S.1 Detailed description of the CAGT procedure

The $k$-medians algorithm used by CAGT is an iterative clustering algorithm. Initially, $K$ of the input signal vectors are randomly chosen as cluster centroids. Each iteration of $k$-medians consists of an assignment step followed by an update step. In the assignment step, given a set of cluster centroids each signal profile $x = [x_1,...,x_L]$ is assigned to the cluster whose centroid is closest to $x$. Then, in the update step, the centroids for each cluster are recomputed as the medians of the signal vectors assigned to the cluster. Using the medians instead of the means (as in the $k$-means algorithm) produces centroids that are more robust to outliers. The assignment and update steps are repeated until an assignment step doesn’t change the assignments of any signal vectors from the previous iteration, or until a maximum number of iterations is reached. The number of clusters $K$ is a user-defined parameter. We empirically found that 40 clusters are sufficient to capture the signal variability in practically all cases, therefore in all our experiments we set $K = 40$. The results of $k$-medians may partially depend on the random cluster initialization. Hence, CAGT runs $k$-medians with several random initializations and finalizes clusters based on the most optimal run (measured by the sum of distances of all input signal vectors to their assigned cluster).

An important limitation of $k$-medians is that the number of clusters $K$ must be given by the user. If $K$ is not large enough, the resulting clusters will be unable to capture all the distinct patterns in the data and may have high variance indicating the presence of distinct subpopulations within the same cluster. On the other hand, setting $K$ to a sub-optimally large value can lead to a redundancy in the clusters i.e. several clusters with similar patterns. In CAGT, we start with a large number of clusters that is practically guaranteed to capture the majority of informative patterns with sufficient support albeit with some level of redundancy. Then, we apply hierarchical agglomerative clustering to merge clusters that are similar to each other. Hierarchical agglomerative clustering is an iterative procedure. At each iteration, the two clusters with the smallest distance between their centroids, i.e. the two “most similar” clusters, are merged, reducing the number of clusters by 1. This process is repeated with the new set of clusters until either there is only 1 cluster left, containing all initial signal vectors, or the distance between the two closest clusters is above a pre-specified threshold. This distance threshold, which is equivalent to one minus the correlation coefficient of the two closest clusters, was tailored for each type of mark (0.4 for nucleosome positioning, 0.2 for DNase, and 0.25 for histone modifications).

When we used CAGT to analyze chromatin marks around TF binding sites, we observed that $k$-medians would often find pairs of clusters with asymmetric patterns that were near identical mirror images of each other (Fig. 1). Hence, we optionally allowed the hierarchical clustering phase to merge mirror images while keeping track of flipped profiles. At each step of
hierarchical clustering, we consider all pairs of the original (or unflipped) centroids, as well as all pairs consisting of one unflipped and one flipped centroid and select the pair with the smallest distance to merge, flipping all signal vectors in each cluster accordingly. The final directionality of a cluster resulting from merging mirrored asymmetric patterns is arbitrary.

S.2 Reproducibility and robustness of CAGT

CAGT relies on k-medians, which is initialized with random patterns from the input data set. A possible concern with such randomized algorithms is that the results might not be reproducible, that is, two different runs on the same data may give different results. However, the hierarchical agglomerative clustering step greatly reduces the discrepancies caused by the random initializations, by identifying and merging similar clusters. The robustness of CAGT is also evidenced in the meta-clustering results (Fig. 7), which demonstrate that similar patterns are shared between different binding proteins, cell lines, and even chromatin modifications. It is extremely unlikely to discover similar patterns in so many diverse data sets, unless these patterns are (a) truly present in the data, i.e. not merely noise and (b) independent from the random initializations for k-medians.

To further test robustness of CAGT, we performed 100 independent runs on the same data set of H3K27ac signal around CTCF sites in K562 that was used for Fig. 1. All runs used the same parameters as described in the Methods, but different random initializations of k-medians. We then used three measures to compare the results of different runs and to assess reproducibility:

1. The number of clusters returned by hierarchical clustering. Large deviations in the number of clusters would suggest non-reproducible results.
2. The mean Adjusted Rand Index (ARI) between pairs of different runs. The ARI is a measure of similarity between two different sets of clusters (Hubert and Arabie 1985). The maximum value the ARI can take is 1 and is achieved when the two sets of clusters are identical. The expected value of the ARI, when data points are distributed at random in the clusters, is 0. Values of ARI higher than 0 and closer to 1 indicate that the two sets of clusters compared are more similar than expected by chance.
3. The correlation between the shapes of the two largest clusters between different runs. A large correlation suggests that the biggest cluster returned is independent of the random initialization of k-medians.

