2 Chromatin patterns at transcription factor binding sites

Both chromatin accessibility and histone modifications have different patterns around different transcription factor binding sites

In Tier 1 and Tier 2 cell types, we identified a mean of 205,109 DHSs per cell type (at FDR 1%), encompassing an average of 1.0% of the genomic sequence in each cell type, and 3.9% in aggregate. On average, 98.5% of the occupancy sites of TFs mapped by ENCODE ChIP-seq (and, collectively, 94.4% of all 1.1 million TF ChIP-seq peaks in K562) lay within accessible chromatin defined by DNaseI hotspots [29]. However, a small number of factors, most prominently heterochromatin-bound repressive complexes (*e.g.*, the Kap1-SetDB1-Znf274 complex[31,32] encoded by the TRIM28, SETDB1 and ZNF274 genes), appear to occupy a significant fraction of nucleosomal sites.

TF binding sites provide a natural focus around which to explore chromatin properties. TFs are often multi-functional and can bind a variety of genomic loci with different combinations and patterns of chromatin marks and nucleosome organization. Hence, rather than averaging chromatin mark profiles across all binding sites of a TF, we developed a clustering procedure, termed the Clustered Aggregation Tool (CAGT), to identify subsets of binding sites sharing similar but distinct patterns of chromatin mark signal magnitude, shape, and hidden directionality [30]. For example, the average profile of the repressive histone mark, H3K27me3, over all 55,782 CTCF-binding sites in K562 shows poor signal enrichment (Figure 3A). However, after grouping profiles by signal magnitude, we found a subset of 9,840 (17.6%) CTCF-binding sites that exhibit significant flanking H3K27me3 signal. Shape and orientation analysis further revealed that the predominant signal profile for H3K27me3 around CTCF peak summits is asymmetric, consistent with a boundary role for some CTCF sites between active and polycomb-silenced domains. Further examples are provided in Supplementary Figures E5 and E6. For TAF1, predominantly found near TSSs, the asymmetric sites are orientated with the direction of transcription. However, for distal sites, such as those bound by GATA1 and CTCF, we also observed a high proportion of asymmetric histone patterns, although independent of motif directionality. In fact, all TF-binding datasets in all cell lines show predominantly asymmetric patterns (asymmetry ratio $>$0.6) for all chromatin marks but not DNaseI (Figure 3B). This suggests that most TF bound chromatin events correlate with structured, directional patterns of histone modifications, and that promoter directionality is not the only source of orientation at these sites.

We also examined nucleosome occupancy relative to the symmetry properties of chromatin marks around TF-binding sites. Around TSSs, there is usually strong asymmetric nucleosome occupancy, often accounting for the majority of the histone modification signal (for instance, see Supplementary Figure E4). However, away from TSSs, there is far less concordance. For example, CTCF-binding sites typically show arrays of well-positioned nucleosomes on either side of the peak summit (Supplementary Figure E1)63. Where the flanking chromatin mark signal is high, the signals are often asymmetric, indicating differential marking with histone modifications (Supplementary Figure E2 and E3). Thus, we confirm on a genome-wide scale that TFs can form barriers around which nucleosomes and histone modifications are arranged in a variety of configurations[63-66]. Further detail is explored in refs[25,26,30].

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 and 5b for YY1 in GM12878 cells and Fig. S10 for all datasets). 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 datasets. To account for the

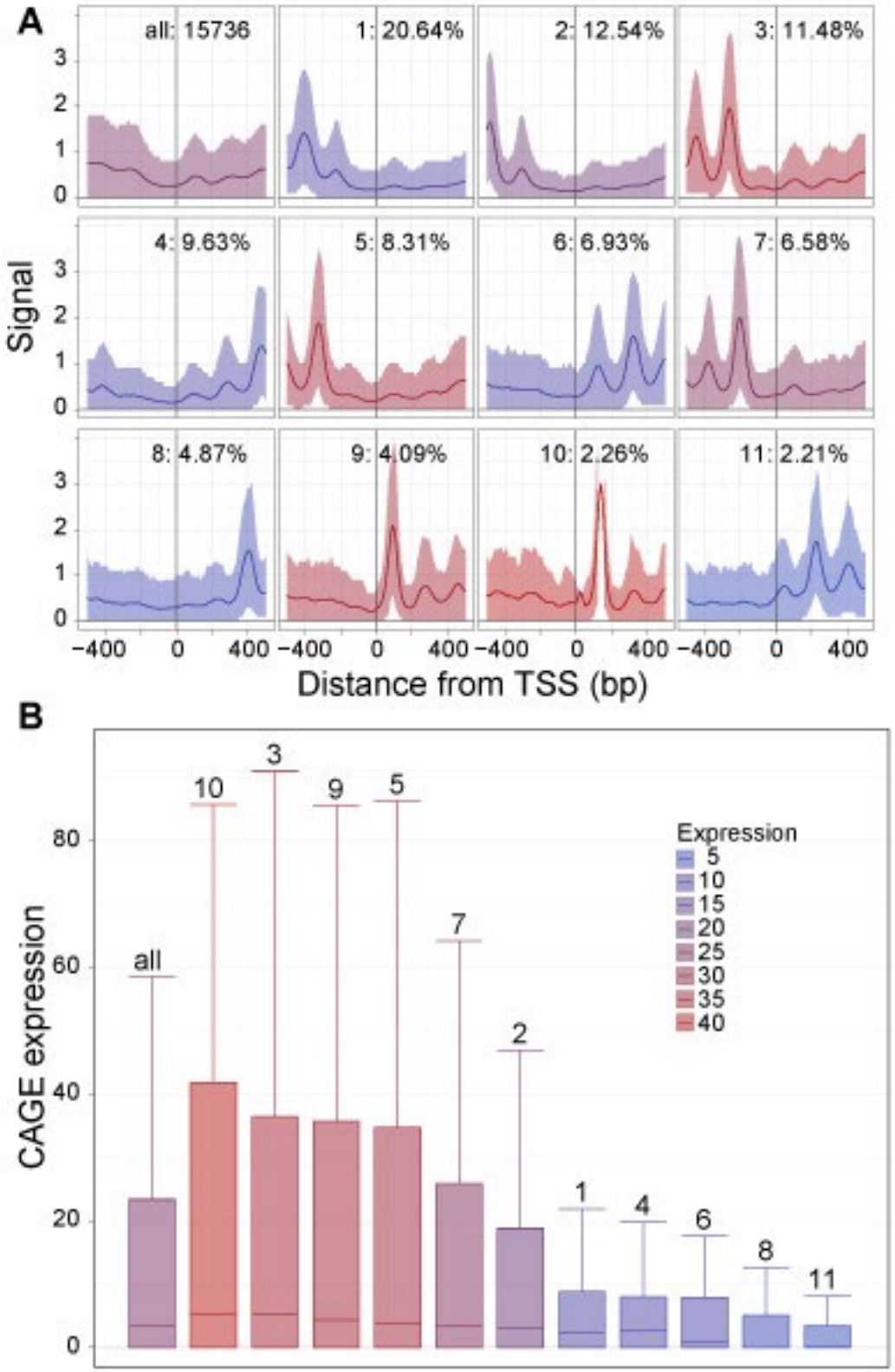


FIG 2

Figure 2 | (**a**) **Nucleosome positioning patterns around TSSs in K562. The** *first* **panel is a traditional aggregation plot of the nucleosome positioning signal in a window of size 1001 bp centered on each of 15,736 GENCODE TSSs. The bold line is the mean signal across all TSSs, while the shaded area around it corresponds to the 10th and 90th percentiles. The rest of the panels show the patterns uncovered by CAGT, ordered by the percentage of TSSs that follow each pattern. Patterns corresponding to <2% of TSSs are not shown. All TSSs are reoriented so that the direction of transcription is from** *left* **to** *right***. Plots are colored according to the third quartile of the expression of TSSs in the corresponding cluster, as measured by CAGE tags. (b) Box-plots of the expression of TSSs following each of the patterns shown in a.**

