifs for four TFs. The sites of these unannotated motifs show strong sequence conservation and high chromatin accessibility, and some motifs show distinct DNase I footprints, suggesting that these are functional motifs.

In a separate study (Whitfield et al. 2012), we performed functional assays to test 455 binding sites for several TFs in promoters in four cell lines. We transiently transfected a construct with an ~1-kb fragment containing a wild-type promoter and a luciferase reporter and measured the luminosity. We also tested a construct with a mutagenized TF binding site. A significant difference between the wild-type and mutant luminosities suggests that the site is functional. At an FDR cutoff of 0.025, the validation rate was 36%–49% in a particular cell line, and ~70% of TF binding sites were validated in at least one of four cell lines, suggesting that most of the motif sites are functional. In that study, we also found that secondary motifs can distinguish functional from nonfunctional binding instances of CTCF and STAT1.

We observed a general agreement between the strength of ChIP-seq signal and motif content among the peaks in the same ChIP-seq data set. Peaks that are ranked highly according to ChIP-seq signal tend to be more likely to contain motif sites, and these sites are more tightly positioned around the peak summits, compared to low-ranked peaks. Thus, the motif sites likely correspond to the base pairs of genomic DNA with which the TF protein forms atomic contacts. Different TFs vary greatly in total numbers of ChIP-seq peaks, from hundreds to tens of thousands. CTCF, CEBPB, FOXA1, and SPI1 are among the TFs with the most peaks (>40,000 peaks in some cell lines); nonetheless, even the bottom-ranked peaks are strongly enriched in motifs (Fig. 1B; Supplemental Fig. S6), suggesting that most of the peaks are bound by the TFs. MacQuarrie et al. and Biggin discussed the biological significance of the vast number of peaks and suggested that binding of TFs may have biological roles in addition to direct transcriptional target regulation (MacQuarrie et al. 2011; Biggin 2011).

Although anecdotal evidence for cooperative interactions between TFs abounds in the literature, it remains unclear if such interactions are a common strategy in transcriptional regulation. High quality ChIP-seq data from the ENCODE Consortium allowed us to examine this aspect of TF function in a systematic manner. We identified noncanonical motifs for the vast majority of the sequence-specific TFs and the non-sequence-specific TFs, revealing a spectrum of cobinding and tethered binding of multiple TFs to genomic DNA. The TFs in some of the predicted pairs may both be components of a large multiunit transcriptional complex without physically contacting each other, and other TFs may bind to neighboring sites that are not close enough for the TFs to form protein–protein contacts. We expanded the analysis by comparing the sites of all discovered motifs, in the same or different data sets, and discovered 92 pairs of motifs whose binding sites showed significant distance and/or orientation preferences. Some TFs prefer to bind to sites with a broad distribution of edge-to-edge distances of <30 bp, suggesting that these TFs interact with each other on the protein level, yet the interactions permit some variation in the distance between their DNA sites. Other TFs prefer to bind neighboring sites positioned in a narrow distribution of distances, and some of these TF pairs show an orientation preference, suggesting more restrictive interactions between these TFs. Taken together, our results indicate that TF-TF interactions are prevalent and can take on a variety of forms.

The majority of the ENCODE ChIP-seq data sets were generated using five cell lines, thus we investigated cell-line–specific TF binding sites and integrated the results with cell-line–specific gene expression using the RNA-seq data in the corresponding cell lines. The results of our systematic analysis support the model that cell-type–specific transcription can be regulated in three ways: (1) Sequence-specific TFs can bind to distinct sites and thus regulate different genes in different cell types; (2) some sequence-specific TF proteins are highly expressed in a cell type, and these TFs bind to the target regions of many other TFs in the same cell type, perhaps because the chromatin at these regions are already accessible; and (3) some non-sequence-specific TF proteins bind to cell-type–specific sequence-specific TF proteins to exert another layer of regulation. There have been many reported examples of TFs and target genes for each mode of regulation, yet an integrative analysis like ours has the power of illustrating all three modes of regulation across a large number of TFs and over multiple cell lines.