The results are summarized in the first row of Supplementary Table 2. CAGT returned 10 clusters in almost all runs, with a mean ARI between clustering runs of 0.4736, much larger than the expected value for irreproducible clustering runs (ARI=0). The biggest cluster identified was almost identical between runs (correlation coefficient of 0.91). Fig. S15A shows the results of a representative run.

To evaluate CAGT’s robustness to noise, we generated two permutations of the H3K27ac signal around CTCF sites by permuting either the signal profile at each CTCF site across all 1001 bp, or the signal values for each bp across all CTCF sites. The resulting profiles are by definition random, so any reproducible cluster identified in them would be a clustering artifact. We ran CAGT on the permuted datasets with 100 different initializations (second and third rows of Supplementary Table 2 and Fig. S15B-C). The clustering results returned by CAGT showed very poor reproducibility, as suggested by the small ARI (close to 0), and the average correlation between the largest clusters of different runs was very poor (-0.0069 and -0.0073 respectively for the two types of permuted datasets). The number of clusters returned was also
large, which is expected for randomized profiles which cannot be summarized by a small number of clusters.

We also generated four other control datasets based on random sampling of genomic locations (last four rows of Supplementary Table 2 and Fig. S15D – E):

1. Dataset “jitter” was created by selecting a random location within a window +/-5000bp around each real CTCF peak.
2. Dataset “matched” was generated by picking a set of genomic locations with the same distribution of distances from the closest TSS as the real CTCF peaks.
3. Dataset “control motif” contains all significant hits of shuffled versions of the true CTCF motif (JASPAR (Bryne et al. 2008) motif MA0139.1). Control motifs were obtained from http://www.broadinstitute.org/~pouyak/encode-motif-disc/ (Kheradpour et al. 2012)
4. Finally, dataset “random” represents random genomic locations.

Not surprisingly, in all four control datasets, the majority of sites had low H3K27ac signal. These low signal profiles are filtered out by CAGT before shape analysis. However, among the high-signal sites, CAGT consistently and reproducibly discovered approximately 10 clusters that looked very similar to the clusters defined by the original CTCF peaks. As demonstrated by our meta-clustering analyses, the H3K27ac clusters around CTCF sites are not specific to CTCF. Contrary to the permuted datasets, the four control datasets contain a few sites that are highly enriched for authentic H3K27ac signal, i.e. these location lie in potential functional regions. By construction, a reasonable number of the sites in the “jitter” and “matched” datasets are expected to be in promoters or other bound regions. In the “control motif” dataset, some of the shuffled versions of the motif happen to lie in the immediate vicinity of other TF motifs. Indeed, in all four randomized datasets, the high-signal sites had significantly higher DNase signal than the low-signal sites (Wilcoxon test p-value ~ 0 in all cases), suggesting that several of these high-signal sites in each of the control datasets are potentially bound by various TFs. Thus, consistent with our global meta-shape analysis, CAGT identifies meta-shape patterns (or shifted versions of these) at random genomic locations that fall within functional regions.

S.3 Effect of peak calling on the patterns discovered by CAGT

In all CAGT runs around transcription factor binding sites (TFBSs), we used the peak calling summits (i.e. the highest-signal location of each peak) as the anchor points around which the signal was extracted. Due to the imperfect resolution of ChIP-seq, the peak calling summit does not always correspond to the exact binding location of the protein. Thus, a reasonable concern is that some of the observed heterogeneity in the CAGT clusters could be attributed to incorrect anchoring, since this would lead to signal patterns that are “shifted” versions of each other (eg. Fig 3B, clusters 8 and 9).

To examine the effects of imperfect anchoring, we selected four TFs with varying sequence specificity: CTCF, REST, E2F4 and E2F6. We used FIMO (Grant et al. 2011) to obtain hits of the corresponding TRANSFAC (Matys et al. 2006) and JASPAR (Bryne et al. 2008) motifs in the ChIP-seq peak regions. For each peak, we then used as anchoring point the middle of the motif hit closest to the summit. For the sequence-specific factors (CTCF, REST) the distance between the peak summit and the nearest motif hit was on average small (<50 bp). The E2F motif is rather degenerate, and hence a large fraction (> 50%) of E2F4 and E2F6 peaks did not have any significant motif hits (Fig. S2A and Supplementary Table 3). In general, we and others have found that a significant number of high confidence, reproducible peaks of TFs do not contain their canonical motifs. These peaks may represent indirect binding events or could
contain non-canonical motifs. All things considered, the peak summit is the best proxy for the true binding location.

