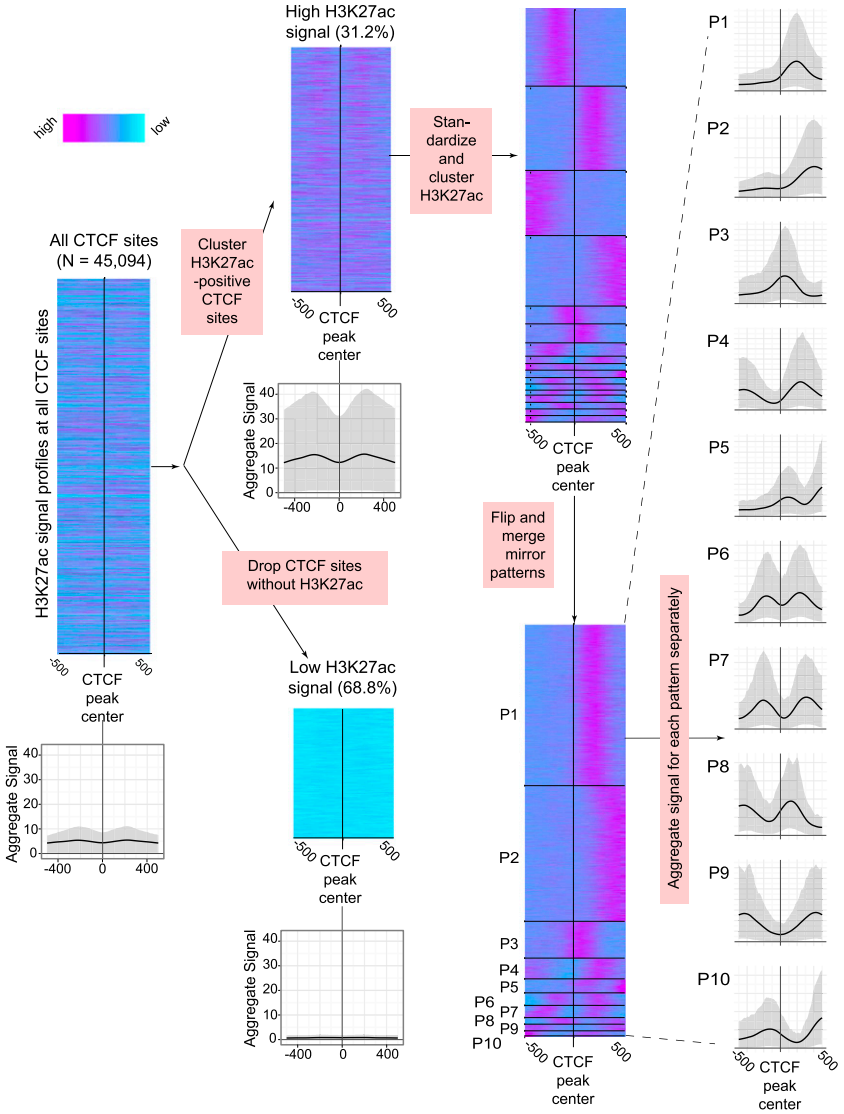

Figure 1. Schematic of the steps followed by CAGT in order to group the signal profiles around a set of genomic features into distinct and coherent clusters. The steps are illustrated using H3K27ac signal profiles around CTCF binding sites in the K562 cell line. (1) We start by extracting the H3K27ac signal intensity profiles in a window (±500 bp) around each feature (CTCF binding site) and aligning all signals at the core of the feature (summit of the CTCF peak). The grayscale plot at the *bottom* is a traditional aggregation plot obtained by averaging all signal profiles. The bold line is the mean intensity, while the shaded area around it corresponds to the 10th and 90th percentiles of the signal. (2) The sites are divided into high and low signals based on the peak intensity of each H3K27ac signal profile around each site. (3) High signal sites are standardized to zero mean and unit standard deviation and clustered with the *k*-medians algorithm. This step typically leads to a large number of compact clusters, some of which may be redundant with similar average patterns. (4) In the final step, similar clusters, as well as clusters that are mirror images of each other, are merged using hierarchical agglomerative clustering, resulting in a small number of distinct, nonredundant, compact clusters (see Methods for details).

transcription, has been well-studied previously (Fu et al. 2008; Mavrich et al. 2008; Schones et al. 2008; Shivaswamy et al. 2008; Jiang and Pugh 2009; Kaplan et al. 2009; Rando and Chang 2009; Segal and Widom 2009; Radman-Livaja and Rando 2010; 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 15,736 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 >2% of the TSSs each and comprised a total of 89.56% of the TSSs studied (Fig. 2A; Supplemental 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).

To reveal correlations with transcriptional activity, we quantified expression levels based on CAGE tags for each cluster (Fig. 2B). The most prevalent cluster, comprising 20.64% of TSSs, had low levels of gene expression as measured by CAGE (Djebali et al. 2012) and exhibited no strong positioning for 900 bp around the TSS, consistent with previous analyses that used standard APs (Schones et al. 2008; Valouev et al. 2011). Other clusters that were associated with low gene expression had no positioning upstream, but pronounced and often somewhat distant positioning downstream from the TSSs. On the contrary, two clusters with high expression that together comprise 19.79% of TSSs (clusters 3 and 5) had strongly positioned nucleosomes ~250–350 bp upstream of the TSS, but much weaker positioning downstream. A similar phenomenon has been observed in yeast, where highly expressed genes often lack a well-positioned +1 nucleosome (Zaugg and Luscombe 2011). Finally, two additional clusters (9 and 10), comprising 6.35% of TSSs, had strongly positioned nucleosomes downstream, at positions consistent with the canonical +1 assignment. Clusters with particularly pronounced nucleosome positioning peaks, either upstream of or immediately downstream from the TSS (clusters 3, 5, 9, and 10) had significantly higher expression than all other clusters ($P < 0.001$).

Most of the clusters that exhibited strong positioning upstream of the TSS (1, 2, 3, 5), as well as cluster 4, which also has a relatively high upstream peak, were significantly enriched in CpG promoters ($P < 0.0001$). The enrichment was more pronounced