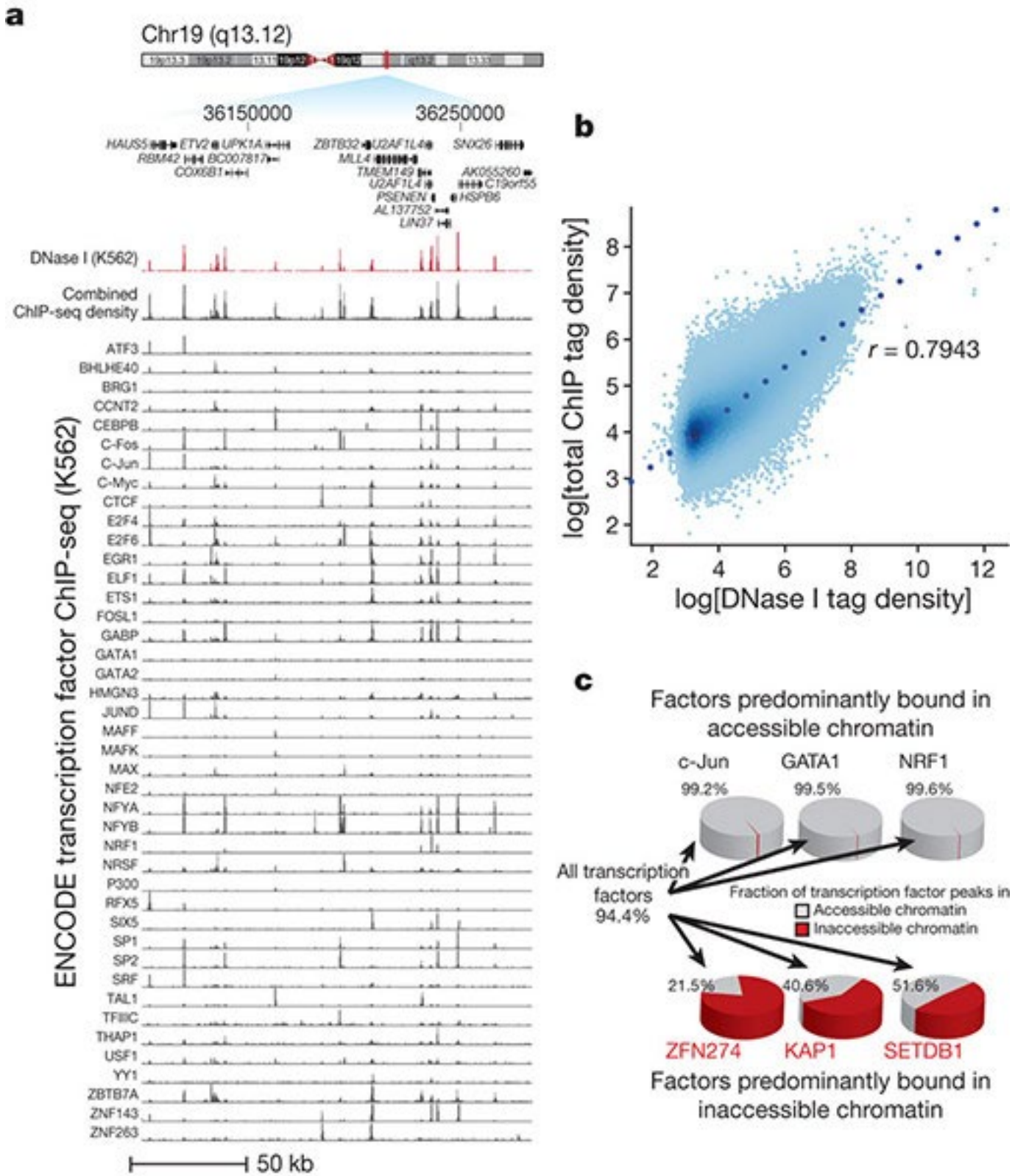


Figure 2 | Transcription factor drivers of chromatin accessibility. (a) DNase I tag density is shown in red for a 175-kb region of chromosome 19. Below: normalized ChIP-seq tag density for 45 ENCODE ChIP-seq experiments from K562 cells, with a cumulative sum of the individual tag density tracks shown immediately below the K562 DNase I data. (b) Genome-wide correlation ($r = 0.7943$) between ChIP-seq and DNase I tag densities ($\log_{10}$) in K562 cells. (c) Left: 94.4% of a combined 1,108,081 ChIP-seq peaks from all transcription factors assayed in K562 cells fall within accessible chromatin (grey areas of pie chart). Top: three examples of transcription factors localizing almost exclusively within accessible chromatin. Bottom: three transcription factors from the KRAB-associated complex localizing partially or predominantly within inaccessible chromatin.

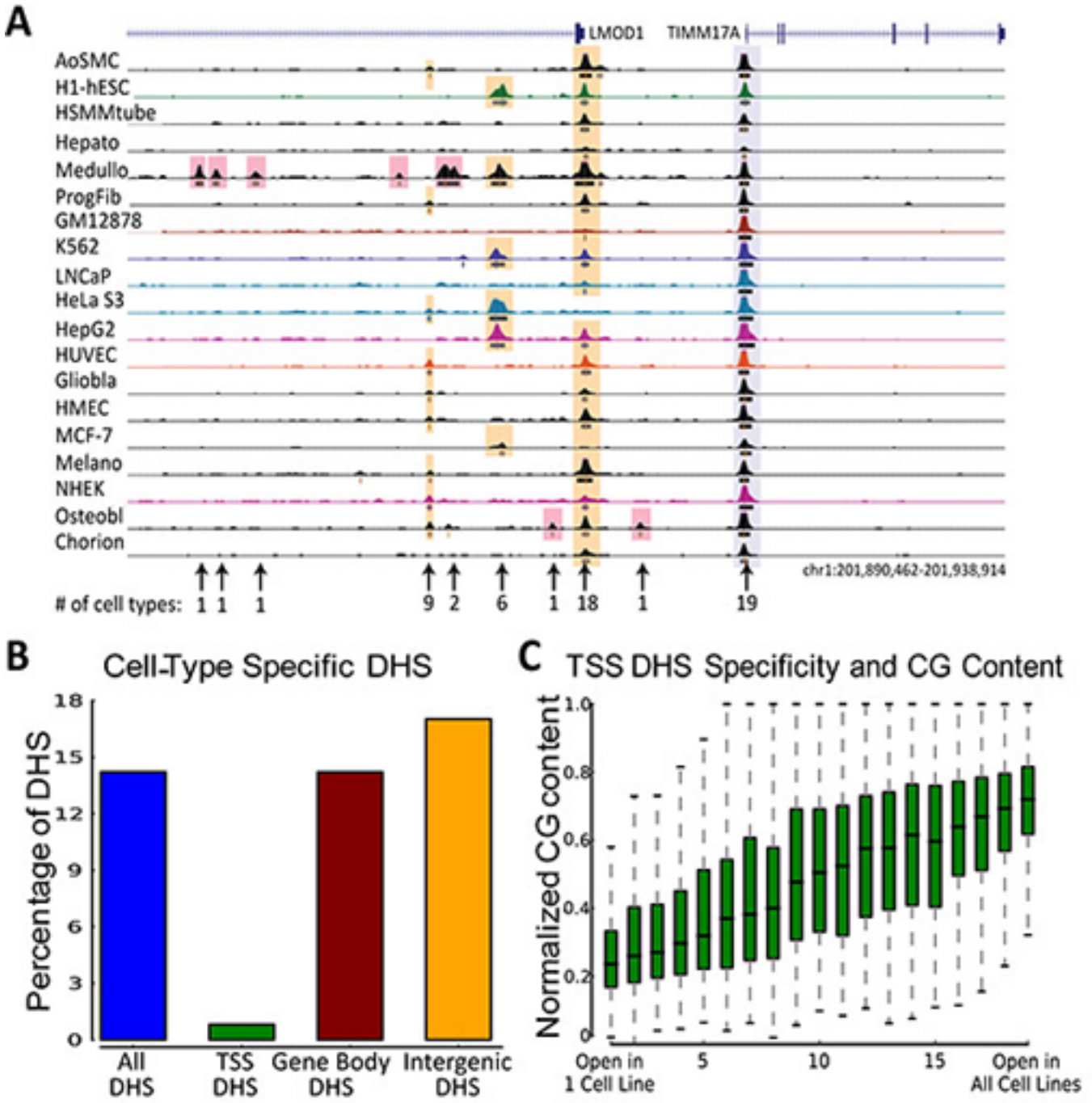


Figure 2 | **Cell-type specificity of hypersensitive regions.** (**a**) Example (Chr1: 201,890,462-201,938,914) showing cell-type-specific DHSs across two cell lines (pink boxes). Note that we called a DHS cell-type-specific if it did not overlap another DHS by more than half in any of the 18 other cell lines. (**b**) Bar graph showing the proportions of cell-type-specific DHSs across different genomic locations averaged across all cell lines. (**c**) TSSs were divided by the number of cell lines that they overlapped a region of open chromatin. For each set of TSSs, normalized CG content in the promoter regions (−900,100) of the TSSs are shown.