We further integrated the ChIP-seq data with nucleosome positioning and DNase I cleavage data in two cell lines (GM12878 and K562) to study the interplay between TF binding and chromatin structure. We found that the ChIP-seq peaks of most TFs correspond to GC-rich, nucleosome-depleted, and DNase I-accessible regions, flanked by well-positioned nucleosomes. We may have underestimated the number of TFs whose binding regions are flanked by positioned nucleosomes, because we simply averaged over all peaks in each ChIP-seq data set. If subsets of peaks are flanked by well-positioned nucleosomes, and the positions of the nucleosomes are offset from each other between the subsets, then averaging may mask the signal. Another ENCODE companion paper clusters peaks by the flanking nucleosome occupancy patterns and reports that subsets of peaks are flanked by positioned nucleosomes for almost every TF (Kundaje et al. 2012). That paper also investigated the positional patterns of nucleosomes with modified histones.

We further investigated the regions that were bound by a TF in GM12878 but not in K562 and vice versa and found that these regions are typically occupied by a nucleosome in the cell line that the TF does not bind, and the increase in nucleosome occupancy is perfectly correlated with a decrease in DNase I cleavage. Consistent with previous findings that GC-rich sequences tend to form nucleosomes (Peckham et al. 2007), we found that TF binding regions show locally elevated in vitro nucleosome occupancy compared to flanking regions, indicating that these regions are intrinsically nucleosomal unless they are bound by TFs. Indeed, He et al. found that androgen treatment dismissed a central nucleosome, which was flanked by a pair of marked nucleosomes, to reveal androgen receptor binding sites (He et al. 2010). Taken together, our results show that a strong correlation between TF binding and positioning of nearby nucleosomes is likely a universal phenomenon for all TFs. The binding of a single TF is unlikely to position flanking nucleosomes (a single TF is thought to have lower affinity for DNA than a nucleosome) (Felsenfeld 1996), but multiple TFs tend to bind to neighboring regions, and they collectively may be able to position nucleosomes. Alternatively, chromatin remodelers may have configured the chromatin structures around TF binding regions in a cell-type–specific fashion to facilitate TF binding. It is also possible that TFs and chromatin remodelers work together to establish the chromatin structure.

Recent work compared chromatin accessibility before and after induction of the *Drosophila* heat shock transcription factor (HSTF) (Guertin and Lis 2010) and the mammalian glucocorticoid receptor (GR) (John et al. 2011); these studies concluded that the chromatin was already accessible prior to induction. Our results go beyond these studies by showing that positioned nucleosomes constitute the chromatin structure around the binding regions of most TFs. We suggest that the GC-richness of TF binding regions may be a mechanism for preventing unintended TF-binding, in that a nucleosome would tend to occupy the region until it is evicted, possibly by chromatin remodelers or by multiple TFs in concert.

Methods

ENCODE ChIP-seq, MNase-seq, and DNase-seq data sets

Members of the ENCODE Consortium generated 457 ChIP-seq data sets on 119 human transcription factors (TFs) in 72 cell lines (Supplemental Table S1; The ENCODE Project Consortium 2011).

In this study, we integrated the following additional ENCODE data sets: the nucleosome occupancy profiles in GM12878 and K562 (Kundaje et al. 2012), RNA-seq data (Djebali et al. 2012), and DNase-seq data (Neph et al. 2012).

De novo sequence motif discovery in ChIP-seq peaks

We constructed a de novo motif discovery pipeline as illustrated in Supplemental Figure S1 and described in detail in Supplemental Methods. Briefly, we separated the peaks in each ChIP-seq data set into independent training and testing sets to ensure the quality of the motif discovery. 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 data set. We asked MEME to report up to five significant motifs per data set. Then, we performed further testing to ensure the quality of the motifs, as described in Supplemental Methods. In total, we identified 1092 motifs. We compared the discovered motifs with annotated motifs in the JASPAR (Bryne et al. 2008) and TRANSFAC (Matys et al. 2003) databases using TOMTOM (Gupta et al. 2007). We manually merged similar motifs and identified 79 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 data sets or were supported by the literature. We further identified another motif, UA12, in the subset of HDAC2-associated GATA1 peaks in K562.