For each dataset, we ran CAGT separately on the set of peaks without motif hits (anchored as before on the peak summit) and on the set of peaks with a motif hit (anchored on the hit) (Fig. S2B-J). In all cases, similar clusters were discovered in the peaks with and without the motif. Clusters that were “shifted” versions of each other were still present, even when peaks were anchored on the motif hit, and even for factors with highly non-degenerate motifs (CTCF, REST). Therefore, although we cannot exclude the possibility that some of the observed heterogeneity is caused by imperfect anchoring, our results suggest that this is definitely not the only factor contributing to signal diversity.

S.4 Case-studies of chromatin patterns at binding sites of various TFBSs

For case studies that would describe the genome-wide chromatin landscape around specific DNA-binding proteins, we chose four types of chromatin modifications: the canonical promoter-associated H3K4me3 (Fig. S6A); the enhancer-enriched H3K4me1 (Fig. S6B); two H3-acetylations, H3K27ac and H3K9ac (Fig. S6C), which are generally considered activating; and two repressive marks, H3K9me3 and H3K27me3 (Fig. S6D). For each type of modification (except H3K4me3) we chose one factor that tended to have binding sites near transcription start sites (“proximal”) and one whose binding sites tended to be further away (“distal”). We defined “proximal” as within 1kb (both upstream and downstream) of a transcription start site, and “distal” as beyond 1kb.

SIN3A in GM12878 is representative of factors that bind at or very near promoters, so the asymmetric H3K4me3 signals associated with SIN3A binding sites are not surprising and as such serve as a positive control. It is worth noting that CAGT identified 11 distinct shapes of H3K4me3 signal around SIN3A binding sites (the top five of which are shown in Fig. S6A).

For H3K4me1, FOXA2 in HepG2 is an example of a distal factor, whereas ETS1 in K562 is an example of a proximal factor. The majority of sites for both factors exhibit asymmetric patterns (Fig. S6B). Notably, 78% of all 32,905 peaks called for FOXA2 are distal sites with a high H3K4me1 signal, eliminating the possibility that promoter architecture drives the asymmetry.

We highlight four factors for H3 acetylation (Fig. S6C): For H3K27ac, we show EP300, the histone acetyltransferase that catalyzes this modification (Creighton et al. 2010; Tie et al. 2009), and E2F4, a transcription factor that in K562 cells is mostly promoter-associated. The majority of EP300 sites exhibit substantial asymmetry, suggesting that the complex in which EP300 acylates H3K27 acts directionally, presumably depending on the context of the site. Strong asymmetries in acetylation of H3K9 are also evident, for both factors shown, NFKB1 (distal) and NRF1 (proximal).

Finally, two components of repressor complexes also exhibit asymmetries. SUZ12, a subunit of the polycomb repressive complex, which catalyzes H3K27 trimethylation (Schwartz and Pirrotta 2007), is found both proximally and distally at about equal proportions, and exhibits moderate asymmetry in the H3K27me3 signal. Similarly, SETDB1, the methyltransferase that catalyzes H3K9 trimethylation (Schultz et al. 2002), while mostly found promoter-distal, exhibits appreciable asymmetry for H3K9me3.
To highlight the prevalence of asymmetric patterns around TFBSs, we chose three TFs with different preferences for TSS-proximal binding (Fig. S7). JUND is an example of a TF mostly found distally from TSSs (Wang et al. 2011). Although only 12% of JUND peaks are proximal, 90% of them show high H3K4me1 signal. Of these high-signal peaks, more than 70% are asymmetric. Clearly, this asymmetry cannot be entirely attributed to the 12% of proximal peaks. E2F1, on the other hand, is a TF with strong preference for promoters. The vast majority of E2F1 peaks are proximal to TSSs and almost all of them (> 98%) have high signal for the promoter-associated modification H3K4me3. The asymmetry in this case is much stronger than in the previous example, with nearly all (97%) high signal peaks exhibiting asymmetric patterns. The last example, MAX, has its peaks almost equally distributed between proximal and distal sites. Of the 88% of peaks with high signal for the activating mark H3K9ac, more than 80% are asymmetric; suggesting again that part of the asymmetry is coming from the distal sites.