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 of the anchor. We further stratified peaks in each dataset as top, middle and bottom thirds according to 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,

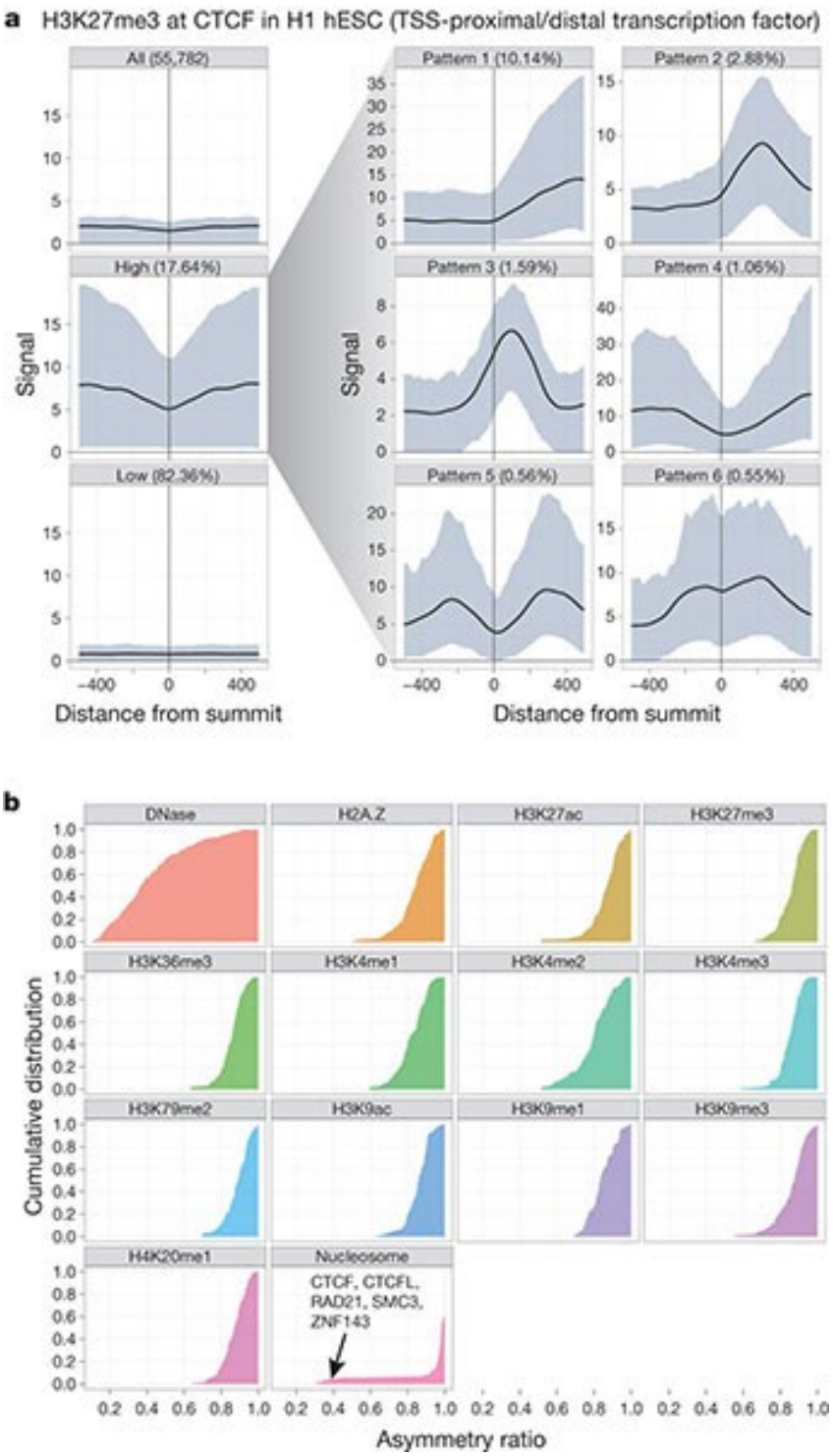


Figure 3 | **Patterns and asymmetry of chromatin modification at transcription-factor-binding sites.** (**a**) Results of clustered aggregation of H3K27me3 modification signal around CTCF-binding sites (a multifunctional protein involved with chromatin structure). The first three plots (left column) show the signal behaviour of the histone modification over all sites (top) and then split into the high and low signal components. The solid lines show the mean signal distribution by relative position with the blue shaded area delimiting the tenth and ninetieth percentile range. The high signal component is then decomposed further into six different shape classes on the right (see ref. 30 for details). The shape decomposition process is strand aware. (**b**) Summary of shape asymmetry for DNase I, nucleosome and histone modification signals by plotting an asymmetry ratio for each signal over all transcription-factor-binding sites. All histone modifications measured in this study show predominantly asymmetric patterns at transcription-factor-binding sites. An interactive version of this figure is available in the online version of the paper.

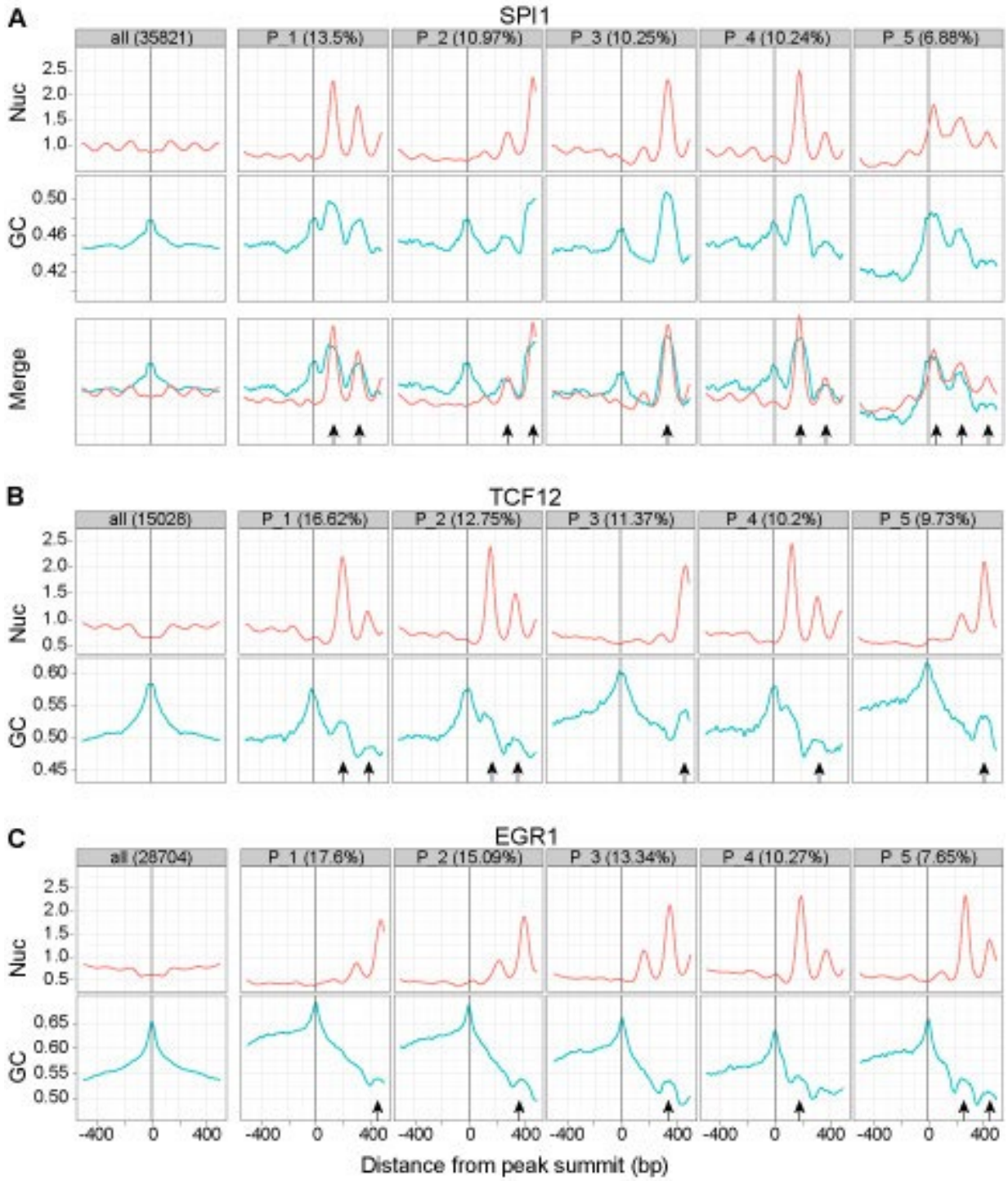


FIG 4

Figure 4 | **Examples of nucleosome positioning clusters around TFBSs and relationship to GC content.** For each TF, the *first* panel of the top row is a traditional aggregation plot, where the signal is averaged over all sites. The total number of sites is shown in the header. The remaining panels of the *top* row show the mean nucleosome positioning signal in the five largest clusters discovered by CAGT, with the fraction of peaks in each cluster shown in the header. Each panel in the *second* row shows the mean GC content of all sites used in the panel *above* it. If a site was "flipped"; during the last step of CAGT (see Fig. 1), then the corresponding GC signal was also flipped accordingly. GC content was computed using a sliding window of 21 bp. The small arrows indicate container sites. (**a**) SPI1 in GM12878; (**b**) TCF12 in GM21878; (**c**) EGR1 in K562.