Distance and/or orientation preferences between motif sites

For all pairs of motifs, we computed the edge-to-edge distance and relative orientation and computed a *P*-value using the Kolmogorov-Smirnov test, as described in Supplemental Methods. The analysis was also applied to peaks in nonrepetitive regions and repetitive regions of the genome separately. 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. All *P*-values were corrected for multiple testing, and an FDR cutoff of 0.025 was applied throughout.

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, as described in Supplemental Methods. 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 profiles, and the output of an FFT is a power spectrum. 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, the more periodic the nucleosome occupancy profile).

Acknowledgments

We thank ENCODE data producers for the rich data. We thank consortium members and Mike Pazin for insightful discussions and critical comments on the paper. We thank John Stamatoyannopoulos for bringing to our attention the distinction between tethered binding and cobinding of transcription factors. This work was funded by NIH grants U01 HG004561, U01 HG004695, and R01 GM103544, and NSF grant DBI-0850008.

Author contributions: Z.W. designed and supervised the analysis. J.W., J.Z., and S.I. performed the majority of the analysis. XY.L. engineered factorbook.org. The remaining authors contributed to the analysis. Z.W., J.W., J.Z., and S.I. wrote the manuscript. Other authors contributed to editing the manuscript.

References

Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, et al. 2009. Diversity and complexity in DNA recognition by transcription factors. *Science* **324:** 1720–1723.

Berger SL. 2007. The complex language of chromatin regulation during transcription. *Nature* **447:** 407–412.

Biggin MD. 2011. Animal transcription networks as highly connected, quantitative continua. *Dev Cell* **21:** 611–626.

Bilodeau S, Kagey MH, Frampton GM, Rahl PB, Young RA. 2009. SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. *Genes Dev* **23:** 2484–2489.

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew J-L, Ruan Y, Wei C-L, Ng H-H, et al. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* **18:** 1752–1762.

Boyd KE, Farnham PJ. 1997. Myc versus USF: Discrimination at the cad gene is determined by core promoter elements. *Mol Cell Biol* **17:** 2529–2537.

Bryne JC, Valen E, Tang M-HE, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A. 2008. JASPAR, the open access database of transcription factor-binding profiles: New content and tools in the 2008 update. *Nucleic Acids Res* **36:** D102–D106.

Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoute J, Brodsky AS, Keeton EK, Fertuck KC, Hall GF, et al. 2006. Genome-wide analysis of estrogen receptor binding sites. *Nat Genet* **38:** 1289–1297.

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, et al. 2008. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133:** 1106–1117.

Clapier CR, Cairns BR. 2009. The biology of chromatin remodeling complexes. *Annu Rev Biochem* **78:** 273–304.

Cohen C, Rebollo R, Babovic S, Dai E, Robinson W, Mager D. 2011. Placenta-specific expression of the interleukin-2 (IL-2) receptor β subunit from an endogenous retroviral promoter. *J Biol Chem* **286:** 35543–35552.

de Nigris F, Botti C, Rossiello R, Crimi E, Sica V, Napoli C. 2007. Cooperation between Myc and YY1 provides novel silencing transcriptional targets of α3β1-integrin in tumour cells. *Oncogene* **26:** 382–394.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi AM, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* (in press).

The ENCODE Project Consortium. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9:** e1001046. doi: 10.1371/journal.pbio.1001046.

Felsenfeld G. 1996. Chromatin unfolds. *Cell* **86:** 13–19.

Frietze S, O'Geen H, Blahnik K, Jin VX, Farnham PJ. 2010. ZNF274 recruits the histone methyltransferase SETDB1 to the 3′ ends of ZNF genes. *PLoS ONE* **5:** e15082. doi: 10.1371/journal.pone.0015082.