S.5 Symmetry correlations between nucleosome positioning and chromatin mark patterns at TFBSs

In section “Spatial relationships among marks at factor binding sites” in the main text, we showed that asymmetric nucleosome positioning clusters may exhibit either asymmetric or pseudo-symmetric histone modifications. Here, we reversed this analysis and chose specific histone modifications as target-marks and nucleosome positioning as the partner-mark. We found that nucleosome positioning in asymmetric H3K27ac patterns at CTCF sites in K562 is symmetric (Fig. S8C). In this specific case, the nucleosome positioning patterns are in fact predominantly truly symmetric around CTCF sites (Fig. S3A), but the nucleosomes on either side of the CTCF binding sites exhibit differential histone modifications. Co-association analysis using nucleosome meta-shapes also supported the observation that histone mark asymmetry is partially correlated with but not entirely explained by asymmetry in nucleosome positioning. Patterns of some marks (e.g. H2A.Z and H3K27ac) show higher correlation with nucleosome positioning as compared to others (e.g. H3K79me2; Fig. S11).

S.6 Symmetry correlations between patterns of different chromatin marks at TFBSs

Using H3K4me1 as the target mark, we investigated its relationships with partner marks at different classes of TFBSs. SIN3A (Fig. S10A), a TSS-proximal factor, and HDAC2 (Fig. S12), a TSS-distal factor, are both part of histone deacetylase complexes. Both showed elevated levels of H3K27ac and H3K9ac around their binding sites. At SIN3A sites where H3K4me1 exhibits primarily asymmetric patterns, co-localized H3K27ac and H3K9ac exhibit partial pseudo-symmetry, that is, they are enriched on either side of the binding site (Fig. S10A). Similarly, at HDAC2, the asymmetric patterns of H3K27ac are associated with pseudo-symmetric patterns of H3K4me1 (Fig. S12). Finally, H3K4me2 was found to correlate with H3K4me1 but anti-correlate with H3K4me3 at SIN3A sites (Fig. S10A). In contrast, H3K4me2 was found to correlate strongly with H3K4me3 at HDAC2 (Fig. S12).

Next, we analyzed the co-association of histone marks around binding sites of EP300, which catalyzes the H3K27ac modification at enhancers (Fig. S10B). As expected, we found elevated levels of H3K27ac at EP300 sites. We also found that H3K27ac patterns at EP300 were predominantly asymmetric and their shapes were strongly correlated with enriched asymmetric patterns of H3K9ac, H3K4me3, and H3K4me2. However, the H3K9ac and H3K4me3 levels at EP300 sites were lower than those around promoter regions. In contrast to the SIN3A context (Fig. S10A), aggregate H3K4me2 patterns were correlated with H3K4me3.
At EP300 sites, H3K4me1 is significantly elevated, consistent with the idea that H3K4me1 can be present at enhancers (Hon et al. 2009). However, we found that the asymmetric H3K27ac patterns were associated with pseudo-symmetric H3K4me1 patterns, indicating a non-trivial relationship between the two enhancer marks that implies functional independence. Multivariate analysis using H3K4me1 meta-shapes as the target mark provided corroborating evidence that on average across all TFs, corresponding H3K4me3, H3K27ac and H3K9ac patterns tracked each other, while exhibiting pseudo-symmetry in the context of highly asymmetric H3K4me1 patterns (Fig. S13).

Finally, we used the repressive marks H3K9me3 and H3K27me3 as target-marks in multivariate analyses to study their relationship with activating (partner-)marks. We focused on SETDB1 and SUZ12 sites, since they are part of protein complexes that catalyze the H3K9me3 and H3K27me3 histone modifications, respectively (Fig. S10C, S10D). Interestingly, the most prevalent H3K9me3 cluster at SETDB1 sites in K562 showed an enrichment of the activating marks H3K4me3, H3K4me1, H3K9ac and H3K27ac compared to the other H3K9me3 clusters (Fig. S10C). The sites in this cluster likely represent bivalent domains. The top two prevalent H3K9me3 shapes were asymmetric and associated with asymmetric patterns of H3K4me3, H3K27ac and H3K9ac, with opposite orientation. For example, a high signal of H3K9me3 on one side of the SETDB1 binding sites is associated with a high signal of H3K9ac on the other side of the binding site. H3K27me3 was enriched around SUZ12 binding sites (as expected), while H3K4me3, H3K9ac, and H3K4me1 were depleted (Fig. S10D). Interestingly, the asymmetry of H3K27me3 was anti-correlated with H3K4me3 and H3K9ac but correlated with H3K4me1. We also observed ubiquitous anti-correlation of asymmetric co-localized patterns of H2A.Z and the H3K79me2 histone mark (Fig. S14).
Legends for supplementary figures