Figure 5

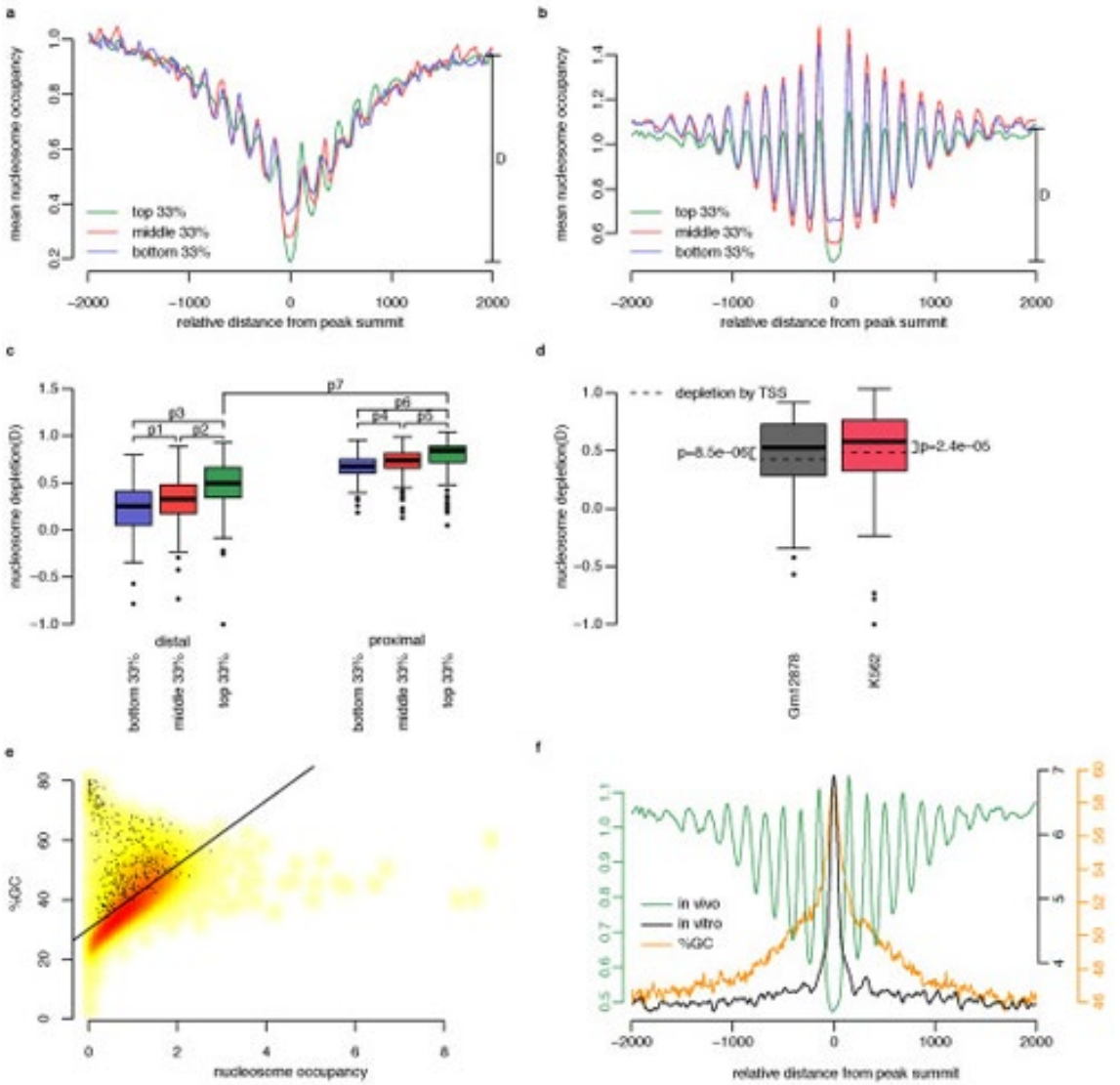


Figure 5 | **Chromatin structure and GC content around TF binding regions.** (**a**,**b**) Nucleosome occupancy profiles anchored on the summits of TSS-proximal (**a**) and TSS-distal (**b**) peaks of YY1 grouped by ChIP-seq signal strength: top (green), middle (red), and bottom (blue) third peaks in 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 $< 2.2 \times 10^{-16}$. (**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.

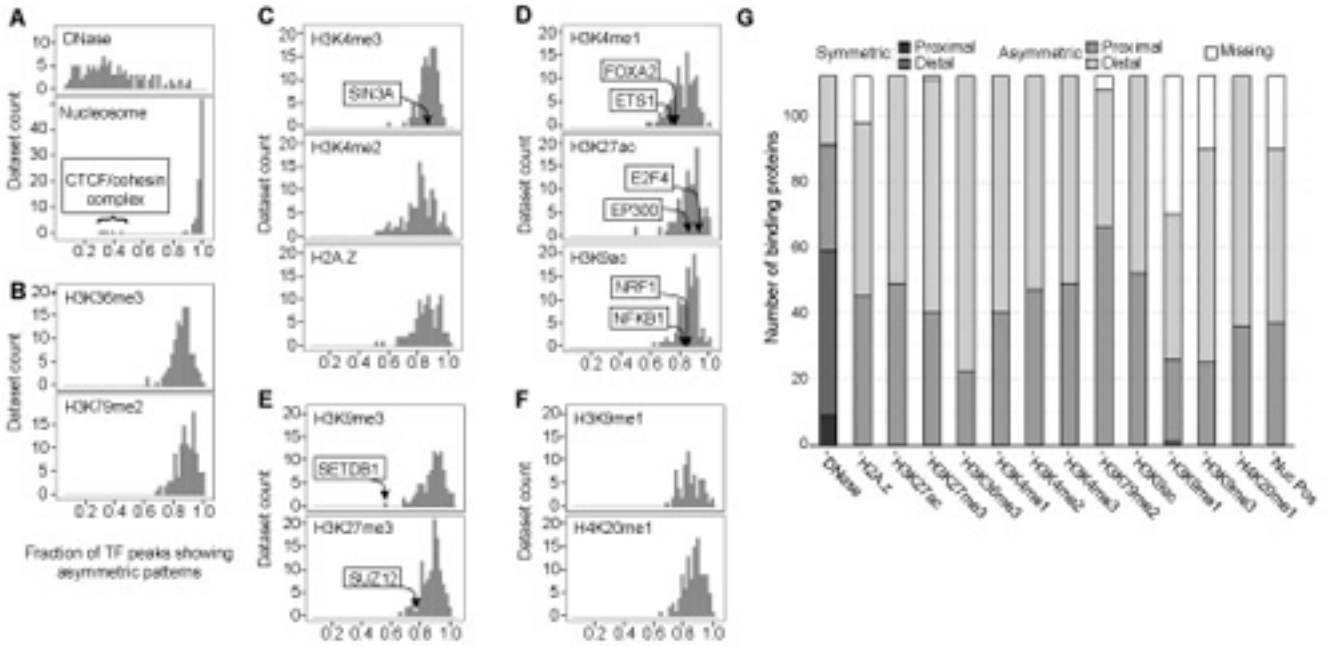


Figure 5 | **Widespread asymmetry of chromatin marks around TFBSs.** (**a-f**) Fraction of TF peaks with asymmetric patterns for each chromatin mark. For each combination of TF and mark, we computed the fraction of high signal binding sites in asymmetric CAGT clusters. Results were averaged over all available data sets for the same TF and mark in all cell lines. Some examples for factors that contribute to the specific data point are shown, with arrows pointing to the asymmetry fraction of the factor. For example, in ~85% of NRF1 binding sites with high H3K9ac signal, the shape of the modification is asymmetric around the binding site. (**a**) DNase and nucleosome positioning and their contrasting asymmetry frequency distributions. (**b**) Gene body marks. (**c**) Promoter-associated marks. (**d**) Enhancer-associated marks. (**e**) Repressive marks that exhibited moderate signal around binding sites. (**f**) Repressive marks that exhibited generally weak signal around binding sites. (**g**) For each combination of TF and mark, we computed the number of proximal and distal binding sites in symmetric and asymmetric CAGT clusters and identified which one of the four groups, symmetric proximal, symmetric distal, asymmetric proximal, and asymmetric distal, contained the largest number of binding sites. Results were averaged over all available data sets for the same TF and mark in all cell lines. The height of each bar shows the number of TFs for which the corresponding group was the most prevalent. The "Missing" part corresponds to the TFs that were not assayed for that mark.