Fu Y, Sinha M, Peterson CL, Weng Z. 2008. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet* **4:** e1000138. doi: 10.1371/journal.pgen.1000138.
Guertin MJ, Lis JT. 2010. Chromatin landscape dictates HSF binding to target DNA elements. *PLoS Genet* **6:** e1001114. doi: 10.1371/journal.pgen.1001114.
Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome Biol* **8:** R24. doi: 10.1186/gb-2007-8-2-r24.
He HH, Meyer CA, Shin H, Bailey ST, Wei G, Wang Q, Zhang Y, Xu K, Ni M, Lupien M, et al. 2010. Nucleosome dynamics define transcriptional enhancers. *Nat Genet* **42:** 343–347.
Hu X, Li X, Valverde K, Fu X, Noguchi C, Qiu Y, Huang S. 2009. LSD1-mediated epigenetic modification is required for TAL1 function and hematopoiesis. *Proc Natl Acad Sci* **106:** 10141–10146.
Huang Y, Myers SJ, Dingledine R. 1999. Transcriptional repression by REST: Recruitment of Sin3A and histone deacetylase to neuronal genes. *Nat Neurosci* **2:** 867–872.
John S, Sabo PJ, Thurman RE, Sung M-H, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA. 2011. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet* **43:** 264–268.
Kardassis D, Falvey E, Tsantili P, Hadzopoulou-Cladaras M, Zannis V. 2002. Direct physical interactions between HNF-4 and Sp1 mediate synergistic transactivation of the apolipoprotein CIII promoter. *Biochemistry* **41:** 1217–1228.
Kashyap V, Rezende NC, Scotland KB, Shaffer SM, Persson JL, Gudas LJ, Mongan NP. 2009. Regulation of stem cell pluripotency and differentiation involves a mutual regulatory circuit of the NANOG, OCT4, and SOX2 pluripotency transcription factors with polycomb repressive complexes and stem cell microRNAs. *Stem Cells Dev* **18:** 1093–1108.
Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: An information aesthetic for comparative genomics. *Genome Res* **19:** 1639–1645.
Kundaje A, Kyriazopoulou-Panagiotopoulou S, Libbrecht M, Smith CL, Raha D, Winters EE, Johnson SM, Snyder M, Batzoglou S, Sidow A. 2012. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res* (this issue). doi: 10.1101/gr.136366.111.
Lemaigre F, Zaret KS. 2004. Liver development update: New embryo models, cell lineage control, and morphogenesis. *Curr Opin Genet Dev* **14:** 582–590.
Levine M, Tjian R. 2003. Transcription regulation and animal diversity. *Nature* **424:** 147–151.
Li X-Y, Thomas S, Sabo PJ, Eisen MB, Stamatoyannopoulos JA, Biggin MD. 2011. The role of chromatin accessibility in directing the widespread, overlapping patterns of *Drosophila* transcription factor binding. *Genome Biol* **12:** R34. doi: 10.1186/gb-2011-12-4-r34.
Machanick P, Bailey TL. 2011. MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics* **27:** 1696–1697.
MacQuarrie KL, Fong AP, Morse RH, Tapscott SJ. 2011. Genome-wide transcription factor binding: Beyond direct target regulation. *Trends Genet* **27:** 141–148.
Maston GA, Evans SK, Green MR. 2006. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **7:** 29–59.
Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, et al. 2003. TRANSFAC: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31:** 374–378.
Mink S, Haenig B, Klempnauer KH. 1997. Interaction and functional collaboration of p300 and C/EBPβ. *Mol Cell Biol* **17:** 6609–6617.
Moqtaderi Z, Wang J, Raha D, White RJ, Snyder M, Weng Z, Struhl K. 2010. Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. *Nat Struct Mol Biol* **17:** 635–640.
Myslinski E, Gérard M-A, Krol A, Carbon P. 2006. A genome scale location analysis of human Staf/ZNF143-binding sites suggests a widespread role for human Staf/ZNF143 in mammalian promoters. *J Biol Chem* **281:** 39953–39962.
Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, Sandstrom R, Johnson AK, Maurano MT, et al. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* (in press).
Oettel S, Kober I, Seifart KH. 1998. The activity binding to the termination region of several pol III genes represents a separate entity and is distinct from a novel component enhancing U6 snRNA transcription. *Nucleic Acids Res* **26:** 4324–4331.
Ohlsson R, Renkawitz R, Lobanenkov V. 2001. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet* **17:** 520–527.
Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, Gregson HC, Jarmuz A, Canzonetta C, Webster Z, Nesterova T, et al. 2008. Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132:** 422–433.
Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, Struhl K, Weng Z. 2007. Nucleosome positioning signals in genomic DNA. *Genome Res* **17:** 1170–1177.
Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20:** 110–121.
Radman-Livaja M, Rando OJ. 2010. Nucleosome positioning: How is it established, and why does it matter? *Dev Biol* **339:** 258–266.
Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, Akalin A, Schmeier S, Kanamori-Katayama M, Bertin N, et al. 2010. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140:** 744–752.
Rhee HS, Pugh BF. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147:** 1408–1419.
Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4:** 651–657.
Roder K, Wolf SS, Larkin KJ, Schweizer M. 1999. Interaction between the two ubiquitously expressed transcription factors NF-Y and Sp1. *Gene* **234:** 61–69.
Runkel L, Shaw PE, Herrera RE, Hipskind RA, Nordheim A. 1991. Multiple basal promoter elements determine the level of human c-fos transcription. *Mol Cell Biol* **11:** 1270–1280.
Sabogal A, Lyubimov AY, Corn JE, Berger JM, Rio DC. 2010. THAP proteins target specific DNA sites through bipartite recognition of adjacent major and minor grooves. *Nat Struct Mol Biol* **17:** 117–123.
Sawadogo M, Luo X, Sirito M, Lu T. 1999. Biological function of the USF family of transcription factors. *Gene Ther Mol Biol* **3:** 447–453.
Schmid C, Bucher P. 2010. MER41 repeat sequences contain inducible STAT1 binding sites. *PLoS ONE* **5:** e11425. doi: 10.1371/journal.pone.0011425.
Sokalski KM, Li SKH, Welch I, Cadieux-Pitre H-AT, Gruca MR, DeKoter RP. 2011. Deletion of genes encoding PU.1 and Spi-B in B cells impairs differentiation and induces pre-B cell acute lymphoblastic leukemia. *Blood* **118:** 2801–2808.
Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. 2006. BioGRID: A general repository for interaction datasets. *Nucleic Acids Res* **34:** D535–D539.
Takahashi S, Onodera K, Motohashi H, Suwabe N, Hayashi N, Yanai N, Nabesima Y, Yamamoto M. 1997. Arrest in primitive erythroid cell development caused by promoter-specific disruption of the GATA-1 gene. *J Biol Chem* **272:** 12611–12615.
Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A. 2011. Determinants of nucleosome organization in primary human cells. *Nature* **474:** 516–520.
Wang C, Rauscher FJ, Cress WD, Chen J. 2007. Regulation of E2F1 function by the nuclear corepressor KAP1. *J Biol Chem* **282:** 29902–29909.
Wang C-C, Tsai M-F, Dai T-H, Hong T-M, Chan W-K, Chen JJW, Yang P-C. 2007. Synergistic activation of the tumor suppressor, HLJ1, by the transcription factors YY1 and activator protein 1. *Cancer Res* **67:** 4816–4826.
Wang Y-L, Faiola F, Xu M, Pan S, Martinez E. 2008. Human ATAC Is a GCN5/PCAF-containing acetylase complex with a novel NC2-like histone fold module that interacts with the TATA-binding protein. *J Biol Chem* **283:** 33808–33815.
Weiner A, Hughes A, Yassour M, Rando OJ, Friedman N. 2010. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Res* **20:** 90–100.
Weintraub H, Groudine M. 1976. Chromosomal subunits in active genes have an altered conformation. *Science* **193:** 848–856.
Whitfield TW, Wang J, Collins PJ, Partridge EC, Trinklein ND, Aldred SF, Myers RM, Weng Z. 2012. Functional analysis of transcription factor binding sites in human promoters. *Genome Biol* (in press).
Xu Z, Huang S, Chang L-S, Agulnick AD, Brandt SJ. 2003. Identification of a TAL1 target gene reveals a positive role for the LIM domain-binding protein Ldb1 in erythroid gene expression and differentiation. *Mol Cell Biol* **23:** 7585–7599.
Yoon H-G, Chan DW, Reynolds AB, Qin J, Wong J. 2003. N-CoR mediates DNA methylation-dependent repression through a methyl CpG binding protein Kaiso. *Mol Cell* **12:** 723–734.
Zhou Z, Li X, Deng C, Ney PA, Huang S, Bungert J. 2010. USF and NF-E2 cooperate to regulate the recruitment and activity of RNA polymerase II in the β-globin gene locus. *J Biol Chem* **285:** 15894–15905.

Received March 25, 2012; accepted in revised form June 7, 2012.