**Figure S1**: Case studies of TSSs with different asymmetry patterns of nucleosome positioning. The left side of each panel shows a region of approximately 1000 bp around the TSS of a gene, as visualized through the UCSC Genome Browser. The right side shows the CAGT cluster from Fig. 2 to which the corresponding gene was assigned. The black arrow indicates the direction of transcription. The small triangles indicate the positioned nucleosomes upstream (red) and downstream (blue) of the TSS. (A), (B) Two example loci with strong nucleosome positioning signal upstream and weaker signal downstream of the TSS. In both cases there is evidence of active transcription (CAGE tags, activation and elongation marks). (C), (D) Two examples with strong nucleosome positioning downstream of the TSS and almost absent positioning signal upstream. As before, there is evidence of active transcription.

**Figure S2**: (A) Distribution of distances between the peak summit and the nearest hit of a TRANSFAC or JASPAR motif for the corresponding TF in four ChIP-seq datasets in K562. (B) – (J) Independent runs of CAGT on peaks with and without a significant motif hit in the four ChIP-seq datasets from (A). (B) – (C) CTCF peaks (D) – (E) REST peaks (G) – (H) E2F4 peaks (I) – (J) E2F6 peaks.

**Figure S3**: Nucleosome positioning clusters around binding sites of (A) CTCF, (B) RAD21, (C) SMC3 in K562, and (D) ZNF143 in GM12878. Only the most prevalent clusters are shown. Contrary to other TFs, which exhibit mostly asymmetric nucleosome positioning around their binding sites, a large fraction of binding sites of ZNF143 and the members of the CTCF/cohesin complex are characterized by symmetric nucleosome positioning.

**Figure S4**: Nucleosome positioning patterns uncovered by CAGT around EBF1 binding sites in GM12878. The first panel is a traditional aggregation plot, averaging the signal over all 27655 EBF1 sites. The rest of the panels show the CAGT clusters in order of prevalence, with the percentage of EBF1 peaks in each of them shown in the header. Clusters containing less than 2% of EBF1 peaks each are omitted from the figure. CAGT uncovers substantial diversity of nucleosome positioning shapes between EBF1 sites, although the standard AP shows a flat signal.

**Figure S5**: Additional examples of nucleosome positioning clusters around TFBSs and relationship to GC content. This Figure complements Fig. 4. 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 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 bottom row shows the mean GC content of all sites used in the panel above it. If a site was “flipped” (re-oriented) 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. (A) REST in K562. (B) EBF1 in GM12878. (C) JUND in K562. (D) GATA2 in K562.

**Figure S6**: Case studies illustrating the prevalence of asymmetric histone modification patterns around TFBSs. For each case, only the five largest CAGT clusters are shown. N: total number of binding sites of the TF. H: percentage of sites with high signal for the mark. P: percentage of TSS-proximal sites. (A) An example for H3K4me3 signal around a promoter-associated factor. (B) Two examples for the H3K4me1 mark. (C) Four examples for acetylation mark signals. (D) Two examples of repressive mark signals.
**Figure S7**: Case studies illustrating that asymmetric histone modification patterns are also present distally from TSSs. For each case study, only the three largest CAGT clusters are shown. N: total number of binding sites of the TF. P: fraction of these sites that are TSS-proximal.

**Figure S8**: Relationship between nucleosome positioning and histone modifications. This figure complements Fig. 8A-B. (A), (B) The top row shows the most prevalent clusters of nucleosome positioning signal identified by CAGT around the binding sites of (A) SP1 in GM12878 and (B) SIN3A in GM12878. The fraction of sites in each cluster is shown in the header. The remaining rows show the signal of other marks averaged over the sites in each of the clusters. (C) Similar to (A) and (B) except clustering was performed on the H3K27ac signal around CTCF sites in K562. N: total number of binding sites of the TF. P: percentage of TSS-proximal sites.