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 (comparing Fig. 5a with Fig. 5b, similarly the proximal and distal panels in Fig. S10; note the difference in 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, and bottom third peaks, which correspond to the peaks with strongest, medium and weakest TF binding, tended to show greatest, medium, and weakest nucleosome positioning (Fig. 5ab and Fig. S10). Thus regions that are more strongly bound by TFs are flanked by better-positioned nucleosomes.

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

Figure 6

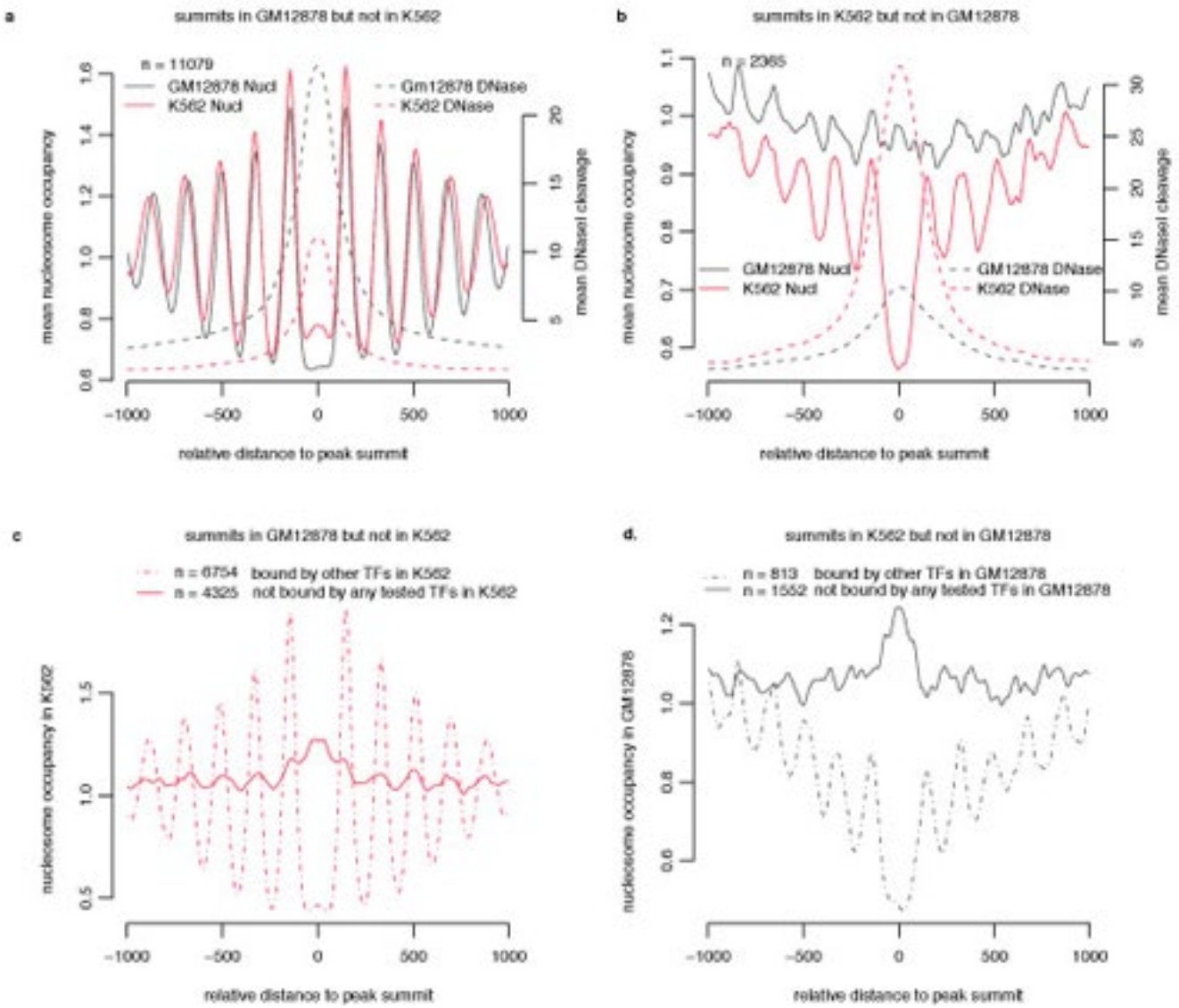


Figure 6 | Chromatin structure around YY1 ChIP-seq peaks occupied differentially between GM12878 and K562. (a) Nucleosome occupancy profiles (solid lines) and DNase I cleavage profiles (dashed lines) anchored on the summits of YY1 peaks in GM12878 but not in K562. Note the average nucleosome occupancy 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.

occupancy and GC content for randomly chosen 250-bp regions of the genome (Fig. 5e; r=0.62 and p-value<2.2e-16). 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

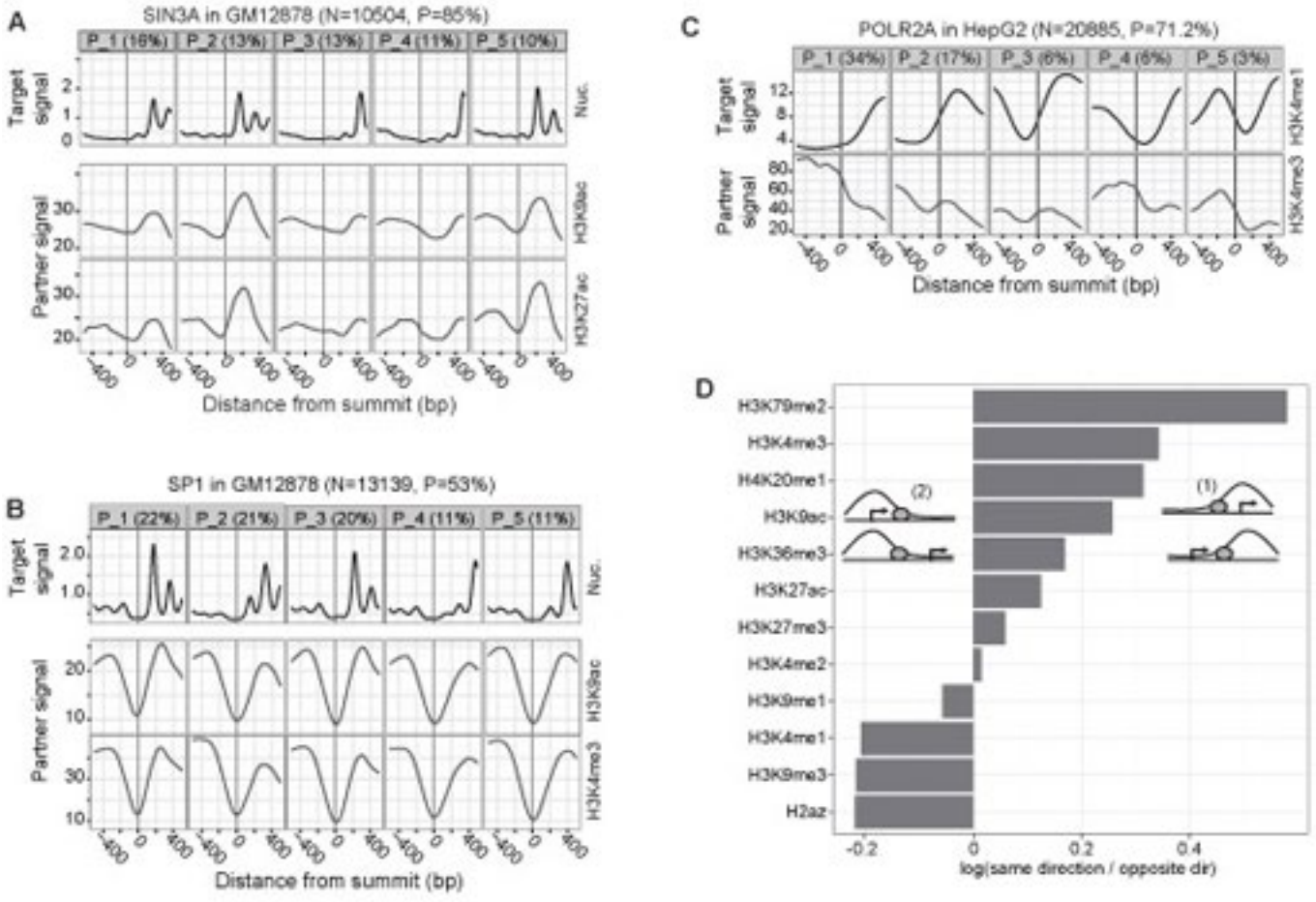


Figure 8 | (**a**,**b**) The *top* row shows the most prevalent nucleosome positioning clusters around SIN3A and SP1 sites, respectively, in GM12878. The remaining rows show the signal of histone modifications, averaged over all sites in the corresponding clusters. TSS-proximal TFs, such as SIN3A, exhibit correlated nucleosome positioning and histone modification patterns. Such correlations, however, are not evident for TFs that tend to bind more distally from TSSs (e.g., SP1). (**c**) Clusters of H3K4me1 signal around POLR2A sites in HepG2 and the corresponding H3K4me3 signal. There is a clear anticorrelation between the two histone marks. (**d**) For all CAGT runs around TFBSs, we considered all TSS-proximal sites that were assigned to asymmetric clusters, and counted how many times the direction of transcription of the TSS closest to a site agreed with (configuration (1)) or opposed (configuration (2)) the direction of the asymmetry pattern (from low to high signal) of the cluster to which the site was assigned. We are showing the log10-ratio of the two counts, aggregated over all CAGT runs for the same mark. Values >0 (corresponding to ratios >1) imply that the mark tends to increase in the direction of transcription, while values <0 imply that the mark tends to increase in the opposite direction.

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 and Fig. S13 for all TFs). Thus TFs or their co-factors (such as chromatin remodelers) prevent the formation of nucleosomes or evict nucleosome at these GC-rich locations of the genome.

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. 6ab for YY1 and 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 (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 (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.

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. 6cd and Fig. S16). The ChIP-seq data we have only cover up to 10% 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.

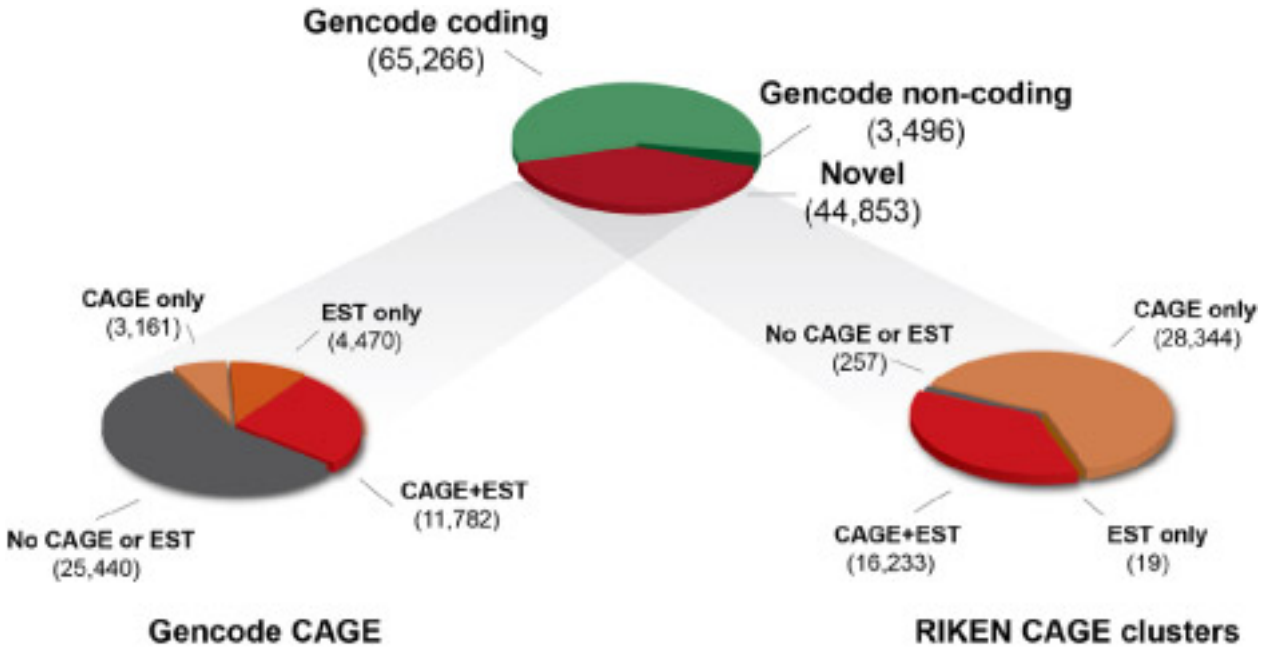
Nucleosome positioning around transcription start sites is highly heterogeneous

We first focused on nucleosome positioning signals around TSSs. Positioning around TSSs and promoters, and their correlation with transcription, has been well-studied previously (Fu *et al.* 2008; Jiang and Pugh 2009; Kaplan *et al.* 2009; Mavrich *et al.* 2008; Radman-Livaja and Rando 2010; Rando and Chang 2009; Segal *et al.* 2009; Shivaswamy *et al.* 2008; Schones *et al.* 2008; Valouev *et al.* 2011). The current consensus on promoter configuration involves a nucleosome-free region upstream of RNA polymerase II, which in turn is bound to the promoter upstream of the so-called $+1$ nucleosome. We used 15736 TSSs from the GENCODE v7 annotations (Harrow *et al.* 2012) as anchor points for CAGT analysis in K562 and GM12878, the two cell lines for which we had nucleosome positioning data. We excluded TSSs of bidirectional promoters to reduce confounding effects on the nucleosome positioning signal (see Methods). Because the results from both cell lines were highly similar, we limit our discussion to K562.

CAGT analysis revealed 17 clusters of distinct nucleosome positioning patterns. Eleven of these clusters contained more than 2% of the TSSs each and comprised a total of 89.56% of the TSSs studied (Fig. 2A and Fig. S1). Broadly, the clusters fall into two categories: those in which there is strong positioning upstream of the TSS, and those that have strong positioning downstream. Surprisingly, no cluster had equally strong positioning on both sides of the TSSs, suggesting that the canonical pattern of a modest but detectable positioning signal emanating bidirectionally from the promoter is an averaging artifact of standard APs (Fig. 2A, first panel).

Highly diverse nucleosome positioning around TFBSs

The richness of the ENCODE ChIP-seq data provides an unprecedented opportunity to understand the relationship between transcription factor binding sites and nucleosome positioning. We extracted 1001 bases of