**Figure S9**: Anti-correlation between H3K4me1 and H3K4me3 signals at POLR2A sites in six cell lines. This figure complements Fig. 8C. N: total number of binding sites of the TF. P: percentage of TSS-proximal sites.

**Figure S10**: Correlation patterns between different histone modifications. In all cases, the top row shows the most prevalent clusters identified by CAGT in the specified data set. The fraction of binding sites in each cluster is given in the header. The remaining rows show the signal of other histone modifications ("partner-marks") averaged over the sites in each of the clusters. The color of the vertical headers specifies the relationship between the partner-mark and the target-mark: pink indicates a positive correlation; blue a negative correlation; and grey refers to a lack of specific type of correlation between the two signals (grey is typically assigned to associations of asymmetric patterns of one functional mark with pseudo-symmetric patterns of another). (A) SIN3A sites in HepG2 (B) EP300 sites in HeLa-S3 (C) SETDB1 sites in K562 (D) SUZ12 sites in NT2-D1. N: total number of binding sites of the TF. P: percentage of TSS-proximal sites.

**Figure S11**: Co-association of nucleosome meta-shapes with all other functional marks and GC content. The first row shows all the meta-shapes (clusters) identified by applying CAGT on nucleosome positioning signal across all assayed TFs in K562 and GM12878. The remaining rows show the mean signal of other functional marks averaged across all TF sites that belong to each cluster and re-oriented based on the CAGT run on the nucleosome signal. GC container sites are visible at strongly positioned nucleosomes. DNase correlates with overall GC content and peaks at TFBS summits while being anticorrelated with overall nucleosome occupancy. Some asymmetric nucleosome positioning clusters, such as P_1, show significant correlation with co-associated histone mark patterns. On the other hand, asymmetric cluster P_4 is associated with pseudo-symmetric patterns of several chromatin marks such as H3K27ac, H3K9ac. H3K79me2 exhibits weak anti-correlation with nucleosome positioning patterns, i.e. on average consistently shows higher signal on the side with weaker nucleosome positioning.

**Figure S12**: The first row shows the five most prevalent clusters identified by CAGT in the H3K27ac signal around HDAC2 sites in HepG2. The remaining rows show the mean signal of four other histone modifications averaged across the set of HDAC2 binding sites in each cluster of the top row. N: total number of HDAC2 binding sites in HepG2. P: percentage of TSS-proximal sites. H3K4me1 shows partial pseudo-symmetry especially for cluster P_1 whereas the other marks are strongly correlated with the asymmetric H3K27ac pattern.
**Figure S13**: Co-association of H3K4me1 meta-shapes with four other chromatin marks. The first row shows the five most prevalent clusters identified by CAGT in the H3K4me1 signal across all assayed TFs in all available ENCODE cell lines. The remaining rows show the mean signal of four other chromatin marks averaged across the binding sites in each meta-shape of the top row. H3K27ac, H3K4me3 and H3K9ac exhibit pseudo-symmetry while being associated with highly asymmetric H3K4me1 patterns.

**Figure S14**: Co-association of H2A.Z meta-shapes with three other chromatin marks. The first row shows the nine most prevalent clusters identified by clustering H2A.Z signal across all assayed TFs in all available ENCODE cell lines. The remaining rows show the mean signal of other chromatin marks averaged across the binding sites that belong to each cluster. H3K79me2 shows strong anti-correlation with H2A.Z. Nucleosomes are well-positioned on either side, indicating differential modifications.

**Figure S15**: Clusters identified by CAGT in (A) the H3K27ac signal around CTCF sites in K562; (B) a randomized version of the first data set where values were permuted within each CTCF site; (C) a randomized version of the first data set where values at each site position were permuted across sites; (D) random genomic locations; (E) randomly picked locations within a window of +/- 5,000 bp from the original CTCF sites. In all cases, only the first 5 clusters are shown. For the robustness analysis, CAGT was run 100 times on each dataset. This figure only shows one sample run per dataset.