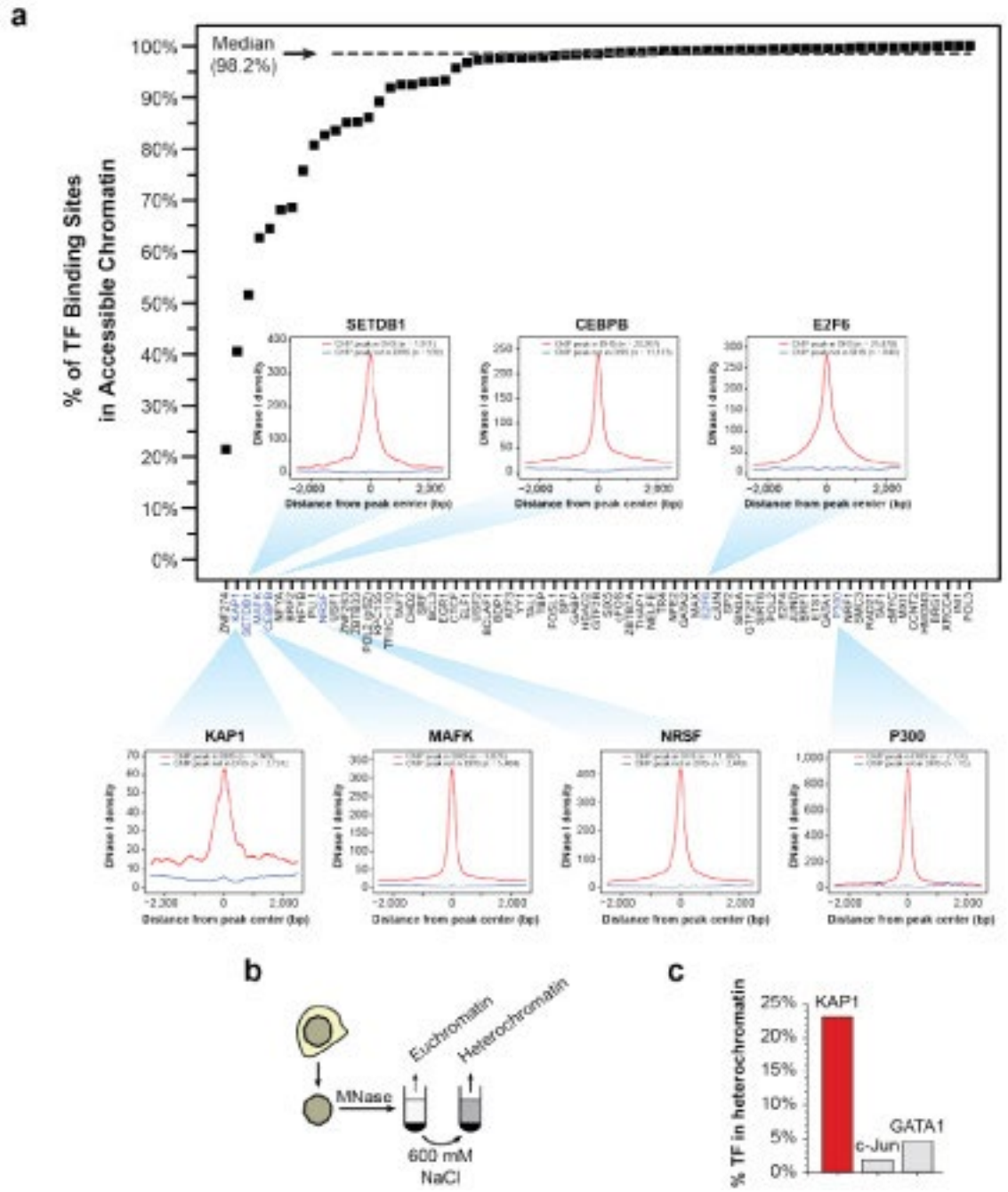


Supplementary Figure 6 | **Quantifying transcription factor impact on chromatin accessibility.** Quantifying the impact of transcription factors on chromatin accessibility. (**a**) As in Fig. 2a, DNaseI tag density is shown in red, followed by normalized ChIP-seq tag density for each of 42 ENCODE ChIP-seq experiments from K562 cells, with a cumulative sum of the individual tag density tracks shown immediately below the K562 DNaseI data; this plot shows a 35-kb region encompassing the beta-globin LCR on Chr11. (**b**) Additive correlation (*y*-axis) of ChIP-seq with DNaseI across Chr19 with increasing numbers of TFs. TFs are ordered alphabetically (*x*-axis). Correlation values for individual factors are shown in red. (**c**) Relative chromatin accessibility (*x*-axis) measured as the mean intensity of DHSs containing the indicated motif (*y*-axis), divided by the mean intensity of all DHSs (using 84 UW DNaseI datasets). Green density plots indicate the distribution of measurements obtained individually across all cell types; values >1 indicate presence of the motif has an average positive effect on chromatin accessibility.

the nucleosome positioning data around the summit coordinate of each peak for all transcription factors (and other DNA-binding proteins, such as RNA Polymerase II, RAD21, etc) that had been assayed in GM12878 and K562 (see Supp. section S.3 and Fig. S2 for a discussion on using peak summits instead of motif locations as anchor points). On each of these 148 data sets, CAGT grouped the nucleosome signal for each binding protein into a small number of shapes. The vast majority of shapes were clearly asymmetric, indicating that around TFBSs nucleosome positioning generally exhibits polarity. The only notable exceptions were the proteins of the CTCF/cohesin complex (RAD21, SMC3, and CTCF) as well as the zinc-finger containing protein, ZNF143, for which 40% to 80% of the fraction of binding sites from these TF ChIP-seq datasets showed roughly symmetrically positioned nucleosomes (Fig. S3). However, even these factors had some sites with asymmetric patterns of positioning. The majority of other factors had very few symmetric positioning patterns.

Positioned nucleosomes around TFBSs occupy container sites

We have previously shown that in vitro reconstituted nucleosomes exhibit particularly strong positioning when they occupy container sites (Valouev *et al.* 2011). Container sites are characterized by a GC-rich core of about 150 bases, which serves to attract a nucleosome, and AT-rich flanks, which repel the nucleosome and therefore lock it into a statistically preferred position, centered on the GC-rich core (Johnson *et al.* 2006; Peckham *et al.* 2007; Segal *et al.* 2006; Tillo *et al.* 2010; Tsankov *et al.* 2011). To investigate whether container sites are present in the vicinity of TFBSs, we investigated the relationship between the cluster-specific shapes of nucleosome positioning and the underlying sequences' base composition.



Supplementary Figure 7 | **Transcription factor occupancies within accessible chromatin.** The occupancies of different transcription factors within accessible chromatin. (**a**) The percentage of transcription factor binding sites within accessible chromatin was calculated for each factor. Accessible chromatin was identified using unthresholded hotspot calls on K562 DNaseI deep-seq data. Transcription factor binding sites were identified in K562 cells using ChIP-seq. Inserts show the aggregate DNaseI density profile ($\pm$2.5 kb of ChIP-seq peak) at sites for six different transcription factors that are within (red) and outside (blue) of accessible chromatin. See Supplementary Methods, section 2.3, below. (**b**) Biochemical isolation of dense heterochromatin. (**c**) Proportion of chromatin-bound protein contained within heterochromatin was measured using targeted mass spectrometry for KAP1 (also called TRIM28), c-Jun and GATA1. Note that nearly 25% of nuclear KAP1 localises to highly compacted heterochromatin, vs. $<$5% for c-Jun and GATA1.

Comparison of the nucleosome positioning signal in the 1 kb around TFBSs with GC content revealed that container sites are a pervasive feature in the vicinity of TFBSs (Fig. 4 and Fig. S5). The first nucleosome immediately flanking the TFBS often occupies a container site, as evidenced by low GC content flanking a

high-GC-content, core-sized (150 bp) region on which the nucleosome peak is centered. Transcription factors that tend to occupy regions with low GC content consistently show the most dramatic correlation between their neighboring nucleosomes and GC content, and the strongest container site characteristics of the ~250 base pairs of the region occupied by the nucleosome core plus the two flanking linkers (Fig. 4A). Transcription factors that occupy high-GC sites show a less pronounced effect, but small local maxima in GC content precisely coinciding with the summit of the nucleosome peak are still evident (Fig. 4B and 4C). This observation is clearer for transcription factors for which many peaks were called as the plots become less noisy with increasing numbers of sites that contribute to a cluster (Fig. S5).

For each of 11 chromatin modification marks, as well as for H2A.Z, DNase, and nucleosome positioning, we calculated the fraction of binding sites of each factor that were assigned to CAGT clusters exhibiting asymmetry. We then made a histogram of the asymmetry fractions over all factors for each mark (Fig. 5A-F). The DNase hypersensitivity assay serves as a control because DNase cuts only at open chromatin right next to the bound factor, and consequently many factors exhibit predominantly symmetric DNase signal around their binding sites. By contrast, the distribution of asymmetry fractions for nucleosome positioning is strikingly different, with >90% of factors exhibiting pronounced asymmetry of nucleosome positioning around >90% of their binding sites (Fig. 5A). The only notable exceptions are the members of the CTCF/cohesin complex, which show predominantly, though not exclusively, symmetric positioning of nucleosomes around their binding sites (Fig. S3). The chromatin marks and H2A.Z also have mostly asymmetric signal around TFBSs. Marks that associate with gene bodies (Fig. 5B), promoter-associated marks (Fig. 5C), enhancer-enriched marks (Fig. 5D), and repressive marks (Fig. 5E,F) all have highly similar distributions, with a mean of 80%-90% of asymmetric sites.

Finally, we used a complementary approach to summarize the asymmetry correlation of different chromatin marks around all TSS-proximal TFBSs. For all CAGT runs around TFBSs, we considered all TSS-proximal binding sites that were assigned to asymmetric clusters of each chromatin mark. At each TFBS, the direction of asymmetry (from low to high signal) of a particular mark can be in the same (configuration 1) or in the opposite direction (configuration 2) as the direction of transcription of the nearest TSS. For each chromatin mark, we computed the ratio of TFBSs in configuration 1 to those in configuration 2. 8 of the 12 chromatin marks showed transcription-oriented asymmetry, with H3K79me2 and H3K4me3 having the strongest positive bias (Fig. 8D). H3K4me1, H3K9me3, H3K9me1 and H2A.Z patterns were anti-correlated with the direction of transcription. Hence, H3K4me1 and H3K4me3 were once again found to be anti-correlated with each other, as were H3K27ac and H3K4me1. Interestingly, the different types of repressive marks, H3K27me3 and H3K9me3 were also found to show anti-correlated behavior around TSSs.