**Figure S16**: CAGT identifies shape clusters with distinct enrichments in TFs co-localized at REST binding sites. (A) Clusters of H3K27ac signal around REST sites in K562. Clusters 1, 6, and 7 are enriched in enhancer associated proteins (EP300, GATA1, GATA2, SMARCA4, TAL1; P < 0.01). Cluster 9, which is similar in shape to 6 and 7, but more asymmetric, is instead enriched in TSS-proximal factors, such as E2F4, SIN3A, YY1, and the TATA associated proteins TAF1 and GTF2F1. (B) H3K4me3 clusters around REST sites in K562. The symmetric shapes 3, 7, and 9 are enriched in the enhancer associated proteins GATA1, GATA2, and EP300 (7 is also enriched in TAL1 and 9 in SMARCA4; 3 is enriched in both). The similar but not symmetric shape 4 is highly enriched in the TATA associated proteins TBP, TAF1, TAF7, GTF2B, and GTF2F2, and in several other proximal factors (e.g. E2F4, E2F6, MYC, MAX). Cluster 6 exhibits similar enrichments. (C) The two symmetric nucleosome positioning clusters around REST sites in K562 from Fig. 3B. Cluster 16 is enriched in enhancer associated factors and in several other TFs including JUN, JUND, and FOS. Cluster 5 shows almost no enrichments, with the exception of a modest enrichment in the member of the cohesin complex RAD21. To compute enrichments we looked at the overlap of REST peaks with other ENCODE ChIP-seq datasets and used a Fisher exact test with Bonferroni correction.

**Legends for supplementary tables**

**Supplementary Table 1**: (Separate supplementary excel spreadsheet) ENCODE data sets used for all figures and supplementary figures. The names of the data files follow standard ENCODE conventions. Columns B-G refer to the ChIP-seq data set that was used to obtain binding sites to use as reference anchor points for CAGT. Columns E-F refer to the data set used to obtain a chromatin signal around these anchor points. The “partner mark” columns are only relevant for the figures that display more than one chromatin mark signal around the same set of sites. In such cases, the “primary mark” is the signal shown in the first row of the figure, that is, the signal based on which clustering was done. The “partner marks” are the marks in the rest of the rows.

---

9
### Supplementary Table 2

Datasets used for evaluating CAGT’s robustness and reproducibility. The datasets are described in the text. For each dataset, we ran CAGT 100 times with different random initializations. 

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#sites</th>
<th>#high signal</th>
<th>#clusters</th>
<th>ARI</th>
<th>shape corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>original</td>
<td>45094</td>
<td>14065 (31.19%)</td>
<td>9.74</td>
<td>0.4736</td>
<td>0.9052</td>
</tr>
<tr>
<td>within row permutation</td>
<td>45094</td>
<td>14065 (31.19%)</td>
<td>40</td>
<td>0.0156</td>
<td>-0.0069</td>
</tr>
<tr>
<td>within column permutation</td>
<td>45094</td>
<td>45094 (100%)</td>
<td>31.55</td>
<td>0.1605</td>
<td>-0.0073</td>
</tr>
<tr>
<td>jitter</td>
<td>45094</td>
<td>8172 (18.12%)</td>
<td>9.79</td>
<td>0.5213</td>
<td>0.9812</td>
</tr>
<tr>
<td>matched</td>
<td>45094</td>
<td>8043 (17.84%)</td>
<td>10.19</td>
<td>0.4788</td>
<td>0.972</td>
</tr>
<tr>
<td>control motif</td>
<td>59478</td>
<td>7360 (12.37%)</td>
<td>9.14</td>
<td>0.4878</td>
<td>0.9893</td>
</tr>
<tr>
<td>random</td>
<td>45094</td>
<td>3175 (7.04%)</td>
<td>9.59</td>
<td>0.4965</td>
<td>0.9561</td>
</tr>
</tbody>
</table>

**Supplementary Table 2:** Datasets used for evaluating CAGT’s robustness and reproducibility. The datasets are described in the text. For each dataset, we ran CAGT 100 times with different random initializations. 

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#peaks</th>
<th>%peaks with motif</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTCF</td>
<td>45094</td>
<td>88.98%</td>
</tr>
<tr>
<td>REST</td>
<td>14144</td>
<td>36.02%</td>
</tr>
<tr>
<td>E2F4</td>
<td>4874</td>
<td>14.77%</td>
</tr>
<tr>
<td>E2F6</td>
<td>26718</td>
<td>7.08%</td>
</tr>
</tbody>
</table>

**Supplementary Table 3:** Fraction of peaks of four ChIP-seq datasets with at least one significant hit for any of the TRANSFAC or JASPAR motifs of the corresponding TF. All datasets are in the K562 cell line.
Supplementary references


Kheradpour P, Kellis M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments (In press)


FIG. S1
Nucleosomes at CTCF sites with motifs in K562

Nucleosomes at CTCF sites with no motifs in K562

Nucleosomes at REST sites with motifs in K562

Nucleosomes at REST sites with no motifs in K562

H3K27ac at E2F4 sites with motifs in K562

H3K27ac at E2F4 sites with no motifs in K562

DNase at E2F6 sites with motifs in K562

DNase at E2F6 sites with no motifs in K562

FIG S2
A  Nucleosomes at CTCF in K562

B  Nucleosomes at RAD21 in K562

C  Nucleosomes at SMC3 in K562

D  Nucleosomes at ZNF143 in GM12878

FIG S3
## EBF1 in GM12878

<table>
<thead>
<tr>
<th>Distance from summit</th>
<th>Nucleosome signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>all (27655)</td>
<td></td>
</tr>
</tbody>
</table>

### P_1 (9.81%)

### P_2 (8.08%)

### P_3 (7.34%)

### P_4 (6.96%)

### P_5 (5.45%)

### P_6 (5.19%)

### P_7 (5.16%)

### P_8 (5.16%)

### P_9 (5.12%)

### P_10 (5.09%)

### P_11 (5.04%)

### P_12 (4.98%)

### P_13 (4.95%)

### P_14 (4.89%)

### P_15 (4.85%)

### P_16 (4.84%)

### P_17 (4.52%)

### P_18 (2.22%)
FIG S5
FIG S6
A
Distal factor:
JUND in HeLa
N = 26,074
P = 12%
Signal: H3K4me1

B
Proximal factor:
E2F1 in HeLa
N = 9,632
P = 88%
Signal: H3K4me3

C
Mixed factor:
MAX in K562
N = 38,963
P = 49%
Signal: H3K9ac
A  Nucleosomes at SP1 in GM12878 (N=13139, P=53%)

B  Nucleosomes at SIN3A in GM12878 (N=10504, P=85%)

C  H3K27ac at CTCF in K562 (N=45094, P=19%)
H3K4me1 at POLR2A in H1-hESC (N=20777, P=70.6%)

H3K4me1 at POLR2A in GM12878 (N=21446, P=70.53%)

H3K4me1 at POLR2A in HepG2 (N=20885, P=71.2%)

H3K4me1 at POLR2A in HUVEC (N=20569, P=83%)

H3K4me1 at POLR2A in K562 (N=18052, P=78.2%)

FIG S9
FIG S10

A. H3K4me1 at SIN3A in HepG2 (N=10562, P=77%)

B. H3K27ac at EP300 in HeLa-S3 (N=18335, P=8.7%)

C. H3K9me3 at SETDB1 in K562 (N=2850, P=38.9%)

D. H3K27me3 at SUZ12 in NT2-D1 (N=4554, P=49%)

Signal

Distance from summit (bp)

Corresponding Signal

Distance from summit (bp)
Coassociation of nucleosome meta-shapes with all other functional marks
H3K27ac AT HDAC2 in HepG2
(N=15440, P=25%)

Corresponding Signal

FIG S12
H2A.Z meta-shapes at all TFBSs in all cell-lines

FIG S14
A. H3K27ac at CTCF in K562
CAGT results using true profiles

B. CAGT results using permuted profiles
H3K27ac signal shuffled across coordinates per profile

C. CAGT results using permuted profiles
H3K27ac signal shuffled across profiles per coordinate

D. CAGT results using H3K27ac signal from random genomic locations

E. CAGT results using H3K27ac signal from randomly jittered CTCF sites

Distance from summit (bp)
A

**H3K27ac at REST in K562**

Distance from summit (bp)

EP300, GATA1, GATA2, SMARCA4, TAL1

E2F4, SIN3A, YY1, TAF1, GTF2F1

B

**H3K4me3 at REST in K562**

Distance from summit (bp)

EP300, GATA1, GATA2, SMARCA4, TAL1

TBP, TAF1, TAF7, GTF2B, GTF2F2, E2F4, E2F6, MAX, MYC

EP300, GATA1, GATA2, TAL1

EP300, GATA1, GATA2, SMARCA4

C

**Nucleosomes at REST in K562**

Distance from summit (bp)

RAD21

JUN, JUND, FOS

FIG S16