Binding of transcription factors to regulatory DNA regions in place of canonical nucleosomes triggers chromatin remodelling, resulting in nuclease hypersensitivity[3]. Within DNaseI hypersensitive sites (DHSs), DNaseI cleavage is not uniform; rather, punctuated binding by sequence-specific regulatory factors occludes bound DNA from cleavage, leaving footprints that demarcate transcription factor occupancy at nucleotide resolution[1,4] (Fig. 1a). DNaseI footprinting has been applied widely to study the dynamics of transcription factor occupancy and cooperativity within regulatory DNA regions of individual genes[5], and to identify cell- and lineage-selective transcriptional regulators[6].

99.8% of DHSs with >250 mapped DNaseI cleavages contained at least one footprint, indicating that DHSs are not simply open or nucleosome-free chromatin features, but are constitutively populated with DNaseI footprints.

A large proportion of TSSs are found in regions of accessible chromatin.

To understand how regions of open chromatin vary between cell types, we inspected the degree to which DHSs were shared in the 19 cell types. DHSs were classified as being specific to a cell line if it was only present in a single cell type or overlapped less than 50% of its length with a DHS from any of the other 18 cell types (Figure

2A). Across all DHSs ~14% were specific to a single cell line (Figure 2B). Intergenic DHS showed the highest percentage of being cell-type specific (~17%). Conversely, TSS DHS were largely not cell-type specific with less than 1% being open specifically in a single cell type. Despite the broad panel of cell lines that vary in expression, the chromatin state at the TSS of these genes was open and largely invariant across multiple cell lines. This is in agreement with a recent study analyzing a subset of the cell types used here (Song *et al.* 2011).

DNaseI hypersensitive sites result from cooperative binding of transcriptional factors in place of a canonical nucleosome[1,2]. To quantify the relationship between chromatin accessibility and the occupancy of regulatory factors, we compared sequencing-depth-normalized DNaseI sensitivity in the ENCODE common cell line K562 to normalized ChIP-seq signals from all 42 transcription factors mapped by ENCODE ChIP-seq14 in this cell type (Fig. 2). Simple summation of the ChIP-seq signals markedly parallels quantitative DNaseI sensitivity at individual DHSs (Fig. 2a) and across the genome ($r = 0.79$, Fig. 2b). For example, the b-globin locus control region contains a major enhancer element at hypersensitive site 2 (HS2), which appears to be occupied by dozens of transcription factors (Supplementary Fig. 6a). Such highly overlapping binding patterns have been interpreted to signify weak interactions with lower-affinity recognition sequences potentiated by an accessible DNA template[15]. However, HS2 is a compact element with a functional core spanning ~110 bp that contains 5-8 sites of transcription factor-DNA interaction *in vivo* depending on the cell type[16-18]. The fact that the cumulative ChIP-seq signal closely parallels the degree of nuclease sensitivity at HS2 and elsewhere is thus most readily explained by interactions between DNA-bound factors and other interacting factors that collectively potentiate the accessible chromatin state (Supplementary Fig. 6b). Given the relatively limited number of factors studied, it may seem surprising that such a close correlation should be evident. However, most of the factors selected for ENCODE ChIP-seq studies have well-described or even fundamental roles in transcriptional regulation, and many were identified originally based on their high affinity for DNA. Alternatively, as originally proposed in ref. 19, a limited number of factors may be involved in establishment and maintenance of chromatin remodelling, whereas others may interact nonspecifically with the remodelled state.

Overall, 94.4% of a combined 1,108,081 ChIP-seq peaks from all ENCODE transcription factors fall within accessible chromatin (Fig. 2c and Supplementary Fig. 7a), with the median factor having 98.2% of its binding sites localized therein. Notably, a small number of factors diverged from this paradigm, including known chromatin repressors, such as the KRAB-associated factors KAP1 (also called TRIM28), SETDB1 and ZNF274 (refs 20, 21) (Fig. 2c). We hypothesized that a proportion of the occupancy sites of these factors represented binding within compacted heterochromatin. To test this, we developed targeted mass spectrometry assays[22] for KAP1 and three factors localizing almost exclusively within accessible chromatin (GATA1, c-Jun, NRF1), and quantified their abundance in biochemically defined heterochromatin[23] against a total chromatin fraction (Supplementary Fig. 7b). This analysis confirmed that factors such as KAP1 show a significant level of heterochromatin occupancy (Supplementary Fig. 7c).

BARs, PRMs and DRMs have strong open chromatin signals (Figures 4A and 4B), consistent with their expected roles as active gene regulatory elements[21,23,42]. PRMs have stronger H3K4me3 signals and DRMs have stronger H3K4me1 (Figure 4C and 4E), which are expected since H3K4me3 is a signature of active promoters while H3K4me1 is an indicator of enhancers [43]. Both PRMs and DRMs have enriched H3K4me2 signals over the whole genome, which is also consistent with previous observations[40]. PRMs have stronger H3K36me3 and H3K79me2 (Additional file 2, Figure S8) signals than DRMs. These histone marks are found in transcribed regions[44-46], and are thus good features for distinguishing between regulatory elements that are close to and those that are far away from transcribed genes.

Both BIRs and LOT regions are depleted of most of the histone modifications relative to the whole genome. BIRs are slightly more enriched for open chromatin and repressive (H3K9me3 and H3K27me3) signals, which suggest that BIRs are more accessible to TRFs but transcriptional activities are repressed, while LOT regions in general have low DNA accessibility.

We observed surprisingly heterogeneous base-to-base variation in DNaseI cleavage rates within the footprinted recognition sequences of different regulatory factors. And yet, the per site cleavage profiles for individual factors were highly stereotyped, with nearly identical local cleavage patterns at thousands of genomic locations (Supplementary Fig. 7). This raised the possibility that DNaseI cleavage patterns may provide information concerning the morphology of the DNA-protein interface. We obtained the available DNA-protein co-crystal structures for human transcription factors, and mapped aggregate DNaseI cleavage patterns at individual nucleotide positions onto the DNA backbone of the co-crystal model. Figure 3a and Supplementary Fig. 8a show two examples: USF1[17] and SRF[18]. For both factors, DNaseI cleavage patterns clearly parallel the topology of the protein-DNA interface, including a marked depression in DNaseI cleavage at nucleotides involved in protein-DNA contact, and increased cleavage at exposed nucleotides such as those within the central pocket of the leucine zipper. These data show that nucleotide-level aggregate DNaseI cleavage patterns reflect fundamental features of the protein-DNA interaction interface at unprecedented resolution.

Many transcriptional regulators are posited to interact indirectly with the DNA sequence of some target sites through mechanisms such as tethering[25]. Approaches such as ChIP-seq detect chromatin occupancy, but cannot by themselves distinguish sites of direct DNA binding from non-canonical indirect binding. We therefore asked whether DNaseI footprint data could illuminate ChIP-seq-derived occupancy profiles by differentiating directly bound factors from indirect binding events. We first partitioned ChIP-seq peaks from each of 38 ENCODE transcription factors[26] mapped in K562 cells into three categories of predicted sites: ChIP-seq peaks containing a compatible footprinted motif (directly bound sites); ChIP-seq peaks lacking a compatible motif or footprint (indirectly bound sites); and ChIP-seq peaks overlying a compatible motif lacking a footprint (indeterminate sites). Predicted indirect sites showed significantly reduced ChIP-seq signal compared with predicted directly bound sites (Supplementary Fig. 10), consistent with lack of direct crosslinking to DNA (and therefore reduced ChIP efficiency). Indeterminate sites exhibited low ChIP-seq signal and were therefore excluded from further analysis (Supplementary Fig. 10).

The fraction of ChIP-seq peaks predicted to represent direct versus indirect binding varied widely between different factors, ranging from nearly complete direct sequence-specific binding (for example, CTCF), to nearly complete indirect binding (for example, TBP; Supplementary Fig. 11). In many cases factors that preferentially engage in direct DNA binding at distal sites show predominantly indirect occupancy in promoter regions and vice versa (Supplementary Fig. 12a, b).

Next, we analysed the frequency with which indirectly bound sites of one transcription factor coincided with directly bound sites of a second factor, indicative of protein-protein interactions (for example, tethering). This analysis recovered many known protein-protein interactions, such as CTCF-YY1 and TAL1-GATA1 [27], as well as many novel associations (Fig. 5). We observed enrichment for NFE2 indirect interactions at promoter-bound USF2 sites, compatible with their known interaction[28]. At distal sites, we observed the opposite, with NFE2 predominantly directly bound accompanied by USF2 indirect peaks (Supplementary Fig. 12a, b), indicating the possibility of a reciprocal or looping mechanism. Notably, directly bound promoter-predominant transcription factors were enriched for co-localization with indirect peaks compared to distal regions (Supplementary Fig. 13a, b). These results suggest that combining DNaseI footprinting with ChIP-seq has the potential to expose a previously unappreciated landscape of complex transcription factor occupancy modes.